

Large AI Models

FOR GERMANY

Gefördert durch:



Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

LEAM:AI



KI BUNDESVERBAND

**Machine
Translated
by Lengoo**

AI made in Germany



FEASIBILITY STUDY



IMPRINT

The feasibility study on LEAM - Large European AI Models was carried out by:

**Akademie für Künstliche Intelligenz AKI gGmbH
at German AI Association**

Schiffbauerdamm 40, 10117 Berlin
+ 49 (0) 30 30 340 388
info@ki-verband.de • www.ki-verband.de

Contracting entity

**German Federal Ministry
for Economic Affairs and Climate Action**

11019 Berlin
+ 49 (0) 30 18 615 - 0
info@bmwk.bund.de • www.bmwk.de

Layout

Dr. Sandra Bütow
KI Bundesverband

Graphics & Design

Victoria Patricia
contact@victoriapatricia.com • www.victoriapatricia.com

The project on which this report is based was funded by the German Federal Ministry of Economics and Climate Protection under grant number 01MT22006. Responsibility for the content of this publication lies with the authors.

© 2023



Jörg Bienert's greeting



Jörg Bienert, President of the German AI Association

In 2002, the New York Times published an article entitled "Google's toughest search is for a business model" (Hansell, 2002). The author of the article was convinced that Google would not stand up to its competitors at the time in the online advertising business and that its core business at the time, licensing the search engine to other web portals, would not bring sufficient growth.

Since then, much has happened. Within a few years, the Google search engine dominated the Internet, the online advertising market, and played a dominant role in many more areas. Today, no online business model can be launched without massive investments in Google Ads; many cars have built-in Google Maps navigation; the most widely used browser, Chrome, collects diverse data about our browsing behaviour; and millions of people worldwide entrust their data to the Alphabet Group through online services such as Gmail, Docs or Drive. Europe has tried unsuccessfully to build an alternative with its own search engine.

We see all of this as problematic but only in a limited way, because Google has managed to build a huge advantage in trust. But what if that changes? Just think of the hypothetical scenario of Elon Musk taking over Alphabet. Would we be able to make meaningful use of the Internet without Google Services? Or has Alphabet already managed to create a lock-in situation through the collection of data and services based on it, from which we can hardly free ourselves?

Artificial intelligence has the potential to change the world just as much as the Internet has. There will be many new functionalities, products, and business models. Process improvements will lead to efficiencies and disruptive changes that we cannot even anticipate today. In 2000, we probably would have just shaken our heads at the prediction that the Internet would lead to the bankruptcy of Germany's largest shipping and department stores.

We are now witnessing the beginning of the second wave of the AI revolution, which began with OpenAI's release of GPT-3 in the summer of 2020. Drawing on vast amounts of data, and using huge amounts of developer resources, money, and computing capacity, OpenAI created a language model that was able to perform at an unprecedented level. In this study, we look more closely at this development and the status quo.

So what will change in the next 2-3 years? We will see a variety of new models and applications. The scope and impact is currently difficult to estimate, but a lot of questions will arise such as:

- When and how will large language models achieve near 100 percent content correctness in the generated texts?
- What impact does this have on companies in all sectors and possible applications in the different areas of the value chain?
- In addition to language models, how good are image and video generators and the combination of these? What does this mean for the creative industry, journalism and education?
- How will the Internet change with a flood of automatically generated content? Will the big language models replace Google search, opening up the online advertising market to other players as well?
- Through the democratization of AI, can we transform the labour market in such a way that the demographically smaller generations can handle the work in a healthy way while at the same time compensating for the shortage of skilled workers, especially in "human" jobs?
- To what extent can the most pressing problems of mankind, such as energy transition, climate change, and health, be solved with the help of large models?

An overarching question is crucial in this regard. Who owns the technology, data and resources to create large models and drive development and revolutionary breakthroughs? Will this remain in the hands of a few large corporations, and will they determine the use, regulation and also the treatment of ethical issues, as is already the case with the Internet, search engines and social networks?

It is precisely here that a major challenge lies for Germany and Europe at the moment. Only those who master the technology will be able to shape its use at their own discretion and positively influence economic and social development. This is far from just about Artificial Intelligence as a platform technology. Large AI models will enable completely new applications and disruptively change existing business models and value creation – across all industries.

If we are not able to develop and deliver this basic technology on our own, German industry will have to shy away from foreign services, lose parts of the value chain and lose competitiveness. We will have difficult discussions around privacy, data security, and the ethical use of AI models. The current controversies around Google and Facebook, in comparison, seem trivial.

In order to prevent this and not lose digital sovereignty in AI as well, we in Germany need to be able to do research at an international level, collect and refine data, train large models, and make them available for use by business, corporations, small and medium-sized enterprises and start-ups.

This is the goal of our initiative Large European AI Models, LEAM for short. Last year, a team of 40 representatives from academia, business and society developed initial ideas for an infrastructure for the creation of large models. This feasibility study, which we present to you here, has now been prepared on behalf of the Federal Ministry of Economic Affairs and Climate Protection.

The core of the concept is to build a dedicated AI supercomputing infrastructure. A team of specialists operates a dedicated hardware infrastructure that is specialized in large AI models. It continues to develop these AI models and makes them available to others. In addition, the team collects and refines the data necessary for operations and applications, and implements software and services around these AI models that simplify the training and tuning of large models, making them easy to use for different target groups.

With LEAM, we are planning a central AI flagship project that will form around a powerful ecosystem of academia, business and start-ups – in close collaboration also with existing activities such as Open GPT-X, Aleph Alpha or Bloom, and as a major player in the European context. Emphasis is placed on taking into account European values and upcoming standards and regulations.

We are very pleased with the broad support from academia, business and politics. Because only together with all stakeholders can we meet the challenge of harnessing the possibilities of Artificial Intelligence for the benefit of people, for staying internationally competitive through powerful research and products, and for securing prosperity in Germany.

KEY RESULTS

AI Foundation models are the future.

AI Foundation models will lead to disruptive applications based on Artificial Intelligence. This development will enable many new applications, platforms and business models in almost every area of the economy and society in the near future. LEAM sets out a roadmap for how Germany can participate in this future technology.

The development of AI Foundation models takes place mainly in the USA.

U.S. models dominate the market. As of 2017, 73% of KI-Foundation models are from the U.S. and 15% are from China. In the US, development is mainly driven by the big tech companies with investments in the billion-dollar range.



Germany faces significant challenges.

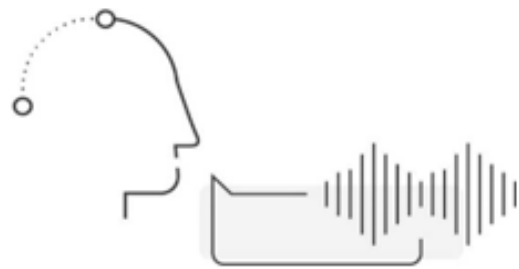
Germany risks missing the paradigm shift in the field of KI-Foundation models and becoming increasingly dependent on American models. The digital sovereignty of KI and downstream applications is under acute threat, and thus the competitiveness of the entire economy in Germany.

The relevance to the economy is high.

Two-thirds of the KI companies surveyed already work with these types of models. More than 80% consider the building of a German ecosystem as well as the development of foundation models based on European values to be relevant. Some of the

large companies surveyed are already experimenting with different applications and evaluate the use for internal processes and their products. The importance of the models will continue to increase in the future.

The need for language models is particularly great. According to the AI Foundation models, language models currently have the greatest relevance in the use and development of all model types. 71% of KI companies focus on this. This is followed by multimodal models (38%) and models based on business and manufacturing processes (34%)



The demand for services is significant.

Business needs applications tailored to its needs and use cases based on freely available foundation models. This requires the provision of the necessary computing capacity, data pools and support for training and consultancy services.

Germany needs a powerful AI computing infrastructure.

A prerequisite for developing AI foundation models is access to a high-performance AI super-computing infrastructure, which does not currently exist in this form in Germany. Building a suitable AI data center in Germany is both hardware and software feasible and will cost around 350-400 million euros.

LEAM connects and expands the existing KI ecosystem. As a unique initiative from companies, K-start-ups and research LEAM brings together relevant players in the field of KI-Foundation models. The initiative is working closely with lighthouse projects and companies such as OpenGPT-X, Aleph Alpha and Huggingface to build synergies.

LEAM enables the construction of trusted foundation models in Germany.

As part of this feasibility study, the LEAM (Large European AI Models) initiative has developed a concept to provide the necessary infrastructure and associated services to develop trusted open-source foundation models according to European ethical standards. This will enable the industry to use the Foundation models developed there for their own applications.

LEAM can act as an AI beacon for professionals in. The emigration of well-trained talent to the US due to a lack of professional development opportunities in Germany is an increasing problem. An AI lighthouse offers the opportunity through state-of-the-art projects to retain talent in Germany and train new professionals.

LEAM can ensure Germany's competitiveness through joint investments. A LEAM AI service center can be successfully built and operated through a joint initiative of business, academics, and politics in Germany. The core financing by the public sector is provided by the private sector! These investments are complementary.



Table of Contents

Jörg Bienert's greeting.....	3
Key results.....	6
Table of Contents	8
Introduction	11
Objectives of the feasibility study	13
Authors of the feasibility study	13
Reading instructions and document structure	16
1. The paradigm of the AI foundation models	19
2. Technological foundations.....	23
2.1 The Importance of the Size of AI foundation models	31
2.2 Areas of application of AI foundation models.....	33
2.3 Trusted AI foundation models	45
2.4 Open research questions, latest developments and expectations	51
3. AI foundation models in international comparison	55
4. Business needs for AI foundation models	67
5. Support for development through research and science.....	76
6. Opportunities and plans for developing European AI foundation models	85
6.1 First European multilingual foundation language models.....	88
6.2 Avoiding misrepresentation, bias and toxicity	97
6.3 Connecting foundation models with large knowledge holdings	99
6.4 Combining language with other modes and media	100
6.5 Questions and developments.....	105
6.6 foundation models in other data domains.....	106
6.7 Summary	107
7. Prerequisites for software and personnel.....	109
7.1 Application Layer: Training & Inference Technologies	112
7.2 Data-Storage & -Loading-Layer.....	118
7.3 System Layer	120
7.4 Framework- & Service-Layer	121
7.5 LEAM as a lighthouse project for the future of the AI ecosystem	123
7.6 Summary	125
8. Building a high-performance AI data center	127
8.1 Definition Data Center.....	127
8.2 Requirements for an AI high-performance data center	130
8.3 Sustainability aspects	136
8.4 Infrastructure requirements in detail	137
8.5 Selection of location	139
8.6 Operation of an AI data center	156
8.7 Summary and recommendation	164
9. The organisational structure of LEAM	167
9.1 Target audiences of the LEAM AI Service Centre	167
9.2 Organizational units of the LEAM AI Service Centre	168
9.3 The LEAM-Board	181
9.4 Summary	182
10. Business management aspects.....	184

10.1 Costs.....	184
10.2 Revenue	189
11. LEAM financing models.....	193
11.1 Public funding	194
11.2 Private financing	195
11.3 Public-Private-Partnership	197
11.4 Legal framework	199
11.5 The impact of legal matters on financing models.....	205
11.6 Final overview.....	218
12. Corporate structure of LEAM.....	220
12.1 Public funding	220
12.2 Private financing	221
12.3 Public-Private-Partnership	221
13. Scenario for a LEAM AI Service Centre.....	223
14. Conclusion.....	227
14.1 Feasibility assessment.....	227
14.2 Forecast	229
I. Source Directory	231
II. Picture Directory.....	240
III. Table List	242
IV. Abbreviations list	243
V. Feasibility study methodology.....	245
Data collection	246

FEASIBILITY STUDY

for setting up and operating a dedicated
high-performance AI data center
for training large AI models in Germany

Introduction

An interview with the Open AI trained model: ChatGPT

Authors

What are large AI models?

Chatbot ChatGPT

Large AI models are AI models that have been trained on very large amounts of data and are capable of solving very complex tasks. They are often used in speech processing to accomplish tasks such as understanding human speech and recognizing objects in images and videos. A well-known example of a large AI model is OpenAI's GPT-3 model. This model has been trained on a huge amount of data and can perform many different tasks in the field of nature and language processing.

Why are large AI models important?

Large AI models are important because they are able to solve very complex tasks that would be too difficult for smaller AI models. By training on very large amounts of data, they are able to make very accurate predictions and classifications and can therefore be used in many areas where accurate results are required, for example in medicine, the financial industry and language processing.

Where are large AI models mainly developed?

Large AI models are developed mainly by technology companies and research-oriented institutions. These companies and institutions have the resources and expertise to source and process the vast amounts of data needed to train large AI models. Some of the well-known companies and institutions involved in developing AI models include Google, OpenAI, and the Massachusetts Institute of Technology (MIT).

Why does Europe need an initiative to remain competitive
in the development of large AI models?

The development of large AI models is an important area in the world of technology and many companies and countries are investing a lot of time and money to make progress in this area. Europe wants to remain competitive in this area in order to continue to play a leading role in the development of technologies and innovation. An initiative in this area could help Europe to remain a leader in AI research and development and to reap the benefits that AI can offer to its citizens and businesses.

How powerful large AI models or AI foundation models ¹ already are is shown by these introductory words. They were written not by one of the many researchers and experts who contributed to this study, but by the GPT-3-based chatbot ChatGPT.

ChatGPT's responses impressively demonstrate what AI foundation models are already capable of today. However, development is still very early and has so far mainly taken place in the United States and China. To prevent a medium-term dependence on foreign technology companies and to withstand competition, Germany and Europe must be able to co-create this next generation of innovative AI technologies.

To this end, the German AI Association has launched the LEAM - Large European Language Models - initiative in 2021, which is supported by over 40 renowned institutions from research and business, as well as other European AI associations. A key requirement of the initiative is to promote a European AI computing infrastructure to be used by academia, industry and start-ups alike, as well as to build a dedicated high-performance AI computing center in Germany.

In order to verify the feasibility of these objectives, the Federal Ministry of Economic Affairs and Climate Change (BMWK) commissioned the German AI Association to conduct a feasibility study on LEAM. This study critically examines the challenges and potentials of AI foundation models for German AI development, identifies and concretizes strategies and tools for implementing the LEAM objectives, and elaborates and compares action options. The findings of the feasibility study are trailblazing for the development of artificial intelligence in Germany and crucial for the innovative capacity of the European AI ecosystem and Germany's digital sovereignty in the future.

¹ The terms "Large AI Models" and "AI foundation models" are synonymous. The term "AI foundation models" is used for this study, which is internationally recognized for this type of AI.

Objectives of the feasibility study

In a very short time, Artificial Intelligence has made huge leaps in development and has thus amazed even technology experts. This is due in particular to the great advances in the area of large AI models. Since OpenAI introduced GPT-3, many applications have been developed based on the large AI models, also called foundation models.

In addition to the tremendous opportunities that it offers for work and life, it also presents some challenges for the German economy, science and society, as Europe is lagging behind in the competition for AI foundation models. Specifically, for Germany, this means that data protection and security are pursuing lower standards, distortions and poor data quality cannot be identified and counteracted due to a lack of transparency, and German companies will become only beneficiaries rather than designers of foundation models. Technologically, Germany is thus slipping into dependence.

A fundamental challenge is: The creation of AI foundation models requires enormous computing capacities and resources, which are not sufficiently available in Germany compared to the US.

The aim of this study is to develop a concept of how to build a dedicated AI computing infrastructure in Germany that allows AI foundation models to be trained and delivered to the economy. Above all, the needs of German companies should be taken into account.

By providing recommendations for the design of an AI data center and a corresponding service company, the study aims to serve as a decision-making basis for the private and public sector to implement LEAM in Germany.

Authors of the feasibility study

The LEAM feasibility study was carried out in cooperation with Alexander Thamm GmbH, the German Research Centre for Artificial Intelligence (DFKI), eco-Verband der Internetwirtschaft e. V., Fieldfisher LLP, Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), Merantix Momentum GmbH, Simmons & Simmons and Ubermetrics Technologies GmbH. The overall project management was carried out by the KI Bundesverband e.V.

Authors



Jörg Bienert
Alexander Thamm GmbH
KI Bundesverband e.V.



Roland Broch
eco – Verband der
Internetwirtschaft e.V.



Patrick Bunk
Ubermetrics Technologies
GmbH



Dr. Aljoscha Burchardt
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Vanessa Cann
KI Bundesverband e.V.



Ben Deetjen
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Alex Dickmann
KI Bundesverband e.V.



Dr. Nicolas Flores-Herr
Fraunhofer-Institut für
Intelligente Analyse- und
Informationssysteme (IAIS)



Laszlo Friedmann
Fraunhofer-Institut für
Intelligente Analyse- und
Informationssysteme (IAIS)



Christopher Götz
Simmons & Simmons LLP



Fabio Haenel
Merantix Momentum GmbH



Dennis Hillemann
Fieldfisher LLP



Kim Lambers
eco – Verband der
Internetwirtschaft e.V.



Lukas Lux
Alexander Thamm GmbH



Dr. Johannes Otterbach
Merantix Momentum GmbH



Dr. Gerhard Paaß
Fraunhofer-Institut für
Intelligente Analyse- und
Informationssysteme (IAIS)



Charlene Röhl
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Dr. Sven Schmeier
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Dr. Philipp Slusallek
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)



Boris Strauch-Rötting
Simmons & Simmons LLP



Oliver Süme
Fieldfisher LLP



Dr. Christian Temath
Fraunhofer-Institut für
Intelligente Analyse- und
Informationssysteme (IAIS)



Jakob Tesch
Ubermetrics Technologies
GmbH



Hauke Timmermann
eco – Verband der
Internetwirtschaft e.V.



Dr. Hans Uszkoreit
Deutsches
Forschungszentrum für
Künstliche Intelligenz (DFKI)

Our thanks also go to the entire team of the German AI Association for the implementation of the LEAM conference and their participation in the study: Daniel Abbou, Stefanie Baade, Dr. Sandra Bütow, Katharina Fischer, Franziska Fink, Phillip Handy, Benjamin Rodatz, Valentin Roth, Julia Sartison and Esther Schragmann.

Reading instructions and document structure

Chapters 1 to 5 are devoted to taking stock. Chapter 1 begins with a section on the key features of the new paradigm, followed by a breakdown of the state of the art in international research and development (Chapter 2). This breakdown also describes multiple applications that have already been realized and evaluated (Chapter 2.2). A special section is devoted to technologies designed to ensure the trustworthiness of AI foundation models and to prevent critical cases of inadequate performance (Chapter 2.3). This is followed by an analysis of international competition, based on a compilation of all AI foundation models published to date (Chapter 3). This analysis also discusses the causes of the identified imbalance between the US, China and Germany or Europe, respectively.

After reviewing the current state of international AI development, we focus on the situation in Germany in chapters 4 and 5. In particular, we highlight the needs of the German economy and show how the research can support the development of the required European AI foundation models. Numerous experts in science and business were interviewed for this purpose.

The sixth chapter draws conclusions from the results of the previous chapters. The investigation has identified the great potential of the technology, but at the same time an immense need for foundation models. This chapter describes which models should and could be developed as a matter of urgency, whether and how to obtain the data for these models, and what options exist for future model developments.

The following two chapters examine the requirements placed on high-performance computing systems (HPC) and specifically on a required AI high-performance data center. Chapter 7 explains what the software architecture needed to do this can look like and how to derive hardware and personnel requirements from it. In doing so, we demonstrate how this architecture can be built using open-source software (OSS) to emphasize the importance of OSS for a competitive AI ecosystem. The technical feasibility is directly linked to the scientific and infrastructure chapters of this feasibility study in order to explain the societal added value of operating a competitive AI high-performance data centre.

In particular, Chapter 8 looks at the technical prerequisites placed by AI foundation models on the infrastructure of a corresponding data center. It will also provide an overview of HPC capabilities in the research sector and the commercial vendors that already offer HPC cloud solutions suitable for AI applications. It also highlights the steps needed to locate and build a high-performance AI data center with the appropriate building infrastructure. Other possibilities such as collocation operation or the possible sourcing of HPC computing services from the cloud are explored using examples. In particular, the energy efficiency of the systems and the sustainability aspect of running an AI data center are important factors to examine in the operation of an HPC data center in view of the evolution of electricity prices and climate protection requirements.

Chapters 9 through 13 shed light on the organizational, economic, and legal feasibility of an AI high-performance data center. We are proposing the establishment of a LEAM KI service centre to offer its services to business and academia. To this end, different organisational models, corporate structures and financing options will be explored and discussed in more detail. In addition, a legal assessment of the options is carried out. The chapter concludes with a scenario for the realization of a LEAM KI service center.

The results are summarised in Chapter 14. The chapter also makes the recommendation to set up a project development company to further the issue.



The paradigm of AI foundation models

.....●

1. The paradigm of the AI foundation models

AI foundation models are writing the latest chapter in the Artificial Intelligence success story.

AI foundation models are large neural AI models that have been pre-trained on gigantic amounts of generic data. The special thing about these models is that the implicit knowledge acquired in the so-called pre-training is suitable as a basis for many different applications. Some applications even require no or minimal additional training. Other applications still require training on application-specific learning data, but often require much less learning data or achieve performance that would not be possible without generic pre-training.

The new paradigm has become known to the general public at the latest since the major media echo on GPT-3, a major language model from the California AI lab, OpenAI. GPT-3 was released in 2020 and is considered a spectacular breakthrough in the field of intelligent speech technology. The AI model can write sophisticated texts, answer questions, complete sentences and translate easily. At present, media attention is focused on the new AI system ChatGPT, a variant of the GPT-3 model, which fascinates and sometimes frightens journalists and test users alike because it speaks so eloquently and gives mostly informative, thoughtful and balanced answers even to difficult questions, and sometimes even formulates print-ready texts for a variety of purposes according to user instructions. And all this in several languages.

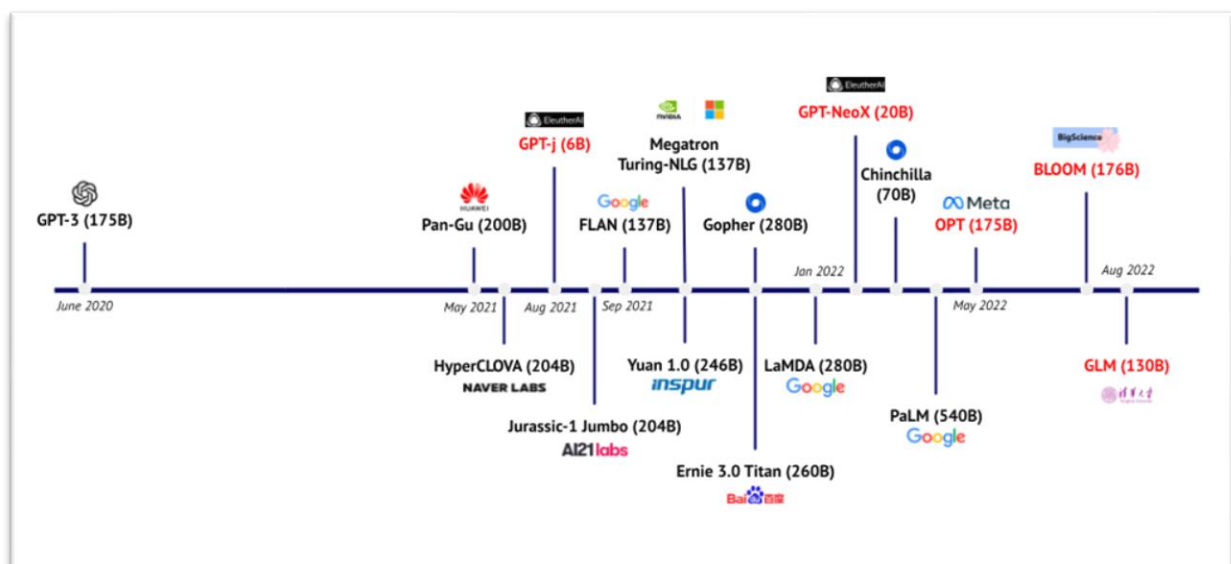


Fig. 1: Number of parameters of large AI language models since GPT-3 (open-source models marked in red). Source: State of AI Report 2022 (Benaïc & Hogarth, 2022)

This successful innovation was made possible by a new architecture for neural networks, called Transformer, which was developed in 2017 in Google research and first successfully tested in automatic text translation. Two AI foundation models followed in 2018, each realizing only parts of the Transformer architecture: GPT, the forerunner of GPT-3 from OpenAI and BERT, an influential bidirectional language model from Google research.

The basic idea of transfer learning is to use already trained networks for new tasks. Instead of training a new network for an application, you use a network that has already been pre-trained for another task. Thus, using the pre-trained network can reduce the need for annotated (labeled) application-specific learning data. If it is now possible to select tasks for pre-training in such a way that either there is a large enough amount of already annotated learning data or the annotated data can be produced completely automatically, then the effort of data annotation will be reduced immensely.

The first AI foundation models were large language models. They differ from previous AI models by their versatile usability. This versatility has been made possible by a key inherent feature of the new core AI technology, called emergence. This has reached a new stage in the evolution of Artificial Intelligence. Previous neural AI models have always been geared towards a specific application. They usually relied on training through supervised or semi-supervised learning. But unlike human intelligence, the acquired skills could not be used for further tasks.

This is why this Artificial Intelligence was also aptly referred to as Narrow AI and compared to the so-called Artificial General Intelligence (AGI).

Now, however, there are many applications that require the combination of tasks. For example, a voice chatbot can consist of a combination of speech recognition, semantic analysis, answering questions and acoustic speech output, each implemented by its own models. There may be inconsistencies here, for example, because one of these models knows words or terms that another has not learned. The homogenization of the models has reduced these inconsistencies. The AI foundation language models can already use the pre-trained implicit language knowledge for multiple applications. It no longer happens that one application or component does not know terms or sentence structures that are dominated by another. The paradigm of the versatile foundation models thus represents the next step towards a broader AI.

The name *foundation models* was only proposed in August 2021 by the new Center for Research on Foundation Models (CRFM) at Stanford University (Bommasani et al., 2021), which is fully dedicated to the new AI paradigm as part of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). The term is more appropriate than the *term large language models*, because foundation models do not always have to be language models, but can also be trained on images, videos or DNA sequences, for example, and then used for a variety of applications. There are also large language models that have been trained with appropriately selected and annotated data only for a specific application, such as machine translation, and are therefore not suitable as a foundation for many different applications.

Because the pre-trained, versatile foundation models are currently dominating the discussion around AI's next breakthroughs, the intuitive new term was very quickly picked up by the international research community. In their initial position paper on the research topic, the Stanford scientists not only outline the application potential of the pre-existing models, but they also argue convincingly that the paradigm of the foundation models will determine the next stage of AI development, in which the models will have capabilities that until recently seemed unthinkable and that surpass humans in many tasks.

But in the same article, they also describe the risks that can arise if this powerful technology exacerbates the concentration of technological and economic power in the hands of a few IT corporations. Mere availability of ready-made models is not enough to safeguard the economic and social interests of society and to effectively prevent the misuse of technology.



Technological foundations

.....●

2. Technological foundations

Language models are among the basic tools of machine language processing (Natural Language Processing). These are mathematical models that can determine whether or not certain phrases or statements belong to language, or how likely they are to do so. Even the earliest versions of Siri, Alexa, or Google Translate used stochastic language models that learned the probability for each sequence of three, four, or five words to occur in exactly that order in texts or spoken words. Such models, which were not yet realized as neural networks at the time, could improve the correctness or naturalness of generated sentences. However, probabilities have also been used to resolve uncertainties in the acoustic recognition of words in the analysis of spoken inputs. These probabilities were determined by syntactic and semantic factors, particularly grammatical regularities and word meanings.

Early on in language processing, therefore, the idea of explaining the meaning of words through the words in the neighborhood was developed. Unfortunately, the spelling of words gives little insight into their meaning. Therefore, the idea of representing the meaning of each word through a long vector, an embedding, arose a long time ago. However, it turned out that many words have different meanings depending on the context. For example, "bank" may be a seating piece or a financial institution. Five years ago, Google scientists developed within the Transformer, which can capture these differences in meaning with context-sensitive embeddings (Vaswani et al., 2017). Transformers are a variant of the deep neural networks that have revolutionized vast parts of AI since 2012.

What's really new about these transformers is that they can calculate context embedding in a comparable way. This was not possible before; the contexts were previously limited to a few words. Similar to humans, the Transformers recognize the relevant words in broader contexts, including across sentence boundaries, and are able to establish the semantic reference.

Since the number of different words in the language is virtually unlimited by combining simple words with compound words, newer language models use a limited vocabulary of tokens (part words and common words) to compose each word instead of the words. The tokenization, like the AI model, is trained on training data and is thus an integral part of the model.

At the heart of the Transformer is the Self-Attention Block, which, in a very flexible way, determines the meaning relationships between different tokens by correlating the respective embeddings, thus constructing new embeddings. With it, language models with billions of parameters were trained. A prime example is GPT-3 (Brown et al., 2020), which can produce syntactically and content-consistent texts of unprecedented quality.

The derivation of context-sensitive embeddings is best explained by the BERT model (Devlin et al., 2019): Each token of the input text is assigned an embedding vector that represents the semantic meaning of the token and another vector that marks the position of the token in the text. These embedding vectors are parameters and are adjusted in the course of training.

The *Self-Attention* procedure can now modify the embedding of "bank" in the sentence "The bank lends money" by considering the embedding of "money" to emphasize the meaning of "financial institution". To do this, it calculates the "correlation" (scalar product) of the embedding of bank with that of all other tokens (Fig. 2). This is done for complementary "aspects" of the embeddings, which are expressed by parameters (*Attention-Heads*). Finally, the correlation-weighted embeds of all tokens are added to create a new embed for the token "Bank" that reflects the importance of "Money". A self-attention layer with multiple attention heads and a subsequent nonlinear fully connected layer of neurons each form an encoder block, which is a basic component of almost all foundation models.

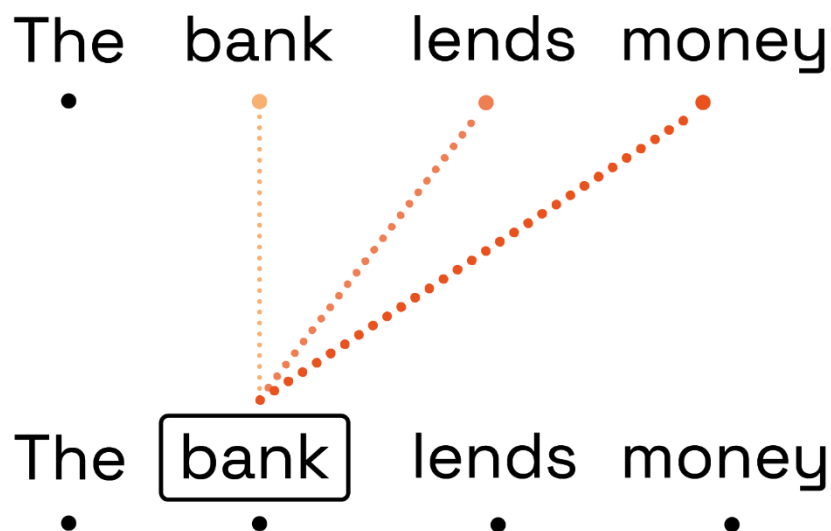


Fig. 2: Clarification of self-attention in the phrase "The bank lends money". The token bank (below) has a high correlation to the token money (above), with the correlation to the other tokens being lower.

The primary goal of the *BERT* model (Devlin et al., 2019) is to determine context-sensitive embedding vectors that contain as much information as possible about the respective tokens of a text. First, the parameters are initialized with random numbers. Further, in the input text, part of the input tokens are replaced by "MASK". For each input token, BERT, as an auto-encoder, calculates the context-sensitive embeds using a series of encoder blocks, which capture the semantic meaning of the token and improve with each layer. The training goal is to predict the highest possible probability for the masked token from the particularly meaningful embeddings in the top layer at the position of a masked token. This is done by optimizing the parameters. Since parts of the observed data need to be predicted and no human annotations are needed, we are talking about self-monitored learning.

The acquired knowledge of language can be evaluated in a second step by training BERT for another classification task. In the simplest case, a context-sensitive embedding is calculated for an additional "CLS" token, from which the probability of the observed class can be predicted using a logistic regression model. An example is sentiment analysis, where the model has to decide whether the input text contains a negative or positive evaluation. This second training task, also called finetuning, requires a human annotated training record. Although finetuning adjusts all parameters of the model, it requires only a small fraction of the training effort for the base model, so the annotated training data for finetuning usually needs to include relatively few examples. For distinction, the first training task with a large general body of text without annotations is also called pre-training. For many semantic classification tasks, BERT can significantly improve the accuracy of the previous model. The transfer of acquired knowledge from one learning problem to another but related problem is called transfer learning.

Language models also use layers of Self-Attention modules. However, they are not trained to predict masked tokens within a text, but are meant to predict the next token for an existing initial text. Here, several layers of encoder blocks are applied to the previously known words of the text. The context-sensitive embedding of the last known word in the top layer then forms the input for a logistic regression model that predicts the probability of the different tokens at the next position. During training, the model is adjusted so that this last embedding contains as much information as possible about the next token and the tokens of the training crowd are given a high probability. Language models such as GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022) are able to continue initial texts syntactically flawlessly and predominantly vocally in content by generating one token after another. The selection of the next token takes place at random according to the estimated probabilities, so that a new text is always created when a token is repeated.

In addition, a language model can also execute instructions. For example, GPT-3 responds to the starting text "Create an outline for an essay about Walt Disney: I: Introduction" with a detailed text about Walt Disney. Thus, without additional training, it is possible for GPT-3 to solve a previously unknown task. Often, the quality of the answers can be improved by additional examples, such as the instruction "English: I do not speak French. French: I do not speak French. English: Where is the restroom? French:". GPT-3 recognizes the input as a translation prompt and provides the French translation. This "k-shot learning" opens up completely new possibilities for using language models without additional finetuning. However, the accuracy is often higher when the language model is trained for the new task by finetuning.

Sequence-to-sequence models (seq2seq) translate one sequence of tokens into another sequence. The most important area of application is the translation of a text into another language. The prototype of this architecture was presented by (Vaswani et al., 2017):

- The **encoder** is a BERT model that computes context-sensitive embeddings of the input tokens.
- The **decoder** works like a language model and is applied to the translation tokens generated so far. Each decoder block contains several self-attentions that evaluate the correlation with the already generated tokens of the translation. On the other hand, the information in the embedding vectors of the input tokens is taken into account through so-called cross-attentions. From the embeds of the top decoder block, the probability of the tokens for the next word position of the translation is calculated.

The training volume contains pairs of input texts and their translations. During training, the parameters of the encoder and decoder are adjusted simultaneously to maximize the probability of correct output tokens.

It is worth noting that the original encoder block with multi-head self-attention is still used by almost all foundation models. Figure 3 shows the similarities in the structure of BERT encoders, language model, and Transformer Encoder decoders.

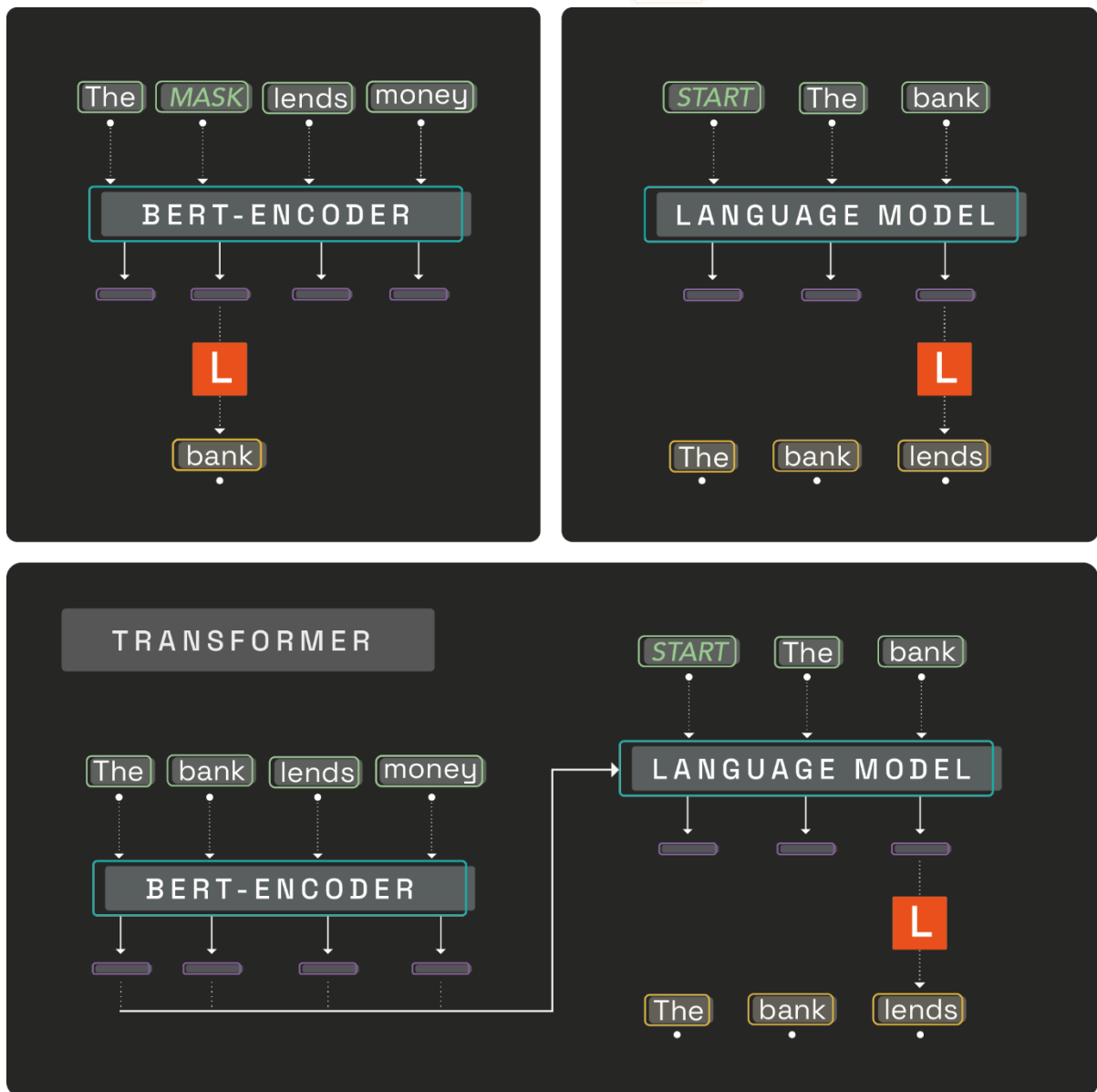


Fig. 3: Layers with self-attention blocks (blue) that compute context-sensitive embedding vectors (violet) of input tokens (green) are central to the foundation models. The logarithistic layer L predicts the probability of the output tokens. During training, the parameters are optimized so that the probabilities of the correct missing or next token (yellow) are as high as possible.

In recent years, a number of improvements to the original architecture have been sought. With Self-Attention, the computational effort and memory requirements increase squarely with the length of the input sequence. Therefore, variants have been developed where the effort only increases linearly. This is a significant improvement that allows much longer texts to be given as an input sequence. This allows language models to refer to text spots later in the text.

Foundation language models can be trained simultaneously on documents in different languages if a common token vocabulary is available. This allows, for example, question-answer systems that can integrate information from different languages into an answer. It turned out that the models can relate the meaning of words in different languages even if each training document is written in only one language (Liu et al., 2020).

Even large foundation models can only store limited information in their parameters. Therefore, different approaches have been used to try to incorporate additional information into foundation models. The main approaches are:

- The additional information will be formulated in linguistic terms. This works for both spreadsheet content (Yin et al., 2020) and knowledge bases. For example, TekGen (Agarwal et al., 2021) verbalizes the complete Wikidata knowledge base using the T5 Seq2seq model. This data set can then be used, for example, as additional training data for a language model.
- The most promising approach is the use of retrieval techniques. Many current language models use a retriever reader module to account for the documents found (Izacard and Grave, 2021).

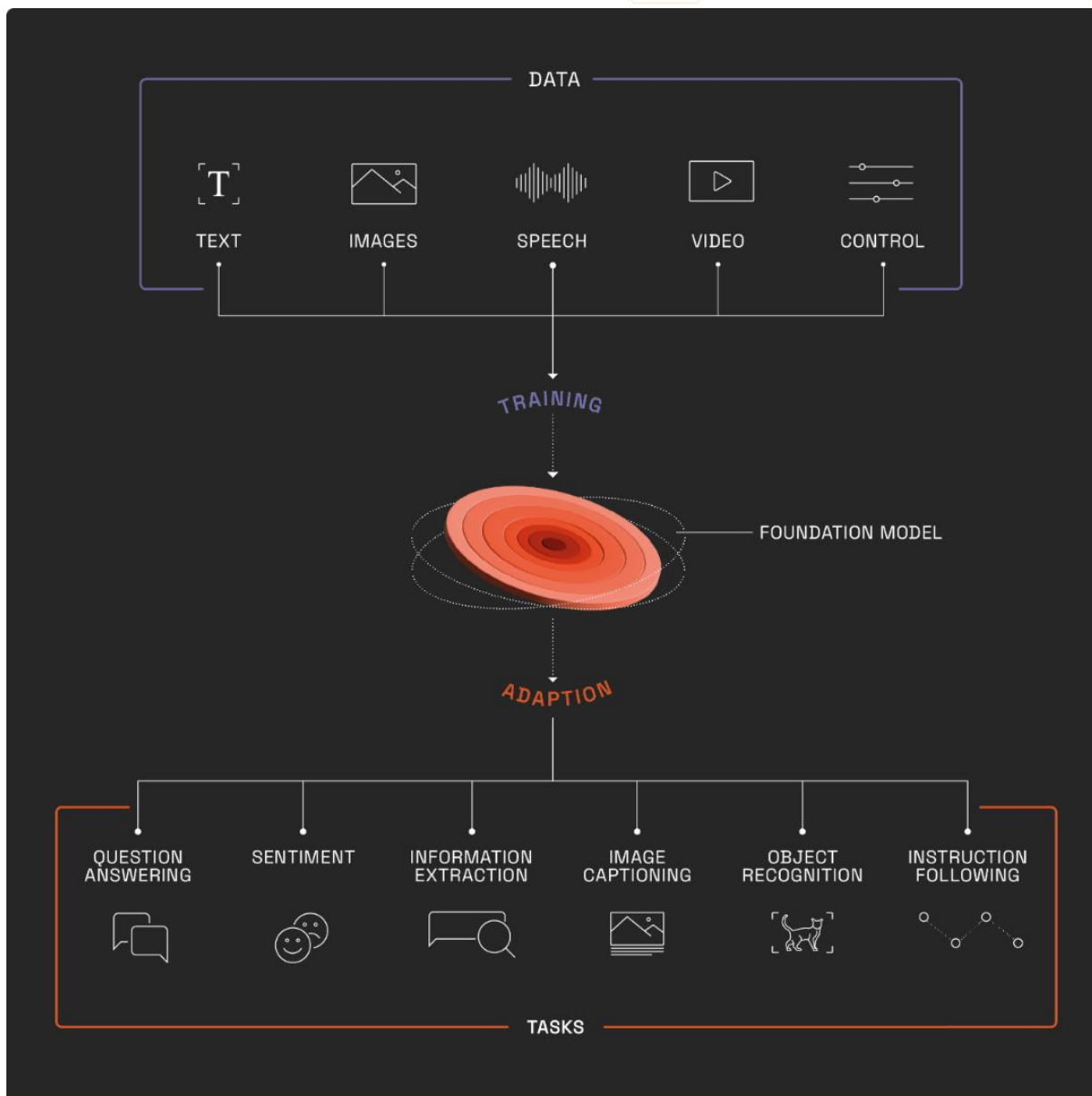


Fig. 4: A foundation model can take into account information from different data sources of different modalities. This one model can then solve a variety of downstream tasks (Bommasani et al., 2021).

It is now possible to represent other media content through tokens as well. For example, an image can be divided into small square pixel areas and thus represented by image tokens. The language models can be applied to these alternative token sequences in virtually unchanged form and are able to match the performance of existing models.

Particularly impressive is the mixture of different modalities by which, for example, a text can be used to create an image. Figure 4 illustrates the pre-training of these models with the same architecture on different media and their application to a wide range of tasks. These models have improved the state of the art for an extremely large number of Artificial Intelligence issues and are therefore referred to as foundation models (Bommasani et al., 2021). They form the basis for a wide range of other AI applications. The monograph by Paaß et al. gives an up-to-date overview of foundation models. (2023).

The salient features of the new paradigm of the foundation models are emergence and homogenization:

- Emergence refers to the fact that a foundation model has skills that are not explicitly constructed, but implicitly learned. An example is the GPT-3 model, which can solve a new task by instructing it through a natural language description of the task, called the prompt. So, the model can solve a problem even though it has never been trained to do so.
- Homogenization is therefore affected by the fact that almost all current models follow the architecture of a few foundation models (e.g. BERT, RoBERTa, T5, GPT-3). This brings about a standardization of modelling for a wide range of applications. Consequently, a model can be used for many different tasks through adaptations. Similar approaches to sequence modeling can be used for text, images, spoken language, video, DNA sequences, etc. to be used. But this also has the potential disadvantage that these AI systems can inherit the same problematic biases or errors of a few basic models.

2.1 The Importance of the Size of AI foundation Models

The size of a model, particularly the number of parameters, has a crucial impact on the performance of the model, its storage requirements, and the computational resources for training. Kaplan et al. (2020) empirically investigated the dependence between the number of model parameters, the scope of training data, and the computational effort for training. They evaluated a large number of models and drew the following conclusions:

- The performance of the models depends largely on these three sizes. Other architectural features such as width or depth have only a weak influence.
- When model size and training data are increased to the same extent, model accuracy grows reliably over a wide range of computational effort. If one of these factors is kept constant, the improvement is less and approaches a barrier.

This also explains the success of large foundation models such as the T5, GPT-3 or PaLM. However, training large models requires an extremely powerful infrastructure.

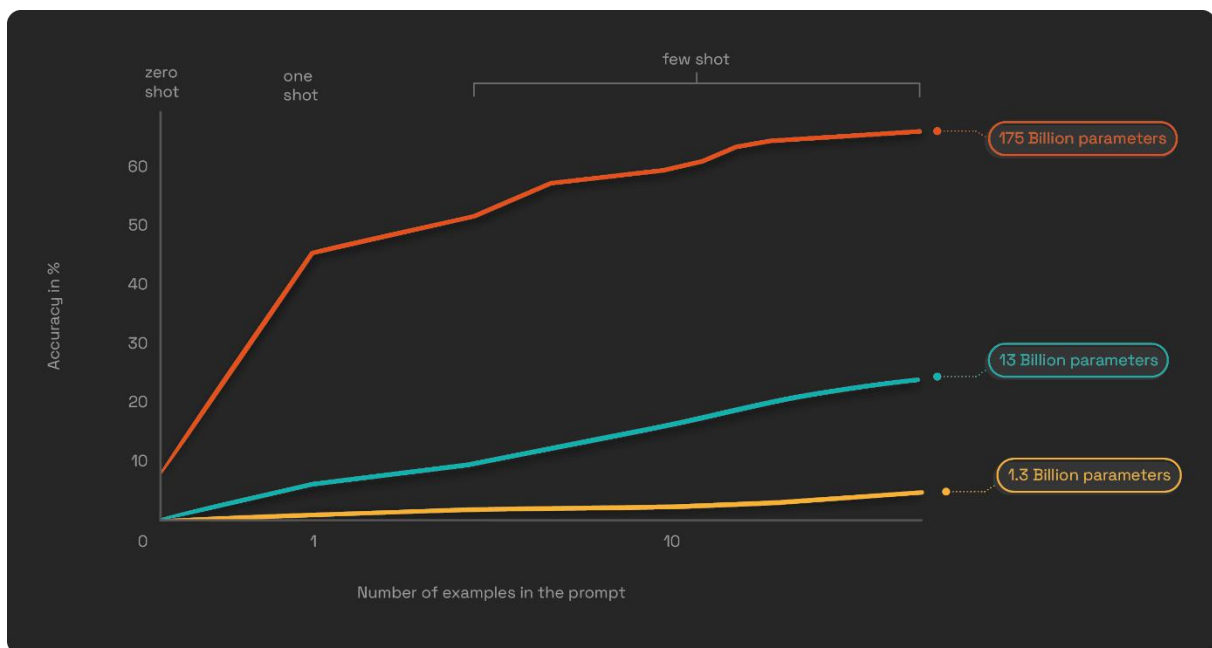


Fig. 5: The accuracy of GPT-3's "few-shot" learning is enhanced by expanding the model size and the number of examples presented (Brown et al., 2020).

Figure 5 shows that important properties of foundation models can only be observed for large models (emergence). While the GPT-3 model, with 13 billion parameters, can respond to natural language instructions with an accuracy of about 20%, with 175 billion parameters, this proportion rises to over 60% (Brown et al., 2020). Apparently, the models need a large intertwining of relationships between terms to respond correctly to natural language prompts. A possible consequence of the emergence is that there are a number of tasks that are unattainable for the current foundation models, but that could soon be successfully tackled.

An easy way to increase the number of parameters without extra training effort is to use a Mixture-of-Experts architecture. It consists of a single gating module and a number of expert modules with identical architecture but different parameters. Each expert module is specialized in only a subset of the data, and the Gating module assigns each input to the few (e.g. 2) expert modules. This assignment is automatically optimized so that the entire model delivers the best performance. A reduction in computing effort can be achieved as only a few expert modules are actually used for an input. The architecture allows for massive models and is particularly efficient for distributed systems where the experts are distributed across different computing units. Examples include Switch, GLaM, and WuDao-2.0.

Since the first versions of successful models are often extremely large, various techniques have been developed to compress and accelerate the models. Knowledge distillation (Hinton et al., 2015) transfers knowledge from a large teacher model to a smaller student model. The advantage of this approach is that the student model can use the internal activations of the teacher model when training. For a number of models, there was a significant reduction in memory and computing effort at nearly identical performance.

Foundation models require a huge amount of data for training. For example, the PaLM model uses 780 billion tokens (25 times more than all the text in Wikipedia) covering a large area of natural language (Chowdhery et al., 2022). It is important that the texts have a high syntactic and content quality, otherwise grammatical and technical errors as well as biases will be reproduced by the model. Low-quality texts can often be found on social media or in user comments on the Internet. They are not suitable for training. Curating data quality is therefore one of the most cumbersome and costly aspects of training foundation models. These aspects are discussed in detail below.

The performance of the first foundation models was checked using a few benchmarks (e.g. GLUE). Meanwhile, AI foundation models are so powerful that they can cover world knowledge from many fields and even draw conclusions by combining facts (Zhang et al., 2022). It has turned out that many benchmarks are saturated after some time and can no longer capture the progress of current models. Therefore, batteries of several hundred benchmarks (e.g., BIG-Bench) are now being used to cover a variety of tasks, including logical thinking, translation, answering questions, math, and others. Meanwhile, with 5-shot prompts, PaLM outperforms the performance of average human processors (Chowdhery, et al., 2022).

2.2 Areas of application of AI foundation models

In this section, we will introduce the many applications of the foundation models.

We begin by discussing applications in the field of language, and then describe applications in the fields of multimedia, life sciences and robotic control. The areas of use listed here are an overview only and are not exhaustive. In many areas, they do not allow us to estimate the potential applications at the moment.

The following sections discuss exemplary relevant application possibilities from very different fields.

2.2.1 Applications in the language field

Application	Description
Information Extraction	Extracts concepts and names from the text and, where appropriate, their relationships
Document Search	Find appropriate texts for a query. Related formulations are also taken into account.
Answering questions	Generates an answer to a question. The basis is the information from the pre-training and the results of a document search. The answer can be explained if necessary.
Machine translation	Translates a text from one language into another. This is possible with the same model for more than 100 languages.
Computer code generation	Generates source code that can be read from a natural language description in very different programming languages
Summary and simplification of documents	Captures the most important statements of one or more documents and writes a simplified version
Generating new texts	Generates a content coherent continuation of a text. In doing so, content requirements may be taken into account.
Text correction	improves and controls texts in terms of spelling, grammar, style, formatting, effectiveness, or terminology
Dialogue systems	Have a longer conversation with a human dialogue partner. In this process, information about the dialogue history is stored and reused.
Opinion and sentiment analysis	Recognizes and classifies opinions and emotional attitudes about products, people, organizations, events, etc.
Discovery of Fake News and Bot Texts	Detects false statements and automatically generated messages

Table 1: A selection of possible applications based on language models

[Language model GPT-3]

The private research company OpenAI achieved a scientific breakthrough with the GPT-3 language model: The model is able to predict the next word in a sentence. A community of developers and users quickly formed around the model, and applications were developed to show that GPT-3 can write compelling essays, create charts and websites from text descriptions, generate computer code, and much more (Tamkin & Ganguli, 2021).

Since the release of GPT-3 in June 2020, more and more competing models, sometimes even offered as open source solutions, have been published. OpenAI offers GPT-3 as an API, allowing users to use the model via the OpenAI platform or build GPT-3 into their own applications. This has resulted in a growing market of tools and applications that today affects many industries and businesses as well as creatives.

The AI foundation model makes it possible to summarize or supplement text in seconds. Chatbots make it easier to control customer interactions. Further, GPT-3 allows the generation of program code by inputting voice commands. Applications of this type are implemented by startups that use the OpenAI API for a fee. Companies developing chatbots and applications for Natural Language Understanding (NLU), especially in healthcare, e-commerce, media and communications, are profiting massively from the further development of large language models.

INFOBOX

Information Extraction

Information extraction is the task of automatically extracting user-relevant structured information from unstructured and/or semi-structured machine-readable documents. These functions are of great importance for the automatic capture of text documents and administrative operations. These include, in particular, the classification of documents according to content criteria, proper name recognition and relationship extraction, whereby the latter extracts relationships between proper names and terms from a text and stores them in a database. foundation models have greatly improved accuracy on these tasks and often surpass human performance.

Document Search

An extremely important application is document search. In this process, both the documents of a text collection and the request are encoded by an embedding and, after calculation, the documents most similar to the request are returned. The advantage is that synonyms and alternative formulations of the same situation are also taken into account. These embedded retrieval methods surpass classic keyword searches and are now used by all Internet search engines.

Frage: when are hops added to the brewing process?

Einer der gefundenen Texte: After mashing , the beer wort is boiled with hops (and other flavourings if used) in a large tank known as a " copper " or brew kettle – though historically the mash vessel was used and is still in some small breweries . The **boiling process** is where chemical reactions take place , including sterilization of the wort to remove unwanted bacteria , releasing of hop flavours , bitterness and aroma compounds through isomerization , stopping of enzymatic processes , precipitation of proteins , and concentration of the wort . Finally , the vapours produced during the boil volatilise off - flavours , including dimethyl sulfide precursors . The boil is conducted so that it is even and intense – a continuous " rolling boil " . The boil on average lasts between 45 and 90 minutes , depending on its intensity , the hop addition schedule , and volume of water the brewer expects to evaporate . At the end of the boil , solid particles in the hopped wort are separated out , usually in a vessel called a " whirlpool " .

Antwort: The boiling process

Von einem Foundation-Modell mit Hilfe von Retrieval gefundene Antwort auf eine Frage im Natural Questions Benchmark. Aktuelle Modelle erreichen eine Genauigkeit (F1) von 80% (Zhang et al. 2021).




Fig. 6: Answer to a question in the Natural Question Benchmark found by a foundation model using Retrieval. Current models achieve an accuracy (F1) of 80% (Zhanag et al., 2021).

Answering questions

When answering questions (QA), a system receives a natural language request and automatically generates an answer in natural language. Advanced systems generally work in two stages (Fig. 6): For a question, an embedded retriever module finds a set of matching documents from a text collection. Documents with similar content that have been expressed differently are also found. A reader then processes the question and the documents found and generates a natural language answer. Retriever reader modules are used by many advanced speech and dialogue models and produce much better responses than very large speech models without this extension. The answer can be explained and substantiated by the documents found.

Machine translation

For machine translation, there are now models, e.g. M2M from Facebook AI (Fan et al., 2022), that allow translation between arbitrary pairs of over 100 languages. Simultaneous training of the encoder decoder with many languages significantly improves translation quality for almost all language pairs, especially for regional languages with few training data.

Computer code generation

The generation of computer code from a textual description is a special translation task (M. Chen et al., 2021), which can now generate relatively reliable running code and has great potential for use.

Document Summary

Automatically summarizing documents can help capture the most important information in documents. Summary models usually use a Seq2seq model, which receives a document as input and outputs the summary. In particular, models with a long input sequence are required. The quality of the summary could be greatly improved with foundation language models.

Referenz-Zusammenfassung: alejandro valverde won ahead of julian alaphilippe and michael albasini. chris froome finished 123rd after a crash during the final 12 kilometres. team sky's sports director gabriel rasch praised froome for finishing. rasch said froome was 'banged up' but expects to ride tour de romandie.

Zusammenfassung durch BRIO: alejandro valverde defended his fleche wallonne title in belgium on wednesday. movistar rider finished ahead of julian alaphilippe and michael albasini. team sky's chris froome fell in the final 12km of the race but finished in 123rd. froome was involved in a crash but finished the race despite being 'banged up'

Zusammenfassung eines Textes von 800 Wörtern durch das Modell BRIO im Vergleich zu der von Experten erstellten Zusammenfassung. (Liu et al., 2022)

Fig. 7: Summary of an 800-word text by the BRIO model compared to the summary produced by experts (Liu et al., 2022).

Text generation

The generation of new texts is the central task of language models. Here, a predefined starting text is continued syntactically without errors and consistently in content. The GPT-3 is the most well-known model here. GPT-3 can be instructed to include a number of points in the generated text through a prompt. These instructions may also contain solution examples, which will help the model to construct the output (k-shot prompts). Because the words of the new text are generated according to their probability, a different text will always be generated on a repeat.

In the meantime, there are also interactive methods for generating text (A. Yuan et al., 2022), in which users can control the layout of the text. PaLM is a multilingual, advanced language model with 540 billion parameters that surpassed the performance of average human processors on more than 150 benchmarks. In addition, PaLM was able to break down complex tasks into simple steps and provide explanations for its conclusions with appropriate guidance (Chowdhery, et al., 2022). Starting texts can, in principle, cause language models to produce offensive statements and fake news. However, after filtering techniques, finetuning and retrieval can largely avoid offensive statements and misinformation.

Dialogue systems

Dialogue systems (chatbots) automatically generate adequate responses to the statements of human interlocutors during a prolonged conversation. They combine techniques for answering questions with the generation of stories and the retrieval of additional information. The status of the discussion and information about the chatbot's "personality" are also stored in a database and reused by Retrieval. An evaluation by human reviewers shows that, for example, the LaMDA system (Thoppilan et al., 2022) is close to human performance in terms of sensitivity, safety and truthfulness. There are many use cases for chatbots in the field of customer contact. The ChatGPT model (OpenAI, 2022b) has a similar architecture but is freely usable on the Internet. It can answer questions, generate code, summarize texts, but also have a coherent dialogue.

[ChatGPT]

ChatGPT is a language model released in an open beta on November 30, 2022, capable of meaningful conversations using text. Trained with large datasets of conversations, it is capable of having realistic conversations, explaining facts, or creating programming code. The most important feature of ChatGPT is that it is able to respond contextually. In this way, further questions can be interpreted and interpreted within a conversation.

ChatGPT has a variety of potential applications, as it is able to generate appropriate responses to a wide range of prompts and conduct conversations in a variety of contexts. Some examples of these applications include the development of chatbots for customer service or information provision, the development of virtual assistants to help users with tasks and answering questions, and the development of interactive tutoring systems. These are just a few of ChatGPT's possible areas of use for developing intelligent and interactive systems that can communicate with humans in a natural and intuitive way.

Since conversational AI models can be used to answer questions directly, a model such as ChatGPT, as well as any future developments of this kind, presents a serious challenge for conventional search providers such as Google. Using conversational AI models, users' questions can be answered directly and further specified, rather than simply referring to further web pages.

Although ChatGPT is still in beta, search engines based on the AI model already exist. This will provide the user with a much more specific search result. It is much easier to find and link content that is directly linked to each other. AI models such as ChatGPT are also conceivable as an internal tool, in the sense of a "single source of truth", as they provide the user with a very authentic and simple search result - similar to a conversation with an employee.

INFOBOX

2.2.2 Applications of multimodal foundation models

AI foundation models are able to model not only tokens of natural language, but also token elements of any other sequences. The following presents applications of the paradigm to different data domains such as spoken language, images, videos, DNA, and proteins, some of which allow for completely new problem solutions.

In particular, linking data from two or more domains opens up additional demanding fields of application. There are models that specialize in just one application, but there are also foundation models such as NÜWA, OFA and GATO that can solve many tasks at once. Because these types of models combine different data domains, they are referred to as multimodal models.

Application	Description
Language to text	Translates spoken language into text for many different speakers
Text to language	Transforms text into spoken language taking into account speech melody and speaker:inner voice
Object detection	Identifies the most important objects in an image and their positions
Detection of actions	Identifies actions in videos and their positions
Image after text	Generates an image description to an image
Text in Bild	Creates a picture that matches an image description
Video by text	Describes the objects and processes in a video through text
Text to Video	Generates a compatible video to an image description
Genomics	DNA sequences are analyzed and the proteins generated from them are predicted
Proteomics	The 2D and 3D structure of the proteins is predicted
Reinforcement learning	Sequences of actions and states for control problems are predicted
Live synchronizations	Translate the language of a video

Table 2: A selection of possible applications based on multimodal models

The application areas of multimedia foundation models are presented in more detail in this chapter.

Spoken language

A field of application is spoken language. The audio signal is often represented by its frequency spectrum (MFCC) for every 10 msec time interval. Wav2vec 2.0 (Baevski et al., 2020) performs unsupervised learning on speech data without transcription with convolution and self-attention layers. Such convolution layers were originally designed for image recognition using convolutional neural networks (CNNs). Similar to the BERT model for text, it learns to predict masked "sound tokens". Transcription from language to text can then be done by a combination of convolution and self-attention layers (Zhang et al., 2020). For example, to transcribe text into spoken language, FastSpeech 2 (Ren et al., 2022) generates the frequency spectrum for small time intervals from the entered phonemes using a Seq2seq model, from which the speech output can be produced directly. Different information such as duration, pitch and energy are taken into account. FastSpeech 2 is judged better by human jurors than competing systems.

Pictures

Images can be broken down into a sequence of pixel ranges that can be used as image tokens. The Vision Transformer (Dosovitskiy et al., 2020) uses 14x14-sized pixel ranges as tokens and performs pre-training with a very large data base of 300 million unannotated images. Masked image tokens are to be predicted. The model is then finetuned to the ImageNet data to classify images into 1000 classes. The Vision Transformer achieved higher accuracy than all CNNs, with much less training at the same time. One of the first models for combining images and their textual description is the CLIP model (Radford et al., 2021). Using separate encoders, it generates an embed each from an image and the associated text. The difference between the two embeddings is then minimized by training. This makes it possible to find the most suitable image for an image caption, and vice versa.

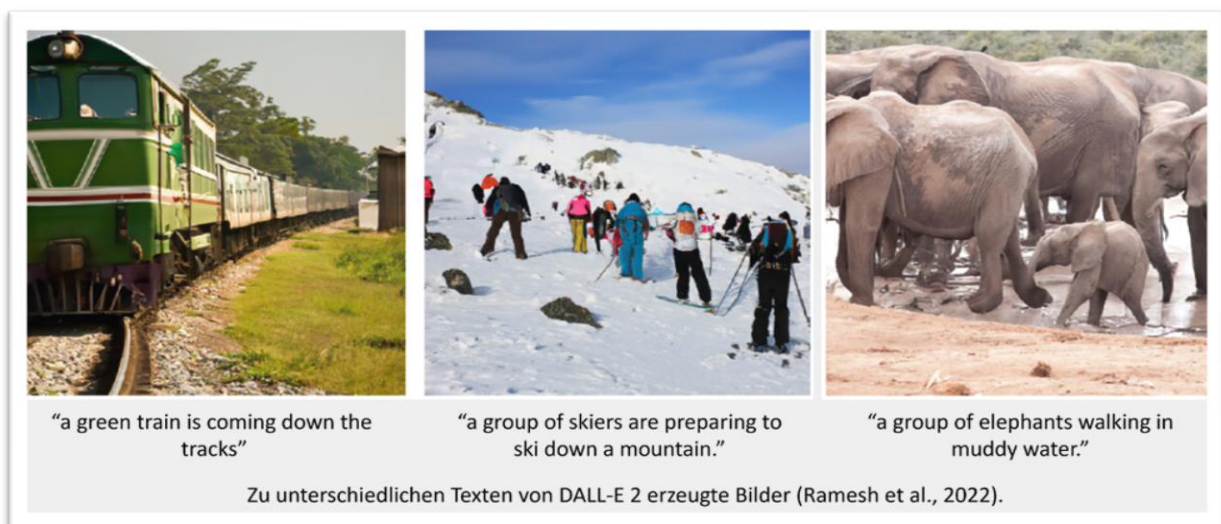


Fig. 8: Images generated on different texts from DALL-E 2 (Ramesh et al., 2022)

Creating images from text often uses CLIP to find matching image embedding to a text. A diffusion model can model the process of degradation of an image through random changes. In this process, the color of individual pixels is randomly changed successively, resulting in a grey area over time. This process can be used in reverse and to reconstruct high-resolution images from the embeds. The most well-known model is DALL-E 2 (Ramesh et al., 2022), which can also modify areas of an image. The OFA model (Wang et al., 2022) can do many tasks simultaneously: determine the position of objects in images, create image captions, answer questions about an image, recognize objects in an image, fill in missing areas in an image, and create images into a text. Many applications can already be used productively today, for example in the editing of photos. Using the Stable Diffusion model (Rombach et al., 2022), it is for example possible to remove unwanted objects from images, to combine several images or to generate completely new images for promotional purposes or as a replacement for so-called "stock photos" based on a textual image description. The images produced in this way are unique. While they reflect the dataset with which the model was trained, they are not copies. Because the images generated are not copies, they can potentially be used commercially. However, the copyright issues have not yet been definitively resolved.

Video

To be able to process videos with foundation models, video tokens are usually used, which describe a pixel range in several consecutive video images. Flamingo (Alayrac et al., 2022) is a visual language model capable of processing sequences of arbitrarily consecutive images, videos, and texts. It uses a large language model for text in the background. On the one hand, the model can answer questions about images or describe images. Secondly, it can describe or classify actions in videos. Finally, it can be instructed to do a new descriptive task by mixing few-shot prompts from text and images/video.



Fig. 9: Videos produced by CogVideo on different texts (Hong et al., 2022)

NÜWA (Wu et al., 2021) is an encoder-decoder model that can generate a video to a text. It uses a special Attention mechanism to capture the relationship of the tokens for both spatial and temporal axes. On the one hand, the model can solve a number of image tasks, such as generating an image to a given text. Further, it can generate a video into a text, predict the continuation of a video, or manipulate videos.

Image-video (Ho et al., 2022) was trained with 60 million image-video pairs and 14 million text-video pairs each from the publicly available LAION 400M dataset (Schuhmann, 2021). The images can also be used to imitate certain art styles and to create, for example, a video in the style of Monet. The Make-a-Video model (Singer et al., 2021) provides the added opportunity for unsupervised learning on video data without textual description to learn realistic movements of objects and scenes. It can also create a dynamic transition between a pair of images in the form of a video. Both models are not only capable of generating videos with high playback fidelity, but the models embrace a high level of world knowledge and are controllable in detail with the ability to generate different videos and text animations in different artistic styles and with 3D object comprehension. Unfortunately, Make-a-Video and Imagen Video are proprietary and the model code is not freely available. The limited length of high-resolution videos of currently five seconds seems to be the limiting factor. Overall, the videos produced are not yet perfect.

Genomics and Proteomics

Decoding the language of DNA is one of the most important objectives of biological research. The genetic code is universal and explains how DNA is translated into proteins. In contrast, the regulatory code that determines when and how genes are expressed varies between different cell types and organisms. This is similar to polysemia and distant semantic relationships in natural language texts. DNABERT has been pre-trained on a large amount of DNA sequences and can improve the state of the art for many specific forecasting tasks through finetuning. These include the analysis of sequence motifs (sections of DNA with biological relevance) and the prognosis of promoter regions (nucleotide sequence that allows for the regulated expression of a gene). MoDNA (An et al., 2022) and GeneBERT (Mo et al., 2021) have similar functionality.

Proteins are linear chains of amino acids that are connected by covalent bonds. Amino acids can be represented by an alphabet of 25 characters. The strings are excellent for many Natural-Language-Processing (NLP) methods (Ofer et al., 2021). AminoBERT is a language model (Chowdhury et al., 2022) that predicts the 3D protein structure from a protein sequence as input. A natural method is also used to describe the polypeptide geometry, which is rotational and translational invariant at the level of the polypeptide as a whole. On average, the model outperforms AlphaFold2 (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021) for orphan proteins and classes of constructed proteins, achieving up to a 10^6 -fold reduction in computing time. There are a number of other models with similar results, e.g., the protein language model ESMFold (Lin et al., 2022). It generates embeddings that can be used in downstream tasks, for example, to capture the structural properties of proteins.

[AlphaFold]

Private research company DeepMind achieved a major breakthrough in protein folding research in 2021 with the AlphaFold software. Using an AI model, the team was able to predict the folding of proteins and thus study their shape and function within the organism more closely. Over 200 million structures of proteins were thus found and stored in a protein database - as opposed to only about 1 million protein structures previously available.

AlphaFold already has a significant, direct impact on human health today. A meeting with researchers from the European Society of Human Genetics highlighted the importance of AlphaFold structures for biologists and clinicians trying to decipher the causes of rare genetic diseases. In addition, AlphaFold speeds up drug discovery by allowing a better understanding of newly identified proteins that are eligible as drug attack points. It helps scientists find potential drugs that bind to these proteins faster.

Ende Infobox

Predicting the interaction between a drug and the target organ is important for active substance design. It is crucial for the discovery of new medicines and the redirection of existing medicines. Yazdani-Jahromi et al. (2021) describe a language model for such applications.

INFOBOX

Dynamic systems control

Reinforcement learning (also known as reinforcement learning) considers a system of states, actions, and rewards at specific points in time. The agent chooses a new action from the given states, while the next states and reward are determined by the environment. The goal is to learn a strategy that assigns an action to each state and maximizes the sum of rewards. Such systems can be used to describe board and video games, but also robot controls and self-driving cars. The decision transformer (L. Chen et al., 2021) is a language model that predicts actions one after another. In doing so, he estimates the optimal sum of all future rewards. The model is trained on a large amount of observed time series. The agent can then select the action that will result in the highest predicted reward amount at a given state. GATO (Reed et al., 2022) is a multimodal model that can process text, images and sequences of values and derive control strategies from them. It achieved good results on more than 600 benchmarks.

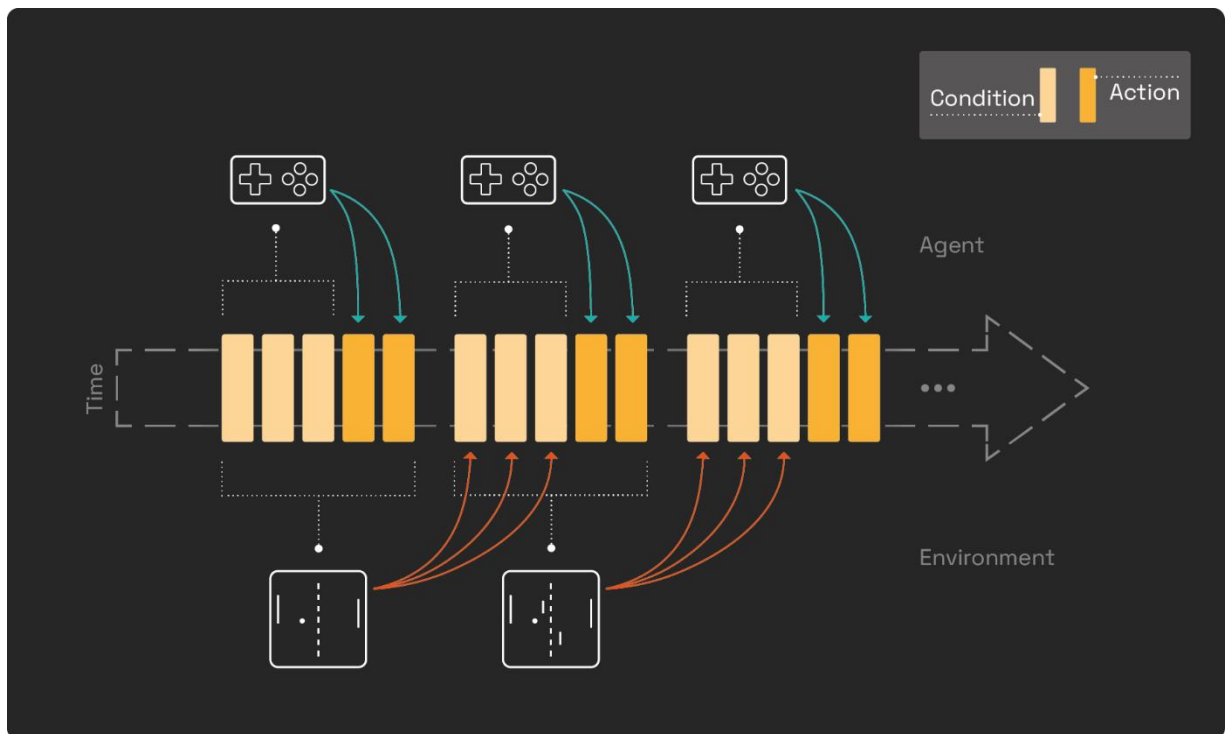


Fig. 10: The Gato model generates new actions (dark yellow) from the current states (light yellow). The environment produces new states, etc. The model can process text, readings, images, etc. (Reed et al. 2022).

2.3 Trusted AI foundation models

The large-scale deployment of AI systems and the associated technological innovation potential are expected to have impacts of broad economic, but also societal, relevance. Within five days, over one million users signed up for ChatGPT to try out the system themselves. This high level of interest from the general public demonstrates how quickly and strongly large AI systems can shape public discussion. The black-box-like nature of the used trained models entails risks that require special measures with regard to the trustworthy use of AI. The wide range of applications of AI foundation models requires a systematic approach to assessing or mitigating the risks that arise. They need to be addressed in parallel with model construction in the LEAM project.

The following presentation is based on the risk-based approach to trusted AI developed in the context of the KI.NRW flagship project "Certified AI " (Poretschkin, 2022) and previous work (Cremers et al., 2019; Poretschkin et al., 2021). This approach aims to systematically capture, make measurable, and sufficiently reduce AI-specific risks, taking into account various dimensions of trustworthiness.

Below, we present six dimensions of trustworthiness and outline current actions to identify and mitigate them.

Fairness

The Fairness Dimension is designed to ensure that the application of AI does not lead to unjustified discrimination. Typical causes of this are unbalanced (biased) training data or statistical underrepresentation of groups of people, which can lead to a reduced quality of AI application in relation to these groups. Because the foundation models train with texts written by humans, these texts often reflect the stereotypes that exist in society. In particular, gender or ethnic groups may be favoured or disadvantaged.

The developers of large AI foundation models are aware of many of the inherent risks and are implementing corrective measures designed to avoid unwanted effects. For example, the developers of DALL-E (OpenAI, 2022a) report that special techniques are used to counter a bias that is actually present in the training data used (e.g. "heroic firefighter" is usually depicted as a male person). Large language models such as PaLM and LaMDA use filter-based techniques with good success to avoid unwanted output ("toxic language"). However, these methods are very specific in nature and are likely to be insufficient in widespread use.

Autonomy and control

This dimension targets two things: firstly, the autonomy of the AI application, and secondly, the autonomy of the human being. On the one hand, it is important to assess what degree of autonomy (Gehman et al., 2020) is appropriate for the application. On the other hand, it will examine whether the human being is adequately supported by the AI application and given sufficient leeway in interacting with the AI application. The specific challenge is that the possibilities of interaction with humans are often defined in the design of the concrete downstream AI application and cannot be controlled by the foundation model per se, which plays a central role in the process. Substantial research is still needed to cover the cross-relationships between the applications and the foundation models in a controlled way.

Similar to interpersonal communication, "toxic" model spending (e.g. insults or bullying) can lead to psychological-emotional impairment of users (Gehman et al., 2020). Beyond hurtful language, (manipulative) communication strategies or the forecasting of false facts can create emotional dependencies, potentially limiting human autonomy. An early approach to addressing this risk is suggested by Glaese et al. (2022). They optimize their conversational AI in such a way that they openly communicate their machine nature whenever necessary. Here, too, we are only at the very beginning, and German AI research could make significant contributions.

Transparency

Aspects of traceability, reproducibility and explainability are subsumed under this overarching term. In particular, the transparency dimension examines whether the basic functioning of the AI application is reasonably traceable to users and experts, and whether the results of the AI application can be reproduced and, if necessary, justified. While the transparency dimension is already addressed in foundation models at the level of documentation and description of the data/models (e.g., through Modelcards (Gehman et al., 2020)), a systematic approach is still needed in relation to the actual use of the AI foundation models in concrete applications.

Like other social media, a chatbot can be made to communicate to the user only in certain aspects through fine-tuning or prompts. They then find themselves in a "filter bubble" in which messages that do not match the opinion expressed are hidden. There are now audit procedures for this issue (Cen & Shah, 2021) that can be used to verify whether the platform uses unwanted content filters, requiring only black-box access to the filter algorithm. In all these areas, we have little control so far, specifically with foundation models, and further research and then probably regulatory efforts are needed.

Reliability

This dimension relates primarily to the quality of the AI component and assesses, among other things, its robustness, i.e. the consistency of its expenditure. An aggravating circumstance is that the production of text is mostly randomly controlled. Factual accuracy of expenditures is also an important aspect, as foundation models cannot distinguish between plausible and correct relationships (Lin et al., 2021). For example, the truth content of "Angela Merkel is Chancellor" has changed over time. In the area of foundation models, various approaches (e.g. safety fine-tuning) are already being explored here to avoid unwanted ("unsafe") spending as much as possible.

Increasing the "factfulness" of language models is currently a highly edited research area in the field of NLU (Natural Language Understanding, see e.g. Glaese et al. (2022), Nakano et al. (2021)). Retrieval-based foundation models such as WebGPT, Retro, and LaMDA can access a large collection of current text documents to enhance the text to be created through relevant retrieved information. Shuster et al. (2021) have shown that the use of retrieval reduces the rate of "hallucinations". Overall, retrieval allows for the targeted use of current knowledge to improve response quality.

Another problem is (deliberate) misappropriation, for example, for the generation of "fake news" or for the production of violent glorifying texts, for which Brundage et al. (2022) to propose a set of measures. More generally, it is difficult to assess the capabilities and thus the potential for abuse of large models, since their "abilities" (skills) are often highly sensitive to the details of the particular model input (prompting) - e.g. Chain-of-Thought Prompting (Suzgun et al., 2022) and Prompt Injection Attacks (Branch et al., 2022) - and cannot be predicted ex-ante. The HELM initiative (Liang et al., 2022) approaches this problem by evaluating and comparing a variety of language models in a standardized manner, while also considering aspects such as calibration, robustness and fairness in addition to performance.

An important aspect in this context is "Trusted-AI": The ways to be able to give assurances about the correctness of the results of the foundation models, as they are important, especially in security-critical contexts. There are two main approaches: Deductive verification, which leads to formal evidence of the correctness of certain aspects based on fundamental theories about AI algorithms, and inductive validation, which derives statements about the properties of AI systems through systematic testing. Both areas are already being used with some success in AI, but much more needs to be done, especially for the very complex models considered here.

Security

This dimension addresses both features of functional security and the protection against attacks and manipulation of the AI application. Since the measures in this dimension primarily relate to the embedding of the AI component, foundation models can be protected here, among other things, by classical methods of IT security. However, there are also open questions arising, in particular, from the objective conflict of increasingly using publicly available training data, while protecting it from attacks that could affect model behaviour.

Dialog systems such as BlenderBot 3 use fine-tuning on "correct" dialogs to improve, for example, the system's ability to respond appropriately to critical remarks (Ung et al., 2022). Nevertheless, many of the well-known challenges of trusted AI for foundation models require better solutions. As an example, only the problem of "visual synonyms" (Mishkin et al., 2022) that can be used to circumvent language-based filtering techniques (e.g., "red liquid" instead of "blood") is mentioned here.

Data protection

This dimension refers to the protection of sensitive data in the context of the development and operation of an AI application. Both the protection of personal data and the protection of trade secrets are addressed. In connection with this, particular attention should be paid to the risks that arise in the context of "model inversion" techniques. For example, sensitive data such as social security numbers or even realistic images (of previously unknown persons) could already be generated through targeted and systematic queries to models.

In addition, so-called training data extractions are relevant to the language models, which are often used generatively, in which the generated model output contains fragments of the training data. Unless the latter has been completely cleaned of personally identifiable or copyrighted information, such "extractions" may violate third-party privacy or property rights (Carlini et al., 2021). Nichol (2022) proposes an ad-hoc mitigation strategy for this.

There are several ways to mitigate privacy concerns in foundation models. One approach would be to filter out sequences from the generated data that are already present in the training data, such as through a Bloom filter. Another approach is training with differential privacy approaches. There is a great need for research here.

Research directions and approaches

Further intensive research and development is needed to systematically cover the requirements from all six dimensions of trusted AI for foundation models described. Existing approaches to addressing specific risks of foundation models need to be expanded and systematically merged. In addition to the complexity and sheer size of the foundation models, the variety of possible applications also poses a particular challenge. Without these approaches, it is usually not possible to use the foundation models developed responsibly in practice.

In addition to the technical measures taken during the development and testing phase, an ongoing process and governance rules are needed to accompany the deployment of a foundation model. These are not present in the previous models from the United States and China or are little controllable. During model operation, monitoring and ensuring that the principles of trusted AI are maintained must continue. In principle, appropriate organisational measures must be taken to respond to situations where, for example, a potential misconduct of a model occurs. This should also take into account the interaction between potentially different organizations involved, such as the developer of the foundation model and the provider of an application based on it. In particular, care should be taken to ensure that errors found to be corrected can also be triggered by new technical measures, such as model improvements or the introduction of additional filters.

Possible and necessary steps include:

- Deeper understanding of the structures and functioning of foundation models
- Procedures for risk assessment and testing along specific proxy tasks
- Establishing benchmarks for trustworthiness of foundation models
- Investigation and testing of semantic properties of latent space
- Validation of models for input or output monitoring
- Identification of appropriate organisational measures to monitor the day-to-day operation

SPOTLIGHT

2txt NLG GmbH

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Johannes Bubenzer,
Founder and CEO of 2txt

2txt is a provider of automatic text generation software based on AI language models. The software is optimized for use cases in the travel sector, the financial industry and e-commerce. 2txt is characterized by particularly simple and fast setup, easy integration into enterprise applications and, most importantly, consistently reliable and very high text quality.

Where do you use foundation models? What is your use case?

We use foundation models with linguistic models to generate descriptive texts for various topics in B2B.

For example, we generate product descriptions, location descriptions or financial reports completely automated. We combine foundation models with classical language models to get the best of both worlds: foundation models make our product scalable and creative, while linguistic models make our products controllable and reliable.

What impact do AI foundation models have on your business model or your projects?

foundation models allow us to scale our product in the shortest time possible. Work steps that used to take months can be done in seconds using foundation models. This saves resources, costs and time and allows us to address a much larger market.

What difficulties and problems do you see that only the U.S. and China are currently providing AI foundation models on a comprehensive basis?

Nowadays: The providers of AI models operate these as digital services. This means that they have access to all incoming data and outgoing texts. This is a nightmare for the privacy and trade secrets of the companies that use it.

For the near future: Developments in the field of AI foundation models are expected to trigger one of humanity's central technological revolutions. Intelligent machines will emerge, and it would be an immeasurable social and economic mistake to leave the race for these technologies to private companies in the US or China.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

Companies can only build sustainable AI business models if they can build and defend a technological edge. That is why it is essential that we be able to train downstream models on free and open foundation models with our own data. Otherwise, any technological advancement in this field would be too easily replicable.

In addition, some of the most interesting applications for AI foundation models are those involving personal data, for example, when it comes to customer communications or business intelligence. Such services are completely unthinkable when the data must be uncontrollably transferred to the US or China.

2.4 Open research questions, latest developments and expectations

The development of foundation models is still in its infancy and there are many open questions and potential risks. At the same time, the opportunities are very promising, especially for the business sector, and foundation models are already shaping the business models and underlying AI solutions of companies that are on the rise.

The following developments are expected:

Extension to additional domains

The strong development and use of foundation models has begun in the NLP field and is expected to spread to all other AI fields, covering more data types: Images, sound, videos, genome and protein sequences, sensor data, and so on.

Multimodality and the limit of text data

Multimodality is an important factor. It is already foreseeable that multimodal foundation models will achieve higher accuracy than unimodal models and can cover an even greater variety of use cases.

In addition, it can be noted that the exclusive use of text data to train foundation models is already reaching its limits, as very large portions of the Internet's usable digital texts are already being used. In this respect, multimodal data collections could be a solution, e.g. the huge amount of videos on YouTube.

Barrier to training foundation models

The existing gap between a manageable number of technology companies with the resources needed to train foundation models and all those who lack the financial means and corresponding computing resources could continue to grow in the future. Initiatives such as LEAM are counteracting this trend. Further influences are being made by aspirations to advance the democratization of AI through distributed learning, or by open source collectives such as Hugging Face, who developed and made public the BLOOM model with the Big Science Initiative.

No monopoly formation, but delayed development

However, despite the limited access to computing resources, a monopoly to develop foundation models is rather unlikely at this time. In the case of the GPT-3, comparable models have been released: A21 Labs' Jurassic-1 models, Meta's OPT, Eleuther AI's GPT-Neo and GPT-J models, to non-English models such as Sber's Russian ruGPT-3, Naver's Korean HyperCLOVA, Tsinghua University's Chinese CPM-1/CPM-2 models, Huawei's PanGu- α and Beijing Academy of Artificial Intelligence's Wu Dao 2.0. Nevertheless, it should be noted that the development of science is as much as two years behind the state of the American economy. Such a circumstance is highly unusual.

Close collaboration of development and application

The line between AI developers and users is likely to become blurred in the coming years. The reason for this is that more and more people without AI expertise will be able to successfully customize foundation models for their own cases. At the same time, an exponential increase in new AI-based products is expected.

Grounding

Foundation language models learn correlations between terms and linguistic concepts on their own. For example, a dog is associated with the terms leash, ears, cat, mammal, leg, fur, tail, toy, baiting, etc. What is missing are aspects such as the dog's three-dimensional shape, the way it moves, the sound of its barking, its dynamic response to cats or humans. With this, laws of physics, such as the permanence and deformability of objects, the effect of gravity, are linked. Therefore, the concept of the dog is best learned when it appears in multiple media, for example, as an image, in words, or in a movie where he is chasing a cat. The use of multimodal foundation models provides the opportunity for such integrated learning of concepts in the world. Yann LeCun says: "However, rather than using language or images, the next generation of AI will learn directly from videos. Meta is currently making great efforts to collect video data from the I perspective for this new generation of AI, but YouTube videos are also suitable as training material" (Schreiner, 2022; Jawahar, 2021). The recently proposed foundation model PLATO is an initial attempt to learn intuitive physics from videos (Piloto et al., 2022).

Planning and simulation

Daniel Kahneman has developed a hypothesis about two different thinking systems based on long-term studies of human behavior (Kahneman, 2011). *Fast Thinking* is fast, instinctive and emotional. Examples include understanding a simple spoken sentence or recognizing an object in a picture. *Fast Thinking* is constantly running, generating impressions, intuitions and quick judgments based on our immediate perceptions. Fast Thinking can be done for the most part today with the existing foundation models.

Slow Thinking is slower, more thoughtful, and more logical. For example, it is responsible for parking in a narrow parking space or solving the 16×34 calculation problem. *Slow Thinking* is only used when problems with *Fast Thinking* occur, that is, when it cannot explain the perceptions well.

Slow Thinking is able to mentally play through problems and simulate results for different edge conditions. It is therefore largely in line with the concept of model predictive control. These capabilities can be implemented by using foundation models such as GATO (Reed et al., 2022). It is necessary to integrate these models with the perceptual models for

different media and to be flexible in their application to new planning problems. According to Yann LeCun, "the ability to construct models of the world is basically the essence of intelligence." These models are needed not only to predict physical movements, but also the behavior of people, economic activities, etc. The great challenge of artificial intelligence over the next decade is to learn predictive models of the world to deal with uncertainties (Fridman, 2022).



AI foundation models in international comparison

3. AI foundation models in international comparison

This chapter illuminates the development of AI foundation models in international comparison. It shows that the US and China are currently leading the way, and explains the reasons why Europe has so far failed to lead the way in the development of large AI models.

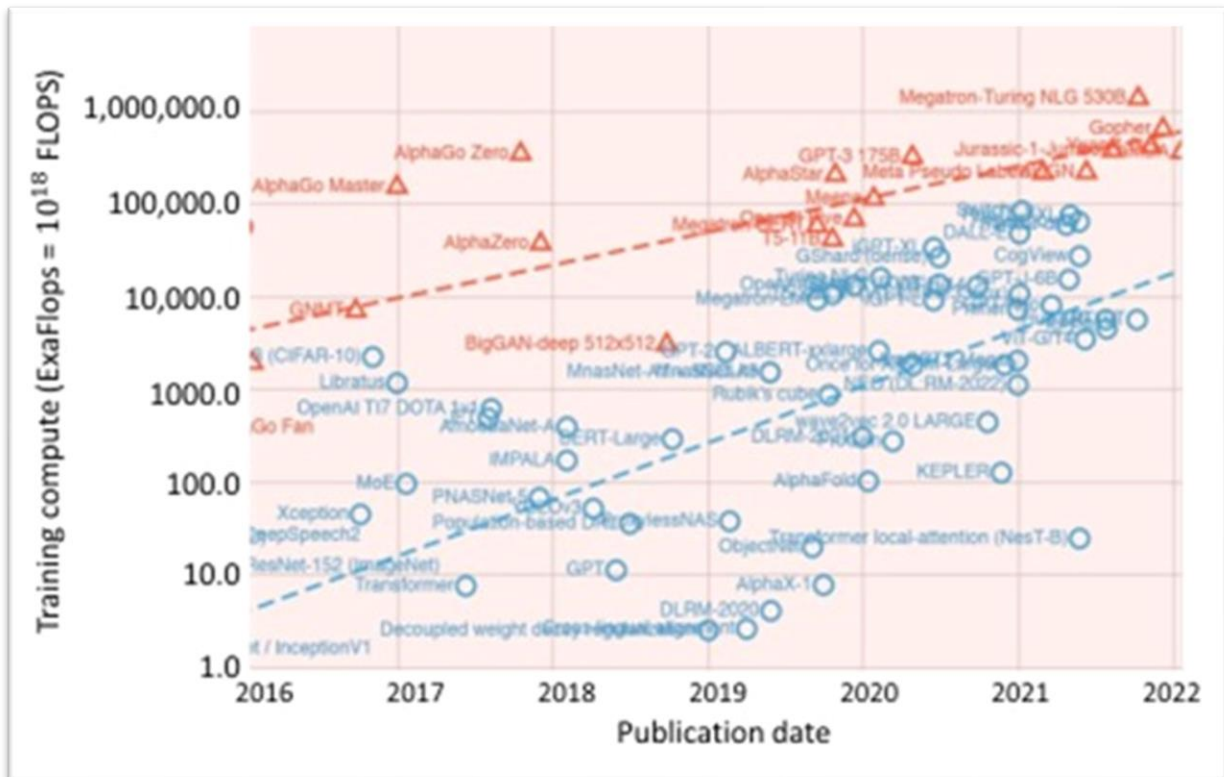


Fig. 11: Around 2016, a new trend to very large models trained by major Internet companies emerged (red). These were able to finance the necessary investments. The bottom blue line illustrates the computational effort of the other models, e.g. of universities (Sevilla et al., 2022).

In international comparison, the following trends stand out:²

- Models developed by the academic community or AI community for public use in text and images processing lag, respectively, about 14 and 15 months behind initial development from the economic sector (Benaic & Hogarth, 2022).
- Most of the published foundation models are from the U.S. and China: As of 2017, 73% are from the U.S., 15% are from China.
- Most of the models developed come from industry, not from the scientific community: Since 2017, 86% of the models have been developed by business sector, and 13% by the academic sector.

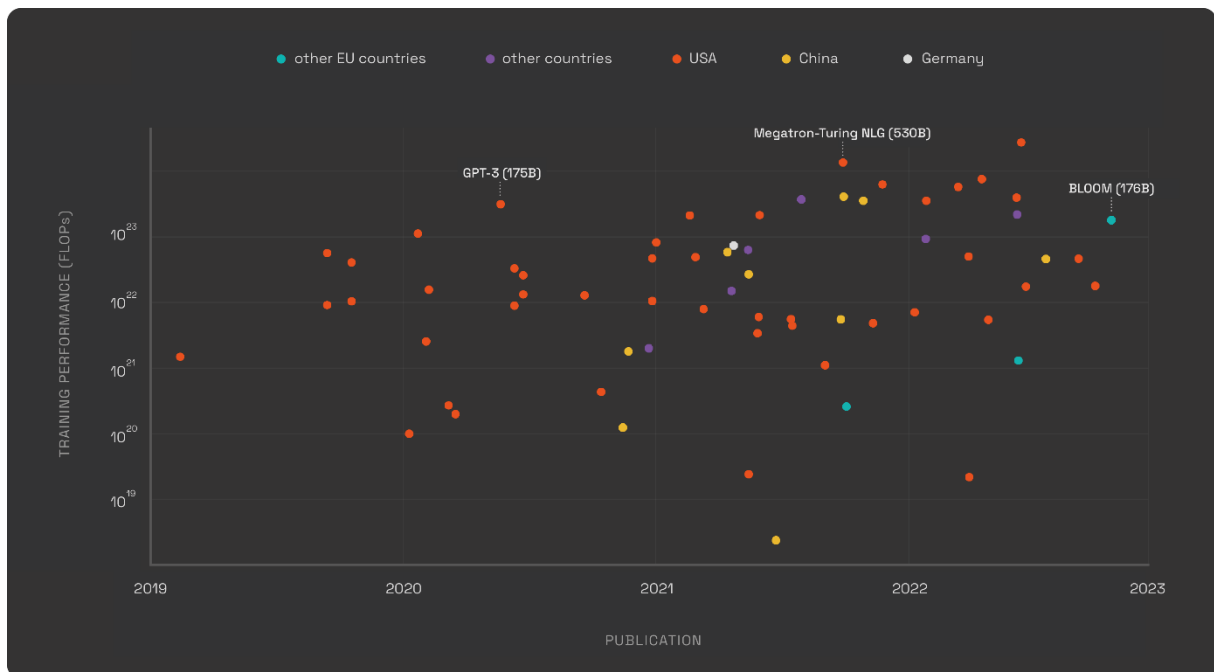


Fig. 12: Training Performance (1 ExaFLOPs = 10^{18} FLOPs)^{2F3} of different foundation models versus Release Year by Country

² For the international comparison, 125 foundation models from the initial development of the Transformer in 2017 were considered. For this purpose, the data set of Sevilla et al. was used. (2022) filtered by foundation models and expanded to include recent releases.

³ Floating-point operations, plural of FLOP (floating-point operation) - here a unit of measure for comparing training effort measured in the smallest operations, not to be confused with FLOPS (floating-point operations per second).



Fig. 13: Training performance (1 ExaFLOPs = 10^{18} FLOPs) of different foundation models over the year of publication by four categories of organizations

The overview graphs (Fig. 12 and 13) clearly show that the technology field is dominated by two countries, the US and China, and that Europe, including Germany, is greatly underrepresented in relation to its size and role in global society. 73% of the models are from the US and 15% are from China.

From Germany and other EU countries, smaller foundation models with a lower training effort have tended to be released. The BLOOM model is an exception and was trained by a academic collective of over 250 institutions on the Jean Zay supercomputer in France. As the only model from Germany, the Luminous model from the German AI company Aleph Alpha appears in Figures 12 and 13.

[Aleph Alpha]

Heidelberg-based Aleph Alpha is an independent German AI research company founded by Jonas Andrulis and Samuel Weinbach. Jonas is a successful AI serial entrepreneur and a former senior manager from Apple's AI research division for secret innovation projects in California. Samuel has 10 years of experience for AI innovation and is currently one of the leading minds for engineering large language models.

Aleph Alpha has developed a GPT3 equivalent based on European data and in five languages, which offers twice the number of parameters compared to OpenAI's best offer at the largest level of deployment. In addition, the team around Jonas and Samuel developed a multimodal extension that not only understands text, but also images in context. By the beginning of 2023, Aleph Alpha, according to his own statements, will have the world's only multimodal offer for large language models. These and other innovations are shared with the community in numerous academic and open-source publications. There is close cooperation with many of the leading scientific researchers.

Aleph Alpha managed to build a team of 50 international experts from the best companies from the US and the rest of the world. In two funding rounds, the team was able to secure a record €28 million of independent financing with the support of some of the top investors from Germany and Europe.

To operate in a sovereign manner even for security-critical application scenarios, Aleph Alpha has built a specialized data center and currently operates the most powerful commercial AI data center in Europe. Aleph Alpha's technology is currently being used in companies of all sizes and industries in the financial sector, health, law, administration and security.

Due to its reputable LEAM infrastructure, Aleph Alpha could also provide support for modern foundation models with high demands in research and open source projects. Even for Aleph Alpha's SME customers without their own data centers, these opportunities are crucial ingredients for entering a new age of technology.

INFOBOX

All the major leaps in the development of the foundation models come from the US and, to a lesser extent, from China. This concerns both the types of models and new functionalities, as well as leaps in model size and the development of new data domains.

In the US, the major breakthroughs have been achieved in a few industrial laboratories or in laboratories that, through private sector investment, have gigantic infrastructures and generous human resources for top researchers. These include the labs of Google, Meta, Microsoft, Amazon and OpenAI, and to a lesser extent AllenAI.

The first major breakthrough was made in China with the Wu Dao 2 model at the Beijing Academy of AI. This laboratory brings together experts from several top universities and works closely with industrial partners (including Xiaomi, Meituan and Kuaishou). A large part of the investment comes from the Chinese state. Other large models are coming from the research labs of Alibaba, Baidu, Tencent, Huawei, Inspur and other Chinese high-tech groups.

From the point of view of the power of innovation, we can roughly distinguish three classes of models. These are:

- i. The best-known models, each representing groundbreaking breakthroughs,
- ii. models that enhance or extend existing model types in size, data, or their functionalities; and
- iii. Models that more or less accurately replicate existing models and are only technologically insignificantly different from their prototypes. This includes scaled-down or simplified imitations.

While models of the classes (i) and (ii) contribute significantly to the evolution of technology, providing models of the class (iii) for the availability of the models in organizations and business models where use was not possible for economic, technical or regulatory reasons.

The development of models of classes (ii) and (iii) has an important side effect: It builds or reinforces technological competence for future technological developments

All models in the class (i) and most models in the class (ii) were created in the United States and China. Most models from other countries copy the GPT architecture, sometimes with minor modifications, and expand the learning data to better cover certain languages (sometimes even applications). These include models from Israel, Russia, Sweden, France and Germany. There are currently a few class (ii) models available, including the GPT-SW3 model by Sweden AI, the Luminous model by German company Aleph Alpha, and the BLOOM model created by the French OpenScience initiative in conjunction with Hugging Face. However, all of these models were not completed until 15-24 months after the release of the GPT-3.

In order to develop foundation language models, three essential prerequisites must be met: Competence, infrastructure and data.

Availability of intellectual competence: Are there enough experts who bring the necessary knowledge and experience for model development, and are these people available?

Availability of infrastructure: Is the computing infrastructure sufficient for model development, i.e. minimal for data processing, training and evaluation?

Availability of data: Are there sufficient amounts of data in one or more languages to generate the emerging capabilities? This is not only a matter of the amount of data, because high diversity and quality can compensate for a lack of mass to some extent.

The evaluation of the expert interviews and surveys (see chapters 4 and 5) showed that both Germany and Europe face significant challenges in all three conditions. The study will focus on the available infrastructure. Nevertheless, it must be stressed that all three conditions must be met for successful European AI foundation models.

The availability of adequate infrastructure has now become a bottleneck for successful participation in technology development and commercialisation. Infrastructures, the new models of the class (i) allow, are far greater than the minimum requirements for the development of models of the class (ii). Researching at the forefront of technology development requires creative experimentation with many candidates for new architectures and learning methods and their many variants. This makes it difficult for universities to participate in this research. 86% of the models were developed from the economic sector and only 13% from the scientific sector. Models developed by the academic or AI community for public use in text and images are about 14 and 15 months, respectively, after initial development in industrial laboratories.

Even the major U.S. universities can no longer compete without cooperation with industry research laboratories. The U.S. government has therefore established the National Artificial Intelligence Research Resource Task Force (NAIRRTF) to significantly improve infrastructures for AI research (*THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE TASK FORCE (NAIRRTF)*, without date).

Attempts to use the quite considerable high-performance computers of scientific high-performance data centers for training foundation models have had limited success. It is possible to train new GPT models with good planning, even with additional data and small modifications. However, there is a lack of continuous availability of large enough computer resources for the systematic optimization and development of models and for the experimentation with new model types. Academia's high-performance computers cannot be used for the development and evaluation of new application classes either.

There is less to worry about developing class (iii) models. The continuous increase in computing power in industry and academia, and the constant improvement of learning processes, will make it easier to copy model types.

While the computational capacity requirements for model training are likely to decrease, even with extensions to the learning data, the computational power required to develop next-generation models is still increasing sharply. The reason for this is the upcoming merger of modalities for acquiring world knowledge and additional functionalities that require the use of large volumes of film data and real-world analog data. Chapter 8 sheds more light on the subject of infrastructure.

[OpenGPT-X]

OpenGPT-X is a collaborative project with partners from academia, business and technology. The aim of the project is to create large AI language models to drive innovative language applications for Europe and according to European values. Using Gaia-X, the secure data infrastructure to drive innovation in Europe, OpenGPT-X will openly provide AI language models and language services across Europe in multiple languages.

For example, AI language models for answering questions in interactive media formats are being developed in the media sector as application examples for future product developments. In the finance domain, the models are used for more efficient document processing in insurance claims, and in mobility, they are intended to contribute to greater safety and comfort as personal assistants during driving.

The project partners are currently training a first model with twelve billion parameters. The focus is on getting answers from the language model that are as content-correct as possible.

The project will be funded by the Federal Ministry of Economy and Climate Protection from January 2022 to December 2024 under the Innovative and Practical Applications and Data Spaces in the Gaia-X digital ecosystem grant programme. The participants include the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS), the Fraunhofer Institute for Intelligent Circuits (IS), 1&1 IONOS SE, the Jülich Research Centre, the Technical University of Dresden, Alexander Thamm GmbH, the German Research Centre for Artificial Intelligence (DFKI), Aleph Alpha, Control Expert, West German Broadcasting (WDR) and the KI Bundesverband.

There is a close exchange between OpenGPT-X and LEAM. The results and experience of the project have been used in the development of the feasibility study and will continue to be used in the planning of LEAM.

INFOBOX

Importance for Germany's technological sovereignty

The international comparison makes it clear: German companies are heavily dependent on the use of American and Chinese foundation models due to the lack of powerful European alternatives. This raises a number of challenges, as the existing models do not meet European quality standards in many respects.

These shortcomings are summarised in five points:

- (1) **Content functionality:** German and European content is underrepresented. This applies in particular to economic, socio-political, legal, regional and cultural issues. The large models are also geared towards the mass market rather than the requirements of applications in business, politics, education.
- (2) **Language functionality:** The existing models are much more oriented towards English than other languages because of the learning data used, but also because of the evaluation data and benchmarks, and in the case of the Chinese models, they are also much more oriented towards Chinese. While languages that are heavily represented on the Internet, such as German, French and Spanish, are also well represented, they are not nearly as effectively covered in measurable functionality. Most European languages are very underrepresented. An important finding of linguistic research is that language influences thinking and even changes fundamental aspects of human experience: Space, time, causation, and the relationship to others (Boroditsky, 2012). Therefore, in order to adequately capture German culture, foundation models for the German language are required.
- (3) **Availability:** The large international models are limited for commercial applications due to their ownership and licensing patterns. This applies to adaptation through modified (extended, corrected, filtered) learning data, additional pre-training through other training tasks, inference (i.e. practical use), few-shot prompting, large-scale fine-tuning, and integration into larger applications.
- (4) **Safety, reliability:** For many European applications, additional measures would be required to ensure better performance in terms of security and reliability, particularly in terms of accuracy, consistency and data protection. However, security requirements also include the confidentiality of input data in testing and inference use. Another requirement is reliability in terms of persistence, i.e. the long-term constant availability of the pre-trained models used and, in particular, the models adapted by elaborate night training. This is particularly the case when they are used in applications which, for economic or safety reasons, need to be operated without failures and interruptions.
- (5) **Ethical Acceptability:** The models' editions sometimes violate the de facto standards for use in business, politics and education through bias (statement/decision based on false bias) and toxicity (use of ethically or stylistically-aesthetically unacceptable linguistic expressions). Therefore, it must be possible for the application to independently define, integrate and adapt corrective and filtering actions into the models.

European research can rightly be proud of its first European foundation language models such as Aleph Alpha, BLOOM or GPT-SW3. Other models such as the Open GPT-X are in preparation. These European models are encouraging examples of how, albeit with some delay, Europe can develop foundation language models. However, they are not yet an indication that European research has been placed in the vanguard of international research on the subject. Moreover, despite great progress, these models do not yet cover the German language or the range of other European languages to the extent that it is already achieved for English today.

One reason for this is the material research conditions. In Artificial Intelligence, as in all computer science, only research groups that can playfully try out many ways of realizing, adapting and optimizing a new technology can succeed in the longer term. If, in the case of the foundation models, the required physical infrastructure for computing and storage is large enough to exclude university research from cutting-edge research, the new breakthroughs will come from the cutting-edge laboratories of industrial companies, which, in addition to the infrastructure, can also build and afford a real research biotope in terms of personnel and culture. One of the hallmarks of such successful structures is the apparent abundance, characterised by redundancy in infrastructure and personnel competence.

The term AI CERN for the infrastructure for European AI research demanded and urgently needed by ⁴ and others is an effective metaphor, but in a sense it is also misleading. The infrastructure needs of particle physics are vastly different from those of AI and software technology research as a whole: While CERN's infrastructure can be booked by basic research projects to answer basic research questions in a very few very large experiments, the AI research groups of large industrial labs use computing resources almost consistently for large numbers of experiments to gradually train large models and to evaluate and modify them. The evolutionary cycles consisting of modification of technology and selection are very short because they do not need to be realized and tested in the physical world as in physics or in classical engineering. The paths from the lab to the applications can also be very short in AI. For example, Google was able to test the major language models in production for search and advertising placement soon after they were completed.

Due to high infrastructure requirements, there are currently only a small number of research laboratories and development centres worldwide that are capable of mastering today's state of the art of AI technology in its full complexity. This means that there are very few organizations that can carry out training and evaluation of large multilingual, multimodal and multimedia foundation models themselves. This complexity increases even more with the multimodal and multimedia linking of data and the combination of the associative models with explicit data and knowledge repositories.

⁴ The Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE) is an association with the aim of strengthening research, innovation and collaboration in the field of AI. More than 1,000 AI experts from across Europe are supporting the efforts.

This complexity naturally increases the entry barriers for new entrants, as in addition to well-trained experts, they will need extensive individual and collective experience in training, evaluating and using large foundation models.

Business needs for AI foundation models

.....●



4. Business needs for AI foundation models

A fundamental question that this study seeks to elucidate is the importance of foundation models to the economy and the demands that companies make on the models.

In order to get to the bottom of this question, two methods were ⁵ These are:

- a survey of small and medium-sized enterprises specialising in the development of AI; and
- Guide-based interviews with the heads of AI, Data, etc. of large German companies.

The companies surveyed can be summarised as follows:

Number of recipients	373 members of the German AI Association 25 contacts from the Merantix Momentum GmbH network
Number of replies	71 participants (18% feedback rate)
Size of companies	90% fewer than 100 employees
Sector of enterprises (Multiple appointments possible)	80% Information Technology 25% services every 11% Industry & Media
Use of AI	90% use AI

Table 3: Overview of the most important information on the Economic Survey

⁵ A more detailed discussion of the methodology can be found in Chapter V.

Interviews were conducted with the following individuals. Minutes of the results of the interviews can be found in Annex B.3:

Title	First name	Name	Company
	Wolfgang	Hauner	Allianz SE
Dr.	Maik	Friedel	BASF SE
Dr.	Marion	Legler	Bayer AG
Dr.	Hans-Jörg	Birds	BMW Group
Dr.	Michael	Fisting	Robert Bosch GmbH
	Jean-Paul	Schmetz	Burda Media
	Mario	Deng	BWI GmbH
Dr.	Corina	Apachițe	Continental AG
Dr.	Matthias	Dorner	DATEV eG
	Stephan	Kaulbach	Deutsche Bahn AG
Dr.	Frank	Cleaning	EnBW Energie Baden-Württemberg AG
Dr.	Sebastian	Kaiser	Ergo Group AG
	Thomas	Wolf	Hugging Face, Inc.
Dr.	Sabine	Donauer	Infineon Technologies AG
	Nico	Kelling	Infineon Technologies AG
	Rainer	Streets	Ionos SE
	Christian	Spannbauer	Lufthansa Group
Dr.	Stephan	Meyer	Munich RE
	Jochen	Kaiser	Mercedes-Benz Group AG
Dr.	Michael	Mr Müller-Wunsch	Otto GmbH & Co KG
Dr.	Armin	Kurrle	Porsche AG
Dr.	Lorenz	Determann	Rewe Group
Dr.	Feigu	Xu	SAP SE
Dr.	Andreas	Wierse	sicos BW GmbH
Dr.	Michael	May	Siemens AG
Dr.	Dirk	Schlesinger	TÜV Süd AG
Dr.	Sebastian	Hallensleben	VDE Association of Electrical Engineering Electronics Information Technology e. V.
Dr.	Patrick	van der Smagt	Volkswagen AG
Dr.	Alexander	Borek	Zalando SE

Table 4: Expert interviewed: women from the business sector

All findings from the survey and the interviews have been incorporated into the LEAM Feasibility Study. In the following, we will briefly illustrate some aspects and concordances.

Analysis of interviews and survey

Approximately 66% of the AI companies surveyed are already using foundation models or intend to use them in the future. This is a significant proportion, considering that foundation models are a relatively new development of artificial intelligence. In addition, many of the large companies interviewed also confirm that they already have AI foundation models in productive use or are currently working on applications. This makes it clear how essential foundation models are already for the entire economy.

The interviewees also indicate that the importance will continue to increase in the next few years. At the moment, it is not yet possible to predict what disruption and new business models will enable AI foundation models in the next few years.

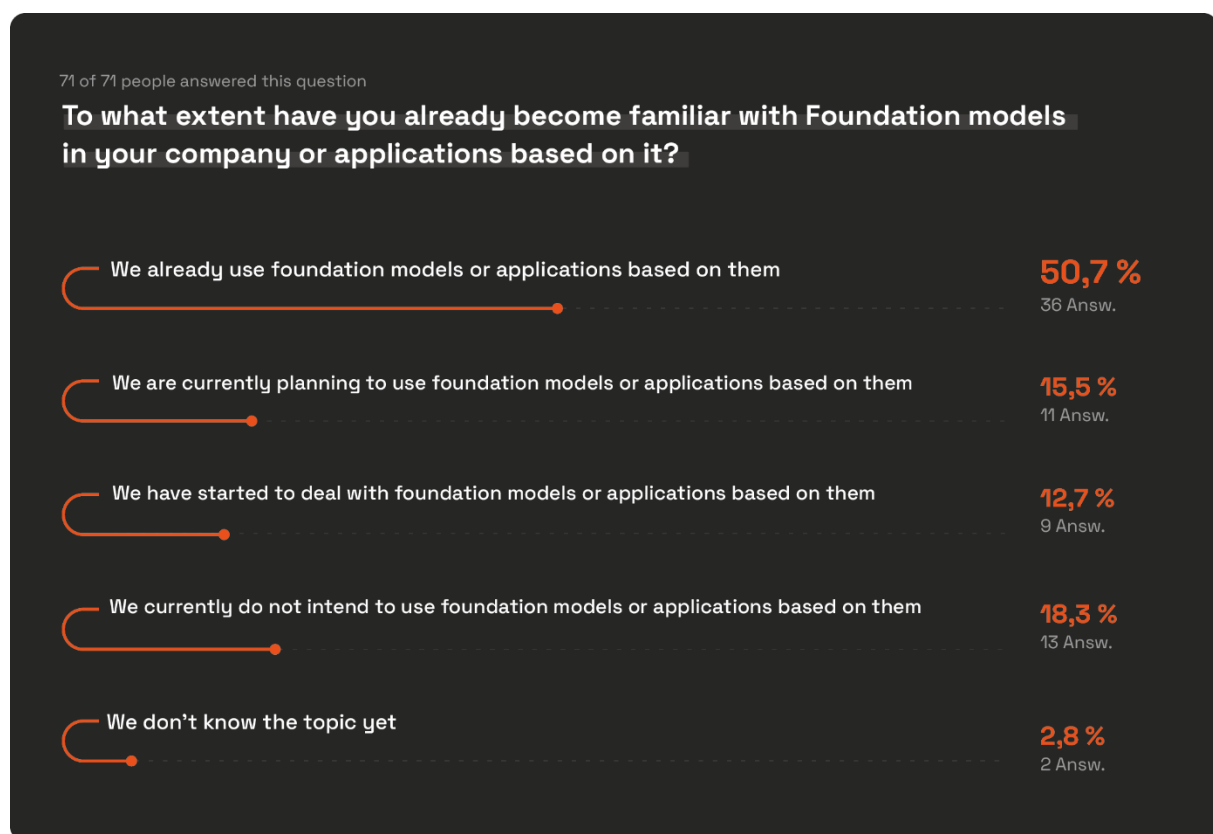


Fig. 14: Results of the survey of SMEs on their use of foundation models

AI companies that are not currently working with AI foundation models provide various reasons that would need to be met in order for them to use foundation models. These included, in particular, lower costs, the provision of open source models and the availability of data, each reported by around 58% of respondents. Other barriers companies cited as reasons for not using foundation models include a lack of computing infrastructure (38%), data protection barriers (33%), a lack of qualified staff (25%) and available European models (25%).

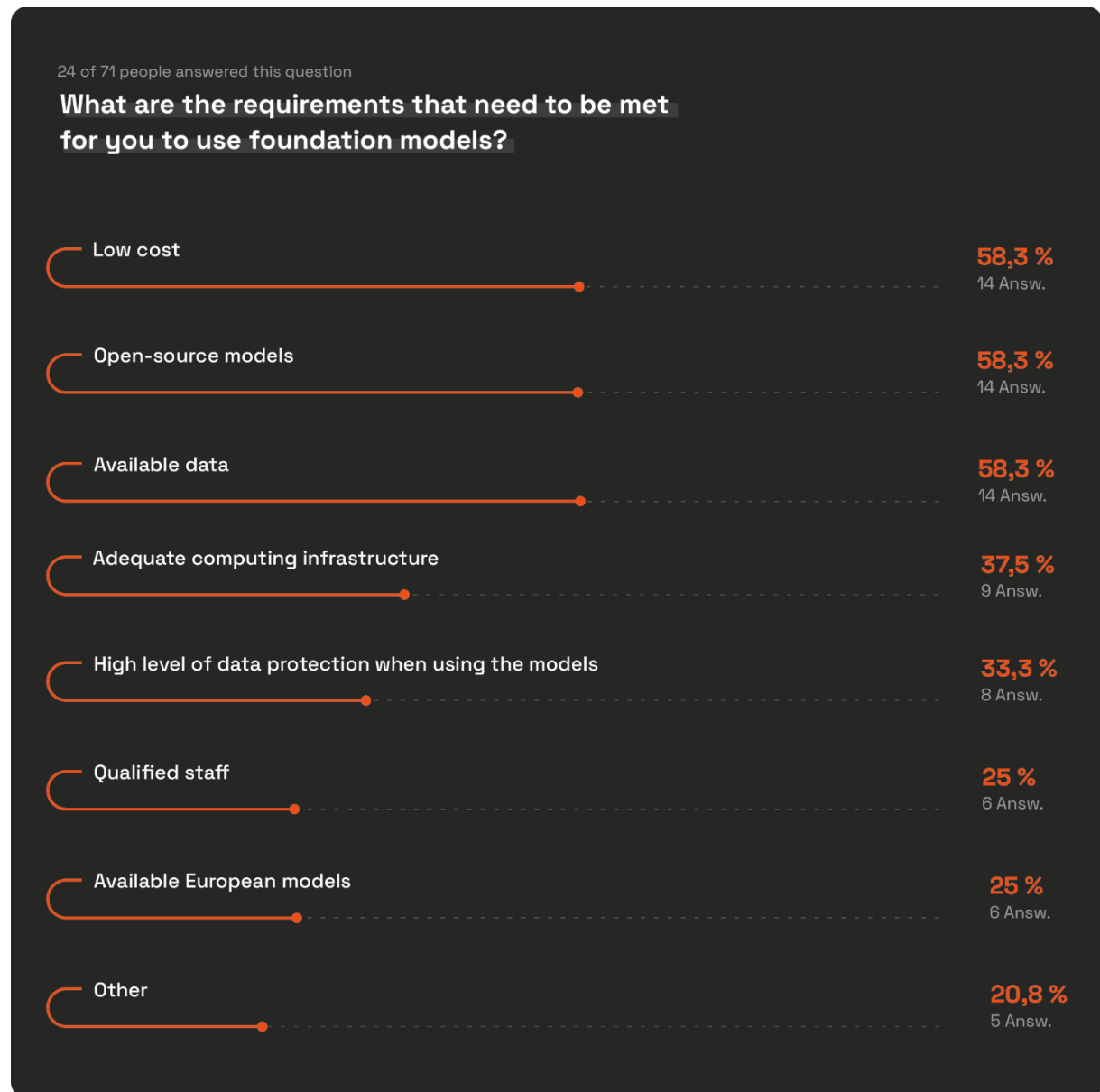


Fig. 15: Results of the survey of SMEs on barriers to the use of foundation models

Large companies also shared these concerns in the interviews conducted. It became clear: Large companies do not aim to develop their own foundation models. Instead, they want to adapt existing models for selected applications. In the current situation, this is often not possible for compliance reasons, as the models are not freely available, but only via programming interfaces, and data must be sent for tuning outside the European Economic Area. Therefore, especially with sensitive data sets, it is not possible to adapt the models for large companies. The large companies surveyed see this as a clear disadvantage in international competition. Even in the highly regulated and nationally defined insurance environment, companies report that new, digital business models emerging outside of Europe pose a threat.

In addition, interviews with large companies showed that general foundation models are often not sufficient to meet the companies' high quality standards. Instead, it needs foundation models that are tailored to the needs of individual industries. So the goal should be, e.g. develop a health model, an industry model and an insurance model, on the basis of which companies can then develop individual applications. Despite existing data sets, this development is not currently taking place, as German companies do not wish to transfer their patient, machine and insurance data to the US. This desire for special models also plays a role, since models trained on the dataset of a particular population are not directly transferable to other populations. For example, the chemical and pharmaceutical company Bayer reports that American patient data is different from European, Asian or African data.

In addition, many companies lack well-trained employees, and it is difficult to integrate AI models into their day-to-day operations.

Overall, the importance of foundation models for the overall economic development of Germany is rated very high at 73%. In particular, building a European AI ecosystem is considered relevant by 82% of respondents, and taking into account values such as transparency, reduction of bias and sustainability by 85% of respondents. Around half (54%) of respondents consider direct cooperation with research to be significant. Cooperation with SMEs and start-ups, on the other hand, is 67%. In addition to the already high level of use of foundation models, these results demonstrate the macroeconomic relevance that they are credited to in the industry.

Number of responses varies

Which of the following do you consider to be of great importance to you?

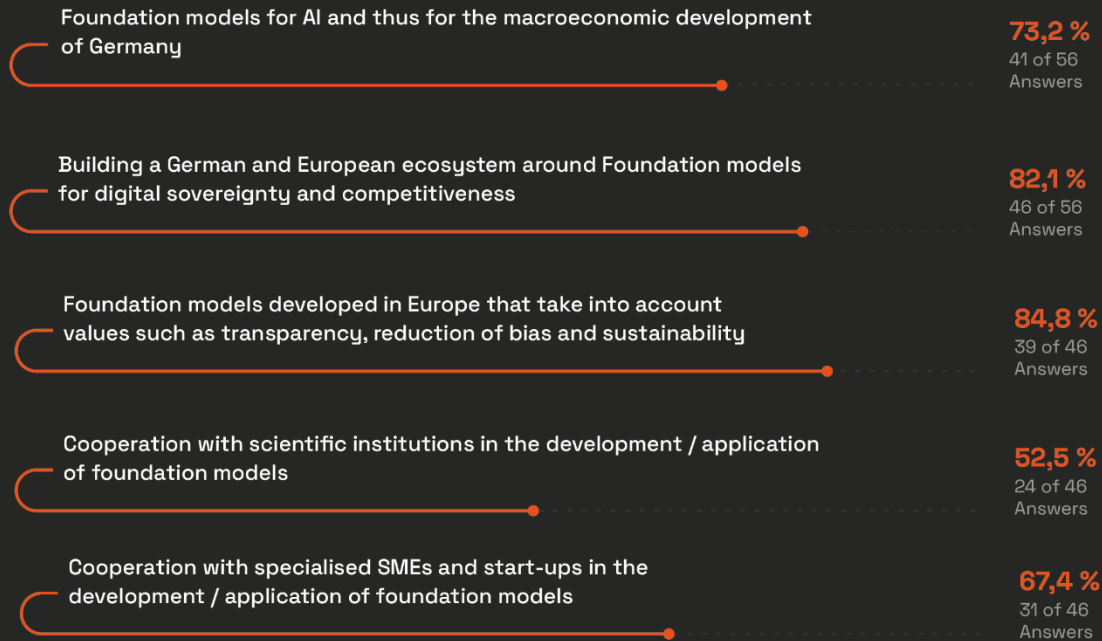


Fig. 16: Results of the SME survey on the importance of different aspects of foundation model development

The interviewees see in AI foundation models a strategic importance for the economic location of Europe. They emphasise the need for their own European models in order to take full advantage of the competitive advantages of the technology, to control the quality of the models extensively and to minimise safety risks. The last point is particularly important in view of Europe's geostrategic situation and existing discussions around 5G and Huawei.

Currently, for the companies surveyed, language models in particular have a high relevance in the use and development of foundation models (referred to by 71% of the companies). This reflects the high success and prominence of language models, such as GTP-3. Multilingual language models (52%) and multimodal models (38%) are also considered important. Both types of models represent logical next steps of language models and are already being successfully developed by companies such as OpenAI with Dall-E 2 and ChatGPT. As indicated above, large companies' assessments go one step further. While the interviews also mentioned mainly language and multimodal models, with which many companies are already experimenting, more specific models for individual application areas are needed. In the current situation, only incremental progress was possible and no disruption was possible.

Business and manufacturing processes as well as robotics are also important areas of attention, with approximately 34% of respondents each. These are dealt with in a little focused way in the current foundation model development. This may be an opportunity for LEAM to occupy these niches in model development.

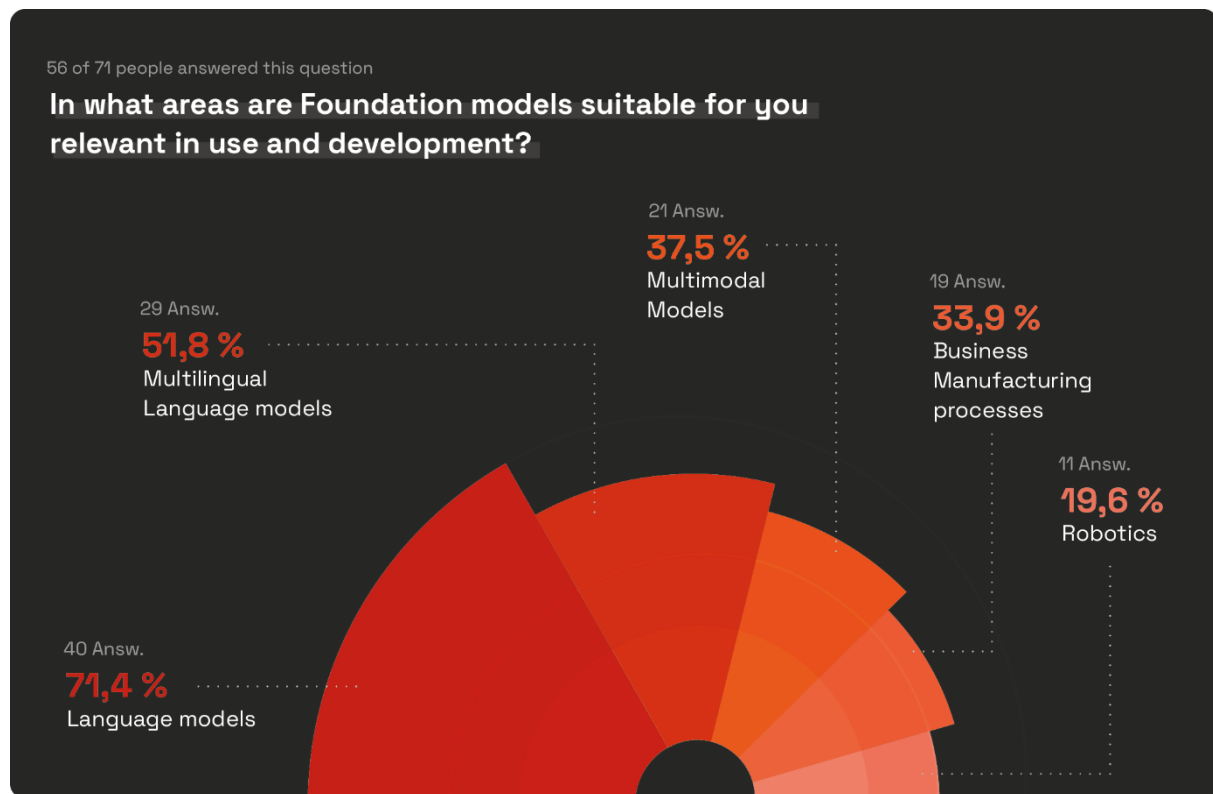


Fig. 17: Results of the SME survey on the relevance of different types of AI models in foundation model development (responses with a response rate of less than 20% were omitted. For full answers, see Annex A.2.)

Overall, the evaluation shows the need for European foundation models. The companies surveyed had the opportunity to leave further comments at the end of the survey. Similar to the Expert's interviews with academia (see Chapter 5), it has been increasingly noted that "everyone except [the] large Internet corporations [are] limited to using existing foundation models" and thus fall into massive dependency. Biases can thus be difficult to avoid in downstream applications. Another common theme in the responses was that the European economy is already lagging behind in international comparisons. These statements are also confirmed by the interviews with the big companies. There is now a need for joint action in the area of infrastructure and data, and there is no denying that competitive European models are preferable to American ones. Particularly in the area of data protection and security, European models would greatly simplify implementation.

Fortunately, many interviewees say they want to support the LEAM initiative and are generally open to involvement in a joint venture or PPP.

The evaluation of the survey and the interviews can be summarised in three key statements:

- (1) AI foundation models are becoming an elementary part of the IT application architecture. Cross-industry, companies are implementing or planning to implement applications based on the models.
- (2) Current models often have the disadvantage that they are not open source and therefore freely customizable. There are significant data protection concerns regarding the use of non-European models. European open-source models are therefore needed as an alternative.
- (3) The task cannot be handled by a single actor alone. There is a need for joint action by companies in Germany. Businesses recognise this need and are open to a commitment.



Support for development
through research and science

.....●

5. Support for development through research and science

Interviews with leading industry and AI companies make it clear that: There is a great demand for European foundation models. Therefore, one of the primary goals that LEAM has set for itself is to provide foundation models for business.

However, the development of foundation models requires extensive research and development efforts. For this reason, it is important to know and think about the specific requirements of research and science in order for the development of powerful foundation models to succeed.

For this purpose, interviews were conducted with leading academics from Germany as part of this study. It highlights the current state of research on AI foundation models in academia and what framework conditions are needed to support business and academia in developing them accordingly.

Methodology of the interviews

The interviews were structured by five key questions:

- Key Question 1 "Relevance of foundation models in science"
- Key Question 2 "International Leading Working Groups"
- Key Question 3 "Scientific and economic deficiencies of the foundation models"
- Key Question 4 "Measures to promote research and economic exploitation of foundation models in Germany"
- Key Question 5 "Special position of the foundation models and future developments"

The interviews were conducted with 21 experts between early October and mid-December 2022 (see table in Annex B.1).

The insights from the interviews have been fed into the planning of LEAM and, where appropriate, into the entire text of this chapter. Below, we will illustrate just a few selected aspects of the interviews along the five key questions.

Evaluation of the interviews

In **key question 1 "Relevance of foundation models in science"**, the experts describe themselves as users of the models and recognize the great importance of the models - including for the economy - as the research paradigm of the years to come. For many experts, there are obstacles to being able to contribute scientifically on an international stage, as they lack access to the necessary technologies and are therefore restricted in their research. foundation models can be made climate-friendly if centrally managed models can be made available to the AI community. The reproduction of the models would thus become obsolete.

In response to **key question 2, "internationally leading working groups"** were identified as U.S. technology corporations such as Microsoft, Open AI, Google, Deepmind, and Meta. In individual cases, mention has been made of universities such as Stanford or Carnegie Mellon University, which are competitive in the smaller language models. The Chinese models, on the other hand, were rarely mentioned. Access to infrastructure, data and talent were highlighted as success factors.

Some interviewees stressed Germany's strength in other areas, such as the presence of well-trained scientists who, however, often do not stay in Germany due to international offers. In this context, the reliance on the predominantly commercial suppliers of the models was also warned.

Key question 3 "Scientific and economic deficiencies of the foundation models" was answered with the lack of adaptability of the foundation models. There is a lack of German and also multilingual vocabulary, expertise, robustness/invariance, and explanations of the results. These obstacles cannot be removed by night training, or only on the financial terms of the hyperscaler. Problems such as bias, privacy and opacity regarding training data and procedures make it difficult to use. The trustworthiness of existing models is thus called into question and legal questions remain unanswered.

Continuous learning and also the connection with domain knowledge, corporate knowledge or application knowledge, such as factual knowledge from knowledge graphs, can be expanded. The question of checking and validating the results is an open research question. Other shortcomings include the lack of business models and the low efficiency of the models. The latter is being actively researched, but the comparison with very large foundation models (e.g. GPT-3) cannot be drawn, which is contrary to scientific practice.

The generalizability of the models was considered insufficient. Currently, most (language) data are used as the only source of knowledge, which may be a limiting factor in future development. Thus, there is a need to improve the curating of training data, enrichment with knowledge, and more complementary data for model development, such as ontologies, sequence data, or images.

Foundation models can be understood as a cultural asset, or a public good for basic research, which should be subject to a centralized and transparent development process to build trust. Otherwise, concentration of power and declining digital sovereignty may result.

In response to **key question 4 "Measures to promote research and economic exploitation of foundation models in Germany"**, the main focus was the promotion of infrastructure and easy access to it for business and academia. Complicated application procedures and too long waiting time for computing capacity are a significant barrier for both sectors. In addition, curating appropriate training data is essential, focusing on European linguistic diversity and multimodality. In addition, time series data and ontologies play an important role. In all cases, legal certainty must be taken into account and ensure that the data strategy complies with ethical principles.

Attractive research conditions are needed as a strong ecosystem in which researchers can collaborate. The aim is to create incentives to make old industries that have not yet been heavily digitized more attractive to young researchers.

The implicit knowledge of the foundation models was highlighted in the theme set of **key question 5 "Special position of the foundation models and future developments"**. The models are thus, among other things, able to learn programming languages, which was not technologically possible until a few years ago.

In Germany and Europe, the potential is seen in supporting language as the most important human means of communication in many different business B2B applications through assistance systems (e.g. shopping, addressing customers, communicating with customers, faster processing of documents, invoices, sales, service, moods, etc.).

Scientific breakthroughs are being made in increasing efficiency (mobile applications and edge computing) and e.g. This is seen in the processing of longer texts, which includes the recognition of cross-documentary relationships. Further potential is also seen in a more controlled generation of texts.

Some experts emphasized the potential of incorporating knowledge into (symbolic) subsystems addressed by the neural system (associative). Here, modularity will lead to adaptation without having to "forget" what has been learned.

SPOTLIGHT

Alexander Thamm GmbH

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Alexander Thamm,
Founder and CEO of the
Alexander Thamm GmbH.

Alexander Thamm GmbH [at] is a leading German consulting firm for data and AI. With 350 employees, [at] has been supporting DAX corporations as well as medium-sized enterprises for more than 10 years and is implementing innovative AI projects.

Where do you use foundation models? What is your use case?

The use of Artificial Intelligence is an important topic for our customers. We develop AI strategies, concepts and implement projects based on the latest scientific evidence. Language processing and foundation models play an increasingly important role in this, and we are investing in the use of the

technology in Germany, including by actively participating in the OpenGPT-X project.

Our teams specialize in multiple areas of AI development, including image processing, natural language processing, forecasting, and anomaly detection. Examples include an AI-driven system to support train disposition at the DB, a robotic system to support elderly care, and novel procedures for autonomous driving.

What impact do AI foundation models have on your business model or your projects?

Foundation models are becoming a central part of AI applications and infrastructure in many areas. Currently, we often develop individual AI applications from scratch using specific customer data. In the future, there will be a shift towards transfer-learning or tuning of existing, powerful foundation models.

At the same time, the use of foundation models will open up new areas of application and we will develop applications for our customers that are still difficult to implement at the moment - especially in the NLP field. The market will grow and we see this as a great opportunity for us, but also, above all, for the competitiveness of the German economy.

This has an intense impact on our business model, especially if we were to rely on the use and licensing of foundation models, which we can only access through APIs and over which we have no direct influence. If we can only obtain these models from non-European suppliers, we will also have to deal with data protection and data security aspects more intensively.

In order for our customers and our company not to fall into a one-sided dependency, it is extremely important that we also have access to foundation models developed in Germany or Europe, and we can not only use them through APIs. At the same time, the consideration of European values, e.g. on the topic of bias, is of enormous importance to us and to our customers.

What difficulties and problems do you see that only the U.S. and China are currently providing AI foundation models on a comprehensive basis?

U.S. Internet companies are currently investing heavily in the development and dissemination of foundation models. By providing through APIs, the first services on the market can simultaneously collect a lot of data, e.g. on the use points. As a result, there is a risk that monopolies, as in the case of search engines, will re-establish themselves and that technological dependence will increase.

If the core AI applications only come from overseas, in the long run, our activities will be limited to designing front-ends and workflows. We have little or no influence on the models, which is especially problematic in terms of quality and bias. This development could also become a potential threat to our current business model – and to our corporate mission to ensure the competitiveness of the European economy in this area.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

European foundation models, which we could use as a basis for our AI developments, would enable us to continue to develop innovative applications in the future. As we would not only access through APIs, but have the models available as open source, we could also intensify our own research activities in many areas and build state-of-the-art AI systems for our clients. This would ensure that we can continue to optimize processes for our customers in the future and enable new products and business models.

Survey methodology

In addition to the 20 industrial and university research experts interviewed, further in-depth opinions from 21 leading AI researchers were sought in the form of a survey for capacity reasons, and in order to give an even more detailed view of the situation. The survey was conducted between early October and mid-December 2022 and included 23 questions (see table in Appendix A.1). Added to this are some more detailed questions that could not be asked in the Expert:Inner Interviews due to limited interview time.

All findings from the survey have been fed into the LEAM Feasibility Study and the text of the entire chapter. Following the guided Expert:In-Interviews, we will illustrate some aspects and concordances along selected key issues.

Analysis of the survey

In the survey, most of the statements made by the interviewees in the Expert:Inner interviews were confirmed.

The vast majority of all AI researchers surveyed are already working with AI foundation models or intend to use the models in the future. Everyone values its importance to academia and business as high.

The type of data used varies considerably. These included publicly available texts and scientific publications, as well as medical and technical data. However, the issue of lack of data is also one of the reasons why foundation models are not yet used everywhere. Other reasons included a lack of transparency in data protection for existing models, a lack of infrastructure, technical limitations, a shortage of skilled personnel, high investment costs and ambiguity in regulation. Nevertheless, there is a desire to use foundation models in the future.

This is also reflected in the diversity of the mentioned potential areas of use, with the nature of the foundation models considered to be significant also playing a role - images, audio and mobility data have been mentioned here. No major role is played by the alleged dangers that might come from the models. Only potential military abuses, as well as the current lack of explanation, have been mentioned here. They also complained about the dominance of individual big tech companies and the lack of responses from academic research. This imbalance was also pointed out more frequently in the Expert:Inner interviews.

In international comparison, Germany is considered to be competitive only in the areas of competence and creativity, as well as training opportunities. Germany is lagging behind the US and China in other areas such as data, hardware and production. Weak digitisation and commercialisation are also seen as missing prerequisites. This is also the result of the increasing exodus of talent, which simply does not find the same conditions in Germany as elsewhere. This involves not only a lack of investment, but also strong regulation of new technologies, strong societal fears, and a general dislike of the progress that can come from AI research.

The respondents rated the relevance of different model architectures as well as scientific questions about data and data processing as "pretty high" on average. The same applies

to the domain agnostic pre-trainings and the subsequent domain-specific fine-tuning. Again, respondents viewed relevance as "rather high". For these basic topics, there is therefore a greater need for research within Germany.

As expected, there is no unanimous opinion on the content areas of use, as respondents come from different scientific backgrounds. The need for foundation models in the applications of the various disciplines was generally rated as "medium" to "rather high". Explainability, building common sense knowledge, including other data sources (other than text), hybrid processing (symbolic and subsymbolic), causation, etc. were also named.

The expected costs of survey participants are in line with the actual known costs of large foundation models. Respondents were already well informed in this regard. However, regarding the regulation of foundation models, the vast majority of respondents felt that they were not sufficiently informed, even though they saw this as a prerequisite for the successful deployment of the technology.

Responses were dichotomous with regard to possible deficiencies in the data, such as bias, discrimination and misrepresentation. Only a small majority of respondents said they were dealing with them and felt they were able to address them appropriately. These results are in line with the results of the interviews, in which half of all respondents pointed to this issue. The majority of respondents saw their knowledge of the regulation of the development of foundation models as a prerequisite for a possible use of the models.

In summary, the findings of the interviews were largely consistent with the results of the survey. Thus, the current leading nations in the field of foundation models are the United States and China. Factors that fostered this development include the data base, hardware, resources and grants. In international comparison, Germany is considered comparable only in the areas of competence and creativity, as well as training opportunities. Factors that still hinder the use of foundation models in Germany include a lack of infrastructure and access, technical limitations, a shortage of professionals, concerns about data protection and a lack of information about the regulation of foundation model development.

Potential hazards of the foundation models were also mentioned, mainly related to the monopolization of technology, as well as potential shortcomings of the models, such as bias, discrimination, toxicity, misrepresentation and explainability. The most common interests in foundation models were related to multimodality, European linguistic diversity, explainability, and the building of common knowledge.

In conclusion, the majority of all AI researchers surveyed are already working with foundation models or intend to use them in the future, and their relevance to academia and business is highly regarded.

Thus, German science and research are in the best position to implement powerful foundation models to meet the demands of the economy.

Opportunities and plans for developing European AI foundation models

.....●



6. Opportunities and plans for developing European AI foundation models

When asked which models are most needed by the economy, a clear picture emerges. Multilingual language models were mentioned most frequently in the interviews with the experts. In second place were multimodal models, which also include language.

Although the economic importance of foundation models in biomedical science, in autonomous control of vehicles and robots, and not least in the management of production and business processes has been highly valued, language models have consistently been cited as the most important area of application.

It is not surprising that major multilingual language models have been given the highest priority. Most of the already successful foundation models are either language models or multimodal models with a high percentage of language data and language functionality. The central role of language in human intelligence and in virtually all subsectors of human society allows language models to be used for a variety of tasks in virtually all sectors of the economy and society.

Therefore, when developing your own European foundation models, we recommend starting with multilingual language models. There are several reasons for this, which we would like to explain in more detail in this chapter.

Immediate needs

Language technologies are already being used in countless applications today: These include chatbots and voice assistants, automatic translation, text generation, text proofreading, text summary, text simplification, semantic search, spelling, tutoring systems, information extraction, fake news detection, automatic classification and replying to emails, language learning software, and forensic text analysis. In all these areas of application, today's language technology is still highly upgradable. The existing deficits in reliability still severely constrain the markets for the language technologies. With the immense performance gains from foundation models, larger markets can be opened up in the very near future, as basic applications with products, services and distribution channels already exist. Germany has hundreds of companies that would benefit from the advancement of technology. On the other hand, the products and services of these companies would be threatened if, instead, only foreign suppliers, especially US suppliers, were to launch more powerful products using the new technology. As SMEs, the German companies concerned are not in a position to train even large foundation models.

Economic potential

The economic potential is not just about improving the quality, functionality and market opportunities of existing applications. Thanks to the added power, many new applications are possible. These include, for example, easy-to-use access for citizens, patients and staff to practical information, knowledge and processes that affect them. This applies to public administrations, health systems, businesses and any other subsystem of society. As a result of digitization, direct passive and active access to all these data and processes is possible, but often too difficult to use. It is only when digital access is as easy as it was at the time of communicating with collaborative human contacts that digitalization can reach its full potential and be accepted by all. Recent chatbots such as Google's LaMDA and OpenAI's ChatGPT demonstrate that foundation language models are powerful enough to implement such interfaces to digital services. Very soon, we will all be able to communicate in our native language, spoken or written, without government appointments, waiting loops or complex input masks, with all the services that shape our daily lives. In business, such systems will revolutionize communication with other companies and end-users. Multilingual models allow organizations to expand to other countries with little effort, acquisition and customer communication. In education, natural language learning or tutoring systems will identify the level of knowledge of learners, identify gaps and target them through appropriate learning materials or personalized knowledge elements and explanations.

Social relevance

Our language is such an important element of human culture that we have to master the technologies that facilitate, influence and use language for a variety of purposes, and adapt them for our purposes. Eventually, language technology will play an even more important role in communication between people and between people and technology in the future.

All societal developments are also reflected in language. This is evident, for example, in the ongoing efforts to create a language that is in line with our ethical values, our understanding of history and our scientific worldview.

Multilingualism is also of high social relevance. Germany sees its own future as part of a multicultural and multilingual European society. This is the only way we can preserve our values in the geopolitical game of power and take an economic role in the world that will allow us to maintain our standard of living. A central component of European integration is the principle of equality among the languages involved. Even though English plays a special role as Lingua Franca in scientific, technical and economic communication around the world, we have guaranteed all members of the European Union that their languages will be preserved and protected, regardless of the number of native speakers.

A major barrier to the European Community's common digital single market is its multilingual nature. Unlike our largest global competitors, the US and China, the end customers of the products and services need to be reached in many languages. While U.S. and Chinese vendors can safely neglect smaller European languages, European vendors cannot. This, of course, does not only apply to digital products, an example being the package leaflets of drug manufacturers.

In 2018, the European Parliament adopted a resolution entitled "Language equality in the digital age" (Parliament, 2018), which not only emphasizes language equality, but also makes demands on language technology to achieve that equality in practice.

Acquisition of knowledge

Foundation language models are distinguished from models in other data domains, such as proteins or DNA sequences, in that knowledge about probabilistic modeling of linguistic texts is learned in many, if not all, subjects. A subject area could only be completely disregarded if it is not represented at all in digital texts. This form of knowledge acquisition also addresses an unresolved problem of knowledge modeling via Knowledge Engineering in symbolic AI, namely the difficulty of establishing the connections between subject areas, e.g. between materials and processes, between materials and market prices, or between synthesizable proteins and their sources of supply. Due to the versatility of the acquired knowledge in classical or newly defined areas of knowledge, the application potential of the foundation language models is particularly high.

Level of development and comparability

The state of development also speaks for the choice of foundation language models. The presence of several years of well-documented language model training experience and extensive and versatile benchmarks makes it easier to engage in cutting-edge research and control one's progress than with newer types of foundation models.

Existing competence

In the fields of neural NLP and multilingual language technology, Germany has a well-established research tradition and a strong scientific community with world-renowned top researchers and well-trained young people. There are also a variety of older language technology companies and new AI start-ups that can integrate the NLP components into marketable products and know the requirements of existing markets.

Available data

Through several successful EU projects to collect language data for European languages, we have access to large amounts of data suitable for European foundation language models. Also so-called minor languages are represented in the data collections. Many of the data are of high quality because they have been provided by public administrations and non-profit organisations in the EU Member States. Among the data are also many parallel bilingual and multilingual text corpora. The language data thus made available increases the likelihood of success for language models that qualify to meet the specific European requirements, and suggests that these foundation models offer clear advantages over the large existing models, at least according to this criterion.

Potential for future developments

More recent developments in multimodal and multimedia language models link together the data of the data domains represented. This means that you can also generate thematically appropriate images or videos with the help of text input. This functionality is called cross-modal (transmodal). The language is of particular importance in this respect, as it is suitable for expressing the concepts and criteria of the users. In the future, the number of data domains will increase, with language expected to always be the semantic basis for combinations. Therefore, new combinations can be made by extending language models.

For these reasons, we recommend starting the development of foundation models with large European language models, as this is where the greatest potential for the German economy lies and the demand is particularly high. Gradually, the foundation models will then be expanded to include more types of data.

6.1 First European multilingual foundation language models

With regard to language models, a variety of questions arise for concrete implementation and further development. In the following, we will explore which architectures and data should be used to support the language models, and which priorities should determine the development and evaluation of the models.

Architecture of the basic foundation models

For the base architecture, there is currently no serious alternative to the Transformer approach. Almost all of the major Foundation language models follow this approach, which has also proven to be viable for image, video and protein data, and is therefore also suitable for the planned extensions to include multimodal and multimedia data.

However, there are now different versions of the original Transformer model, which roughly fall into three classes: (i) encoder-decoder models that are more or less similar to the original model, (ii) Encoder models; and (iii) Decoder models.

The most successful models to date are in Class (i) and (iii), with the very largest models more likely to fall into Class (iii), but this is certainly because the training effort for large decoder models is much less.

However, the majority of the professional community assumes that class (i) models, i.e. full-transformer models, offer advantages for certain challenging tasks, which can take advantage of the deeper analysis of the input and the abstract representation of semantic relationships that is generated. This includes answering questions and other tasks that require conclusions, as well as automatic translation.

Until now, decoder models have been able to offset the disadvantages of simpler architecture due to size, a larger amount of learning data, and a higher number of parameters. The performance of ChatGPT, the latest model in the GPT-3 class, in particular, impressively demonstrates that the decoder model (along with appropriate night training procedures) enables functionality that requires a high degree of semantic abstraction.

For the first generation of models, therefore, it is recommended to use the decoder architecture of the GPT models. This offers the following advantages:

- rapid skill building through the use of the simpler architecture
- better comparability with the latest GPT models and with European models (BLOOM, Luminous, GPT-SW3)
- comparability makes it easier to evaluate the contributions of the additional data
- shorter time until the first usable results

Versions of the first models should also differ from the existing models by the choice of learning data. In addition to the crawled web data used so far, specialised European corpus data will be used to better cover both European languages and business-relevant subject areas. We expect a particular effect for the desired multilingualism from the use of parallel bilingual and multilingual corporations.

In addition, the training data is also extended to include knowledge data from large knowledge graphs.

6.1.1 Learning data

Learning data for foundation language models are large amounts of digital text data that best represent European languages.

The central component of training data for the largest language models are very large volumes of text curated by web crawling. The non-profit organization Common Crawl has been collecting and archiving large portions of the World Wide Web since 2011. These are made available free of charge for research, development and other uses, in WARC web archive format since 2013.

Building on the datasets of Common Crawl, the web crawl corpora that underlie the largest foundation language models have been compiled and cleaned up. A selection of these data sets will be presented here.

C4 - Colossal Clean Crawled Corpus

This is in particular the C4 corpus (Colossal Clean Crawled Corpus), which exists both exclusively for English but also in multilingual variants (Raffel et al., 2020). The "Clean" attribute refers to several data cleanup procedures used to separate real monolingual texts from mixed data.

The purge procedures also included the use of blocklists, lists of words that were expected to detect obscene, racist and otherwise offensive texts. Examples include the Lists of Dirty, Naughty, Obscene, and Otherwise Bad Words (LDNOOBW) offered on GitHub for English and about 25 other languages. The C4 corpus now filtered out all web pages that contained at least one of the offensive words in the respective language. It was assumed that there would be fewer problems with the large amount of data if this rough criterion was sometimes unfairly filtered, rather than by including offensive content in the AI models.

However, it was then shown by critics that the deletion of texts with sexual references or slang expressions meant that important areas of society were underrepresented in the texts thus "cleansed", including parts of the LGBTQ community or ethnic minorities (Dodge et al., 2021). Fortunately, however, there are also "noblocklist" versions of the C4 corpus to which these lexical filters have not been applied.

mC4

For multilingual foundation language models, the webcorpus mC4 was created, a special multilingual version of the C4 corpus, the benefits of which could be demonstrated by the performance of the mC4-trained model mT5 (Xue et al., 2021).

The mC4 body contains 27 TB of text data for 101 languages. The best-represented language is, of course, English with 10,401 GB, while West African Yoruba forms the final light with only 0.158 GB more in this corpus.

The English language accounts for 5.67% of the data, with 2,733 billion tokens. German comes fourth with 1,404 GB or 347 billion tokens, after Russian and Spanish, representing 3.05%. Smaller European languages such as Lithuanian, Estonian and Latvian are still

represented with 11, 6.9 and 7.9 billion tokens respectively. The smallest European languages present in the corpus are Luxembourgish with 1 billion tokens and Irish with 0.5 billion tokens. In contrast to Luxembourgish, Irish has even been one of the working languages of the European Union since 1 January 2022.

6.1.2 European projects

Fortunately, there have been a number of projects in Europe since 2010 curating and making European language data available for research and business. These projects, supported by the EU in various programmes, should enable European AI, in particular language technology, to make European linguistic diversity technologically manageable, thereby removing competitive barriers to the European economy and the common digital single market.

These include, in particular, the ParaCrawl and European Language Resource Coordination (ELRC) projects supported by the Connecting Europe Facility (CEF) programme and the European Language Grid (ELG) supported by the Horizon 2020 programme.

ParaCrawl

ParaCrawl curated parallel texts in European languages through web crawling in three phases from 2017 to 2021. The result is a parallel dataset with 41 language pairs that mainly associate English with another language (39 of 41). In addition to European languages, ParaCrawl also contains data for nine resource-poor non-European languages.

ELRC - European Language Resource Coordination

The ELRC infrastructure project, which will be completed in January 2023, focused on strengthening European translation technology. The European economy and society should also be able to overcome language barriers independently of Google Translate and the other major translation services of the American hyperscale, without suffering the position of European languages, not even the smallest and smallest ones. Therefore, ELRC's focus has been on collecting or generating bilingual and multilingual parallel corpus, which can then be used for supervised training of machine translation systems. One of the challenges was to curate a high-quality and economically/socially relevant corporation in EU countries, and that of organisations in all EU Member States.

ELRC has collected and compiled a total of over 200 billion words of high-quality language data in Europe and has established a pan-European ELRC network in all EU Member States. The data contains 5600 resources: bilingual and multilingual parallel corpus as well as monolingual text corpus. There are more than 880 parallel corpora with more than 1 million words, including more than 230 with more than 10 million words and more than 50 with more than 100 million words. The European national languages, as well as the languages of the largest minorities, are represented in the data. Over 950 resources contain German data, and even small languages such as Irish and Maltese are still fairly well represented with over 300 and 200 resources, respectively.

ELG - European Language Grid

The ELG European Language Grid project has created a platform under its name that makes tens of thousands of language technology resources available for use, most of them on non-commercial licensing terms. These resources also include thousands of text corporates, including very many for European languages, which are underrepresented in the large web corporates.

EFNIL - European Federation of National Institutions for Language

The European Federation of National Institutions for Language (EFNIL) is the umbrella association of national language institutions of European countries. These institutions include language institutes such as the German Language Institute in Mannheim or the Czech Language Institute in Prague, but also government institutions such as the Danish Language Council or the General Delegation for the French Language. Many of these institutions manage and maintain national corporates, large, well-maintained and more or less representative collections of text for their respective language or languages. In the past, this data was usually not accessible to language technology because of the current ownership or copyright rights in the way of such use.

OPUS

Alongside these EU projects, the Nordic Natural Language Processing Lab's OPUS initiative has been in place since 2004 and now provides over three million parallel texts with over 100 million sentences freely available.

eTranslation

Most European organisations in business, politics and civil society are reluctant to use the translation services of large multinational companies for translation, search and interpretation for security reasons.

For pure text translation, the European Commission created the eTranslation translation service back in 2017, which provides European organisations with free translation between all EU working languages.

This service of the European Commission's Directorate-General for Translation already employs neural translation systems for many language pairs, usually always translating between English and another European language, and all other translations are done indirectly via English as an intermediate language. The European Commission does not have any foundation models.

Although eTranslation is not comparable in quality to Google Translate, it is used by many organisations for raw or information translations, not only because of the security requirements of user organisations, but also because of the strengths of the European system in terms of language and content specific to Europe. Approximately 300 million pages are translated each year.

6.1.3 The importance of parallel language data

It is an open research question what role parallel language data will play in the world of foundation models in the longer term.

If parallel language data already exists, it can be incorporated as translation pairs directly into the data for pre-training, and the model will then learn on its own that they are pairs of sentences of equal meaning in the respective two languages. However, it is also possible to integrate supervised translation skills training into pre-training without additional costs for data production or annotation. The parallel data can then of course also be used as monolingual data for the complementary or replacement tasks of the self-supervised pre-training.

However, it has been observed that the ability to translate becomes an emerging functionality once there are only sufficiently large volumes of text for the individual languages in the pre-training data. Text translation thus becomes a zero-shot or at most a few-shot application of the multilingual foundation model. For languages that are not sufficiently represented in the learning data, the model would then have to continue to be re-trained through supervised learning. This reduces the need for parallel corporations, which are limited in their existing quantity and expensive to produce.

In their research report *nmT5 - Is parallel data still relevant for pre-training massively multilingual language models?* show Kale et al. (2021), however, that parallel corpora, due to the scarcity of data for many languages, are still essential to the applicability of the models in those languages and to the quality of the translations.

This unique feature has great economic potential that goes far beyond applications for text translation and multilingual parallel text production. Another area of application is cross-lingual search, both for searching documents and for searching for information and knowledge content. It makes life easier for citizens or customers, especially those who are not native speakers of English, to search for multilingual content in their own language. Another application is to support bilingual and multilingual conversation, such as by providing a simultaneous interpretation function for deliberations, meetings and negotiations. Zoom already offers, based on German translation technology,

almost simultaneous translation between nine major languages, including five EU languages, but an extension to the breadth of European languages is not in sight because of the lack of appropriate training data.

6.1.4 Recommendation

Despite the fact that the European languages, particularly the so-called minor languages, are not yet sufficiently represented in the corpora, the existing corpora C4 and mC4, respectively, should be used first for the training of the multilingual European foundation models. This allows for comparability with other large language models and avoids missing data that have contributed to the performance of the known models. On the other hand, it allows for a better assessment of the contribution of specific European data to relevant applications.

To date, there has been no investigation into the extent to which the data in the corporation provided by the European projects overlaps with the web data of the major American foundation models. However, since Common Crawl does not collect data that can only be accessed after registration, and also respects the crawling rules (nofollow, robots.txt) of the original websites, it is likely that large volumes of additional data can be used for training through the European projects.

The ELRC and the ELG are coordinated by the DFKI. The maintenance and provision of data for both projects is the responsibility of the project partner ILSP in Athens. The project coordinators and the director of the ILSP have expressed their willingness to support the development of European models for data curating from their holdings. In September 2021, at the EFNIL Annual Meeting, national language institutions were presented with the possibilities and intentions for European foundation models. Initial communication with EFNIL member organizations revealed their willingness in principle to provide their corpus data under controlled conditions for the training of foundation models.

The Institute for the German Language has declared its willingness to make available the Institute's high-quality body of text of some 40 billion words for the training of European models and to provide scientific support to the project on aspects of the representation of the German language in the European foundation models, if necessary.

In addition, the OpenGPT-X project is collecting data for the training of a European language model. The experience gained and the data used there can also be applied beyond the project.

These data, together with the presence of parallel language data for European languages, allow the training of the first European foundation models.

By using the parallel data in pre-training, there is an opportunity to achieve several goals:

- (1) meeting the specific European requirements for overcoming language boundaries
- (2) Better consideration of smaller European languages
- (3) the creation of a special unique feature of the models over the existing foundation models

SPOTLIGHT

Bayer AG

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Dr. Marion Legler, Head of Decision Science & Advanced Analytics, Bayer Pharma.

Bayer is a global company with core competencies in the life science areas of health and nutrition. Through its products and services, the company aims to benefit people and save the environment by helping to solve fundamental challenges of an ever-growing and aging world population.

Where do you use foundation models? What is your use case?

We use AI foundation models to automatically process large amounts of text and help our experts find relevant information quickly. For example, the models:

- They convert the texts written by doctors during clinical trials into standardized medical diagnostic codes, which simplifies the subsequent manual review by our experts.
- Read all the medical literature and clinical trial reports to identify those that are particularly relevant to Bayer's therapeutic areas and treatments.
- Thousands of documents from licensing agencies (EMA, FDA, etc.) are reviewed and the topics in each part of the documents are automatically classified, making the information easy to find. Scan messages received from patients for adverse events.
- Grouping the findings written by Bayer representatives after meetings with physicians to discover emerging topics of discussion.

We also use pre-trained computer vision models for use cases where limited annotated images are available, such as those associated with specific cancer tumors.

What impact do AI foundation models have on your business model or your projects?

Most of the use cases made possible by these models were not possible before, especially those involving the detection of large volumes of documents. In other cases, such as coding in clinical trials or detecting adverse events in text, AI models support the manual work of human experts, saving about 50% of the time they spend on the repetitive tasks.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

Because most major language models are only available in English, we often need to translate from other languages into English first and then use the models. The content and subtleties of the text are certainly lost in the translation.

Open-source models covering all European languages could help ensure that patients, doctors or authorities speaking in different languages are considered equally. While such models already exist (BLOOM), they are currently hardly used because they are trained on general, non-medical texts. The challenge of ever having a workable multilingual model for health applications is that the scientific literature and documents used to train medical language models were originally written in English only.

6.2 Avoiding misrepresentation, bias and toxicity

As described in Chapter 2.3, several forms of inadequate behavior were observed in neural models trained on large amounts of web data. In order for the foundation models to meet the criteria for applications in business and society, such behaviour must be prevented or minimized. Therefore, the prevention of misconduct should be a key priority in the design, training and evaluation of models.

False statements

The most obvious form of misconduct is erroneous statements. The model will make such statements when, despite pre-training, knowledge gaps remain or knowledge inconsistencies arise due to conflicting training data. Moreover, foundation models cannot distinguish between factually correct statements and plausible incorrect statements, which are also often found in the field of literature. The methods of fixing are better data selection and the addition of knowledge from non-linguistic knowledge stocks (see Chapter 2.3).

Bias

Imbalances, including biases, misgeneralizations, and misjudgments with ethically reprehensible consequences, are more difficult to detect and more difficult to prevent. These are often subsumed under the collective term bias.

Frequently commented forms of bias are biases about gender and ethnic or social origin that we also find in humans. (ref)

However, not every bias needs to be negative, for example, in the case of a product, the more frequent mention of rare hazards compared to the mention of the benefits of the product may lead to statements suggesting cautious use.

It is also not possible to predict all forms of negative bias because they are determined by certain societal developments of problem awareness and time-dependent sensitization.

Attempts to preemptively erase all sources of negative bias from training data are therefore unrealistic. Instead, developers try to override already known forms of bias by appropriate response behavior through targeted night training. A similar approach has been taken by the developers of ChatGPT, which has been effective in responding to criticisms of previous GPT models.

Adequate response behaviour also includes a plurality and balance in the naming of alternative responses. The complementary tasks of pre-training are not enough to achieve such behavior. If a model is not to simply choose the one statistically preferred by the learning data for multiple possible responses, a language model must learn the appropriate response behavior through dedicated training. Again, ChatGPT has shown how this goal can be achieved through night training. OpenAI has achieved these improvements essentially through reinforcement learning, which in this case is the systematic correction of the weights through the response of test users.

Annotated learning data can be generated from the test input, together with the answers and reactions of the test users, so that later models can be monitored and post-trained, thus saving the cost of manual performance.

In addition to night training, a proven, simpler method for improving response behavior is embedding user input into prompts that specify in one or more sentences what form of response is expected. This method is also widely used to adapt the content and form of expected responses to the requirements of specific applications.

Toxicity

The phenomenon of toxicity ranges from the use of obscene or ethically offensive expressions to statements that may be interpreted by people as expressions of hatred or contempt or otherwise perceived as offensive or hurtful.

Attempts to achieve such spending by censoring learning data, i.e. by automatically filtering out texts that contain certain offensive words or phrases, are not the appropriate means to prevent toxicity. Many words that have been included in the lists of offensive words (blocklists) also have uses that are quite acceptable. Filtering out the harmless uses would weaken entire subject areas on the data page. On the other hand, not all insults or contempt of the words used can be recognized alone. However, it appears that the newer language models themselves learn an abstract mapping of linguistic expressions to linguistic styles and registers. Many obvious forms of toxicity can be independently avoided by the model after reinforcement learning. The language usage of ChatGPT seems very cautious, almost conservative, compared to previous foundation models.

However, it is an open question whether it is possible to reliably avoid all statements or presuppositions that may be perceived by sensitized user groups as offensive or offensive. This topic is also a challenge for human text producers. However, the problem will only arise in very special applications anyway.

Recommendation

For all forms of inappropriate response behaviour: Even if it is not immediately possible to 100% exclude all forms of this misconduct for all applications, it is essential for the success of the foundation models and their acceptance in the economy and society to empirically determine or predict the nature, degree and frequency of potential misconduct.

It was therefore necessary to assemble a test battery of input prompts for all known types of bias and toxicity, capable of generating expressions of misbehavior and measuring their frequency, already during the development of the first models. The constant evaluation and improvement of models in relation to these problems should be a central research theme.

6.3 Connecting foundation models with large knowledge holdings

In addition to knowledge derived from texts, explicitly formalized knowledge can also be made available for foundation models. Explicit knowledge can be encoded in databases, ontologies, or knowledge graphs. Nowadays, knowledge graphs (English: knowledge graphs) are mainly used to represent large amounts of knowledge. Knowledge-Graphs, short KG) are used because they combine the benefits of ontologies and databases and have proven themselves in high-volume use.

The use of KGs is intended to solve or reduce several problems of the foundation models:

- **Incorrect answers in the event of insufficient knowledge or confidence:** Neural models sometimes give completely wrong answers in such situations. Even without being asked directly for them, models can sometimes claim facts in issues that do not correspond to the truth. This belongs to the realm of the so-called hallucinations.
- **Dynamics of knowledge:** Changes in knowledge or completely new facts are often not yet represented in the training data or statistically underrepresented compared to outdated knowledge.
- **Gaps in detailed knowledge:** KGs also contain details that cannot be found on the openly accessible Internet due to lack of general relevance, e.g. certain parts of product specifications, member lists, metrics, etc.

There are three promising approaches to using knowledge repositories:

- (1) The incorporation of the knowledge stock into the training data (e.g. KELM)
- (2) The model's access to the knowledge pools as part of the inference
- (3) The Consideration of Knowledge in Large Text Corpora by Retrieval

In the first approach, the knowledge elements of the KG, called RDF Triple, can be automatically converted into sentences of a natural language, with English being offered because of the naming in the KGs. Alternatively, the RDF triples can also be added to the multilingual training data in the RDF syntax, such as sentences from their own language.

The addition of knowledge to the training data has the advantage of improving the world knowledge of the model already during pre-training. It is also hoped that the projects will be able to learn to translate between simple linguistic statements and the corresponding knowledge graph fragments similar to translation between natural languages.

In the second approach, the model learns to access necessary knowledge through SPARQL queries and to take this into account to calculate the final answer. This approach has the advantage that all new knowledge in the KG is immediately available without night training.

Both approaches are currently being tested in the Open GPT-X project. Follow-up projects will be able to build on the findings and practical results of this forerunner project in the realization of this objective.

Suitable candidates for knowledge repositories include DBPedia and WikiData, two large generic knowledge graphs maintained by large knowledge communities and available free of charge. In these resources, however, the knowledge of European business enterprises is not complete, nor is it always up-to-date for all companies. Therefore, the inclusion of specialized knowledge sources such as OpenCorporates or alternative commercial offerings should also be considered.

A third approach is retrieval procedures to use additional large text data sets. These may be current texts (e.g. search engine results) that have not yet been fed into the training data. On the other hand, this can also be internal data (e.g. service reports) that can be valuable as additional information for the model. Through an embedding-based retriever-reader model, they can feed into response generation, thus reducing errors in the texts generated and using up-to-date information. Multilingual retriever reader models can use the information in different languages.

6.4 Combining language with other modes and media

When asked about the intrigue of foundation models, Prof. Andreas Dengel: "Of course, from the research point of view, multimodality is very exciting." In particular, images, videos, spoken language, audio, and 3D models that can be combined with text can be considered as additional media. Yann LeCun, head of research at Meta, goes a step further: "However, rather than using language or images, the next generation of AI will learn directly from videos. Meta is currently making great efforts to collect video data from the I-perspective for this new generation of AI, but YouTube videos are also suitable as training material" (Schreiner, 2022).

European business and academia should take this development seriously. Implementing and training these models for European languages and visual content, as well as extending them over a longer period of time, is a worthwhile goal for European models.

The successful use of the models requires the solution of a number of problems:

- Extending the duration of videos by about an order of magnitude.
- Possible use of an autoregressive architecture.
- Efficient use of existing models for single image production (e.g. Stable Diffusion).
- As a rule, a simple sentence is no longer sufficient to specify the content, instead a multi-point storyline must be stated and taken into account.
- New strategies to be able to process longer input sequences, e.g. following the pattern of S4.
- Advanced approaches to temporal and spatial disaggregation in video generation by diffusion models.
- Incorporation of other modalities, such as spoken language and sounds.
- Incorporation of existing images and 3D models that can be animated with the video techniques.
- Simultaneous handling of different objects of the same type.
- Consideration of camera movements, morphing, and changing scenes.

There is a huge scope of application for synthetically generated videos:

- In education and teaching, relationships can be directly visualised. An example is: "Show how the Indian subcontinent overtook the Himalayas."
- In language lessons, scenes and processes can be visualized, which the student must then describe.
- For the personalized promotion of a new product: "Show Karl Müller driving his new electric car across the Gotthard Pass".
- Instructions for using a product can be created "on the fly" for a new environment.
- Animated films can be produced in a cost-effective way.

Learning text-video models usually requires a dataset of video clips with subtitles entered manually. However, the creation of such data sets is costly and time-consuming and is therefore difficult to achieve on a large scale. Instead, these models can be trained using videos with natural language annotations in the form of automatically transcribed speech data. One example is the HowTo100M database of 136 million video clips from 1.22 million annotated educational videos on the Internet (Miech et al., 2019), showing people performing and describing over 23,000 different visual tasks. However, this approach also has some potential drawbacks. On the one hand, the texts generated by speech recognition are not error-free and the temporal mapping is not perfect. It is also the restricted domain of educational videos. The more recent data sets are more extensive and cover a wider subject area. The HD-VILA-100M contains 100 million videos in 720p resolution that simultaneously contain audio data, subtitles, and video frames, lasting an average of 13.4 seconds (Zellers et al., 2022). The authors show that audio data

significantly improves training success. (Nagrani et al., 2022) transfer captions from image-text data to video clips without additional manual effort. Using this pipeline, they create fuzzy annotated audio-video data with millions of paired clips and captions. They show that these data can be used to train very powerful models for video search and video captioning.

Overall, there are many millions of videos available with audio and some with subtitles on Youtube, Shutterstock, Dreamstime and Reddit. Another valuable resource is the archives of the broadcasters, who must routinely produce transcripts and subtitles in order to access their programs without barriers.

SPOTLIGHT

Continental Automotive Technologies

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Dr. Corina Apachițe,
Head of AI, Continental
Automotive Technologies

Continental develops pioneering technologies and services for the sustainable and connected mobility of people and their goods. Founded in 1871, the technology company provides safe, efficient, intelligent and affordable solutions for vehicles, machinery, transportation and transportation. Continental had sales of €33.8 billion in 2021 and currently employs more than 190,000 people in 58 countries and markets.

Where do you use foundation models? What is your use case?

We use foundation models in particular in the field of "Natural Language Processing", i.e. the analysis of texts. There, these models are the measure of all things at the moment. Our applications include the analysis of "requirements documents to support our developers, or in the field of "Conversational AI", i.e. generally for human-machine interfaces. Another application that may play a role in automated driving in the future is the analysis and formalization of traffic rules using speech models. In the area of image understanding or image generation, foundation models can also be used to create and analyze image datasets, for example, to automatically annotate training data for AI models.

What impact do AI foundation models have on your business model or your projects?

foundation models are essential for high-quality applications in the area of language comprehension. Thus, it is only the AI foundation models that enable us to implement our applications. In the field of text analysis of requirements documents, technology saves us considerable time, makes analysis easier and more efficient, and thus lightens the burden on our employees. The usability of chat bots depends primarily on the quality of the AI models. Future foundation models, such as multi-modal variants that understand both text and image data, will also be of interest for a variety of other tasks that are so far difficult to implement.

What difficulties and problems do you see that only the U.S. and China are currently providing AI foundation models on a comprehensive basis?

In general, AI models have a direct dependence on the data with which they have been trained. The quality and function of the models depends primarily on these data. Thus, models must be trained primarily on data from the corresponding regions. European specificities, such as linguistic diversity, regulatory or cultural differences in road transport, may be overlooked by non-European AI models. Quality and functionality therefore require regional solutions.

In addition, the high standards of data protection law in Europe ensure that sensitive data can hardly be sent to American or Chinese AI models. Furthermore, the models remain closed systems ("black boxes"). More detailed analysis or further development of the models is not allowed.

Another difficulty is the dependence on the owners of the AI models in commercial terms. Barriers to access could be set up or specific features of the models could be designed that are not in the interest of European users. And last but not least, it's also about location disadvantages by migrating talent to where the "best" foundation models are created.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

Each of these aspects would be a major advantage for using AI foundation models. Publishing as open source would fully allow for the study of the content of foundation models. There would be more possibilities for design, especially in view of the high diversity in Europe. foundation models can be customized and modified in any way, such as tailoring models to regional specificities, to the needs of different industries or application domains.

We see a competitive advantage in high standards of privacy that, together with transparency, will significantly increase the acceptance of foundation models.

6.5 Questions and developments

Especially for multimodal models, Prof. Dr. Wrobel: "The biggest deficit is a lack of understanding of skills and limitations." This is done by Prof. Schütze in his interview supports and concretizes: "It would be another paradigm shift if other modalities could be fully integrated into the models and if real grounding could actually be taken into account."

Foundation models, which combine text with images or videos, have an additional dimension of complexity compared to purely language models. Therefore, approaches should be developed to explain the model results and to estimate the reliability of the results for specific inputs.

Videos contain many different types of information that can be integrated into the analysis and generation process:

- OCR can be used to detect characters that are visible in the video. For example, you can integrate a text recognition model that provides text or token embeddings.
- Spoken language can be acquired through speech recognition that produces texts or token embeds.
- Faces and the persons associated with them are particularly important features for understanding a video. Here, facial recognition models can be incorporated into the video pipeline.
- Audio, for example, engine sounds, can be used in the interpretation of videos. For example, it is possible to use an audio detection system that was trained on the YouTube-8M data, where objects were annotated in the videos.
- Scenes (e.g., trees, mountains, cemetery) can be recognized using a scene model that has been trained with the PLACE365.
- Objects, such as those annotated in ImageNet.
- Movements that can be detected by action recognition models that can be trained with the Kinetics data.

An early model in this direction is Merlot Reserve, which incorporates audio, subtitles, and video (Zellers et al., 2022).

The same scene can also be captured by multiple videos, e.g. binocular cameras or multiple cameras with different field of view. The task here is to integrate these processes and achieve a uniform interpretation. Such an arrangement facilitates the reconstruction of the 3D scene.

However, video production models can also be misused to create, for example, fake, slanderous, degrading or malicious content. These hazards can be reduced, for example, by internal filters for the input texts and for testing the video content produced. However, there is a great need for research here to secure the models and to filter out hard-to-

recognize social prejudices and stereotypes. Google and Meta cite these dangers as the reason why they are not releasing their models yet.

There is already an American dominance in this area of AI foundation models. European developers need to be able to use video data to develop advanced AI foundation models. This is the only way to ensure that European values and standards are also respected in these models. For this, developers and researchers must have the best possible conditions.

6.6 Foundation models in other data domains

The proposed model development strategy of gradually expanding the multilingual foundation language models to include knowledge stocks and multimodality also provides a good starting point for the development of foundation models in very different data or scientific domains. Successful applications of models in genetics and proteomics, for example, are based on the insight that patterns in symbol sequences can also be learned in these areas, allowing new insights into the properties, effect or change of molecules. It is very likely that more such applications will be found in chemistry, materials science, life sciences or other natural and engineering sciences.

Large future applications of the technology lie in business administration and all other areas of society where processes need to be designed, controlled, optimized, automated and verified. Because processes are sequential sequences of sub-processes and individual actions, the use of transformer models to design, improve and monitor such processes is common. The digitisation of companies and administrations will provide sufficient data on the actual processes in business, production and administrative processes from which foundation models can gain a general understanding of the characteristics and elements of these processes. Such an approach to obtaining neural process models is likely to become multimodal very quickly, incorporating the linguistic naming and descriptions of process elements, as well as the information objects of the processes, such as forms, data transactions, geolocations, and transport pathways, into the learning data. GATO (Reed et al., 2022) is the first multimodal model that can process sequences of text, images and readings and derive control strategies from them. It achieved good control results on more than 600 benchmarks.

The AI data center discussed in this study would greatly facilitate the entry of the German economy and research into such new model types.

6.7 Summary

So far, the study has shown that the economic potential and societal relevance of AI foundation models is immense. But in order to realize the full economic potential, European developers need to be empowered to develop their own foundation models. Otherwise, there is a risk of becoming dependent on American models or not using them at all. Both cases represent a significant competitive disadvantage for the economy.

To train high-quality foundation models, bias, toxicity, and misrepresentation must be reduced, or at best eliminated altogether. Although there is already great progress here, only working with and researching AI foundation models helps to address this complex of issues. German and European societies should do their utmost to ensure that these developments are not left to the Americans or the Chinese.

The focus should first be on language models. Europe's multilingualism is a challenge, but at the same time a great opportunity for European language models. An adequate representation of this multilingualism in the functionality of the foundation models increases their acceptance and economic benefits. However, it is also of great relevance to the overall development of technology, as most of the world is more similar in terms of linguistic diversity to Europe than to the US or China.

Several projects are already preparing and making available multilingual datasets. Behind this is already the next generation of foundation models in the launch holes, which could produce videos, for example. Here the need for development is even greater, the potential is immense and many questions remain unresolved.

German and European developers and researchers must now be empowered to work and research AI foundation models to the best standards. The first step should be to build a competitive infrastructure for training on AI foundation models.

The next chapter now gives an insight into how this can technically succeed.

The background is a solid red color with a subtle pattern of light red hexagons. There are several abstract shapes: a horizontal bar at the top, a diagonal bar on the right, and a curved bar at the bottom right.

Prerequisites for software and personnel

.....●

7. Prerequisites for software and personnel

The development and training of AI foundation models envisaged in LEAM (Bommasani et al., 2021) presupposes an integrated and powerful hardware and software infrastructure not yet available in public investment plans in Germany and the EU. While the basic structure of this infrastructure can be derived from established structures in high-performance computing (HPC) systems, working with AI foundation models places special requirements on their organization and the software used. This chapter explains both these particular AI elements and how they interact with the basic HPC architecture, as well as the prerequisites and investments needed to make LEAM possible.

This requires simultaneous consideration of the technical prerequisites and possibilities of the LEAM initiative as well as its opportunity for the AI ecosystem and German and European society. While this chapter can show that the challenges facing LEAM are solvable on the software side, its social and economic added value is also highlighted. One objective of LEAM should be to break down the monolithically simplified professional image of "IT expertise" to differentiate and strengthen the diverse, specialized and interdependent roles necessary for a competitive AI ecosystem. As a technological, social and economic flagship project, LEAM can provide the right impetus for Germany and Europe to become a global AI competitor if the opportunities of the AI ecosystem are properly assessed and exploited.

This chapter outlines a variant for the software stack of an AI HPC system and the highly specialized occupations and skills required for it. It is possible to build structurally on existing and proven HPC systems and to add AI-specific components.⁶ Particular attention is being paid to building the stack using open source software (OSS), as the global AI ecosystem is built on non-proprietary software on the one hand and thus avoids dependence on software corporations on the other. This aspect of independence also promotes the sovereignty and resilience of the AI ecosystem, as subcontractors in the software sector can be omitted to the fullest extent, as, for example, the Federal Government's Digital Strategy provides (Federal Ministry of Digital and Transport, 2022). In actual implementation, the details may vary, but for this feasibility study, it is crucial that LEAM can be implemented at the software level with pre-existing technologies.

When looking at this infrastructure, we can distinguish four overlapping layers (layers) that interact with individual modules to enable the development and training of AI models (Fig. 18). These layers each require different specialized occupations in order to provide their complex functionality. These individual hardware and software layers are optimized for specific tasks to organize dedicated, compute-intensive applications in an efficient computing process. In the graphics, the AI elements necessary for LEAM and special for

⁶ In this chapter, we prefer to consider centralized and homogeneously organized infrastructures, although B. Yuan et al., (2022) prototypically demonstrated that AI models can also be trained decentrally and heterogeneously on networked computers. Centralized infrastructures have the advantage that the deployed software stack can be kept simpler and less error-prone, data sources are more reliable and available faster, and user rights and security measures (DMZ, VPN) are easier to handle. In addition, a centralized high-performance AI center can provide more reliable statistics on energy consumption and efficiency.

working with AI foundation models are bundled in the training and application layer, which is located on the system and framework layers of the high-performance computing infrastructure and can be controlled by the users through a service layer.

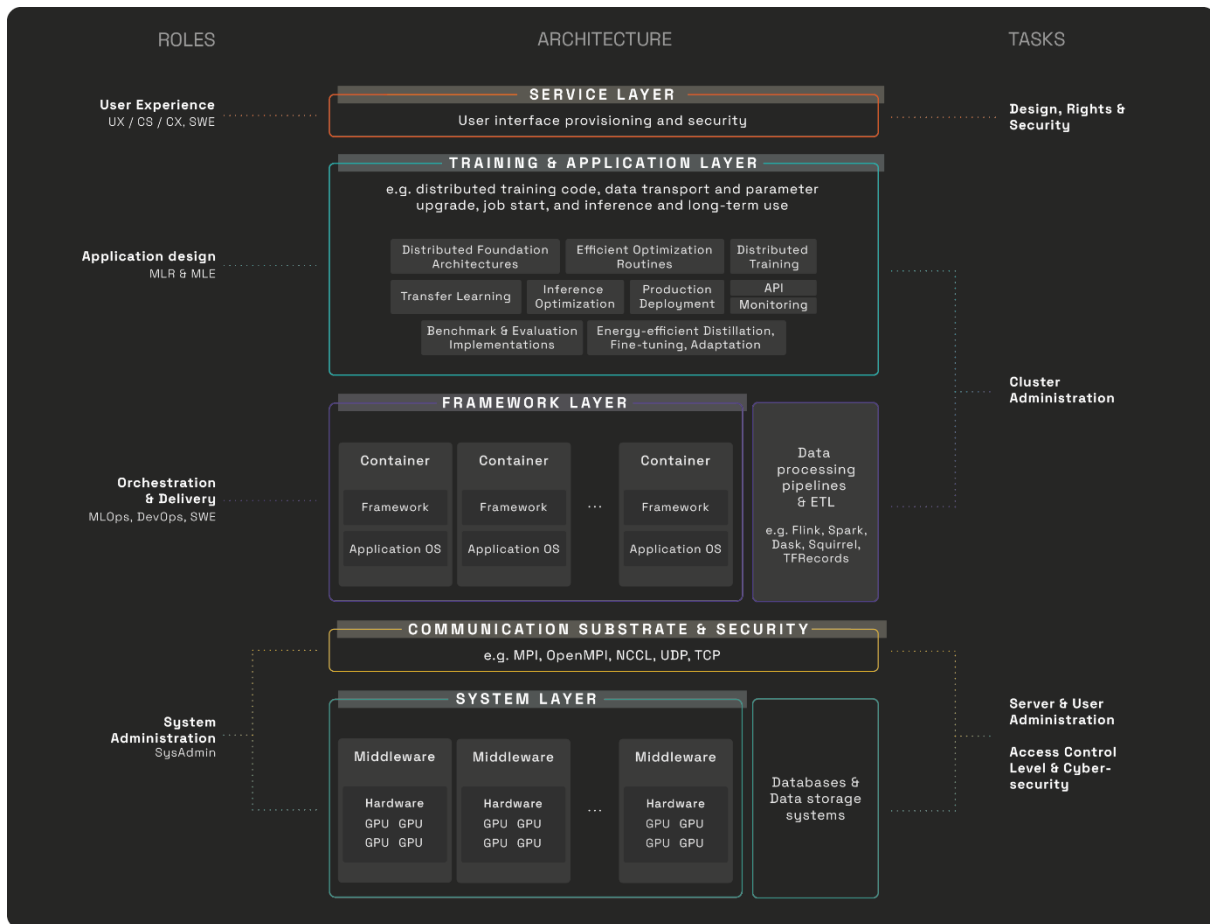


Fig. 18: Simplified presentation of the hardware and software infrastructure of HPCs

The system layer and its computing units form a hardware foundation, which is bundled into operational software systems by the framework layer. The training and application layer uses these systems to perform AI applications, while it is supplied with data from the Data Storage & Loading module. The key difference to building traditional HPC systems here is that application-specific AI accelerators such as GPUs, FPGAs, Cerebras chips, and others (see System Layer) are used instead of CPUs to develop AI models. These are necessary to enable the compute-intensive training and inference applications built on the system and framework layers to be completed within acceptable timeframes.⁷ While CPUs could be rented for data processing via a cloud solution, the several thousand GPUs and supporting CPUs that would be necessary to run LEAM should be available locally. Chapter 8 calculates a number of 4480 GPUs for this feasibility study.

⁷ With a single NVIDIA V100 GPU, it takes 355 years to train GPT-3 (Li, 2020). CPU-based frameworks are at least 5-10 times slower than GPU-based frameworks. So without the use of GPUs, it takes several thousand years of computing time to train GPT-3.

Pre-Processing	Training	Tuning	Inference
10 - 20 k CPUs Cloud hired	3.5 - 4.5 k GPUs available locally	30 - 50 GPUs Cloud / local	30 - 50 GPUs ...
Data storage space: several petabytes			30 - 50 GPUs ...
			30 - 50 GPUs

Table 5: The operation of LEAM requires thousands of GPUs for training, tuning and inference.

Therefore, LEAM is essential for the development and expansion of the innovation site in Germany and Europe, in order to build this complex and coordinated network of hardware and software and make it available to academia and business. This massive computing infrastructure is necessary for the development of AI foundation models. For example, training large AI models requires a variety of powerful and optimized AI accelerators that run both data and models in parallel, which requires a significant increase in the performance of local data centers. The reusability of these compute-intensive models is crucial in order to amortize the resources used and the costs invested.

This performance improvement can be achieved by meeting technological prerequisites, such as loading and storing data efficiently and quickly. But at the same time, it is also dependent on social changes, such as the targeted attraction of professionals to sustain the growth of the AI ecosystem. Here, LEAM should be both a thrust and an inspiration to invest in technologies that will strengthen the innovation and economic position of Germany and Europe, and to find ways to recruit and train professionals in a more differentiated and successful way.

A survey of 71 participants from the German AI ecosystem conducted for this feasibility study found that the professions of DevOps, MLOps, Machine Learning Engineer and Machine Learning Researcher are particularly important for the development and training of AI foundation models. It is also needed for the operation of LEAM Software Engineers and System Administrators, while Site Reliability Engineers were considered less important among respondents. Fifty-six per cent of respondents said it would cost more than €25 million to develop an AI foundation model, while 37 per cent expected these costs to be amortized within five to ten years.

This chapter of the feasibility studies thus demonstrates that the challenges to software for developing AI foundation models in LEAM can already be solved with the technologies and resources available today if Germany and Europe are willing to invest in the integration of LEAM. The use of open-source software in conjunction with a few selected proprietary applications also allows large AI models to be developed and trained in this country, although most of the hardware required for this is from the US and non-European countries.

One of the biggest challenges and opportunities for LEAM is the specialized and re-engineered professionals required to do so. A team of about 20 dedicated experts alongside the roles described above will be required to train the AI foundation models envisaged in LEAM. While the AI ecosystem in Germany and Europe is already run by highly qualified experts, demand still significantly exceeds supply (Streim, 2022). LEAM can help to improve the quality of training and the attractiveness of the local AI ecosystem and, above all, can make AI professionals want to stay in Germany and Europe, because LEAM will help them find conditions that currently exist only outside Europe.

7.1 Application Layer: Training & Inference Technologies

The training and development of large AI foundation models has some peculiarities compared to classical, distributed systems without naïve parallelism: Deep-learning-based AI developments are uniquely capable of designing and training AI foundation models because their algorithms can be switched in parallel. This means that a variety of data points are processed simultaneously, rather than being processed one after the other, i.e. sequentially. This feature provides a significant advantage over classical HPC applications, allowing specific algorithmic advantages to be exploited to enable training of large AI foundation models at all only in finite time.

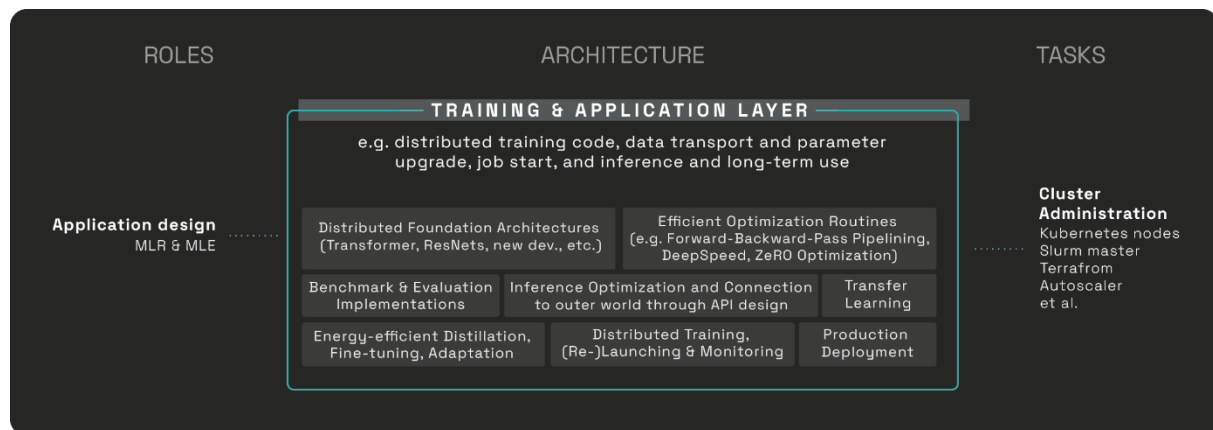


Fig. 19: The architecture of the training & application layer in detail

Data & Model Parallelism

These properties can be found in the "data" or "data" section. Summarize "model parallelism" and explain why it requires a massive computing infrastructure (Hugging Face, o.D.). While in data parallelism training data can be divided equally across the AI accelerators, the number of parameters to be calculated in AI foundation models is so large that even in model ⁸ the models need to be divided across multiple accelerators. This means that not every accelerator computes an identical copy of the model in parallel and iteratively transmits its results to all parallel accelerators, but the data to be computed must be transmitted sequentially from one accelerator to the next. In order to update the model parameters, the last accelerator in a row transmits the error signal (loss) back to

⁸ Narayanan et al. (2021) describe model or pipeline parallelism in detail.

the first accelerator. Data and model ⁹ can be used simultaneously, further increasing the complexity of development.

However, with such circular sequencing, most AI accelerators remain unused when they do not perform any operation in the row circuit. This lactation time can be minimized in model parallelism through refined programming (Kosson et al., 2021) and through optimization systems that ensure maximum utilization of all accelerators (see Chapter 8.3). These prerequisites make programming and supervising the training of AI foundation models complicated and challenging (Bommasani et al., 2021).

⁹ Mudigere et al. (2022) describe best practices in training large Recommender models.

Training Management, Evaluation & Benchmarking

The application layer also controls the reliable management of individual training jobs in order not to lose results in the event of a hardware failure. Unlike training smaller AI models, training AI foundation models often involves a failure of individual AI accelerators.¹⁰ A reliable launch and relaunch system prevents loss of result through continuous monitoring in order to continue training from the same location where the error occurred.

Furthermore, the evaluation and benchmarking of AI models are also implemented in this layer. This includes both monitoring during the training process and testing the models against application-relevant datasets in order to determine their performance and later translate them into Data & Model Card documentation (Pushkarna et al., 2022). Connection to the ETL ¹¹, which is implemented, tested and versioned here, is also required for training and evaluation.

Reusability in deployment infrastructure

Due to the compute-intensive training of AI foundation models, the reusability of costly trained AI models is central to amortizing the resources and costs required to do so. This includes the use of distillation mechanisms, adaptation or fine-tuning processes (see Chapter 2), which in turn require application-relevant benchmarks.

The development and maintenance of an optimal deployment infrastructure plays a central role in this, as it allows for significant efficiencies. The focus here is on an implementation to enable the AI foundation models to operate in the best possible way with the underlying hardware and to respond quickly and in a timely manner to external requests via API.¹² Complex aspects of production must be taken into account in order to ensure a smooth process even under stress. For example, the deployment infrastructure should be resilient to a cluster of requests, be equipped with a stable API, and include robust access restrictions and security checks.

¹⁰ This is a well-known phenomenon from the datacenter industry: The more hard drives you run, the more often they fail. Accordingly, the same applies to HPC accelerators.

¹¹ ETL stands for Extract-Transform-Load and describes the aggregation of raw data from a production database into a format that can be used to analyze the data.

¹² Examples include freeware libraries such as TensorRT and Triton, which are provided by NVIDIA.

Resource efficiency of training

The resource-intensive training of AI foundation models can become a problem if the training process is not continuously monitored and energy consumption is reduced with appropriate measures. Sustainability must therefore be considered from the outset when designing a LEAM AI high-performance data center and when developing and training AI models.¹³

The resources consumed in AI development can be captured at the software level through the regular and ongoing measurement of CO₂ equivalents. The measured CO₂ equivalents on three different components are checked to map them to the correct source (Hintemann, 2020).

During training and inference, the AI accelerators require energy to perform the computational operations, called MACs.¹⁴ These can be detected at the application and framework levels (Bannour et al., 2021; CodeCarbon, 2020). Due to the high volume of data, the Data Storage & Loading layer also consumes a significant share of energy consumption for the provision, processing and transport of data. The reliable cooling of the entire system and other smaller components also require an uninterrupted power supply. Chapter 8.6 explains how the high-performance centre itself can be designed to be resource-poor/climate-neutral.

The various measurements can be passed directly to the service layer to ensure that the user has insight and control over the energy consumption of the overall system and the training process. Uncontrolled overconsumption can be prevented by warning signals. Some of these control data are already being made available by hyperscales.

Prerequisites

The Lingua Franca for the development of modern AI systems is Python (van Rossum, 1995), with AI applications increasingly being developed in other programming languages. Many languages are involved in the development of the service layer and are not intended to be listed here. The software libraries used in training and inference are largely, if not entirely, covered by open-source software. Examples at the framework level include PyTorch (Paszke et al., 2019), TensorFlow (Abadi et al., 2016), Keras (Keras, 2015/2022), and Jax (Frostig et al., 2018).

When orchestrating, the application layer sometimes uses the same software as the framework layer, e.g. Ray (Moritz et al., 2018), Slurm (Yoo et al., 2003), kubeflow (*Kubeflow*, 2017/2022), hydra (*Hydra*, 2019/2022), Abseil (Abseil *Python Common Libraries*, 2017/2022), and others (see Chapter 7.4). Numerical libraries such as BLAS ("An Updated Set of Basic Linear Algebra Subprograms (BLAS)", 2002), NumPy (Harris et al., 2020), and others will also be used. In ML-specific monitoring, libraries such as AimStack (Arakelyan et al., 2020) or MLFlow (MLflow, 2018/2022) can be used. This list is purely exemplary and

¹³ See e.g. (CSTB *Releases Report Fostering Responsible Computing Research*, 2022; Patterson et al., 2022) and the references contained therein.

¹⁴ MAC stands for Multiply-Add Compute, the central computing operation of computer processors.

is only intended to illustrate the importance of open source software for the development of modern AI applications.

The tasks in the application layer require a high degree of specialization, which is usually carried out by computer scientists and software developers in shared roles. These can be distinguished into four role profiles: Machine Learning Researchers design and develop new algorithms, while Machine Learning Engineers work on optimizing training and inference code and implementing distributed learning. Together, they form a team to scale AI models.

Full-stack and back-end engineer develop the service layer and work with system administrator and dev-ops expert, among others, to pair the user interface with the various functionalities of the overall system. Data Engineers and Data Scientists are working on the data storage & data loading layer and its infrastructure. This includes designing the databases or other storage units, developing the ETL pipelines, as well as controlling the data loading pipelines and correctly designing the training records.

SPOTLIGHT

Fyrfeed GmbH

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Fyrfeed's founding team:
Ehud Alexander Avner,
Dr. Thomas Lindemann,
Benjamin Zengler

Fyrfeed creates content for B2B businesses through a combination of expert writers and Artificial Intelligence.

As a result, clients only need to invest 5 minutes per month and save 80% of costs compared to agencies.

Where do you use foundation models? What is your use case?

In addition to other technologies and tools, we also use large, general-purpose language models to help expert writers create high-quality content - such as social media posts, blog articles or whitepapers. This approach, where humans and AI work together, is called Human-in-the-Loop.

What impact do AI foundation models have on your business model or your projects?

AI foundation models allow us to deliver high-quality texts quickly and at a low cost. Instead of having to re-train models from scratch on our own, which involves significant cost and time, we can use general, pre-trained models right away. This makes further product development and testing of new applications many times easier.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

As we serve customers from all over the world, it is conceivable that we will use different models for different languages. Although existing state-of-the-art language models are multilingual (and we do so very successfully), the quality can vary between languages (and disciplines). Models that specialize in European (and non-European) languages - while also offering all the benefits of open source - would be a big win.

7.2 Data-Storage & -Loading-Layer

The success of many AI foundation models depends substantially on the efficient and rapid processing of the data to be loaded. Both in the ETL and during the training of the models, efficiency is crucial. This gives the data-storage & loading layer a special importance, which is often not the case in other HPC systems.

When training AI foundation models, data loading is accelerated using a multi-layered process: The ETL process, as well as the filtering and processing of the raw data, are performed sequentially, so that the data can be loaded with as little latency as possible and distributed in the data parallelism. For ETL, large, classic CPU servers are often used. They need to be able to process peta and exabytes of data, for example, when scraping web pages and map-reduce jobs (Dean & Ghemawat, 2004) and similar, massive operations. This involves the use of specialized ¹⁵ in the data infrastructure, such as Spark (Apache *Spark*, 2014/2022), Flink (Apache *Flink*, 2014/2022), or Dask (Dask, 2015/2022).

The raw data provided by ETL is then stored in either a database or a cloud-based storage layer. It is possible to generate a pre-filtered dataset from these raw data and also to store it locally or in the cloud. Pre-filtering requires similar tools and resources as ETL, so these are already in place. This type of data processing will later speed up the training process, but will require an increased amount of memory.

While training AI foundation models, it is important to transport the data to the AI accelerator efficiently and as much as possible without redundancy. This requires not only an appropriate network architecture (see Chapter 8), but also a dedicated software stack that has been optimized for loading data in distributed systems. It is essential that data, because of its size, cannot be loaded and stored on individual machines, but must be provided through an efficient streaming system, which makes development significantly more difficult.

The data is periodically streamed over and over in random order (Nguyen et al., 2022) when requested by varying AI accelerators. The software stack must be able to read from both local databases or hard drives and different cloud storage systems to meet any user requirements.¹⁶ Such systems are distributed, for example, by Amazon, Google and Microsoft, or are privately operated by groups. Likewise, loading requires support for filtering functions, data augmentations, and transformations, as is common in Deep Neural Networks training. This is required to train models robustly and generalizably. The software stack must also be able to combine multiple records together to avoid new ETL jobs and thus save resources.

¹⁵ These and the following lists do not claim completeness and only exemplary elements of each category are mentioned. The exact choice of tools will ultimately depend on the implementation details and is not intended to be determined at this time.

¹⁶ Experience shows that expectations of data sovereignty are very heterogeneous. In order to achieve broad acceptance and use, the system must therefore be compatible with as many scenarios as possible.

The data storage and loading layer must also be compatible with local user rights as well as applicable data law. As a library, the data infrastructure must take into account Access Control Levels ¹⁷ to secure access to data and thus prevent (un)intentional access to data by third parties. This is controlled via an interface to the service layer where user management is organized.

International data law requires that the storage and processing of data must be based on strictly regulated practices when it relates to the processing of personal data, such as those governed by the General Data Protection Regulation (*Data Protection in the EU*, o.D.). In addition, certain certificates are also required or expected by the market in different geographical areas, such as SOC-2/SOC-3 (*System and Organization Controls*, o.D.) in the US and North America, or ISO 27001 (ISO - *ISO/IEC 27001 and Related Standards - Information Security Management*, o.D.) or ISO 27017/27018 (*ISO 27017 and ISO 27018 Certification* / *DEKRA*, o.D.) in the European area. Data with certain characteristics, e.g. Health or credit information is subject to other locally regulated regulations, such as the HIPAA or FCRA regulations in force in the United States.

In addition, it may also be necessary for the data layer to be able to access the data within a Demilitarized Zone (DMZ) or in a Virtual Private Network (VPN) if the data cannot be temporarily stored within the high-performance center. However, this should be avoided in practice, as the associated latency can severely affect the speed of loading.

Prerequisites

Special skills of Data Engineers and Machine Learning Engineers are required to develop and maintain the data systems to ensure ETL and stream-based, distributed loading of the data to the accelerator nodes. Embedded software includes, for example, Spark, Flink, or Dask for ETL or torchdatasets (Maske, 2019/2022), Squirrel (Sohofi et al., 2022/2022), Deep Lake (Deep *Lake*, 2019/2022), ffcv (*FFCV*, 2021/2022), and others (Ofeidis et al., 2022) for distributed data loading. These are usually free-to-use open-source software, which is not a challenge to obtain. Classical distributed databases such as HDFS (Apache *Hadoop*, 2014/2022) or GlusterFS (GlusterFS, 2011/2022), cloud-storage systems, similar to Google Cloud Storage, Azure Blob Storage, or NAS systems can be used to store the data.

¹⁷ A distinction is made between Read, Write, Execute and Discoverable. So the permission to read, write data, run applications, or - less known - the ability to find information about the presence of data without being allowed to read it.

7.3 System Layer

The system layer organizes the actual AI accelerators and the hardware structure needed to deploy them to provide high-performance, low-loss computing power. AI accelerators ("AI Accelerator", 2022) are typically specialized processors that have been optimized for the development and training of AI models (Reuther et al., 2022). The use of these processors is essential to manage the time-consuming computing effort for AI foundation models as well as to remain competitive (Khan & Mann, 2020). The world's leading manufacturers of high-performance processors include US companies NVIDIA (Campa et al., 2020), AMD (AMD, 2021) and Intel (Intel, 2022), but also start-ups such as Tachyum (Tachyum, 2022). Edge and embedded systems, such as NVIDIA Jetsons or FPGAs, are not considered here as they are not suitable for training AI foundation models.

These processors are controlled using specialized libraries, often developed by hardware manufacturers themselves. Examples include CUDA (NVIDIA Developer, 2013) and ROCm (AMD, o.D.), which are usually provided as ¹⁸ with the hardware. The above server operating system controls physical hardware using these libraries, almost always using Linux operating systems, such as distributions of the Debian (Ubuntu), Redhat (Fedora, CentOS) or Arch families (Joseph et al., 2022). The individual server units, which bundle multiple AI accelerators, are then incorporated into the communication substrate by means of corresponding middleware, as described in detail in Figure 20.

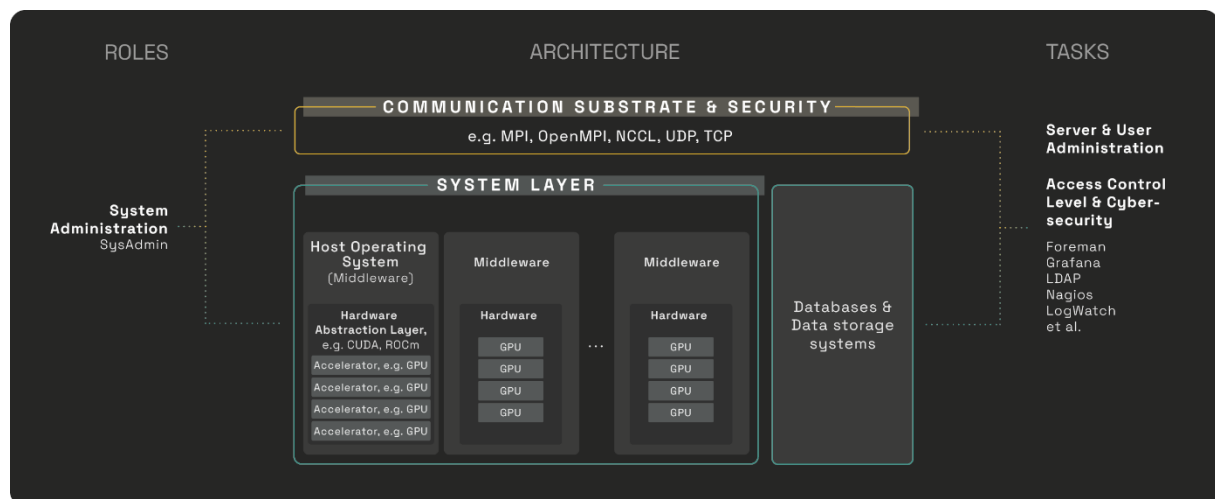


Fig. 20: The architecture of the system and data storage & loading layers in detail

¹⁸ Compared to OSS, freeware libraries only have the compiled binaries available, not the entire human-readable source code.

Communication Substrate

The communication substrate is used to seamlessly connect system and framework layers, but also to provide databases and data stores to AI accelerators. In this way, the communication substrate can link several server units to each other in order to efficiently process large amounts of data. The communication libraries typically used for AI applications are based on specialized technologies, such as MPI (Message Passing Interface, 2022), OpenMPI (Open MPI, 2022) or NCCL (NVIDIA, o.D.).

The exact topology of the substrate is given by the actual physical wiring of the hardware and must be considered when building a high-performance center. At the outset, it must be made clear which applications will be used to calculate the AI accelerators.

Prerequisites

Trained professionals as system administrators are required to design, build and maintain the system layer and communication substrate. Specialized control software, which can consist of proprietary and open-source (OSS) software, is used for the maintenance. In general, the hardware and software infrastructure for AI high-performance computing applications is predominantly comprised of OSS, which highlights its importance to the AI ecosystem (Sonnenburg et al., 2007).

In addition, the system administrators are also responsible for granting rights and controlling users, e.g. through appropriate access control levels using LDAP or AD systems, but also server administration and monitoring to ensure the permanent health of the system. In addition, the administration of DMZs and VPNs as well as cybersecurity measures also fall within the remit of the system administrator.

7.4 Framework- & Service-Layer

Framework-Layer

The specialized AI technologies are used in the framework layer above the communication substrate (Fig. 21). For this purpose, AI accelerators, grouped in the system layer as server units, are again abstracted by container solutions such as Docker (Merkel, 2014) to normalize and homogenize the environment to the corresponding AI systems (Carpintero, 2021). This allows the development of a wide variety of AI systems, although the underlying hardware remains the same.

The container environment uses another proprietary Linux-based operating system, equipped with high-performance libraries such as BLAS or cuBLAS ("Basic Linear Algebra Subprograms", 2022), as well as programming languages such as Python, C, and C++, with their compilers, necessary for the development of AI models. Furthermore, it also contains the frameworks that are specific to deep learning applications and with which AI applications can be programmed. These include, for example, PyTorch, TensorFlow, Keras or Jax (Gopani, 2021).

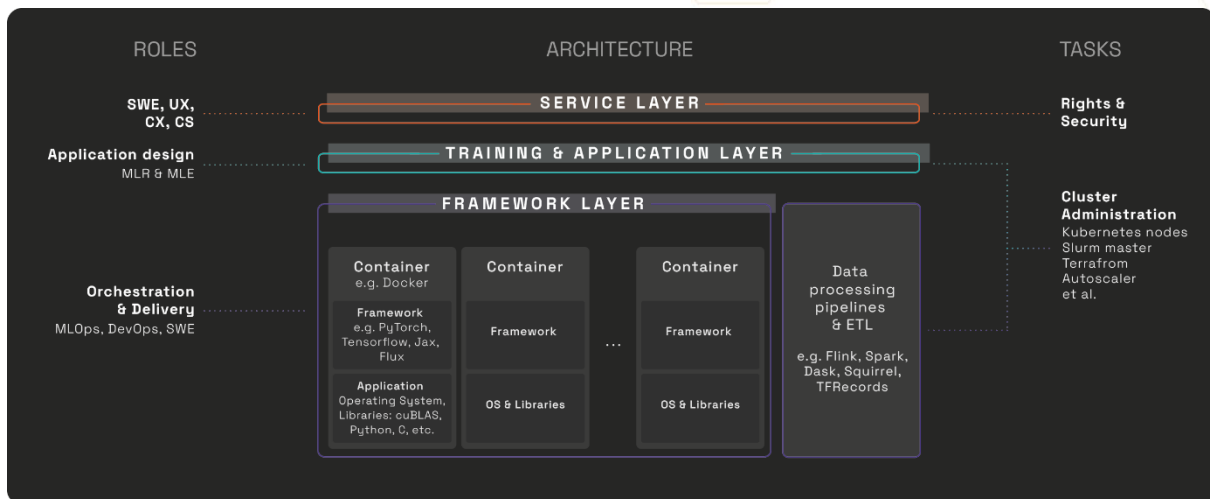


Fig. 21: The architecture of the Framework & Service Layer in detail

Service Layer

Through the hardware and software infrastructure required for the operational development and training of AI foundation models, some of the services included are mapped in the service layer via user interfaces. For example, unchanged AI models are used for their own use-cases, or AI models are automatically adapted to the data sets provided. However, the service layer can also handle administrative processes such as controlling users or requesting additional resources.

Prerequisites

The framework layer controls cluster orchestration in order to launch tasks in the system on a targeted basis and to use available system resources efficiently. A job management system is used to do this, typically based on SLURM, Kubernetes (Kubernetes (*K8s*), 2014/2022), Terraform (Terraform, 2014/2022), DockerHub, or other components (Mujkanovic et al., 2020). This allows orders to be lined up and prioritized individually. In addition, the framework layer also includes a component specific to deep learning applications to accurately and effectively load data from the data storage systems into the computer environment (see Chapter 7.2).

This wealth of specialized tasks requires professionals trained in the partially automated operationalization of development (DevOps) and machine learning (MLOps) environments. In doing so, it is necessary both to ensure the continued health of the environments, and to keep informed of the development and re-development of software solutions for the framework layer, in order to make improvements to the system upon in-depth review. Above all, software engineers and DevOps experts are needed to orchestrate and prioritize clusters.

7.5 LEAM as a lighthouse project for the future of the AI ecosystem

However, a high-performance AI software environment necessary for the design, training and maintenance of AI foundation models requires not only technical and infrastructural requirements, but also many and diverse requirements for well-trained personnel, which are currently barely or not met in Germany and the EU. For example, studies on the skill shortage in the IT services industry show a shortage of specially trained professionals, but also a lack of differentiation between individual highly specialized roles that necessarily need to work together to develop and train AI foundation models (Hickmann & Koneberg, 2022). This makes it difficult to address and advertise the offspring accurately.

In practice, work on AI in general and AI foundation models in particular is organized into teams that bring in different skills and training. Specialized teams must master various programming languages as well as mathematics, data science, and data engineering, but also combine in-depth knowledge of computer science, statistics, and knowledge of software, hardware components, and architectures: "Machine learning specialists work in teams - with ML specialists at the top, software developers in large numbers at the grassroots level, and data scientists and engineers in between - this is the typical staffing structure of advanced technology clusters such as Silicon Valley" (Philippe Lorenz & Kate Saslow, 2019, own translation).

Today, however, these diverse teams are understaffed in Germany, so that "nine out of ten vacancies [in computer science] did not last time have adequately qualified unemployed" (Hickmann & Koneberg, 2022). Bitkom President Achim Berg said: "The worsening shortage of IT specialists is growing into a very real threat to Germany's major transformation tasks." (Bitkom e.V., 2022a).

In a three-part study, the Center for Security and Emerging Technology (CSET) at Georgetown University identified how rapidly the need for AI experts is growing: Between 2015 and 2019, the share of total workers directly involved in the U.S. AI ecosystem grew from six percent to nine percent – to 14 million workers (Gehlhaus et al., 2021). Within the next ten years, this occupational group is expected to grow even twice as fast as the federal average (Gehlhaus et al., 2021). But unlike in the United States, bureaucratic barriers to hiring and linking foreign professionals to the labour market can be much higher in the United States.

In schooling and training, there is a need to differentiate more and train the different roles in the AI ecosystem in a more targeted way. Here, in addition to universities, especially technical colleges and colleges of applied sciences can invest more in attractive offerings for specialized fields, while universities should have more funds available to train ML specialists (Wannemacher & Bodmann, 2021).

The education offer and the US labour market are several steps ahead of us. In addition, CSET further noted that many career paths in the AI ecosystem do not require several years of university education or even a doctorate. In the study, "Training Tomorrow's AI Workforce", CSET advocates a new strategy to better empower non-university educational institutions to train AI experts, promoting them at the federal and regional levels and

better linking them to locally based businesses (Gehlhaus & Koslosky, 2022). In this way, students and employers are better placed to operate regionally integrated AI ecosystems.

In addition, CSET advocates a shift away from far-reaching, long-term courses in computer science towards a greater variety of modularly structured continuing education courses in order to specialise more quickly and thus be available to the labour market. According to CSET, this policy shift must be managed out of the US government by creating an AI expert training staff in the Federal Cabinet (Gehlhaus & Koslosky, 2022). An equally centralised, empowered staff office would also be an important signal in Germany and the EU to be able to properly assess the shortage of skilled personnel in detail and to improve the supply of AI training on a mass and in-depth basis.

It is precisely this economic development of the AI ecosystem towards a central cross-sectoral industry that permeates all other sectors of the economy that underlines the importance of LEAM as an enabler of German and European science and business. In addition to its economic potential, LEAM can also be an economic policy lighthouse for training, recruiting and long-term engagement of experts, as well as for inspiring young people to learn IT professions. Thus, LEAM can not only set a strong mark on the competitiveness of the innovation location in Germany and Europe, but also fulfill an educational mission in society that is necessary for the sustainability of our economy beyond its practical application.

7.6 Summary

The challenges of the LEAM-proposed high-performance AI computing center for developing and training AI foundation models are technically solvable at the software level. The necessary technologies have already been developed and are mostly available as open source software. So the main challenge is not to develop the software, but to operate such a high-performance AI data center, which requires a large number of highly specialized experts to work together in teams.

If successful, LEAM can launch scientific breakthroughs from Germany and Europe by robustly scaling modern, massive hardware AI models with complex software already available today. In this way, LEAM could not only make German and European science and business future-proof and competitive, but also provide added value for our society through its actual, technical implementation in connecting professionals and solving societal problems.

Building a high-performance AI data center

.....●

8. Building a high-performance AI data center

The experts interviewed for this study cited the lack of high-performance computing (HPC) infrastructure for computing such models as one of the main reasons why regular training of large AI foundation models à la GPT-3 is not possible in Germany and Europe respectively. Because AI foundation models place special demands on computing capabilities in particular. Unlike traditional HPC systems that operate on the basis of a central processing unit (CPU), in the field of artificial intelligence, and specifically in the computing of large AI foundations models, exceptionally high computing capacities of up to 4500 graphics processing units (GPUs) are required, which, despite this high capacity, require very long runtimes to handle the necessary computing tasks. This type of concentrated computing capacity is still unparalleled in Germany and the EU so far. As industry and academia lack access to these computing capabilities, breakthrough innovations in AI foundation models are severely hampered.

8.1 Definition Data Center

A data center (RZ) is an infrastructure facility that brings together computers, servers, storage systems and other technology components to process and store a large amount of data and applications (Hintemann & Clausen, 2018). Data centers typically serve as a central resource for processing and storing data and applications in companies, organizations, and institutions. They can also be used for a variety of purposes, from processing transactions and providing IT services to running AI applications.

There are different types of data centers, which differ in size, purpose and technical equipment (see Table 6).

Data Center	Description
Enterprise data centers	Enterprise data centers are data centers operated by companies and organizations to enable the processing and storage of data and applications for their internal purposes. Enterprise data centers can come in a variety of sizes and shapes, from small rooms with a few servers to large facilities with thousands of computers and servers.
Collocation data centers	Collocation data centers are data centers that are operated by third parties and leased to companies and organizations. Collocation data centers provide tenants with space for their own servers and other computer components, and provide the necessary infrastructure such as power, cooling, and network connections.
Hyperscale Data Centers	Hyperscale data centers are data centers operated by hyperscale companies to enable the processing and storage of data and applications for their customers. Hyperscale data centers are often very large and offer high computing power and storage capacity to process and store large amounts of data quickly and efficiently.

Table 6: Examples of data centers

In addition to the business model, data centers can also be distinguished by their size. The following dimensions can be used as a reference here (*The role of data centers in an interconnected world*, o.D.):

Art	Sizes	Performance recording	Focus
Micro Data Center	From 1 server rack upwards; matching configurations for a container	100 kW	EDGE applications that benefit from proximity to the IoT source
Small data centers (also EDGE data centers)	approx. 500 m2	1 MW	Often an in-house data center for critical data
Medium Collocation/Hosting Data Centers	approx. 10,000 m2	10 MW	Multi Tenant Collocation RZs as well as focus on hosting and managed services
Large collocation data centers	approx. 50,000 m2	50 MW	Multi and Single Tenant Collocation as well as large hosting/managed service providers with an international focus
Hyperscaler	100,000 m2 and more	100 MW and more	Large, globally operating cloud providers, often operating in multiple locations (10-200) worldwide

Table 7: Data center sizes

Delimitation of AI high-performance data centers

LEAM plans to compute large AI foundation models, which requires special hardware. The high-performance AI data centers required are distinguished primarily by their focus on processing AI applications. Other data centers are mostly geared towards the processing of general data and applications, and can be used for a variety of purposes. An AI high-performance computing center, on the other hand, is specifically designed and equipped to process AI applications, providing the computing power and environment required to train AI models and algorithms quickly and efficiently.

AI high-performance data centers also differ from other data centers in terms of their technical equipment and operation. For example, AI high-performance data centers may have higher computing power and special hardware such as graphics processors (GPUs) that are particularly suited to processing AI applications. You may also have dedicated software tools and environments that allow you to develop and train AI models and algorithms.

8.2 Requirements for an AI high-performance data center

Overall, therefore, AI high-performance data centers are tailored to meet the specific requirements of AI applications and provide the infrastructure and resources required to effectively process AI applications. The following describes the concrete requirements for an AI high-performance data center to compute large AI foundation models.

Processors

Graphic Processing Units (GPUs) are suitable for computing AI applications. GPUs are specifically designed to handle graphics tasks. They have a variety of computing cores, which allows them to process large amounts of data quickly and in parallel. This allows them to quickly handle tasks such as rendering 3D graphics or training models for machine learning and artificial intelligence. However, many data centers use CPUs (Central Processing Units) instead of GPUs. CPUs have a smaller number of computing cores (6-12) and can handle a wide range of tasks, but are not particularly well suited for tasks that require a high level of parallel processing.

Network requirements

In addition to the performance of the processors, the connection technology and bandwidth used between the GPUs and the bandwidth to which the individual nodes are connected is crucial to the calculation of large AI foundation models. It is imperative to distribute the workload on a large number of GPUs. According to the expert interviewed, this requires an InfiniBand or fast Ethernet interconnection and a strong bandwidth within the cluster, but also between clusters and storage. InfiniBand is particularly well suited for use in the HPC field and is up to 10 times more powerful than the common interconnect PCIe. In addition, storage environments with low latency should be included (Hensel & Ostler, 2020).

Power densities

Power density is understood as the electrical absorption of the IT components per server rack or per m² of whitespace (area required per rack). According to the interviewee, power densities per rack in the High Performance Computing sector are steadily increasing and currently range between 20-30 kW per rack and occasionally higher. The power density required to calculate foundation models is between 30 and 45 kW (e.g. NVIDIA Superpod) per rack, and thus in the High Performance Computing range. The power density also influences the choice of cooling used in the systems.

Cooling

Nearly 100% of the electrical power absorbed by the server hardware is converted into heat. The higher the power density of a server rack, the higher the amount of heat delivered. However, computers must not exceed the operating temperatures allowed for them in order to ensure safe operation. Manufacturers of IT equipment therefore usually prescribe the temperature ranges and humidity conditions to be observed, to which in turn the warranty services are linked. Many of the climate engineering requirements for operating IT equipment in server rooms are reflected in the ASHRAE Data Center Power Equipment Thermal Guidelines and Best Practices, which are followed by IT manufacturers (ASHRAE, 2016). In order to keep temperatures within the allowed range, there are various cooling options that are used in data centers. In general, a distinction is made between methods of air cooling and liquid cooling (water cooling).

Because coolants are physically denser media than air, they can usually carry off significantly higher heat loads.

The ever-increasing power densities in the racks and the resulting amount of waste heat make it more challenging to cool systems efficiently with air cooling, especially in the HPC area. According to the expert interviewed, the air cooling of a rack with a power density up to a maximum of 20 kW per rack can be implemented. Thus, 300 m³/h of air corresponds to a water flow rate of 90 l/h (Dürr, 2018).

Only direct-to-chip liquid-cooled or liquid-cooled racks are suitable for LEAM's 36 kW per rack power density. Liquids are a more efficient heat transfer medium than air and are therefore better suited to high power densities. However, according to experts, very few data centers have been equipped with liquid cooling or hybrid solutions of air and liquid cooling. The building infrastructure is now being planned for new data centers by data center operators and collocation providers to allow the installation of liquid cooling.

Cooling method	Description
Air cooling	<p>The most common method of air cooling is room-based double-floor cooling in a cold and hot-drive configuration. In this process, cool air is passed through outlets in the double floor into the so-called cold corridor in front of the racks and sucked into the server cabinet through the fans of the servers. The heated air is discharged from the back of the rack via the heater and is fed to cooling again.</p> <p>There are also concepts where a double floor is avoided. In this case, the cold air is blown into the server room from the side and the heated exhaust air is drained out of the room via a special "double-floor ceiling".</p>
Direct-to-Chip/Direct-to-Plate <i>Liquid cooling</i>	<p>In direct-to-chip cooling, a cooling fluid is piped into small tubes or copper tubes to cooling plates located right next to the components to be cooled (e.g., CPUs and GPUs). The liquid thus heated is discharged through a heat exchanger and further downstream cooling circuits.</p>
Immersion Cooling <i>Liquid cooling</i>	<p>In so-called immersion cooling, the electronic components, usually the complete board of a server, are immersed in a container containing a special non-electrically conductive liquid, which is the coolant. The liquid absorbs the heat and is then discharged through a cooler to dispense the heat.</p>
Cooling via back doors <i>Liquid cooling</i>	<p>Water-cooled back doors of server cabinets offer the advantage that conventional server technology can be built into the cabinets. The heated exhaust air conveyed by the fans is absorbed through heat exchangers into the rear rack door. The connected piping drains the heat from the server room. As this technology uses both an air stream to transport the heat to the heat exchangers of the back doors and a water circuit to further transport it from the server room, these systems are a hybrid approach.</p>

Table 8: Overview of cooling facilities in data centers

Building infrastructure

The building infrastructure requirements of an AI data center are not fundamentally different from those of other data centers. However, it should be noted that before a new data center is built, the cooling technologies will be planned and the construction of the building will be adapted accordingly. The computing resources required for the LEAM project and the power densities used of approximately 36 kW/rack can only be cooled by liquid cooling. Therefore, the building infrastructure should be equipped for the use of liquid cooling. This should be taken into account when building a data center as well as when choosing a suitable collocation/housing provider.

Area requirements

The power density per rack has a decisive influence on the area required. When the power density per rack is increased at a predetermined total data capacity, the number of racks required decreases and the area required decreases. On the other hand, if the power density per rack is reduced, more racks will be needed to achieve the desired overall power, and thus the area requirement will also increase. According to the expert interviewed, the area required per rack (so-called whitespace) is usually between 3.0 and 3.5 m² gross (the net standing area of a rack is about 1 m²). Thus, for the liquid-cooled 140 racks (e.g. 4 NVIDIA DGX H 100 nodes/rack) calculated by the LEAM initiative, with a total output of approximately 4.0 MW to calculate and train the AI foundation model, the area required is 345-525 m². The use of older nodes (e.g. NVIDIA DGX A 100) in air-cooled racks results in a higher number of racks and a correspondingly larger area, as fewer nodes can be installed per rack.

Power supply

According to the interviewed experts, the basic requirements for AI high-performance data centers are comparable to those for the power supply of other data centers. However, high-performance data centers are characterized by a higher power density (kW/rack) and consequently a significantly higher waste heat load per server room. The power supply of a data center usually consists of several components that work together to ensure that the data center is supplied with sufficient power.

These components usually include:

1. **Grid supply:** This is the point where the power needed to operate the data center is delivered. As a rule, electricity is sourced from the public grid. Micro and small data centers are usually supplied with 400 V AC power via the normal home connection. In the case of larger take-off volumes, the input is usually 10 kV or 20 kV (medium voltage).
2. **Power distribution:** The main power supply is distributed through cascading gradients from MSHV (medium voltage head distribution) and NSHV (low voltage head distribution) to the respective power supply level connection points in the data center. The power distributors distribute the power to the different areas of the data center, ensuring that the power reaches the required locations.
3. **Uninterruptible power supply (USV):** In order to ensure that the data center can continue to operate even in the event of short-term and long-term power outages, a USV must be installed according to the BSI INF.2. The USV provides a continuous power supply to the data center through battery buffers or swing mass systems until the main power supply is restored. Other tasks of an UPS are the smoothing of voltage shocks (surge; < 4ms), the relief of upper vibrations, or the galvanic separation of the internal circuit from the external circuit. The use of a USV is mandatory in Germany.

The sustainability of electricity supply is discussed in Chapter 8.6.3.

Energy Consumption & Efficiency Parameters

The Power Usage Effectiveness (PUE) is a technical indicator introduced by the industry consortium *The Green Grid*, which puts the energy used by IT in the data center in relation to the total energy consumption. In theory, the optimal PUE value is 1.0 (total power consumption is equal to pure IT consumption without any other energy costs such as cooling or the loss of power). The average PUE of newly built data centers was less than 1.5 in 2015, while the average PUE of air-cooled stock data centers in Germany was 1.8 in the same year. In 2010, the average PUE of German data centers was still 1.98 (Stobbe et al., 2015).

The PUE value of a data center is primarily influenced by the climate zone in which the data center is operated, as well as the cooling technology used, followed by the energy costs involved in running an UPS. In warm climates (e.g. Mediterranean), the energy expenditure for cooling is inherently higher than in colder climates (e.g. Scandinavia). In terms of electricity price developments, lowering the PUE is the primary way to reduce operating costs (Lamonica, 2014). Air-cooled data centers average a PUE of between 1.5 and 1.2, while liquid-cooled data centers can reach a PUE of up to 1.06 (PUE values in Google Data Center). German data center operators confirmed this in the survey and indicated that the PUE value for new water-cooled systems is already up to 1.1. Currently, the use of waste heat is being concentrated here in order to become climate-neutral overall.

The calculation method KPI4DCE (Key Performance Indicators for Data Center Efficiency) developed by the Federal Environment Agency is a holistic approach to calculating the

energy efficiency of data centers, which also includes the life cycle of the IT equipment and the technical supply structure. This calculation method is in theory more meaningful than the PUE value alone, but it is also more elaborate. Not all data centers provide the required readings. The goal is automated measurement (Schödwell et al., 2018). However, at EU level, the *Climate Neutral Data Centre* initiative is considering monitoring data centres with these indicators in the future (*Climate Neutral Data Centre Pact - The Green Deal Need Green Infrastructure*, o.D.).

Scalability / Modular construction

The scalability of an AI high-performance data center is subject to the same legalities that apply to the scalability of other data centers. Scalability is largely influenced by three factors: expandable vacant areas, the maximum power density per rack, and the maximum guaranteed power output (as opposed to current consumption). For example, if a data center has already fully utilized its existing space, the power density in the racks can be increased by checking the existing cooling and power supply. If the available server rooms are not fully utilized, the overall performance of the data center can be increased by commissioning additional rooms. Since 2021, there has been a slight decrease in the total IT area of 2.1 million m² in Germany due to an increase in power densities. These consolidation efforts are most evident in the reduction of space in the traditional, often corporate-owned, data centers. A migration of installed IT performance to cloud and edge operating models is evident here (Hintemann et al., 2022). The IT area as a measure for describing the development of data center capacities is therefore only somewhat meaningful.

According to the expert interviewed, a modular data center structure could also be conceivable with a corresponding land area. Modular construction means that not the entire area of the property is first built, but only the areas that are currently needed. Area will be reserved for foreseeable future growth of the data center, provided that this is provided for in relevant building applications.

Latency

The term describes the time interval between the moment a request is made to a system and the moment the system's response is received. Latencies are measured in milliseconds or microseconds.

According to the experts interviewed, latencies are not of great importance for training an AI foundation model, as it is done locally on a suitable data cluster. For the operation and development of inference applications that will later emerge based on the foundation model, a data center should have a bandwidth of at least 100 Gbit/s, depending on the use case, and meet latency requirements of less than 10 ms RTD (Round Trip Delay).

Certification, data protection and compliance requirements

AI-based high-performance data centres are comparable to other data centres in terms of their certification requirements and the General Data Protection Regulation (GDPR). According to legal experts, the most common certifications include ISO 27001 (certification based on basic IT protection), ISO 9001 (quality management system) and ISO EN 50600 (construction and operation of secure data centers). The BSI's Criteria Catalogue C5 describes the minimum requirements for secure cloud computing and must be taken into account.

8.3 Sustainability aspects

Waste heat

The use of the generated waste heat is one of the most essential sustainability aspects in data centers and is heavily debated due to a number of challenges. To date, the largely untapped potential lies in the supply of CO₂-free waste heat from data centers to local and district heating networks. The existing cooling technology is the decisive factor in how energy-efficient the waste heat generated from a data center can be used further. If air cooling is installed in the data center, according to Expert, the waste heat reaches temperatures of 30-35°C inside, and in special cases up to 50°C. However, these temperatures are too low to feed the waste heat directly into a heat grid. This means that a heat pump must be used to adjust the temperature to the requirements of the heating network before the waste heat is passed on. When water cooling is installed in the data center, the waste heat can be supplied directly to a fourth-generation heat grid at temperatures of 60-70°C. Technologically older heating networks, however, require higher feeding temperatures. Due to the high compaction of the waste heat, which is too low, additional energy costs are generated by the operation of the heat pumps. Modern methods of water cooling are still rarely used due to their novelty (cf. cooling).

Another challenge is the local absorption of waste heat via local heating networks. Although it would be possible to use waste heat to heat surrounding buildings, there is often a lack of on-site buyers (e.g. residential and office and industrial buildings) that can absorb sufficient waste heat, as well as the policy and regulatory framework (Bitkom e.V., 2022b). So far, only small parts of the waste heat from the data centers can be used.

Electricity

Powering a data center with green electricity is another aspect of sustainability. The decisive factor here is the consistent and successful implementation of the energy transition, which will accelerate the expansion and in particular the availability of electricity from renewable energies in Germany (Bitkom e.V., 2022b). Green electricity is produced from renewable energy sources such as sun, wind, water or biomass.

In the future, the Federal Ministry of Economic Affairs and Climate Change (BMWK)'s ESCADE research project will use world-leading hardware and software technologies to examine how to improve the sustainability record of AI applications.

8.4 Infrastructure requirements in detail

The calculated infrastructure requirements to calculate the AI foundation model were based on the public information on OpenAI's GPT-3 language model.

Compute requirements for calculating a foundation model	
Number of GPUs	4480 GPU (e.g. 560 NVIDIA DGX H 100)
Time per run, in hours	694
Total passes	4
Total maturity, in hours	2777
Internet bandwidth	>100 Gbit/s
Interconnect	InfiniBand or similar
Storage	2000 - 5000 TB
Preprocessing	
CPU	10 - 20 k CPU-Cores
Time	840 - 1000 hrs.
Total IT performance	approx. 4,0 MW
kW per rack	36
Number of racks (for 4x DGX H100 / Rack)	approx. 140

Table 9: Compute requirements for calculating a foundation model

According to the expert, the following computing resources are required for the training of the LEAM foundation model. For example, the calculation is based on the use of 560 of the most powerful GPU nodes (e.g. NVIDIA DGX H 100, AMD Instinct MI200 or Intel Ponte Vecchio Data Center GPU). The system requires 694 hours for one training course, with a total of four courses and 2,777 hours or 115 days, respectively, for training the foundation model. The installation of the system would require 140 racks, each with 4 DGX nodes and 8 GPUs. Waste heat of up to 9 kW per rack is generated per DGX, thus reaching a total of 36 kW per rack. The required cooling at these power densities can only be achieved by liquid cooling or direct-to-chip cooling. The latest generation of GPU systems (e.g. NVIDIA DGX H100) offer the option of direct-to-chip cooling.

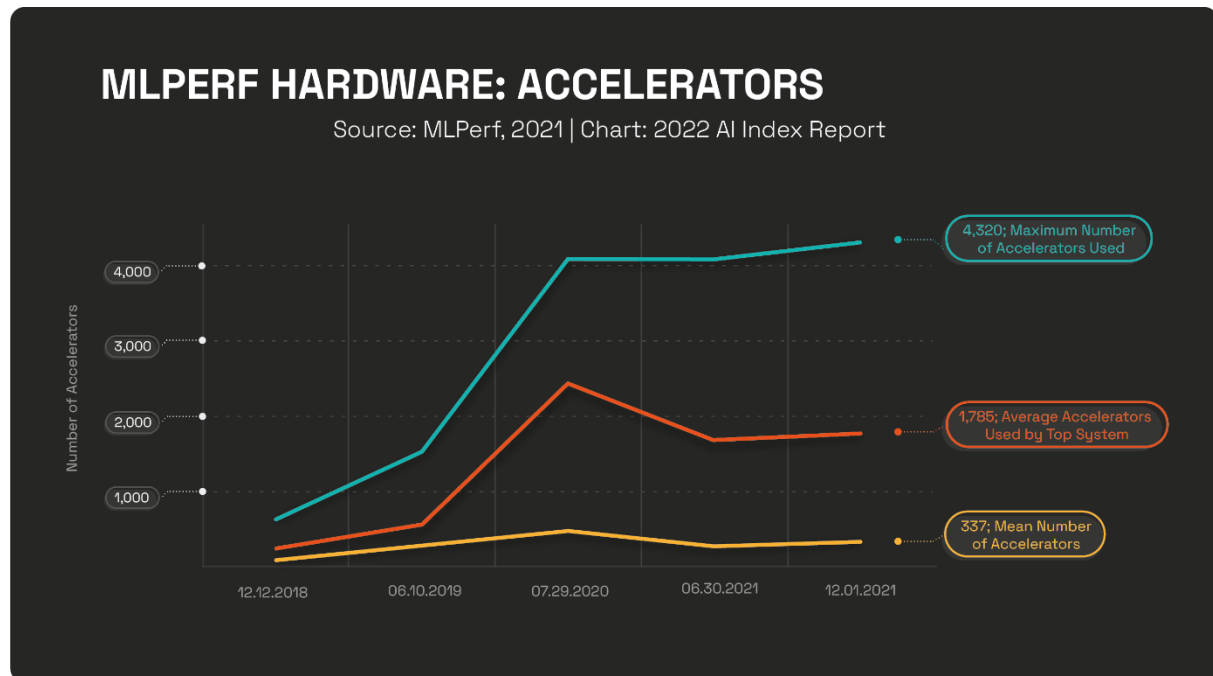


Fig. 22: MLPerf hardware: accelerators (Zhang et al., 2022, p.18)

Stanford University's 2022 Artificial Intelligence Index Report describes the evolution of the number of GPUs used in the top HPC systems for training machine learning algorithms, and shows that the fastest AI algorithms used for training are computing clusters with a very high number of GPUs (Zhang et al., 2022). The maximum number of GPU accelerators used has probably increased again since the January 2021 survey.

8.5 Selection of location

What are the parameters for a qualified selection of a location for an AI data center? In addition to regional localization, it examines what needs to be considered in terms of energy supply as well as connectivity to existing infrastructure.

8.5.1 Existing HPC capacities in Germany

Research area

The offer of high-performance computing for science is structured in Germany according to the performance of HPC centers at levels 0-3.

There are three level 1 high-performance data centers in Germany. The Gauss Centre for Supercomputing (GCS) brings together the three most important high-performance data centers under one name. The total output of the Jülich Supercomputing Centre (JSC) in North Rhine-Westphalia, the Leibniz Data Center (LRZ) in Garching near Munich and the High Performance Data Center Stuttgart (HLRS) is 130 peta-FLOPS (as of November 2021). The HPC centers have different orientations and can thus meet different user requirements. In addition, they can act together or as part of their work.

Level 2 includes 12 cross-regional high-performance data centers at research institutions and universities. Eight of the 12 university level 2 HPC centers have joined together to form the National High Performance Computing (NHR) group. These include:

- IT-Center - RWTH Aachen
- Zuse-Institut Berlin - Berlin University Alliance
- University Computing Centre (HRZ) - Darmstadt University of Technology
- Centre for Information Services and High Performance Computing - Technical University of Dresden
- Erlangen Regional Data Center - University of Erlangen-Nuremberg
- Society for Scientific Data Processing mbH Göttingen - University of Göttingen
- Steinbuch Centre for Computing (SCC) - Karlsruhe Institute of Technology
- Paderborn Center for Parallel Computing - University of Paderborn

Level 3 includes regional HPC centers and institutions with their own clusters for applications that require lower performance. Two of the most powerful HPC centers are exemplified here:

Jülich Supercomputing Centre (JSC)

The Jülich Supercomputing Centre at the Jülich gGmbH Research Centre (JSC) is an institute of the Helmholtz Association of German Research Centres and is funded by the federal government (90%) and the state of North Rhine-Westphalia (NRW) (10%). Among other things, JSC has the JULES supercomputer with 86 peta-FLOPS power, one of the most powerful computers in Europe at present. The JUWELS multi-petaflop supercomputer has a booster module equipped with NVIDIA Ampere GPUs, making it suitable for computing large AI foundation models. Access to HPC computing resources is provided through biannual project calls selected through an objective peer-review process. The FZ Jülich is a public law non-profit company with 50% funding from the EU through the organisation PRACE - Partnership for Advanced Computing in Europe and 50% through the Federal Ministry of Education and Research (BMBF) and the Land of NRW through the Gauss Centre for Supercomputing (GCS). Projects (usually simulations in the field of climate research and quantum physics) have a maximum calculation time of 24 hours. Only in exceptional cases and for larger projects are job chains and reservation possible. Starting in 2023, we started here with the installation of the first European exascale calculator, JUPITER, which will be used, among other things, to train the latest generation of compute-intensive machine learning algorithms (Jülich Research Center, 2022). The JSC is primarily aimed at the scientific community and awards the computation time twice a year in a competitive procedure following a peer-review procedure.

FZ Jülich is also a partner in the project OpenGPT-X, sponsored by the Federal Ministry of Economy, in which a language model is being trained on the JUWELS supercomputer. However, the project currently uses only about 320 GPUs. This value will certainly go up, but it is not comparable to the 4480 GPUs proposed here.

Stuttgart High Performance Computing Centre (HLRS)

The Stuttgart Highest Performance Computing Centre (HLRS) is a central institute at the University of Stuttgart, which operates the Hawk supercomputer with 26 peta-FLOPS, among others, and has been available to academia and industry for 25 years. The HLRS is a member of the German Gauss Centre for Supercomputing (GCS), which gives it partial basic funding from the Federal Ministry of Education and Research (BMBF). The other part of the basic funding is provided by the Land of Baden-Württemberg. In addition, HLRS is funded through research funding (project funding) and revenues from the use of HLRS HPC computing capabilities by businesses and industry. Resource use by the private sector is limited to about 10% of computing capacity and accounted for about 2% of third-party revenue in 2021. HLRS is characterized by the so-called Solution Centers, which, as external companies, organize the transfer to science and business and promote access to high-performance computers.

Use of HPC computing capabilities from the field of research in the example of HLRS*:

The HLRS can currently offer a **maximum of 192 GPUs (24 GPU nodes)** at a time for the calculation of a foundation model. Here, the equivalent of **16,192 knot hours would be required** for a training course. This corresponds to about **675 days of continuous operation of 24 AI nodes with 8 GPUs each**.

Exclusive use of all GPU nodes in this form would not be feasible at HLRS at the moment and would not be compatible with the time requirements of the innovation cycles in the development of a foundation model (see 1.3 Requirements for an AI high-performance data center).

For GPU/CPU computing power, the following costs are incurred to calculate a foundation model:

1. GPU usage

The 24 AI nodes of the HLRS require about $23 * 694$ node hours for calculation ($560/24 = \text{about } 23$) --> 16,193 node hours = 4 runs then yield 64,772 node hours.

64,772 AI node hours x EUR 19.50 per AI node hour corresponds to a price of EUR 1,263,054

For data preprocessing, which will take up to 20,000 CPU cores and an estimated maximum runtime of 1,000 hours, the following target costs would be incurred for utilizing the CPU clusters at HLRS:

2. CPU usage

20,000 CPU cores for 840 hours of runtime --> 168,000,000 core hours
The current HAWK nodes on the HLRS have 128 CPU cores. Since the calculation is per node hour, the result is:

131,250 knot hours x research prize (as at: 2022) at EUR 1.13/node hour = approx. EUR 80,000

A surcharge of between **10% and 30% is also estimated** for industrial use at HLRS.

* The cost estimate was made on the basis of a publicly available fee schedule in 2022.

The required GPU hardware resources are not sufficiently available for LEAM at all the HPC computing sites studied, so they are not eligible as an infrastructure provider. They are more likely to be users of specialized AI computing infrastructures that will be available in the future. Co-operation is also possible here.

Commercial providers

In addition to research facilities, there are also commercial providers providing HPC capacities in the market. Of these, the so-called hyperscalers are particularly relevant. Hyperscalers are companies that operate very large data centers and provide cloud services to their customers based on highly scalable infrastructures. Computing power, storage capacity and other resources can be made available on demand almost without delay, and billing for the resources used is usually after actual usage. Hyperscalers are distinguished from other providers by the sheer size of their data centers (some well above 100 MW) and their ability to process large amounts of data and computing power quickly and efficiently, and to provide services at a high scale worldwide. To cater to customers' increased demands for low latency, they typically operate at least one data center on the relevant continents and offer guaranteed data storage and processing across different geographic availability zones.

The largest and most well-known hyperscalers are Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP) from the American space, and increasingly the Chinese companies Tencent Cloud and Alibaba. In addition, there are other large companies such as IBM, Oracle or HPE that provide HPC capabilities, and some smaller, highly specialized SMEs and start-ups such as Lambda. As German companies offer e.g. IONOS, Northern Data and also Aleph Alpha HPC solutions.

Generally, all of these companies provide dedicated systems designed for the intended purpose of processing AI applications and providing access to high-performance GPUs. However, there are differences in the hardware used and the deployment. For example, NVIDIA's A100 GPUs are available from most vendors. The use of the latest generation of H100 GPUs is currently being evaluated and should soon be available, for example, on Microsoft Azure. Another difference arises in the way in which the requested resources are provided. As cloud specialists, by their very nature, all companies offer the possibility of virtualized systems. In this case, one or more virtualized systems are running on a host system. Some vendors, such as IBM and Microsoft Azure, also offer access to dedicated systems that can be configured even more flexibly. In the case of Microsoft Azure, even a dedicated supercomputer from manufacturer Cray can be integrated.

Even if, in principle, it is possible to build a suitable cluster, at least in Europe, there are not enough resources available in the form of GPUs. According to the experts interviewed, there is no vendor that can provide the required number of approximately 4,500 latest-generation GPUs from a data center. The distributed use of resources from multiple data centers currently fails at the necessary bandwidth of the internal network. In this context, however, it should be mentioned that the providers would probably be willing to build up the corresponding capacity in the case of long-term user contracts. However, the costs for this are likely to be higher than for the other operating models shown here. An exemplary business comparison for different HPC offerings for sourcing computing power from the cloud is provided in Chapter 10.1.

While the use of capacities outside Europe would be conceivable, it is contrary to the objective of building a German or European AI ecosystem for the development, deployment, operation, integration and validation of highly powerful AI models.

Another point that argues against building a cloud-based system using a hyperscaler comes from the target audience of SMEs. Although the level and standards of data security among the hyperscaleers are more likely to be above average, there are still concerns about the protection from unauthorized access to their own data, particularly by government agencies. The more sensitive one's own data is assessed, the greater the skepticism. This could lead to the failure to provide necessary training data. In addition, many hyperscalers use so-called lock-in effects for customer loyalty, which make it easy to get started and more difficult to switch to another provider later.

SPOTLIGHT

Merantix Momentum GmbH

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Das Leadership-Team:
Dr. Johannes Otterbach,
Nicole Büttner-Thiel,
Dr. Thomas Wollmann.

Merantix Momentum is a Berlin-based AI startup that specializes in the deployment and scaling of AI-based solutions across various industries. With an experienced team, we, as an AI service provider, solve our customers' challenges through bespoke machine learning solutions, ensuring their future competitiveness in digital and data-driven markets. With our own research department, we also actively support the transfer of machine learning methods into productive application in companies and organizations in Germany and Europe.

Where do you use foundation models? What is your use case?

Merantix Momentum specializes in the implementation and implementation of risky AI innovation projects. In doing so, we support our customers and partners from the initial use case development, through the data strategy, through to the development and production operations of the AI solutions.

Foundation models offer the advantage of developing new deep learning solutions quickly and efficiently on limited data sets of our customers. In doing so, it turns out to be beneficial that foundation models have been trained on a broad database that is indirectly transferable to the customer data. With a wider availability of different foundation models, that is, trained on various Data, thus, allows the application of modern AI methods to previously unexplored applications to be realized.

What impact do AI foundation models have on your business model or your projects?

AI foundation models support us in project development at several locations. On the one hand, they help with use case development, as they enable rapid prototyping and thus unleash the creativity of our customers and partners, but also of our developers. On the other hand, they also offer the opportunity to develop small and efficient models through distillation, which would not have been possible by training a new model from scratch, as the data base is often insufficient. In this way, AI foundation models do the same twice: In the innovation phase by supporting the creative process and subsequent development that would not otherwise be possible.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

First of all, it should be made clear that certain biases are wanted. These are the basic ethical and moral attitudes of the models, based precisely on European values. This is due to the different cultural circles in which the models are developed. Apart from this, European-based models are helpful in bringing applications into practice faster. The European bias allows us to expect to be able to adapt the models faster with less data. On the other hand, we need to be less concerned with the output of the models and their examination for unwanted output or possible vulnerabilities, as a certain minimum line of operation can be assumed, which is not the case with non-European models.

8.5.2 Required location parameters

Geopolitical and data protection location

In the European context, the so-called FLAP markets (Frankfurt, London, Amsterdam and Paris) - often supplemented in recent years by Dublin as FLAP-D - represent the historically grown centres of gravity for Europe's major Internet hubs. Similar to the development of the historic Silk Road, these digital Internet hubs run the global data transmission lines for global Internet traffic. With the exception of London, these centres are within the scope of the European Union and thus provide a secure and EU-compliant legal framework for their users in terms of data protection law.

Historically, this development follows the development of international telecommunications hubs in the first half of the 20th century. In the second half of the 18th century, the development of the first industrial revolution had already taken place. ==References== From an economic point of view, this compact settlement of digital infrastructure elements is to be located in the area of the so-called "Blue Banana", a densely populated chain of agglomerations from Manchester, Greater London, Amsterdam's Randstad, the Ruhr, the Rhine-Main region, to the industrial sites of Mannheim, Ludwigshafen and Basel, and finally to the industrial centres of northern Italy with Milan and Turin.

With a transmission capacity of more than 50 Tbps, the West-East data route from Dublin via Amsterdam to Frankfurt alone is one of the largest trans-European routes for IP routing, along which large volumes of Internet traffic travel. This is more than five times the throughput of today's world's largest Internet node in Frankfurt (Simons & Frese, 2021). Another data corridor (north-south) connects the Scandinavian data center settlements with European and American content providers via Stockholm, Copenhagen, Düsseldorf, Frankfurt and Paris.

Along these routes, a digital economy with all aspects of the value chain has emerged over the past 20 years. Digital ecosystems from software development, content, AI or IT services have often been built around these large data center settlements. The gravitational anchor for these settlements is often short latency, broadband connections to the transcontinental backbone tracks, compacted metropolitan structures with a corresponding number of potential users, and a sufficient market for well-trained professionals.

At both European and national levels, the growth trend in the data center market continues unbroken. Between 2016 and 2021, capacities increased by 30% when measured in terms of IT connectivity (Hintemann et al., 2022). The beneficiaries of this trend are primarily the data center providers with installed power from 5 MW and well beyond. Smaller installations below 5 MW are currently in a stagnant or declining trend.

In terms of importance, the Rhine-Main region remains a top location for data centers in Germany. In addition to Frankfurt, Berlin will play an increasingly important role in data center developments in the future. Measured in terms of IT connection capacity per inhabitant, Hessen is more than three times higher than the city states of Hamburg or Berlin. In addition to Berlin, locations such as Munich, Hamburg, and the regions of Cologne/Düsseldorf and Leipzig/Dresden will become increasingly important for data center developments in the future (Hintemann et al., 2022).

In the field of High Performance Computing (HPC), the following important locations in Germany can be mentioned:

Stadt	Establishment	Peak Performance	Memory	CPU-Cores
Eggenstein-Leopoldshafen	Steinbuch Centre for Computing - ForHLR - bwUniCluster	1171 TFlop/s 444 TFlop/s	136 TB 86 TB	34800 Cores 18304 Cores
Frankfurt	Center for Scientific Computing - LOEWE CSC - FUCHS	823 TFlop/s 41 TFlop/s	70 TB 18 TB	18960 Cores 6456 Cores
Offenbach	German Weather Service - Cray XC40 - Cray CX40 II	1073 TFlop/s 1073 TFlop/s	125 TB 125 TB	29552 Cores 29952 Cores
Hamburg	German electron synchrotron - Location Zeuthen - Maxwell	16 TFlop/s -	- 402 TB	2288 Cores 26732 Cores
Hamburg	German Climate Computing Center Mistral (HLRE-3)	3590 TFlop/s	266 TB	101196 Cores
Kaiserslautern	Competence Center HPC - Beehive - Stagnation - Ocean 1	67 TFlop/s 35 TFlop/s -	14 TB 6 TB 23 TB	3224 Cores 1584 Cores 11600 Cores
Kaiserslautern	Reg. University Computing Center - Elwetritsch - Elwetritsch II	134 TFlop/s	17 TB 53 TB	5624 Cores 10520 Cores
Cologne	Reg. Uni Cologne Data Center - CHEOPS	100 TFlop/s	36 TB	9712 Cores
Göttingen	G.S. f. knowledge. Data processing - Scientific Compute Cluster - Emmy	2883 TFlop/s 8261 TFlop/s	92 TB 498 TB	16640 Cores 116152 Cores
Darmstadt	University Computing Center - Lichtenberg II	3148 TFlop/s	251 TB	61824 Cores
Stuttgart	High-performance data center - Hawk - Cray CS-Storm - volcano	26000 TFlop/s - -	1 TB 9 TB 102 TB	702896 Cores 608 Cores 13856 Cores
Aachen	IT Center RWTH - CLAIX-2016 - CLAIX-2018	678 TFlop/s 4965 TFlop/s	88 TB 251 TB	16152 Cores 62736 Cores
Jülich	Jülich Supercomputing Centre - JUWELS	12000 TFlop/s	286 TB	123088 Cores
Berlin	Konrad-Zuse Centre - Lise	7907 TFlop/s	455 TB	110016 Cores
Garching	Leibniz Data Center - Super MUC - Super MUC NG	3580 TFlop/s	197 TB	86016 Cores
Garching	Max Planck Computing & Data Facility - COBRA	12720 TFlop/s	530 TB	136960 Cores
Paderborn	Paderborn Center for Parallel Comp. - XCL FPGA Cluster - Noctua 1 - Noctua 2	- 835 TFlop/s 7100 TFlop/s	512 GB 53 TB 355 TB	32 Cores 10960 Cores 143488 Cores
Mainz	Data Processing Centre - Mogon - Clover - Mogon 2	379 TFlop/s 106 TFlop/s 3125 TFlop/s	90 TB 10 TB 194 TB	35760 Cores 5120 Cores 52248 Cores
Dresden	Center for Information Services / HPC - Taurus - Alpha Centauri	2621 TFlop/s -	279 TB 35 TB	64536 Cores 1632 Cores
Pineapples	Erlangen National Center for HPC - Emmy - Meggie - TinyGPU - TinyVec - TinyARM - TinyFAT - Alex - Fritz	232 TFlop/s 511 TFlop/s - - - - - - -	36 TB 47 TB 5 TB 96 TB 128 GB 22 TB 65 TB 242 TB	11088 Cores 14560 Cores 1392 Cores 12 Cores 64 Cores 2484 Cores 8960 Cores 67968 Cores

Table 10: HPC locations in Germany

The data center locations for HPC applications listed above can be located in close proximity to universities and research institutions and do not necessarily follow the pre-described development logic of commercial collocation and hyperscale data centers along historically developed data routes. In the scientific and university fields, these data centers have usually connected with each other through their own networks (e.g. the scientific network X-WiN, which is operated by the German research network DFN). The X-WiN connects to external networks at dedicated locations via local Internet hubs, such as the DE-CIX in Frankfurt am Main and Hamburg, the ECIX in Düsseldorf and the BCIX in Berlin.

Regional location: Area requirements, cubatur, building infrastructure

In the regional and local location of data centers, different criteria are at the forefront than in a geopolitical or national view. General criteria for the location of data centers include:

- The properties are not in the immediate vicinity of or in residential areas.
- It is easy to reach, including by public transport or road connections.
- Exclusion of threats/hazards impeding settlement (see also Seveso II/III areas): Proximity to airports, chemical industries, freight routes, electromagnetic exposure, vibration sources, etc. However, potential hazards caused by natural events (e.g. high water, active seismic zones, proximity to coastlines, etc.) should also be avoided.
- Approaches exist for local uses of the data center, including a diverse range of data-centric companies and research institutions; ideally within a 50 km radius.
- A reliable, redundant and scalable power supply is provided. A spatial proximity to the converter plant is sought here. In the case of redundant supply from two transistors, the geographic centre between two transistors is often preferred.
- There are good connections to trans-regional fibre-optic routes. Ideally, there should be at least two cross-regional (so-called Longhaul) fiber-optic route providers that can implement a total of three grinding-free and cross-tracked routes with at least 20 usable fiber-optic pairs per route to the route, with a maximum distance to the route of 5 km, in relation to a potential settlement area.
- There is access to an Internet node provider that can implement a variety of enterprise-level interconnection services using a non-discriminatory, distributed platform concept.
- An important factor in the future will be the proximity to potential waste heat collectors (e.g. quarters, swimming pools, vertical farming, etc.) or the possibility of direct feeding into a local or district heating network.

Area requirements

For large data centers and hyperscaleers, scalable area scalability plays a crucial role in the respective location concept. Settlements of medium-sized operators are competing with correspondingly smaller land sizes. Basically, the following are to be distinguished:

- "Solitary" data center: Building a single, self-contained data center (often for in-house use). Order of magnitude: 0.5 to approx. 3 ha of land area, power intake usually up to 10 MW
- Campus Data Center: There are several self-contained data center operating units on a larger data center campus. The operating model is often oriented towards collocation, shell & core, or hyperscale. Order of magnitude: 10 ha of land and more, performance class: 50-120 MW
- Large Hyperscale Campus: From approx. 15 ha upwards. The currently largest settlement project in Europe is located near Esbjerg, Denmark, covering approximately 200 hectares (the size of Monaco); power classes: > 100 MW to 300 MW and beyond.

Kubatur / Building Indicators

As an indication of the intensity of land use, the measure of building use is an integral part of the public building law and is thus an important urban planning instrument of the BauGB. From a building law point of view, a number of requirements must be complied with when building data centers. Among other things, the base area number (BSA) specifies the percentage of land that may be overgrown. The cubature describes the reconstructed space of a building. Similar to increasing the power density in a server rack, the total number of servers that can be operated can be increased as the number of shells (number of floor areas) of a building increases. Finally, the number of building mass (BMZ) indicates how many cubic metres of building mass are allowed per square metre of land area.

These requirements are enshrined in the local building law accordingly and cannot be changed by the builder. In Germany, multi-storey data center buildings are common in industrial areas, but examples of data centers in high-rise buildings such as *60 Hudson Street* in Manhattan or the *Royal Bank of Canada Data Centre* in Toronto in Germany have not yet been found.

Power supply: Connectivity and security of energy supply

Data centers require a redundant and well-thought-out power supply concept for safe (failure) operation. In order to achieve the required redundancy, separate feeding can be used, where possible, via two converters. If this is not the case, a ring feed is recommended via two separate feeds of a transducer.

If the data center is supplied with a medium voltage (10 kV or 20 kV), a customer-owned or power-supplied medium voltage station is usually installed on the property. Transformer feed and measuring equipment are usually integrated into these systems. Casting resin transformers are commonly used in data centers to lower the medium voltage accordingly. Compared to conventional dry transformers, higher partial discharge,

short circuit and voltage resistance is achieved, and sensitivity to environmental stresses such as moisture and dust is also lower (Dürr, 2018). In some cases, oil transformers are also used. These are less loss-bearing due to their construction and have a higher energy efficiency. However, oil transformers have other fire protection requirements, which may require more space. Uninterruptible power supplies (UPSs) and replacement power systems (NEAs, e.g. diesel generators) are also necessary components of energy security.

The task of a SSP is primarily twofold: (a) the short-term bridging of power outages by switching to battery storage or swing mass storage; and (b) the elimination of voltage fluctuations and spikes, as well as frequency variations. Based on the principle of operation, three basic types of UPS systems can be distinguished:

- VFD (Voltage and Frequency Dependent): The current is passed directly from the inlet to the outlet in this type of UPS. There is no "galvanic decoupling" with the forward load. The battery storage is continuously powered through the input of the UPS. However, switching to battery-powered mode in the event of a power failure is intermittent and may take up to 10 ms. This is a value that may already be noticeable with sensitive IT hardware.
- VI (Voltage Independent): Type VI USVs operate with an AC/DC converter as the central component, which acts both as a power converter and is responsible for charging the batteries. Unlike the above type, the UPS is "active" even when the power is available over the mains voltage. However, voltage spikes are filtered out that could damage the IT hardware. The switching time is between 2.5 and 10 ms.
- VFI (Voltage and Frequency Independent): In addition, unlike the two types of VFI listed above, a VFI type ensures that there are no undesirable frequency deviations in addition to peak voltage and voltage fluctuations. The current at the output side is completely decoupled from the current at the input side. As the UPS of this type operates in continuous operation, there are no additional switching times.

When utilizing the UPS systems, care must be taken to ensure that the efficiency of the system decreases continuously as the utilization rate decreases. The increased loss power is thus negatively felt in terms of energy efficiency. However, the task of a system of AC and battery storage is always only to bridge a relatively short period of time (a few minutes) until a power supply replacement system (NEA; = emergency power diesel) is in operation for longer-term emergency operation. Emergency power supply systems are standardised according to ISO 8528.

Diesel-powered emergency power replacement systems must undergo regular testing to ensure their operation. Different types of test operations are distinguished:

- Grid-parallel load-testing operation: The NEA is started and synchronized with the network power sine wave. The generator switch is then coupled and the NEA is operated in parallel with the grid. After prior consultation with the energy supplier, even load can be fed back into the grid.
- Load sampling operation in the island operation: The NEA is started in parallel operation as described above. However, when the load is high, the power switch is disconnected, so that the NEA can now deliver the full load for the servers to operate. This method is preferable to the previous one, as it can better represent the real conditions in the event of a grid failure.
- Back Building Test: This is a "real" network test. The central power switch is disconnected before the NEA starts, so that the USV and NEA have to jump in spontaneously. Many operators are afraid of this type of network testing, because they overestimate the risk of failures and a "real" data center outage.

Customers in the data center are usually notified of upcoming tests of USVs and NEAs in advance by the operator. Test runs of diesel units must be approved by the competent authorities (including for emission requirements). Typically, an hourly quota per year (e.g. 30 h/a) is approved for test operations. The diesel generators are electrically preheated in standby mode to ensure the shortest possible start-up times until full load is reached in the event of use. Thus, NEAs also consume a certain amount of electrical (heating) energy offline.

Availability classes

Data centers are divided into so-called availability classes (VK1 - VK4) according to the degree of existing redundant components. The following are distinguished:

- Availability class 1 (N): If you designate individual components of a data center (e.g. an NEA, an AC, an air conditioner) with the variable "N", there is no additional redundancy in this concept. If a component fails, it must first be serviced/repaired before returning to service.
- Availability class 2 (N+1): The failure of a single component does not result in the failure of the complete supply path here, as an additional replacement component (+1) may jump in. Example: 5 air conditioners are needed to cool the server room. The sixth device installed in the room will jump in in the event of the failure of another device.
- Availability class 3 (2N): In this concept of redundancy, all supply pathways are "double" designed. Example: All servers are equipped with two power supplies, which are powered through two different power phases. By shutting off a complete power phase (e.g. during maintenance work), operational operation in the server room is still guaranteed.
- Availability class 4 (2N+1): This represents the highest availability class. In contrast to the 2N concept, the failure safety of a single component can also be ensured in the case of maintenance of a complete supply path.

The availability classes are assigned corresponding maximum downtime per year:

Availability class	Name	Minimum availability	Max. Downtime per month	Max. Downtime per year
VK 0	No guaranteed availability	--	--	--
VK1	Normal availability	99,0 %	< 8 hours	< 88 h
VK2	Increased availability	99,9 %	<44 min	<9 h
VK3	High availability	99,99 %	<5 min	<53 min
VK4	Maximum availability	99,999 %	< 26 sec.	< 6 min

Table 11: Availability classes (VK1 - VK4)

With higher redundancy interpretation and the associated availability class, investment costs also increase. Since the additional components are often not locked in situ in the event of maintenance or failure, but rather run (with) in the so-called semi-blast parallel operation, the efficiency values are also lower with higher redundancy than in operation without redundancy. However, this is deliberately taken into account by the operators in order to ensure a higher degree of failure safety. LEAM requires 99% availability for the inference applications.

Fibre supply: Backbone networks, redundancies, non-discriminatory and carrier-neutral access, proximity and access to Internet exchange nodes

Similar to how power is routed from the power provider's external transmission infrastructure to the server power supply via mid-voltage and low-voltage distribution on the data center site, data connectivity of a comparable distribution structure also occurs.

- **Transfer point:** External data connection to the data centers via internal Meet-Me Rooms (MMR) is ensured through national and international carrier connections. Meet-Me rooms are the central location within a collocation data center where telecommunications companies and carriers, as well as the collocation operator's customers, can physically connect and exchange data. Often, the high-performance routers of a decentralized Internet node are located in the Meet-Me rooms of the data center partners ("Enabled Sites"), allowing access to the node operator's 'Connected Networks'.
- **Location distribution (SV):** From the Meet-Me room as a central transfer point to the external network, data is delivered to the individual floors/server rooms via primary/campus cabling.
- **Building wiring (GV):** If necessary, the network cabling is divided into individual floors via corresponding floor distributors (vertical distribution).
- **Tertiary wiring (EV):** The wiring is then routed to the server cabinets in the individual server rooms via a tertiary or floor distribution.

Two concepts can be distinguished when distributing at the server barrier level:

- **End of Row (EoR):** The first and/or last cabinet in a rack series is equipped with the access switch for connecting all other cabinets in the respective rack series. The cabinet with the EoR switch has to accommodate a large number of patch cables for horizontal wiring through its patch panels. The advantage of this arrangement is the simplified change management, as all the patch cables of a rack series run together in this central location.
- **The Middle of Row (MoR) concept** is similar to that of the EoR, except that the central cabinet with the access switches is positioned in the middle of the rack row.
- **Top of Row (ToR):** With this concept, each cabinet (mostly at the top) has its own switches. Thus, the patch cables can usually be kept very short. However, many smaller (edge) switches are required for high rack numbers. Cross-cabinet change-requests are not so easy to implement with the ToR concept. The concept is more costly, as more switches are usually needed.

Need for protection/risk analysis

Many data centers provide system-relevant services. Collocation- Cloud and hosting data centers with a contracted power of more than 3.5 MW are subject to the Critical Infrastructure Determination Regulation (BSI-CriticalV). Regardless, all data center operators have more or less stringent security policies for their facilities. This includes the establishment of a zoning concept. The building parts and sections of a data center (campus) are divided into different zones and security areas according to their criticality. From the outside to the inside, the following five zones/areas can be distinguished according to an "onion-shell principle":

- Zone I: The property or company site is either open access or secured with a fence and video surveillance system.
- Zone II: Semi-public area within the premises, e.g. with normal office workplaces for employees:
- Zone III: IT sub-rooms and administrative control. At the latest, this area is reserved only for a certain group of people. There are often access points in the form of locks and separation facilities.
- Zone IV: Direct access to the data center's technical facilities. Often, access in this area is reserved exclusively for technical maintenance personnel.
- Zone V: Access to the server rooms, the actual "heart" of the data center. In high-security data centers, the access to the server rooms as well as the climate/electricity engineering rooms are designed in such a way that the relevant employees move in separate corridors and a direct encounter is excluded. This aspect is particularly important when external maintenance and technical personnel are used.

In addition to the zoning concept, the DIN EN 50600-1 defines four different classes of protection:

Type of protection	Type of access
Protection class 1	Public or semi-public domain.
Protection class 2	Area accessible to all authorised persons (employees and visitors).
Protection class 3	Area reserved for designated staff and visitors. Other persons with access to Class 2 protection must be accompanied by persons with access to Class 3 protection areas.
Protection class 4	This area is reserved for designated employees who have a demonstrated need for access. Other persons with access to Class 2 and 3 protection must be accompanied by persons with access to Class 4 protection areas.

Table 12: Protection classes according to DIN EN 50600-1

The different protection classes 1 - 4 must be adequately secured by appropriate measures and processes. The most important points in this regard include:

- **Protection class 1:**
 - Identifiable physical barrier at the external (land) border.
 - Doors, windows, grids must meet resistance class 2 (DIN EN 1627:2011).
 - Physical separation of access from pedestrians and Class 1 and 2 protection vehicles.
 - Designated parking spaces for unauthorized vehicles (visitors inside).
- **Protection class 2:**
 - Identifiable physical barrier on the boundary to protection class 2.
 - Doors, windows, grids must meet resistance class 3 (DIN EN 1627:2011). Windows and doors shall be so constructed that they cannot be opened when closed outside of protection class 2.
 - Physical separation of access from pedestrians and Class 2 and 3 protection vehicles.
 - Measures to detect and prevent unwanted and unnecessary access.
 - Any opening of an emergency exit door shall trigger an alarm to trigger an appropriate response.
- **Protection class 3:**
 - Identifiable physical barrier at the external border.
 - Doors, windows, grids must meet resistance class 4 (DIN EN 1627:2011).
 - Confinements in the area of protection class 3 shall not be located locally together with those of protection class 1.
 - Constraints in the area of protection class 3 placed together with boundaries of areas of protection class 2 shall be equal to the sum of the resistance for protection classes 2 and 3.
 - The following measures must be in place to detect and prevent the following events: Unwanted or unnecessary access between Class 3 and Class 4 areas; unauthorized access from a Class 3 area into Class 4 areas; detection of all persons and materials and devices (e.g. IT equipment) entering or exiting Class 3 areas.
- **Protection class 4**
 - Identifiable physical barrier at the external border.
 - Doors, windows, grids must meet resistance class 4 (DIN EN 1627:2011).
 - Confinements in the area of protection class 4 shall not be located locally together with those of protection class 1.
 - Constraints in the area of protection class 4 placed together with boundaries of areas of lower classes of protection must be equal to the sum of the resistance for all classes of protection.
 - All breaches of the physical boundary must prevent access by unauthorized persons. This includes, for example, pressure relief valves for gas extinguishing systems.

Access control

In order to ensure controlled access in and out of the data center area, access control systems are installed in the data centers with corresponding processes behind them. Basically, online and offline access systems can be distinguished. Online facilities are permanently connected to an access control center. All access movements are sent directly to a central point. For offline facilities, the access permission is stored on appropriate access cards and can be transmitted daily, e.g. via the time recording terminal or online reader (Dürr, 2018).

Different systems are available on the market for identification and authorization. The following basic types can be distinguished:

- Physical identification: by means of readers and ID cards or key hangers.
- Logical recognition: by entering numerical codes. (This type of access is not personalized and poses the risk of unauthorized sharing of codes.)
- Biometric identification: Unique identification of an accessed person, e.g. by fingerprinting, iris scanning, hand back recognition, vein recognition.

In order to increase access security, the methods described can usually be combined.

8.6 Operation of an AI data center

In general, there are three options for running an AI data center. Working with a collocation provider, building your own HPC data center, and using a distributed infrastructure. The three options are discussed in more detail below.

Collocation

A collocation operator is a company that provides space and infrastructure to house data centers and other IT facilities. Collocation operators offer their customers the option of hosting their data centers and IT facilities in spaces specifically equipped to operate data centers (see Table 6). These rooms are typically equipped with power systems, cooling technologies, network connections, and security measures to ensure that data centers and IT facilities can operate properly. Collocation operators also offer services such as maintenance and support for IT facilities to their customers. They thus help companies to operate their IT facilities safely and reliably without having to worry about the necessary infrastructure and equipment maintenance. The normal situation, however, is that the maintenance and installation of the racks will be carried out by the customer or staff appointed by the customer on the premises of the collocation provider.

When choosing a collocation provider, meeting cooling system requirements also plays an important role. Other criteria should include the purchase of green electricity and a coherent approach to the use of waste heat. The collocation provider should therefore have the appropriate infrastructure for water cooling. The requirements for training foundation models have already been explained in Chapter 2.2.

[GSI - Helmholtz Center for Heavy Ion Research]

The GSI Helmholtz Center for Heavy Ion Research in Darmstadt operates one of the world's leading particle accelerator systems for research and the Green IT-Cube, an advanced data center with a capacity of up to 12 MW. This has high CPU compute capacities of more than 300,000 cores and 400 GPUs. It serves as a test centre and has an area of 4,645 m² on six floors, accommodating 768 19" racks at a height of 2.2 metres (4 MW and 256 racks in the first phase of expansion). Thanks to a special cooling system, the Green Cube is particularly energy- and cost-efficient. The Power Usage Effectiveness (PUE) here is less than 1.07 and is achieved through passive water cooling in the back doors of the racks as well as heat exchanger evaporation cooling towers. The absence of indoor air cooling allows for high spatial power densities, which are required for HPC systems with many GPU nodes. In an AI cluster, up to 4 NVIDIA DGX H100 systems are typically installed per rack. The resulting waste heat of 36 kW/rack can, according to the GSI expert, be managed with the Green Cube's cooling system.

The Green Cube has an N+1 redundancy and offers the ability to provide rackspace and data center services (collocation). According to the operator, the planned expansion stages could provide sufficient space for a 4 MW HPC supercomputer and corresponding cooling (water cooling in the rack and prospectively also direct-to-chip cooling) as part of a collocation model. As an additional possibility, valuable data and scientific evidence on the operation of such a large GPU-based HPC system could be collected and published here as part of the HPC cluster's test operation. The Green Cube is sourced from 100% green electricity and, among other prizes, was awarded the 2020 Blue Angel Eco-label by the federal government. The waste heat from the racks is used to heat the office and canteen building on campus. The technical feasibility of a collocation operation of the LEAM AI data center in the Green Cube has already been confirmed by the operator GSI.

INFOBOX

The European collocation market had sales of approximately \$10.7 billion in 2021 and is dominated by a large number of globally operating providers, which operate a variety of data center locations across Europe and around the world. An annual market growth rate (CAGR) of 14.8% is currently forecast for the period from 2021 to 2027 (Research and Markets Ltd, 2021). For example, in the Rhine-Main region, the capacity of collocation data centers added annually has more than doubled from 62 MW in 2020 to 139 MW in 2021 (Hintemann et al., 2021).

Some of the leading companies in this industry include:

- **Digital Realty / Interxion maintains** collocation data centers in over 300 locations worldwide. In Europe, Digital Realty operates a total of 114 data centers in 15 metropolitan regions, spread across 13 countries
- **Equinix** operates 240 data centers worldwide in 31 countries on 6 continents. In total, the company maintains 2.6 million m2 of space worldwide. The average availability of the data centers is 99.9999%.
- **NTT Global Data Center maintains** 600,000 m2 of RZ space in over 20 countries with an IT capacity of 1,500 MW. In Europe, the company operates in 15 metropolitan regions with collocation services.
- **Cyxtera operates** more than 60 data centers in over 30 markets. In Europe, the company operates at its locations in London, Amsterdam and Frankfurt.
- **Cyrus One maintains** 50 collocation locations in North America and Europe. In Europe, the company is represented in Germany, the UK, Ireland, the Netherlands and Spain with 14 collocation data centers.
- **KDDI/Telehouse operates** over 45 collocation data centers worldwide. In Europe, the company has a presence in London, Paris and Frankfurt with 9 data centers.
- **Vantage Data Centers operates** 24 campus locations on 5 continents. In Europe, the company has 8 locations in Berlin, Frankfurt, Milan, Warsaw and Zurich and Cardiff.
- **Iron Mountain maintains** over 370,000 m2 of gross RZ in a total of 21 locations on three continents. In Europe, the company has a presence in Amsterdam, Frankfurt, London and Madrid.
- **Global Switch operates** 13 data centers on 2 continents. There are a total of 9 data centers in Europe, located in London, Amsterdam, Frankfurt, Paris and Madrid.
- **COLT Data Center Services operates** 14 data center locations in Asia (Tokyo, Osaka and Mumbai) as well as in Europe (Frankfurt, London, Paris, Rotterdam).
- **Penta Infra operates** data centers in the Netherlands, Denmark and Germany. The company has a presence in Germany, with data centers in Berlin, Hamburg, Düsseldorf, Cologne and Leipzig, among others.
- **North C Datacenters operates** data centers in a total of 14 locations in Germany, Switzerland and the Netherlands. In Germany, the company is represented in Nuremberg and Munich.

The collocation operators operating primarily in the German-speaking area continue to include (selection):

- **noris network AG** has seven data centers in five locations in Nuremberg, Munich and Hof.
- **StackIT (Black IT) with** two locations in Ellhofen and Ostermiething (AT).
- **Data Center One with** locations in Düsseldorf, Leverkusen and Stuttgart.
- **MyLoc Managed IT with** 3,500 m2 of space in six locations in Düsseldorf.
- **Plusserver** operates its own data centers in Cologne, Düsseldorf and Hamburg.
- **ScaleUp Technologies operates** a total of seven high-performance data centers in Hamburg, Berlin and Düsseldorf.
- **Maincubes One operates** three locations in Frankfurt, as well as one in Berlin, in addition to a Dutch location in Amsterdam, Germany.
- **Akquinet** operates four data centers in Hamburg, Norderstedt and Itzehoe.
- **CollocationIX operates** a high-security collocation data center at its Bremen location.
- **Grass Merkur operates** 3,500 m2 of RZ at the Hanover site

In addition, many Internet service providers also offer collocation services in addition to their classic IPS portfolio (e.g. M-net, Pfalzkom, Telemaxx, NetCologne, Docom, EnviaTel).

According to the expert interviewed, the cost of collocation offers in the planned range is between 100-120 EUR/kW/month. Add to this the electricity, which in the example is priced at 20 cents/kWh. This corresponds to the scale of approximately 4 MW required by LEAM with an estimated monthly cost of approximately EUR 400,000 to 500,000 for running operations.

Build and operate your own HPC data center

Another option is to build and operate your own HPC data center with the corresponding building infrastructure. The cost structure for the construction and operation of data centers is primarily based on the required availability class. Land costs often represent a minor role in the Total Cost of Ownership (TCO) view. The following example calculation is intended as a rough guideline and may vary due to local building conditions and special technical requirements:

500sqm Whitespace	approx. 4 MW IT performance	36 - 45 KW / Rack Power / Rack	115 Number of racks	UK 3 Availability	1,2 PUE
Total performance: 4 MW * 1,2 = 5 MW			35 ct / kWh Base of calculation Energy costs		
Invest / Construction: EUR 15,000 per KW installed = EUR 75,000,000					

When utilization of	60 % $4.000 \text{ KW} \times 0,6 \times 8.760 \text{ h/a} = 21.024.000 \text{ kWh/a}$ $21.024.000 \times \text{EUR } 0,35 = \text{EUR } 7.358.400$ IT electricity costs / year plus 1 MW Utility: EUR 3.066.000 = EUR 10.424.400 electricity costs / year	70 % $4.000 \text{ KW} \times 0,7 \times 8.760 \text{ h/a} = 24.528.000 \text{ kWh/a}$ $24.528.000 \times \text{EUR } 0,35 = \text{EUR } 8.584.800$ IT electricity costs / year plus 1 MW Utility: EUR 3.066.000 = EUR 11.650.800 electricity costs / year
	80 % $4.000 \text{ KW} \times 0,8 \times 8.760 \text{ h/a} = 28.032.000 \text{ kWh/a}$ $28.032.000 \times \text{EUR } 0,35 = \text{EUR } 9.811.200$ IT electricity costs / year plus 1 MW Utility: EUR 3.066.000 = EUR 12.877.200 electricity costs / year	100 % $4.000 \text{ KW} \times 1,0 \times 8.760 \text{ h/a} = 35.040.000 \text{ kWh/a}$ $35.040.000 \times \text{EUR } 0,35 = \text{EUR } 12.264.000$ IT electricity costs / year plus 1 MW Utility: EUR 3.066.000 = EUR 15.330.000 electricity costs / year

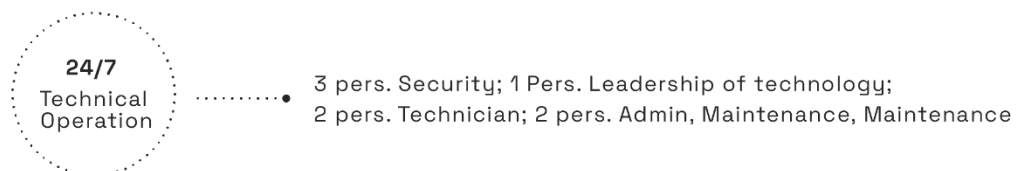


Fig. 23: Example invoice Building and operating your own HPC data center

For the commissioning of a data center with its own building infrastructure, the experts interviewed expect two to three years, depending on the length of the building permitting process, which can vary considerably depending on the municipality responsible. The view taken here is intended to exemplify the various dependencies and to provide guidance for the planning of a data center.

Use of a distributed infrastructure

It is questionable whether distributed computing of a large AI foundation model with multiple interconnected HPC data centers is possible. Basically, distributed computing is a possible approach to creating a large AI foundation model, but it is still largely untested. In addition, distributed computing places increased demands on infrastructure, network, latency (failure safety) and security. Initial studies on "Decentralized Training of Foundation Models in Heterogeneous Environments" come from Stanford University (B. Yuan et al., 2022). However, these systems are still very new and have not yet been sufficiently tested in practice. Thus, such a distributed computing model for LEAM would be associated with a very high risk, since the training process should not be interrupted when calculating a foundation model. In particular, computing on distributed infrastructure often fails to reach the necessary bandwidth when networking individual GPU nodes, which are switched with each other and with storage at bandwidths up to 900 Gbps.

8.6.1 Electricity price development and contract design

According to Béla Waldhauser, CEO of Telehouse Deutschland GmbH and CEO of KDDI Deutschland GmbH, electricity costs now account for 50% of the costs for data center customers in Germany (Weidmann & Krüger, 2020). Depending on their size, data centers buy the necessary electricity either from the local energy supplier or directly from the energy exchange. Here prices are currently volatile, ranging from EUR 0.20/KWh for normal operation to EUR 0.50/KWh for peak-load purchases. By comparison, electricity prices for suppliers in Norway, Finland and Sweden are around EUR 0.05-0.10/KWh. In addition, the lower Nordic temperatures provide additional advantages in terms of data center cooling. Fluctuations in the price of electricity can have a significant impact on the economics of the business model. This is also confirmed by the results of a survey by the Borderstep Institute, in which respondents ranked electricity price development as the biggest risk to the data center market (Hintemann et al., 2022).

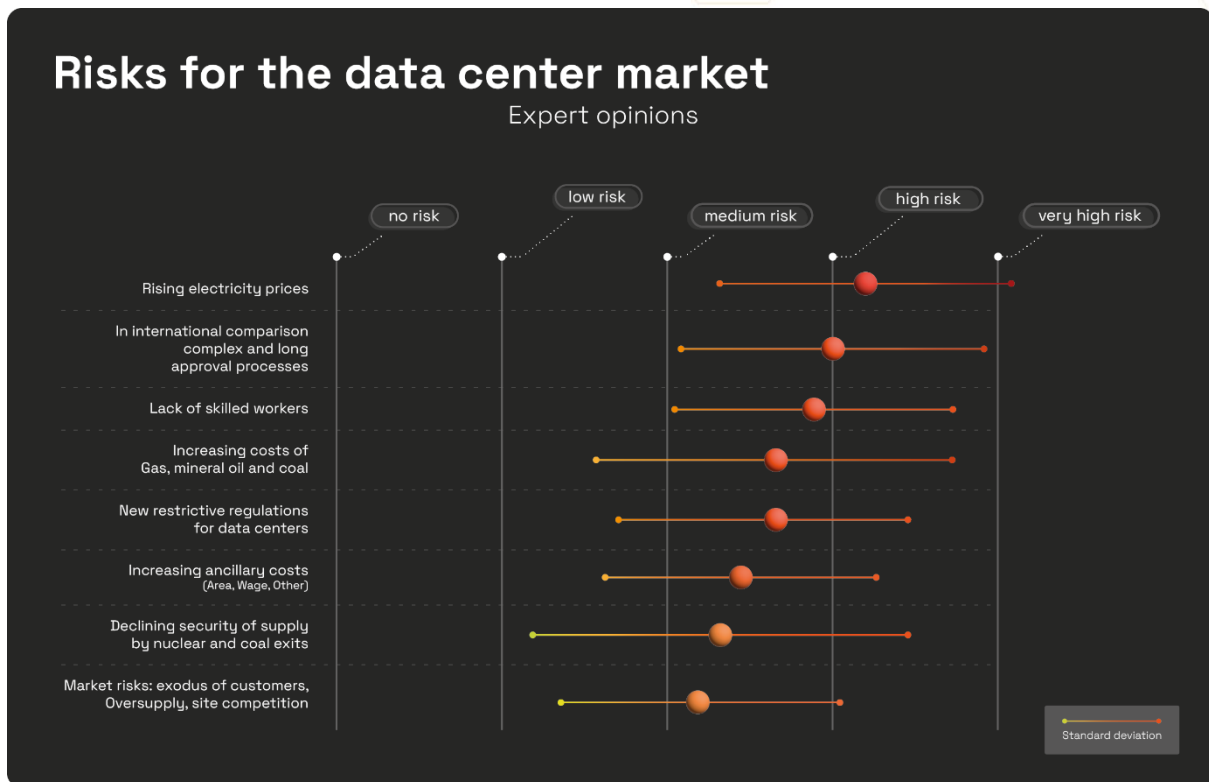


Fig. 24: Delphi query: How do you assess the following risks for the development of the data center market in Germany? (Hintemann et al., 2022, S. 37)

Different timelines are usually agreed between data centres and their customers when designing contracts for an electricity price cap. In doing so, the data center operator ensures that its collocation customers receive electricity at a fixed cost rate within the term. Since the data center operator has to purchase this electricity from the market at fluctuating prices, larger fluctuations pose a risk and may jeopardise the economic viability of the provider's business model.

8.6.2 Availability and period of procurement of the required IT resources

According to the interviewed expert, for example, delivery times for NVIDIA DGX H 100 systems are currently less than six months for larger systems. Manufacturer NVIDIA currently sees supply shortages more likely for the network components, which are around three to six months.

8.6.3 Aspects of sustainability

Last but not least, the aim of this study is to address the sustainability of data centers on the basis of the example of an important initiative. Under the *Climate Neutral Data Center Pact*, as part of a self-regulatory initiative, numerous associations and some 80 companies at European level are committed to creating standards for the sustainable operation of data centers (*Climate Neutral Data Centre Pact - The Green Deal Need Green Infrastructure*, o.D.). As part of a self-commitment, the members agreed on the following objectives:

- **Energy efficiency**

By 2025, data centers with 50 KW and more energy requirements in cool climates are expected to achieve a PUE of 1.3 under full load. Data centers have a score of 1.4 in warmer zones. New efficiency metrics for data centers will also be developed.

- **Green Energy**

In the future, data centers will meet their electricity needs by purchasing green electricity. 75% carbon-neutral or renewable energy by the end of 2025; 100% by the end of 2030.

- **Water use**

By 2025, all data centers operating at full utilization in cool climates in water scarcity areas that use drinking water for cooling will meet a maximum WUE (Water Usage Effectiveness) value of 0.4 L/kWh. Existing data centers installing new cooling systems will meet the WUE targets by 2040.

- **Circular economy**

Reusing, repairing and recycling servers, electrical equipment and other electrical components is a priority for data center operators. Here, high standards are set for recycling and the aim is to recycle 100% of the server equipment used in the future.

- **Use of waste heat**

The data centres are planning to expand the supply of waste heat to the general power grids and to other customers. This should be done in the most environmentally friendly and cost-effective way possible.

In order to further advance the development of data center sustainability while reducing costs, the data center industry is also currently monitoring and evaluating the subject areas of Refurbished IT, Remanufacturing and Re-Use, as well as the use of low-carbon building materials or the reuse of concrete (Bitkom e.V., 2022b).

8.7 Summary and recommendation

In conclusion, the HPC resources needed to calculate large language models are currently unavailable in Germany/Europe in the short term. The high demands on the GPU numbers (approximately 4500) and the corresponding rapid networking of the individual GPUs with each other cannot be guaranteed today or would require computing times that would completely deplete the most powerful German HPC centers for almost two years and would not be able to guarantee the required short innovation cycles. The relevant cloud services from the US and China can provide sufficient AI computing capabilities in part. However, they are difficult to exploit while respecting digital sovereignty and European data protection requirements. In addition, lock-in effects are usually effective here, which make it difficult to switch to another provider later.

Due to the many factors mentioned, a site recommendation can only be given in relation to the specific use case. This case-by-case assessment cannot be carried out in the context of the study. However, discussions with representatives of various federal states and regional initiatives have shown that there is a general willingness in various federal states to build a data centre.

Building your own data center would generally be significantly more expensive and would prevent a rapid start to development activities. Including the required planning, approval and installation phases, it would take up to three years before a corresponding AI data center could become operational. This is too long because of the need for action already described. Therefore, for reasons of time and cost, it should be avoided to set up an AI data center with its own built infrastructure.

A collocation model is a short-term feasible option for operating your own computing capabilities. As demonstrated in the example of the GSI Helmholtz Center for Heavy Ion Research, there are already providers offering short-term green electricity and sustainable waste heat-use capacities that meet requirements for the required water cooling and enable scalability. Offers from operators in Norway, Finland, Sweden and Iceland offer an additional operating cost advantage over the electricity price due to the lower average outdoor temperatures. These countries also follow European data protection standards (Sweden and Finland are EU Member States, while Iceland and Norway are members of the European Economic Area (EEA)) in terms of law and, in particular, data protection law, and therefore receive a recommendation from the authors. However, depending on the use case, latencies may be required for inference applications that Norwegian, Icelandic and Swedish providers may not be able to meet due to large geographical distances.

Another option is to use newly created HPC capabilities that could be provided by a vendor, for example, by billing for GPU hours. This possibility is discussed in more detail in Chapter 10 in the context of business considerations.

According to the expert interviewed, the relevant European and international data centers and collocation providers mentioned above are already evaluating major investments in HPC infrastructures in Europe, specifically in concentrated AI-friendly capacities with very large numbers of up to 20,000 GPUs. Different cooperation opportunities are evaluated in parallel. The Open GPT-X project is already exploring the development of large language models in collaboration with the Jülich Supercomputing Centre (JSC). In close collaboration with the project, LEAM will incorporate these findings into its further infrastructure planning.



The organisational structure of LEAM

.....●

9. The organisational structure of LEAM

The previous chapters have shown that building an AI high-performance data center is crucial for the technology and industrial location of Germany. The set-up and operation are feasible. The question of the economic feasibility of such a centre is open. The economic feasibility is discussed in more detail in the following chapter.

Based on interviews with data center operators and industry, the authors developed the idea of a LEAM AI Service Center (LKS). The specific organisational structure behind the idea of the LKS will be explained in more detail below.

9.1 Target audiences of the LEAM AI Service Centre

The aim of the LEAM AI service centre discussed in this study is to develop AI foundation models based on European values and particularly to make them usable for established industry and young technology companies. The LKS has two objectives:

- (1) Provide AI High Performance Computing Center capabilities available from business and academia for training in-house models.
- (2) To offer open source AI foundation models that can be tailored (tuned) to the individual needs of industrial companies with appropriate consulting and service services.

The capabilities and services are to be offered to four target groups. The target groups are:

- **Economy:** The business target group includes private sector organisations that use the services of LKS for the development of models, applications or products.
- **AI startups:** As with the target group for business applications, these are private sector organisations. However, they differ from the first in two respects:
 - (1) AI startups are predominantly working on Artificial Intelligence and
 - (2) they are young companies.

Start-ups should be supported specifically through specific funding projects. An example is an AI-Compute voucher designed to provide startups with access to the services of the LKAS.

- **Public:** Public institutions, authorities, ministries and services can use the LKS services to optimize administration or to gain specific security-related insights based on large amounts of data. The use of LKS Services by public institutions is subject to specific requirements regarding data security, data protection and general confidentiality policies.
- **Science:** The target group for research and development includes universities, non-university research institutes and government research institutions that use the services of the LKS for research and, in particular, support the creation of foundation models.

9.2 Organizational units of the LEAM AI Service Centre

The LEAM AI Service Centre (LKS) is structured in horizontal and vertical organizational units (Fig. 25).

The horizontal organizational units form the infrastructure and capacity base of the LEAM AI Service Centre. They ensure the operation of the infrastructure and the provision of the core services. These include:

- Housing
- Infrastructure-as-a-Service
- Training-as-a-Service

The vertical organizational units build on the horizontal organizational units and are divided into four different core services for customers and users. These include:

- KI-Foundation-Model Development
- AI Model Tuning
- Inference
- Consulting

In addition, the "Coordination" organizational unit, as the management unit, is responsible for the development of the LKS, maintaining close contact with the LEAM Board (see Chapter 9.3) and other political, academic and business stakeholders. Depending on the type of company chosen, it may also assume the functions of a body required by company law, such as the management or the board of directors.

The individual organizational units can be mapped within a company or divided into different legal units according to a governance concept.

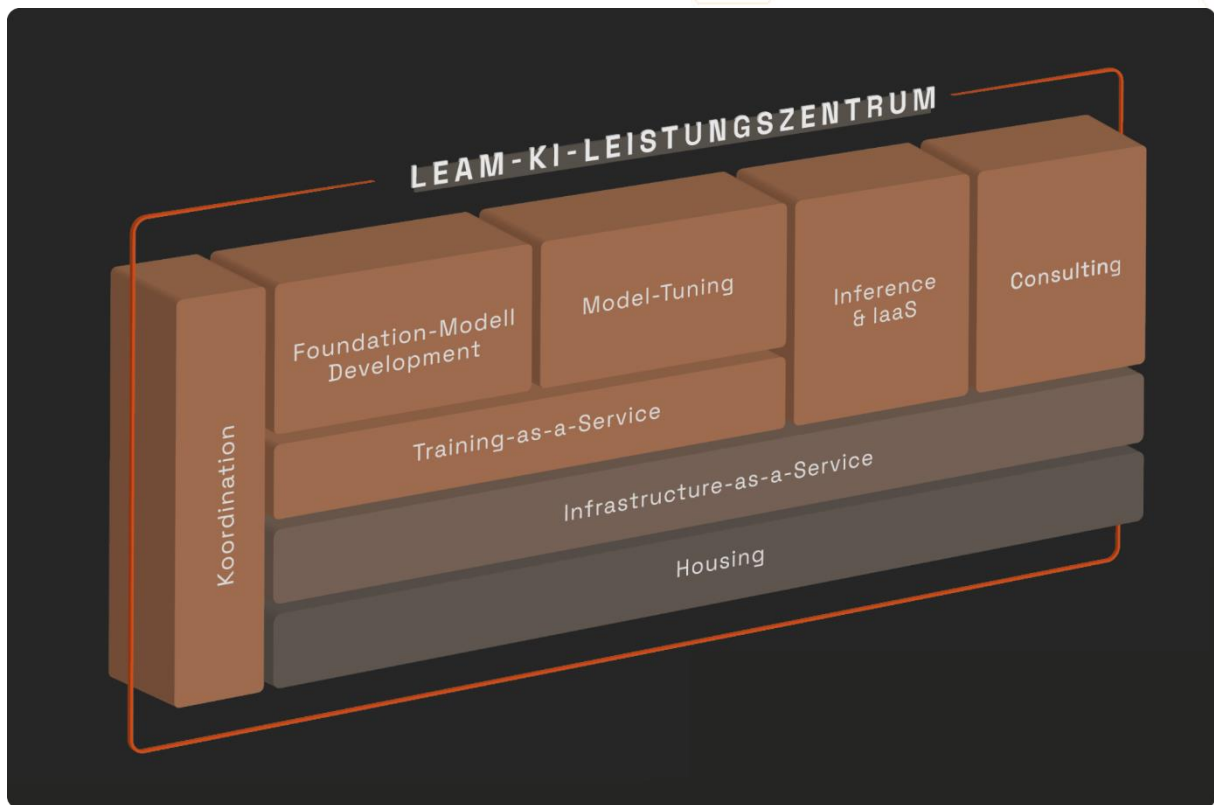


Fig. 25: Organizational units of the LEAM-KI Service Centre

9.2.1 Coordination

The OE (Organizational Unit) Coordination serves as the management unit for the entire LKS. It ensures the continuous construction and operation of the LKS and expands it within Germany and the EU.

The OE is a point of contact for interested people from business and academia and keeps in touch with politics. It coordinates the interests of stakeholders, establishes and manages the horizontal organizational units, conducts market studies and acquires funding and investment funds. As the controlling unit of the LKS, it supports the other services and takes over the marketing of the LKS.

This area should initially be built together with the consulting area, regardless of the chosen organisational scenario.

9.2.2 Housing

OE Housing provides the basic infrastructure for the operation of the data center. These include e.g. building infrastructure, electricity and telecommunications, as well as a cooling concept. Chapter 8 summarises the specific requirements for equipment.

As the investment in the required infrastructure is high and it will take several years to rebuild a data center, it should be avoided to build a new one at this stage. Instead, various organizations offer housing services that can be purchased or rented. A suitable housing partner must be found to build the data center.

Governance:	Outside LEAM
Organizational interfaces:	Coordination; Infrastructure-as-a-Service
Cost Collocation HW:	EUR 1,344,000 per year

Table 13: Overview of OE Housing

9.2.3 Infrastructure-as-a-Service (IaaS)

The building and operation of an AI supercomputer is a central component of OE IaaS. The computer requires the services of the housing infrastructure to operate. The purchase of the computer is one of the largest items in the overall budget and is in the three-digit EUR million range (see Chapter 10). For more information on the technical requirements and other aspects of the computer, see Chapter 8.

There are basically two scenarios for the organizational structure of OE IaaS: On the one hand, the acquisition and operation of the computer by the LKS is conceivable, on the other hand, the infrastructure of the LKS can also be provided as a service by an external company.

Due to the high relevance of the chosen scenario for LKS investment and operating costs, as well as organisational circumstances, these two scenarios are discussed in more detail below.

9.2.4 IaaS within LEAM

In the first scenario, the LKS is responsible for the acquisition of the computer as well as the operation and provision of the infrastructure as a service. This has the following advantages and disadvantages.

ADVANTAGES OF HAVING YOUR OWN INFRASTRUCTURE

Extensive control and design possibilities

When purchasing an AI supercomputer, individual components, components, and associated frameworks and software can be selected independently. This allows for a more flexible design and alignment to the future use of the data center.

Support for a self-contained business model

Investing in your own AI supercomputer makes it possible to continue using it after the depreciation period. The dependence on possible cost fluctuations and external service providers is limited to the costs necessary to operate.

No data sharing with third parties

When using your own data center, no data about usage etc. is passed on to third parties. This simplifies possible questions regarding third parties (e.g. security services). Similarly, this way, cybersecurity requirements (for example, BSI requirements) that must be demonstrated to be met in order to participate in certain public funding programs or public tenders, can be better demonstrated when using an in-house data center. In particular, in the case of public funding, the following should also be taken into account: The purchase of the services of an external AI supercomputer is subject to a public tender procedure. A corresponding purchase contract may be drawn up for a longer period, but must be re-invited periodically. This may limit the willingness of an external service provider to invest.

DISADVANTAGES OF HAVING YOUR OWN INFRASTRUCTURE

High investment costs

The purchase of an AI supercomputer involves a very high initial investment. The depreciation period is assumed to be four years.

Complex financing

The financing and construction of the AI supercomputer raises complex questions of financing, ownership and rights of use.

Establishment of its own business unit

As the data center operator, there are high costs for staff and software. The management and maintenance of an AI supercomputer is complex and needs to be monitored by appropriate staff around the clock. In addition, it is necessary to be able to establish adequate IT security. This entails considerable administrative burdens and high costs. Particularly in the case of funding projects or public administration contracts, there may be additional requirements and associated burdens in demonstrating cybersecurity standards.

9.2.5 IaaS through an external partner

In this scenario, the LKS purchases the computing capabilities from an external partner. The partner will undertake the complete installation and operation of the Compute infrastructure, using its experience in the hosting sector from existing business models.

ADVANTAGES OF EXTERNAL INFRASTRUCTURE

Low investment costs for own staff, hardware and software

When providing an AI supercomputer through an external partner, it is also responsible for the availability and functionality of the facility. As a result, there are comparatively low personnel, hardware and software costs, only support functions need to be taken over.

Flexible scalability of the AI supercomputer

Depending on current availability and needs, capacities can be booked or cancelled flexibly if agreed. This makes it easier to adapt to the current needs and catches any spikes.

Access to professional know-how and established processes

An expert from an external partner can, if necessary, provide technical expertise to assist with questions and thus simplify the use of the services for customers. Furthermore, the billing and operation can be based on established processes and experiences - e.g. in the area of IT security standards - the third-party vendor is used.

DISADVANTAGES OF EXTERNAL INFRASTRUCTURE

Sharing of business-related data with third parties

When performing services on the infrastructure of an external provider, usage data may be passed on to them. In this context, data security considerations must be taken into account and provisions for sharing data must be carefully reviewed.

Dependence on external suppliers

The use of the infrastructure of an external provider is and is coinciding with the availability of the capacity of that provider. If this is not the case, even in the short term, sales losses and compliance difficulties can be expected. In order to avoid this, external suppliers must be able to guarantee the availability of capacity and to contract with the customer.

Market-dependent rolling costs

The cost of using the infrastructure of an external vendor continues to rise and is higher than the cost of running your own data center. Costs may vary, depending on contractual arrangements, due to market movements.

Risk of takeover or insolvency

The infrastructure of the external provider is under its control and could either be relocated to another country, potentially adversely affecting data security aspects, or purchased. There is also the possibility that the infrastructure, e.g. is no longer available after insolvency

Recommendation

The experts from industry and academia interviewed consider it desirable to build their own high-performance and flexible supercomputing infrastructure, since the private and scientific infrastructure available so far does not meet these criteria. This has been shown by the interviews in the framework of this feasibility study, but also by the recommendations of scientific initiatives, e.g. of the European Language Equality Project.¹⁹

A decision-making basis for financing the infrastructure (own hardware or external infrastructure) needs to be further analysed and prepared, also taking into account the high investment costs.

9.2.6 Training-as-a-Service

Based on the IaaS and the computing capabilities that it provides, processes are established to provide LEAM users and customers with extensive higher-level services. In this way, the LKS allows potential users to use the infrastructure to train their models.

The training of AI foundation models requires extensive measures to build and operate the corresponding processes and software infrastructures (see Chapter 7).

The activities can be categorized into different levels (layers). Activities include:

- System Layer
 - Management of AI accelerators
 - Provision of (open source) software and software libraries
- Data Layer
 - Ensure efficient and stress-resistant data storage and loading routines
 - Compliance with legal and regulatory requirements regarding privacy, data quality and data security
- Training and Application Layer
 - Build an effective optimization system to optimize the utilization of GPUs during the training process
 - Building a system for managing training jobs
 - Implementation of evaluation and benchmarking processes
- Framework & Service Layer
 - Normalizing the training environment by providing containerized environments
 - Implementing user-friendly interfaces for developers and administrators

¹⁹ "Current LT research also requires flexible access to High Performance Computing (HPC) facilities in the form of clusters of high capacity GPUs. There are many EU initiatives offering HPC: EuroHPC JU, PRACE, national computing facilities, etc. However, it is unclear if these initiatives are ready to provide the computing support that the European LT research community currently needs for developing state-of-the-art language models for all languages, domains, tasks and modalities." (ELE Consortium 2022, p. 23).

In order to offer this service, an expert team consisting of Data Science, Data Engineering, Machine Learning Engineering and DevOps Expert must be built. This is merged under the Team Services and forms interfaces to the Team Consulting.

Governance:	Within LEAM
Organizational interfaces:	Coordination; Housing, Training-as-a-Service, External Clients
Cost of office space:	EUR 180,000 per year
Cost of Team Services:	EUR 3,000,000 per year

Table 14: Overview of the Training-as-a-Service

9.2.7 AI foundation model development

The AI foundation model development service is the core service of the LKS. Under this service, all activities directly related to the development of new AI foundation models are summarized.

LEAM will offer the AI foundation model development service to private, public and academic institutions. However, due to the high complexity and investment costs, the basic AI foundation model Development is expected to be predominantly used by research and business collaborative projects. Interviews with the research have shown that there is a strong interest in developing new and competitive AI foundation models. In addition, they will use the service to answer scientific questions.

If the LKS is (partially) publicly funded, at least 80% of the models must be made available open source. A maximum of 20% can be used by companies with exclusive rights. In principle, an opportunity, because our survey of AI companies revealed that the lack of open source AI foundation models is one of the main barriers for AI companies to be able to use them productively. This, together with the availability of data and high costs, was identified as a barrier by 58% of the companies surveyed.

Research projects can then apply for computational time using a process to be defined. This will draw on the experience of the data centers organized at the Gauss Centre for Supercomputing. However, in order to meet the specificities of LEAM, a few points should be taken into account:

- ✓ Interested projects will be given the opportunity to apply for projects on an ongoing basis. This gives research projects the flexibility they need to do time-consuming research in fast-paced AI research. In addition, this is a way to mitigate spikes in infrastructure usage, as not all projects start calculating their models at the same time.
- ✓ The application and evaluation procedures are to be made as simple and flexible as possible. A common criticism of the current HPC landscape is that the application phase is taking too long and binding important resources.
- ✓ Priority should be given to application-related research projects over basic research.

In the area of AI foundation model development, the following services are offered:

- **Creation and provision of general training data sets:** A general training data pool will be set up, maintained in accordance with privacy policies and quality criteria, and made available to interested organisations.
- **Provision of basic algorithms:** A repository provides code bases of existing (open source) programs, necessary auxiliary tools, and other frameworks.
- **Management and provision of trained foundation models:** The trained foundation models are stored and managed in a repository for further use.

In order to provide the best possible service to interested organizations, a number of conditions must be met:

- **Recruitment of employees:** Staff are recruited and trained to accompany and supervise the training.
- **Allocation of computing resources:** To enable users to use the AI supercomputer to train foundation models, LEAM staff coordinates the management and optimal distribution of compute resources to users.
- **Development of a billing model:** A billing model will be developed for the training of the models. The focus is on existing services in the HPC field or in the commercial cloud services environment.
- **Management of the training data pool:** Training records are collected in repositories and made available to users to develop AI foundation models. Employees: assist the LKS in the collection and maintenance of relevant data sets.
- **Frameworks support:** For the training of AI foundation models, relevant frameworks are collected in repositories and made available to users. LEAM staff assist users in the application of these in the context of the AI supercomputer.

9.2.8 Model Tuning

In addition to developing AI foundation models, resources and infrastructure are provided for tuning models. This is necessary in order to add domain-specific knowledge to the general AI foundation models in order to use and optimize them for specific applications.

Most of the companies surveyed are interested in concrete applications based on AI foundation models. 51% of AI companies surveyed are already working with AI foundation models and 18% are planning to use AI foundation models in the future (see Chapter 4).

Existing models are to be expanded through model tuning and made usable for specific purposes. Currently, only 27% of the companies surveyed that deal with foundation models do so. Thus, it is foreseeable that the demand for model tuning will increase in the near future. LEAM can substantially simplify this for German industry by pooling expertise, models, data and infrastructure and allowing companies to perform model tuning without much extra effort. A particular focus here should be on providing computing capacity to AI startups wishing to further develop their models for industrial applications.

In addition, scientific institutes are also interested in using the model-tuning service. Similarly, the combination of expertise, models and data provided by the LKS is of interest to science.

The following services will be offered in the tuning area:

- **Tuning-as-a-Service:** Customers can order the tuning of models from the data center. In cooperation with the customers, the employees of the data center are adapting the foundation models.
- **Advising and providing expert support** in the selection of models, data and algorithms.

In order to provide customers with the best possible service, a number of conditions must be met:

- **Recruitment of employees:** In order to advise and assist users in the operation of the infrastructure and in the selection of the model and the data sets, LKS hires staff.
- **Allocation of computing resources:** To enable users to use the AI supercomputer for tuning foundation models, LEAM staff coordinates the management and optimal distribution of compute resources to users.
- **Development of a billing model:** A billing model for tuning the models needs to be developed. For example, a GPU/hour model is available.
- **Management of the training data pool:** Training records are collected in Reopsitorys and made available to users in order to develop AI foundation models. Employees: LEAM collects and maintains relevant data sets.

SPOTLIGHT

SAP SE

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Dr. Feiyu Xu, Vice President
and Global Head of AI, SAP

SAP SE is a publicly traded international software group headquartered in Walldorf, Baden-Württemberg. As a leader in business software, SAP helps companies of all sizes and industries achieve their goals in the best way possible: SAP customers generate 87% of all global trade.

Where do you use foundation models? What is your use case?

foundation models will play a very important role in enterprise AI, leading to disruptive innovations in process management. These include new applications for business process mining, business process prediction, business process composition, and process optimization.

In addition, the adapted large language models will also contribute to better human-machine interaction via chatbot or digital assistants, and will also significantly improve the semantic processing of business documents, particularly information extraction and entity linking.

Another field of application is the generation of program code by the foundation models. This increases the efficiency of programmers. Because users can then also have new tools or extensions of software programs created by AI, this also leads to a democratization of software development.

The potential use cases can be classified according to the data structures required for their I/O:

1. Text2Text or Speech2Speech:
 - Chatbots and Digital Assistants are important applications for SAP that allow users, such as employees of a client company, to ask natural language questions and get answers about facts or transactions. Digital assistants can be used for customer services.
 - Translation and localization of software, services and business documents
 - Summaries of text documents or meeting transcripts
2. Text2Process and Process2Text
 - The integration of BPMN (Business Process Modelling Notation) into the language models allows for natural language requests for the generation, validation and execution of the business processes. In the other direction, textual process descriptions that are needed for training or certifications, for example, can be generated to existing process models.
3. Process2Process
 - This includes process composition, process modification, **process** validation and process optimization.

4. Text2Code: low-code/no-code

- The automatic generation of program code increases the efficiency and productivity of programming

What impact do AI foundation models have on your business model or your projects?

On the one hand, foundation models result in large incremental improvements, as better results are achieved with significantly less (if not at all) training data. But they also have the potential to enable new and modularised end-to-end business processes, revolutionizing the entire enterprise resource planning landscape.

We will give a first example from the field of business document processing. Adapting existing deep learning models currently requires a significant amount of effort.

Customers need to provide large amounts of training data for their specific document formats. The localization of the models in other languages is also not feasible without training data. AI foundation models have the potential to significantly reduce the effort in both areas. Through foundation models, we can make a new type of offer to customers: Processing their specific formats in the languages relevant to them, with little training data or even out-of-the-box.

What difficulties and problems do you see that only the U.S. and China are currently providing AI foundation models on a comprehensive basis?

Europe needs its own AI computing infrastructure to avoid becoming dependent. In addition, access to data, including multilingual content and explicitly coded knowledge, is essential. AI in Europe can only be successful in the long run if we manage to train and retain AI talent. An AI-friendly policy is also an important framework condition for Europe.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

Language coverage is essential to us, as is transparency on the data used. There is a need for an infrastructure that allows AI models to be generated and made open-source, while ensuring transparency and adhering to European data protection standards. Bias is a critical issue in many business areas, such as human resources and application processing. Here, as well as in other AI applications, AI models can only be used if they comply with relevant regulations such as the AI Act of the European Union.

9.2.9 Inference

The LEAM KI Service Centre will also be used to provide the models to potential customers. The service will be offered at market prices, with around 5-10% of the compute infrastructure to be used for the delivery of the models.

Companies are particularly interested in the productive use of applications based on AI foundation models (see Chapter 4). In particular, companies that do not have an independent AI department and therefore cannot develop or adapt their own models will make use of this possibility. Since currently only 50% of the companies surveyed that use AI also use foundation models, there is a relevant target group for this service. This service is of less importance to science seeking to answer fundamental questions about AI foundation models.

Based on the Inference Services, AI companies can develop and offer their own AI products and applications and make them available to their customers, for example through an API and specific billing models.

Various services are offered in the Inference area:

- **Hosting-as-a-Service:** Customers can order the deployment of applications from the data center. About 5 to 10% of the total infrastructure is reserved for this service. Employees: In the data center, they support and coordinate the projects.

In order to provide customers with the best possible service, a number of conditions must be met:

- **Recruitment of employees:** LKS hires staff to set up and operate Inference APIs.
- **Allocation of computing resources:** To enable users to use the AI supercomputer for tuning foundation models, LEAM staff coordinates the management and optimal distribution of compute resources to users.
- **Development of a billing model:** A billing model must be developed for the provision of the models. It offers a method of calculating the fee based on the amount of tokens used (single requests to the model).
- **Training of users:** In order to be able to use the infrastructure, potential users must be trained. For this, training must be prepared and offered.

9.2.10 Consulting

In the area of consulting, a consultancy service is provided to assist clients from different target groups in the development, optimization and implementation of AI applications. This service is independent of the provision of data center infrastructure services.

The target audience for Consulting Services is all scientific institutions, start-ups and companies that want to develop their own foundation models or adapt existing foundation models to their specific requirements.

Close collaboration with business consulting firms and their ability to provide services in the context of AI foundation models is also part of the activities of this OE.

In the area of consulting, the following services are offered:

- **Advice:** Customers interested in AI applications can seek advice. This will support organizations in the development, optimization and implementation of AI applications based on AI foundation models.
- **Training & Workshops:** Interested companies as well as research institutions will be specifically prepared for the realities of the development of foundation AI models using the LEAM AI supercomputer.
- **Training by external consulting firms:** In order to familiarize external consultants with the specific circumstances and technology, they must first be trained. This is achieved through close cooperation between external consultancy and the in-house staff of the LKS or external companies that are already consulting.

In order to provide customers with the best possible service, a number of conditions must be met:

- **Establishment and operation of an advisory team:** In order to offer this service, a team of expert consultants must be available. This should consist of subject matter experts for data science and machine learning. Alternatively, external advice may be used.
- **Practice Work:** The LKS consulting team is constantly developing its own skills in the development of foundation AI models in order to always cover the most relevant areas and technologies in the consultancy.

9.3 The LEAM-Board

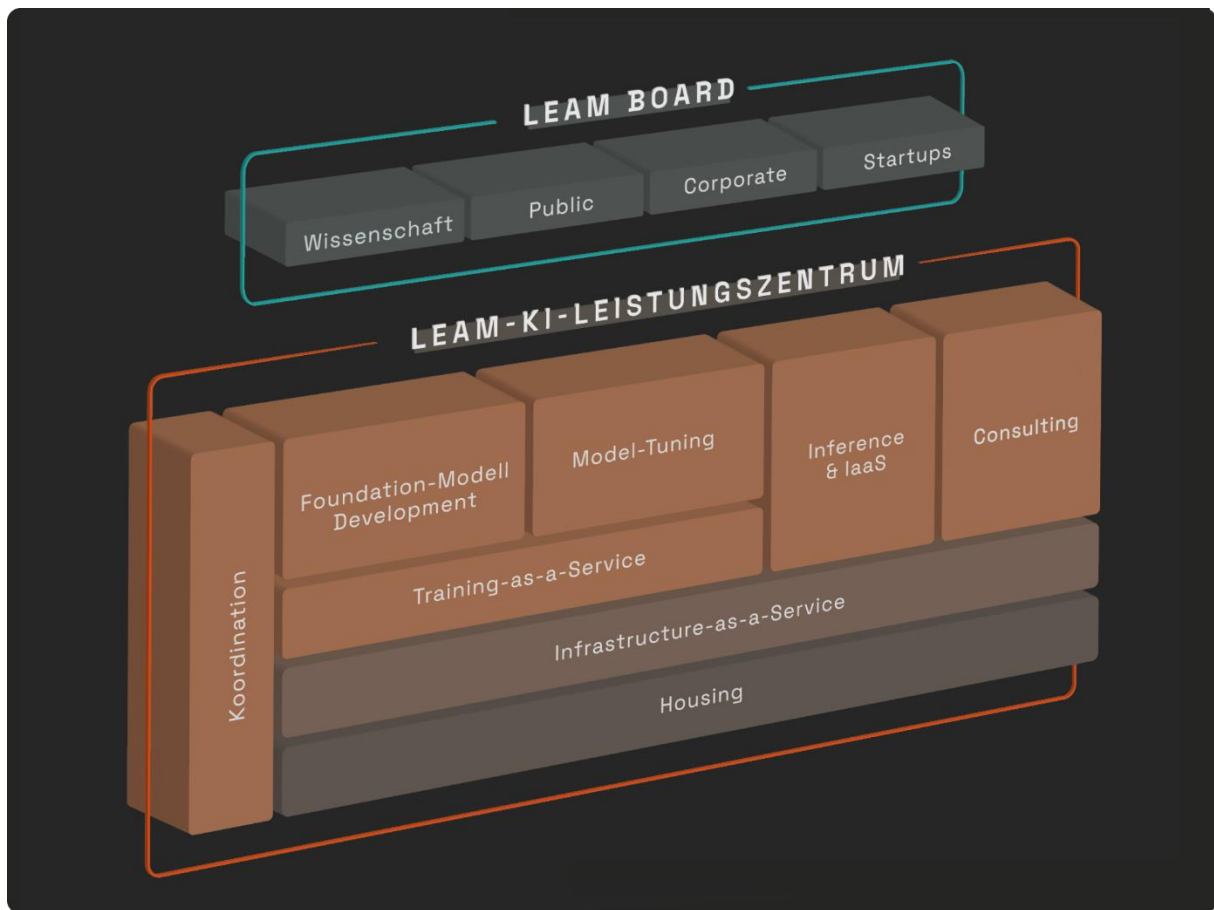


Fig. 26: The LEAM Board as the central governance unit of the LKS

The LEAM board monitors and controls the strategic direction, objectives and further development of the LKS. The LEAM Board includes decision-makers and experts from academia, politics, business and start-ups. The LEAM Board then decides on investments and the allocation of computing resources for research and development.

9.4 Summary

The LKS aims to offer customers four core services, which are particularly targeted at businesses. The AI foundation model development is of the utmost importance. The goal must be to optimally support clients in the development and productive use of AI foundation models. The Service Consulting assists with the development of an AI model, up to the productive use of the Inference Service. However, interested companies and research can also use individual services such as model tuning, inference or infrastructure. This means that the target audience of the AI data center is large and that the AI supercomputer can be used flexibly. A team of LEAM staff will provide advice and support to help companies develop AI foundation models.

These services are based on the organizational units Housing, Infrastructure-as-a-Service, and Partly Training-as-a-Service. The housing unit will be transferred to external partners for the most part. This saves the LKS high investment costs and reduces the time to launch the LKS. The Coordination Unit supports the other units and coordinates the operation of the LKS.

Business management aspects



10. Business management aspects

The LEAM KI Service Center will generate unique growth momentum through the trained and provided foundation models, achieving huge efficiency gains across the industry. After the initial investment required, either as public, private, mixed full funding or, in the case of external IaaS, with the help of a GAIA-X compliant infrastructure operator at best, the computing infrastructure will be made available to significant parts of academia and companies developing in the open-source process to develop foundation models. The developers use this infrastructure to train the latest generation foundation models and establish ways to standardize fine-tuning processes. The models, as well as the procedures and processes for training the models, are made available either open source or at the lowest possible cost.

Established businesses as well as start-ups can quickly incorporate these technologies into their products, as they come from reliable partners and have been developed to European standards. By making it available at a cost price, business companies are thus independent of American HyperScalers. As a result, the barrier for business to incorporate these models into their products or to offer new products on this basis will be reduced. Businesses also benefit from the developed standardized blueprint processes for training, tuning and delivery. By pooling know-how within the LKS, business and start-ups can learn and receive expert advice from experienced experts along the entire AI foundation model lifecycle (development, training, tuning, deployment, application).

In this way, LEAM makes a significant contribution to attracting rare talent through the building of experience and the expected spillover effects. At the same time, there is a transfer of knowledge into the business about the key technology "KI", which can use LEAM services in finetuning and consulting expertise in implementation in the Inference area before building their own teams.

10.1 Costs

The funding needs for the LEAM KI Service Centre consist of

- Investitionskosten (CapEx)
- Operating costs (OpEx)

The annual cost situation depends, among other things, on the depreciation period of investment costs, which is assumed to be four years.

In the case of building the infrastructure through an external partner and purchasing data centre services, the costing should be adjusted accordingly.

The costs incurred depend on a variety of parameters, which can only be estimated in this feasibility study.

The following table shows the parameters relevant to costing, the assumptions made and the possible ranges beyond them:

	Unit	Adoption	Comments - Range
Hardware			
Depreciation period	Years	4	Generally 3-5 years
Hardware (incl. Storage, Netw. Support, SW)	kEUR	260.000	Model NVIDIA Selene
Number of DGXA100	Number	560	Alternative HW: Cerebras, Graphcore, Intel, AMD, Nvidia H100
DGXA100 per Rack	Number	2	gem. Info by NVIDIA
Operating costs Infrastructure			
Rental Colocation HW	[EUR/Rack Months]	400	Average
Rental of office space	[EUR / qm]	30	15 - 40 EUR/m2
Office space	qm	600	Average 10 sqm / MA
Peak power consumption	kw	4.200	gem. Info by NVIDIA
Electricity price	EUR/kwh	0,35	a high level of uncertainty due to the current geopolitical situation
Energy efficiency	Factor	1,50	relatively high value, may be lower in a Green Data Center
Continuous load of max current intake	%	65	50 %-90 %
Team Operations	FTE	20	Administration Hardware and SW Infrastructure
Costs Organization and Services			
Team Business Administration	FTE	20	Governance, administration, sales, marketing
Team "Training-as-a-Service" & Consulting	FTE	20	Implementation of operations, processes, services
Staff costs	kEUR/FTE/year	150	Average

Table 15: Overview of the costs of the LEAM KI Service Centre

Note: The NVIDIA A100-based architecture underlying this computational model will be replaced by the newer, more powerful H100 line by manufacturers in 2023. This will also lead to an improvement in the performance and price ratio.

This gives the following overall cost structure on the basis of an operation and a four-year depreciation period:

Invest	Year 1 kEUR	Year 2 kEUR	Year 3 kEUR	Year 4 kEUR	Total kEUR
Infrastructure	88.732	88.732	88.732	88.732	354.927
HW depreciation	65.000	65.000	65.000	65.000	260.000
HW Collocation	1.344	1.344	1.344	1.344	5.376
Electricity consumption	19.316	19.316	19.316	19.316	77.263
Team Operationen	3.000	3.000	3.000	3.000	12.000
Office room 1/3	72	72	72	72	288
Organization and Services	6.144	6.144	6.144	6.144	24.576
Staff	6.000	6.000	6.000	6.000	24.000
Office room 2/3	144	144	144	144	576
Total	94.876	94.876	94.876	94.876	379.503

Table 16: Total cost structure of the LEAM KI service centre with a four-year depreciation period

Investment costs

The central part of the investment costs will be absorbed through the purchase of an AI supercomputer.

For the calculation of these infrastructure costs, the following scenario is assumed:

- Building Operation of an AI supercomputer on the order of magnitude of the NVIDIA Selene (Wikipedia Contributors, 2022).
- The duration of a training course for a GPT-3 model is approximately 1-1.5 weeks on Selene.
- At the heart of Selene is the NVIDIA Superpod architecture based on the DGX A100.
- The size of an NVIDIA Selene is 506 DGX A100 nodes with 8 GPUs each, in total.
- The acquisition cost is in the range of 260 million euros.

On this basis, an example calculation is only carried out here. Alternative solutions (e.g. Graphcore, Cerebras, AMD, Intel) or NVIDIA's latest architecture (H100) are analysed and evaluated at the time of the infrastructure purchase or tender.

Operating costs

The operating costs of the AI supercomputer consist of collocation, energy requirements, rental costs and personnel costs.

- **Rental Collocation HW:** To operate the AI supercomputer, appropriately equipped premises and server racks are needed. These must be externally hired and equipped with the AI hardware to be procured. Two DGX A100s can be installed per server rack. Rental costs of EUR 400 per month per rack result in an annual cost of EUR 1,344,000.
- **Energy costs:** The energy demand of the plant is approximately 4,200 kWh. It is highly dependent on the utilization of the plant. An average continuous load of the system of 65% is to be assumed. In the current situation, electricity costs are highly volatile. Although the electricity price for industrial customers is currently even lower, it is currently planned at 0.35 EUR/kWh. This results in an annual cost of EUR 19,316,000.
- **Team infrastructure:** The Infrastructure team takes care of the operation of the hardware and SW infrastructure. A team size of 20 FTE (Full-Time Equivalents) is required. A flat-rate wage cost of EUR 150,000 per FTE is foreseen.
- **Office rental:** In order to provide LEAM employees with a place to work, appropriately equipped office space must be rented. It is currently planned with an office of 600 m² at EUR 30 per square metre. This corresponds to an annual cost of EUR 60,000 .
- **Team Koordination:** The Business Administration team deals with governance, sales, marketing and administrative activities around the LEAM project. A team size of 20 FTE (Full-Time Equivalents) is required. A flat-rate wage cost of EUR 150,000 per FTE is foreseen.
- **Team Services and Consulting:** This team is dedicated to the implementation of processes and services around the services that LEAM offers. A team size of 20 FTE (Full-Time Equivalents) is required. A flat-rate wage cost of EUR 150,000 per FTE is foreseen.
- **Office rental:** In order to provide LEAM employees with a place to work, appropriately equipped office space must be rented. It is currently planned with an office of 600 m² at EUR 30 per square metre. This corresponds to an annual cost of EUR 60,000.

Alternative cost situation when purchasing GPU-RZ features

When purchasing computing power, the initial investment costs for LKS are almost completely eliminated by the lack of need to procure AI hardware. A prerequisite for this is the willingness of a cloud service provider to invest in the infrastructure, taking into account the following framework conditions:

- It must be ensured that a sufficiently large compute cluster is provided (approximately 4500 GPUs).
- LKS will guarantee a minimum amount of computing capacity (e.g. 60%) over the four-year period.
- Purchases of computing power or GPU hours are made at market prices (see above).
- The details of the calculation and the design of the contract must be defined.
- The selection of a cloud service provider may take place in the context of a public call for tenders, the details of which have to be defined.

In this scenario, the following cost items occur, as shown in Table 17:

	Adoption	Annual cost in kEUR
Purchase of data center capabilities / GPU hours	60% of the cost calculated for the own operation	53.239
Organization, Training-as-a-Service and Consulting	According to calculation with own RZ	6.144
Total		59,383

Table 17: Cost of the LEAM-KI service center when purchasing the GPU-RZ power

10.2 Revenue

Different sources of income can be defined by the services offered by the LKS:

- ✓ Sales of data center capacity (GPU hours)
- ✓ Services for training foundation models (Training-as-a-Service)
- ✓ Services for tuning tailor-made models (training-as-a-service)
- ✓ Inference Service (GPU hours)
- ✓ General consulting activities (Consulting)

The ways to generate sales must be differentiated according to the target group:

- **Economy:** Corporates and SMEs are offered services at market prices when developing models for private sector operations. In the case of research projects that are open-source, the economy receives services free of charge or at low prices. This depends primarily on the governance and funding of the LKS.
- **Academia:** Projects that work in the open-source process receive services free of charge or at low prices. This depends primarily on the governance and funding of the LKS.
- **Public Sector:** Public establishments receive services free of charge or at affordable rates. This depends primarily on the governance and funding of the LKS.
- **Start-ups:** Young technology companies can apply for government-provided quotas (e.g. AI-Compute Vouchers) to use LKS services.

Sale of data center capacity

The basis for the AI foundation model development, tuning and inference services is the Use of data center capacity by GPU hour.

The cost of one GPU hour based on the o.a. The cost scenarios, assuming full load, are calculated as follows:

Annual costs of infrastructure management gem. Table 16: EUR 88,732,000	
Number of GPUs	4.480
Number of hours per year	8.765
Capacity utilisation	
100 %	

Cost per GPU hour	EUR 2.25
--------------------------	-----------------

This is within the range of current market prices (see Annex C).

Assuming that 20% of the computing capacity can be sold in the market at these prices, a possible annual turnover (cost-covered, without margin) is:

Price per GPU hour	2.25EUR
20% of 4480 GPUs * 8765 h	7,848,960h

Annual turnover is approx. EUR 14.464 million

This rough calculation will need to be adjusted when the scenario is developed in detail, especially in terms of parameter utilization, availability, margin and adjusted infrastructure costs.

AI foundation model training and tuning

The AI foundation model development and tuning services are a combination of consulting and the use of computing capabilities. The consulting activities, as well as the actual development and tuning, are provided by the LKS's Team Services and Consulting. Usual project-specific daily rates between EUR 1,200 and EUR 2,500 per consultant/developer are charged.

	Adoption	
Available capacity	10 FTEs at 80% Chargeability & 200 days	1,600 days/year
Average daily rate	between 1,200 & 2,500 per day	1,600EUR
Total		EUR 2.56 million/year

Table 18: Overview of revenues from the model training

Advice

In the consultancy sector, consultancy services are to be charged at market prices. These are calculated on the basis of person days and the seniority and expertise of the consultant. Depending on the requested service to the LKS, a different level of advice will be required.

	Adoption	
Available capacity	10 FTEs at 80% Chargeability & 200 days	1,600 days/year
Average daily rate	between 1,200 & 2,500 per day	1,600EUR
Total		EUR 2.56 million/year

Table 19: Overview of the revenue generated by the consultancy

Inference Service

The inference service gives companies the opportunity to use AI models with the LKS's compute resources. Billing here is based on the amount of computing resources used, similar to renting GPU hours. Additional costs may be incurred insofar as the Team Services or Consulting Models must first prepare a model for use using LKS computing resources. These additional necessary activities are calculated analogously to the AI-Foundation-Model-Development.

LEAM financing models

.....●

11. LEAM financing models

There are three different models available to finance the construction and operation of the LKS: public and private financing and public-private partnership financing. The exact funding structure needs to be further evaluated, depending on potential funding and associated legal constraints on LEAM. The following is an overview of the possible financing models. Chapter 11.5 discusses the specific legal considerations relating to the various possible financing models.

	Public funding	Private financing	Public-Private-Partnership
Sources	<ul style="list-style-type: none"> • Bottom • Countries 	<ul style="list-style-type: none"> • Corporates • Venture Capitals • Foundations 	<ul style="list-style-type: none"> • Combination of public and private sources
Vehicle	<ul style="list-style-type: none"> • State-owned company • Funding (start-up funding, institutional funding, project funding) • Guaranteed decrease in computing power • (Credits) 	<ul style="list-style-type: none"> • Equity • Debt • Guaranteed decrease in computing power 	<ul style="list-style-type: none"> • Combination of the above
Evaluation	<ul style="list-style-type: none"> + Maximum independence + Consensus between public. Interest and EU-compliant development of AI models - Rigid structures and processes - Limitations in agility, staffing, etc. 	<ul style="list-style-type: none"> + Flexibility in the design of the company + Easy and flexible recruitment of staff - Complex fundraising and possible governance - Collaboration with science is difficult due to high costs 	<ul style="list-style-type: none"> + Public as an anchor investor + More flexible operational business while complying with the framework conditions for LEAM - Restraints under competition law - Complicated procedure for setting up
Commentary	<ul style="list-style-type: none"> • Consideration of EU state aid law with privileges for open source AI projects for business and academia 		<ul style="list-style-type: none"> • Example: DFKI

Table 20: Comparison of the three financing scenarios for the LKS

11.1 Public funding

There are a number of publicly funded data centers in Germany's science landscape, some of which are among the world's leading (see Chapter 8). As a rule, the federal government and the associated federal state share the financing of investment and running costs. Funding will be provided for a staff base to maintain the operation, as well as research positions. Typically, these data centres receive a significant proportion of their funding from third-party funding projects, i.e. personnel and in-kind costs for specific research projects, which are recruited in competition with funding providers (DFG, Federal Government, Länder, industry). A distinction is usually made between institutional support and project support: Institutional support is provided on a recurring basis on an annual basis; it may vary in amount depending on various factors, but it provides a secure basis for the economic operation of the institution. In addition, the above-mentioned project funding is available to applicants themselves (often in competitive competition with other institutions).

Benefits of public funding

The advantage of public funding for LEAM lies in the close interconnection between public interests of a foundation model based on European values. At the same time, there is a high level of reliability and planning certainty for publicly supported projects following a commitment to funding. This would allow the project to be reliably profitable over a period of time and would be independent of the current market situation. The corresponding incentives for start-ups could also be offered in this form of financing.

Disadvantages of public funding

As a rule, purely publicly funded data centres are subject to restrictions on use by private operators for EU state aid, competition or tax reasons. Thus, for example, the HLRS of the University of Stuttgart, which offers deciduous computing time available to industry, only a fraction of the possible computing time available (see Chapter 8.5). Thus, the use of a publicly funded LKS by the more economically organised LEAM Services would be subject to restrictions and there might not be enough computing time available for legal reasons. This also applies to startups, which also compete for the scarce computing resources with the other economic entities. This could inhibit the building of economic applications on the foundation models. At the same time, purely publicly-funded projects are subject to extensive approvals and conditions, and thus usually to lengthy voting processes, as well as to changing political circumstances with changing political majorities. Public companies tend to be tariff-bound or rely on collective agreements, which can make it difficult to have the flexibility to attract the highly skilled staff required for operations and services.

11.2 Private financing

There are two main scenarios for private sector financing of LEAM:

- Joint ventures of large companies (cloud service providers and users). A general willingness and interest for this was signalled in the interviews conducted by various companies.
- Financing through venture capital through classic venture capital (VC) companies or private equity investors.

Since the LKS's purely private-sector business model is associated with high risks (including due to the dynamics of technological developments and their impact on the market), 100% financing by private companies and investors is not very likely.

In this context, the public sector can use tools such as seed funding, loans, etc.e. help and mitigate financial risk. In return, it can be agreed that, for example, a certain part of the infrastructure will be reserved for non-economic open source projects by companies or academia. The compatibility of this variant with a partial or even predominant economic use also from the point of view of subsidy law needs to be examined.

In this scenario, academia and collaborating companies willing to collaborate in the open-source process would use the computing time to train the open-source models at the LEAM AI Service Center in the context of contract research or a project grant. In the case of commissioned research, both LEAM itself and the public authorities could be considered as the contracting authority for science.

Advantages of private financing

The advantage of private funding is that training and utilization can be organized in a society without the regulatory requirements of public funding. The investment costs could be spread over many shoulders depending on the type of financing, which would minimise the individual risk of the shareholders.

The company would also be able to generate unlimited profits and would be free to choose investments, taking into account the interests of the investors. The company would also be more flexible in recruiting appropriate personnel through freely-selectable remuneration and could thus gain an advantage in building know-how.

Disadvantages of private financing

Private financing presents the challenge of finding financiers for a significant amount of investment. It is unlikely that a single company will take the risk of taking on the heavy financing burden.

Multi-company joint ventures would have to be formed, which could make the design of the company and its governance complex and time-consuming.

In the area of venture capital financing, the amount of funding required, which in some cases significantly exceeds the funding practice of most German and European funds. For example, the High-Tech Founder Fund (HTGF) provides only single-digit million start-up funding.

Another disadvantage of private funding is the cooperation with the scientific partners, who are not directly involved in LEAM in this model, but only through contracts and projects. The training of the foundation models, which themselves do not generate a profit but require a lot of computing time, would in this case be in close competition with the core business of the operator company, Inference and Tuning. Computing time would thus be used as a priority for LEAM's commercial products, which would make it more likely that the training of the foundation models would not be of an experimental nature. In addition, companies and/or science willing to work in the open-source model would have to pay market prices for their use.

The biggest drawback, however, is that LEAM's goal of providing AI foundation models in the public interest would compete with the organization's profit drive. Thus, training data for the foundation models would also be selected more strongly from the perspective of economic feasibility in this scenario, which is why rare languages would be similarly poorly represented as the models previously offered on the market due to poor scalability in the market and the greater effort involved in obtaining training data.

11.3 Public-Private-Partnership

In the case of a public-private partnership, the financing is provided by the public and private companies in shares to be defined.

An example of a successful PPP in the field of artificial intelligence is the German Research Centre for Artificial Intelligence (DFKI). DFKI was founded in 1988 as a non-profit public-private partnership (PPP). It maintains locations in Kaiserslautern, Saarbrücken, Bremen, Lower Saxony, laboratories in Berlin and Darmstadt, and outposts in Lübeck and Trier. Funding is provided through grants from public donors such as the European Union, the Federal Ministry of Education and Research (BMBF), the Federal Ministry of Economy and Climate Change (BMWK), the Länder and the German Research Association (DFG), as well as through development contracts from industry.

When designing PPPs, it is important to maximise the respective benefits of public/private funding and minimise its disadvantages.

The challenge of implementing LEAM as a PPP lies in the nature of LEAM as an infrastructure facility and the corresponding high investment costs required to build it. In this scenario, the company would acquire the infrastructure from public funds (Federal and, where appropriate, the Land). Ways and means should be found to divide the funds for the necessary infrastructure between public and private partners. In principle, the more economically oriented services of training-as-a-service, consulting and inference/tuning can also be better represented in this scenario together in a society with the infrastructure.

Benefits of a Public-Private Partnership

A PPP is a conceivable model in which the interests of business in optimising the use of an Open Source foundation model are reconciled with the interests of politics and society in a digitally sovereign Europe, since both sides would be involved in a society. In addition, the project will be easier to implement financially, as the investment costs will be shared. At the same time, the public budget would be relieved, as the private sector also bears part of the investment costs for infrastructure. The corresponding incentives for start-ups could also be offered in this form of financing.

Disadvantages of a public-private partnership

The disadvantage of a PPP is that the company would not be able to act in the same private way as a commercial enterprise, as the federal government is subject to competition law restrictions on its holdings. In addition, there will be an intensive negotiation process between representatives of the private sector and the public sector, both with regard to the precise organisational structure of PPPs and the specifications of the foundation models.

SPOTLIGHT

TUI Germany

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Henning of Roon,
Common Analytics Capabilities
Lead, TUI Deutschland

The TUI Group is one of the world's leading tourism groups. The group includes over 400 hotels and resorts and 16 in-house cruise ships, as well as leading European operator brands and online marketing platforms, five airlines and over 1,000 travel agents. In addition to expanding its core business with hotels, cruises through successful joint ventures and activities in the holiday destinations, TUI is increasing its focus on expanding digital platforms.

Where do you use foundation models? What is your use case?

The TUI uses pre-trained models for image recognition and text classification. These models allow us, for example, to optimize the presentation of products.

What impact do AI foundation models have on your business model or your projects?

By using AI foundation models, many business processes can be automated, scaled up or improved more easily. Especially when it comes to personalizing offerings for our millions of customers a year, we can add great value with AI foundation models without massively increasing development costs. Personalization on this scale would otherwise be hardly feasible.

What difficulties and problems do you see that only the U.S. and China are currently providing AI foundation models on a comprehensive basis?

Our high standards of data protection and security in Europe are often not met by models provided outside the EU. Europe must therefore develop its own AI foundation models in order to maintain privacy and data security while reaping the economic benefits of using AI foundation models. In addition, transferability is often not given, as the

Training data do not reflect European languages and behaviors reflect.

In addition, dependencies could arise, resulting in unfavourable pricing and licensing.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

Open source models provide transparency and allow for a better interpretation of the results. They also have the advantage that we can develop them according to our specific needs.

European models are subject to local standards from the outset and reflect the local realities much better, which makes the use more attractive for us.

11.4 Legal framework

The implementation of the financing of the LEAM AI supercomputer raises a number of legal issues that cannot be fully or definitively addressed in this feasibility study. A more in-depth evaluation of the realities and measures in relation to the legal framework will therefore be required in the aftermath. However, for further consideration, three specific areas of law have to be distinguished in principle, which affect the planned structure and operation of the LKS and must be taken into account in structural planning. These are EU state aid law, (EU) procurement law and public service and remuneration law. The following is intended to give an overview of these legal subjects. The following is intended to present the legal subjects and their impact on the three fundamentally distinct financing models in view of the specificities of the LEAM project, in order to summarise the advantages and disadvantages of these in a comparative legal view.

The importance of EU state aid law

State funding for the LEAM project and the operation of the LKS must be compatible with EU state aid rules. This is binding and is subject to German law. State aid rules - Art. Articles 107 to 109 of the Treaty on the Functioning of the European Union (TFEU) - are among the most important European competition rules. The aim of the State aid rules is to prevent distortions of competition within the European internal market caused by aid granted by Member States to undertakings. Therefore, art. Article 107(1) TFEU provides for a general prohibition of aid, but there are exceptions to it.

EU State aid rules are examined in a simplified way as follows:

At the first level, it is examined whether there is any concept of aid - which, according to what has just been said, is in principle unlawful. Often this is not already the case, with very favourable legal consequences, as a result of the clever design of a state subsidy. If there is a concept of State aid, it will be necessary to assess, at a second level, whether it is permissible under certain EU rules, without the EU having to expressly approve the aid in particular. In particular, aid may be permissible where the rules of the General Block Exemption Regulation (GBER) apply. If such generalised exemptions are not relevant, the aid must be approved by the EU Commission (so-called notification) at the third stage. This is often lengthy and complex and should therefore be avoided for projects that are particularly focused on speed. The German state is not allowed to grant aid that does not comply with European law and is obliged to recover any funds granted. Therefore, EU State aid law must be compliant.

What is important for the LEAM project is to first examine whether there is "conceptual" state aid at all. One such is given when the following four characteristics are cumulatively met. Aid is:

- a state measure;
- which is a beneficiary,
- represents a particular enterprise; and
- the result is a distortion (at least potential) of competition and an effect on intra-Community trade.

State support for the infrastructure of the LEAM project, without any specification, constitutes a State measure which gives rise to an advantage and may also have an impact on the European market, since such a large infrastructure can already have its meaning and purpose in terms of pan-European importance.

However, irrespective of the legal form, the question arises whether the LEAM project also constitutes a "company" within the meaning of EU state aid law. The concept of enterprise in EU State aid law does not depend on the legal form of an entity, but is related to activity. The question of whether the entity carries out an economic activity within the meaning of State aid law is always relevant. According to settled case-law of the ECJ, an activity is considered to be economic if it consists of offering goods or services in a given market,²⁰ irrespective of its legal form, the way in which it is financed and any intention to make a profit.²¹ It is therefore irrelevant whether the entity is operated by a private or public entity. Examples include research infrastructures: Simply put, if they do research for the public, they are not a company within the meaning of EU state aid law. On the other hand, if they provide commissioned research for companies other than the client, they are a company under state aid law and the limitations of EU state aid law apply. The same entity can therefore be both "company" and "non-company" within the meaning of EU State aid law, depending on the activity carried out.

In summary, this also means for the LKS:

- In so far as it carries out non-economic activities within the meaning of EU State aid law, the State aid or the use of the State aided resources do not constitute aid.
- If the LKS carries out economic activities within the meaning of EU State aid law, it is a company and the restrictions of EU State aid law apply to the state support or the use of the state-sponsored funds.

The distinction between non-economic activities and economic activities which, as in the present case, are related to science, has particularities. The advantage is that the LEAM project intends to provide a large part of the capacity for the development of AI

²⁰ Cf. ECJ, judgment of 16.06.1987, Rs. 118/85, Rn. 7 - Commission v Italy; of 12.09.2000, Joined Cases C-180/00 and C-200/00, Rsen and Others. Cases C-180/98 to C-184/98, paragraphs 119 to 120. 75 - Pavlov et al. and of 01.07.2006, Rs. C-49/07, Rn. 22 - MOTOE.

²¹ Judgment of the Court of Justice of 23 April 1991 in Case C-41/90, ECR 1991, p. 1991, I-1979, paras.

²² - Höfer and Elser/Macroton; von Wallenberg/Schütte, in: Grabitz/Hilf/Nettesheim, The Law of the European Union, stand: 67. EL June 2019, Art. Article 107 TFEU, paragraphs 1 and 2. 39.

foundation models to be made available to the general public (or non-discriminatory to business) through the open source process. In particular, the "Union Framework for State Aid for Research, Development and Innovation" of 19.10.2022, 2022/C 7388 (update of the previous version of 21.05.2014, 2014/C 198/01; hereinafter the "R&D Framework") is instrumental in the interpretation and application of EU State aid law in the field of the development of AI models for free use. It sets out the main distinctions between economic and non-economic activities in the context of science, as well as the privileges for the use of infrastructure in science that can benefit the LEAM project.

In a simplified way, the following distinction can be made:

To the extent that the LKS cooperates with science, open source companies or consortia from both sectors, and in particular develops AI foundational models, which in turn are made generally available to science and business in the open source process, these may generally be the (European) science and business enrichment activities and thus non-economic activities within the meaning of the R & D framework (cf. Rz. 20 of the R & D Framework). AI has found an explicit mention in the new 2022 R&D framework: The term "experimental development" is used in Rz. 16 lit. (k) defined as (abbreviated) the acquisition, combination, design and use of existing scientific, technical, economic and other relevant knowledge and skills with the aim of developing new or improved products, processes or services, including digital products, processes or services, in any field, technology, industry or industry. Here, the development of artificial intelligence is explicitly mentioned as an example of application. Thus, the AI foundation models (as well as other AI models) are likely to fall under the concept of development under the R & D framework. In addition, the development of software with the aim of making it available as open source software is explicitly classified as a non-economic activity and therefore not aid-related (figure 10). 20 of the R & D Framework).

As is well known from the German funding landscape, companies can apply for such projects, even in conjunction with science, if they are willing to make the software available in the open source process at the end. The LEAM project can choose several ways to achieve an aid-compliant design, such as developing the software in an open-source manner with companies or academia, with the essential contribution being the provision of infrastructure and technical support (see, for example, Rz. 20, lit. (a) (ii) the R & D framework, "effective cooperation"), or more generally, enable companies to develop such AI foundational models through bidding/non-discriminatory procurement (see the logic in Rz. 20, lit. (a) (iii) and lit. (b) the R & D framework, according to which even the public procurement of services by companies in the context of such projects does not affect the non-economic nature). Companies can therefore participate in the development of AI foundations models together with the LEAM project, provided they work on an open source basis and thus agree to make developed software available openly and without discrimination (at least in the EU). How this is ultimately shaped in practice depends on the desired form of organisation.

For the LEAM project and its objectives, this may constitute a significant aid privilege and encourage public funding of the project. In simplistic terms, this means: In any event, state funding of the LEAM project infrastructure is possible under state aid law if the infrastructure is provided in a non-discriminatory manner, largely to the scientific community and/or to cooperating open-source companies. In the context of

implementation, it would still be necessary to clarify whether there must be a mandatory obligation for the science to use, developed AI models, or for the use of AI.-In turn, to make open source solutions available to the general public - this would probably be the "safest" way in any case from the point of view of state aid law. In the case of developing companies, there is likely to be a strong case for making open-source use mandatory under the Framework, with details to be determined in the context of the specific design.

Even if, for example, generated models or knowledge are disposed of as part of the knowledge transfer to the economy, this may be a non-economic activity, provided that the profit from the disposal benefits the non-economic operation (also Rz. 20 of the R & D Framework). This means that, for example, the transfer of licenses for developed software or AI models by the LEAM project for a corresponding fee does not jeopardise public support within the meaning of EU State aid law if the profits from these licenses in turn benefit the non-economic sector of the LEAM project (i.e. more computing time or better infrastructure for science or companies working in the open source model). In this respect, there is no problem of state aid law.

On the other hand, other planned activities of the LEAM project, such as consulting or infrastructure leasing, are economic activities. They are subject to the limitations of EU state aid law.

If the public financing of the LEAM project is subsequently subject to EU state aid rules, the second level is first for exceptional cases which are explicitly regulated and which do not require the approval of the European Commission. Exceptional circumstances cannot be determined in advance in detail, but depend on the specific nature of the project and the funding sought. While the de minimis regulation will not play a role (the aid limits are set at max. 200,000 euros too low to be relevant to the project), the so-called General Block Exemption Regulation may play a role.

The General Block Exemption Regulation (EU Commission Regulation (EU) No 652/2014 of 17.06.2014, OJ L 175, 19.6.2014, p. 1). L 187/1 of 26.06.2014; "AGVO") contains, to a greater extent, situations in which a concept of aid to a company is compatible with state aid law even without notification (approval) by the EU Commission. The rules are very detailed and must be complied with so that no notification is required. They are to be presented below, where appropriate, for each model.

In addition, measures in favour of undertakings engaged in so-called "services of general economic interest" may be taken on the basis of the Altmark Trans case-law of the ECJ and the EU Commission's exemption decision (Commission Decision (2012/21/EU) of 20.12.2011, OJ L 176, 27.6.2012, p. 1). No. L 7/3 of 11 January 2012) should be exempted from the EU ban on state aid. However, it is unlikely that the LEAM project (despite its potential significance for the German economy) can fall under this, as this case law is more concerned with social needs (public transport, health care, etc.).

If rules such as the General Regulation do not apply to justify the aid, then it is necessary to notify the aid (and therefore an approval procedure) to the European Commission. This is legally complex and can take a long time, and the EU Commission usually also has requests for changes to the funding structure or other aspects of the implementation of the planned structure. Details of such a procedure cannot be presented at this point and would go far beyond the scope of this study. In addition, the rules under which the EU Commission notifies also largely depend on how the funding structure of the project is

designed and what its core objectives are intended to be. If the notifiable aid can fall under the R & D framework after the purpose of the LEAM project, the EU Commission will examine whether the State aid for R & D can be considered compatible with the internal market. To this end, the Commission examines whether the aid measure facilitates the development of a particular economic activity and whether it alters trading conditions to an extent contrary to the common interest (considerations 119 to 120). 38 of the R & D Framework). The decisive factor here is whether the aid brings about a substantial improvement that the market itself cannot bring about, especially when there is a market failure. On the other hand, the Commission requires that the amount and intensity of aid be kept to a minimum, which is regularly and extensively examined in practice. These are just some summaries - for an overview - of the EU Commission's audit programme under a notification procedure.

The importance of public procurement law

Unlike private companies, when public authorities procure goods and services, they are not free to choose, but are subject to the restrictions of public procurement law. This is intended to improve the efficiency and transparency of state action and to ensure fair competition for public contracts. To this end, the procurement law sets out the procedure to be followed in the award of contracts and how to choose between potential contractors.

In particular, contracting authorities within the meaning of Sections 98 and 99 of the Anti-Competition Act (GWB) are obliged to tender for public contracts. This includes, according to § 99 (2) lit. (a) GWB may include, inter alia, legal entities governed by public and private law, established for the specific purpose of carrying out non-commercial tasks of a general interest nature, provided that they are predominantly financed by local authorities, i.e. the federal government and the states, through participation or otherwise. Depending on the specific financing arrangements, this may apply to the LEAM project. This question is discussed below in relation to each of the different financing models.

With the exception of minor transactions with a value of EUR 500 or EUR 1,000 (net) (depending on the state or the federal government), public contracting authorities have to call for tenders either at the federal level or, if certain thresholds are exceeded, at the European level, depending on the estimated volume of contracts. Currently, the threshold for supplies and services for contracting authorities is EUR 215,000 (net). The invitation to tender must be published and bidders may submit a bid. The contract is awarded to one bidder after the tender selection process, which is subject to certain specifications.

It is essential to comply with the requirements of public procurement law when it is applied, as non-compliance with this law can have significant financial consequences and significantly delay the timing of projects. The requirements of public procurement law, which are subject to state control, make the procurement of goods and services by contracting authorities much less free in terms of time and content than for purely private companies.

The importance of public service and remuneration law

When deciding on a financing model, public service and remuneration law considerations must also be taken into account. In the civil service, there are specific requirements for recruitment and remuneration provisions. However, restrictions in this area may also apply to private companies benefiting from government funding.

For recruitment in the civil service, Art. 33 para. 2 of the Basic Law (GG), which guarantees equal access to public office. Therefore, recruitment in the civil service must be based on the principle of the best selection, which imposes specific requirements on the design of application procedures and the selection of staff. Therefore, public employers are also less free to recruit than private companies.

Federal employees are also covered by the collective agreement for the public service (TVöD), while employees in the Länder outside sectors with their own collective agreements are covered by the collective agreement for the public service of the Länder (TV-L). Salaries for highly skilled workers under these collective agreements fall far short

of their earning potential in the free economy, making them rather unattractive financially.

Recipients of large amounts of aid are often subject to the so-called "improvement ban". In this case, the grant approval statement is accompanied by a requirement that the beneficiary may not pay its employees better than comparable employees of the grantor. For example, in such a case, the recipient of federal aid may not pay its employees better than would be the case under the TVöD.

11.5 The impact of legal matters on financing models

The effects of these legal issues on the three main financing models are set out below.

The Public Funding Model

In advance, a note on the principle of admissibility: In principle, the federal government has the power to set up research infrastructures under its own responsibility or to participate in their financing. In addition to the necessary approval of the budget by the Bundestag, the concrete design of the structure is important in order to be able to answer the underlying constitutional questions appropriately.

In art. Article 91b of the Basic Law (GG) provided constitutional protection for federal and state cooperation in the promotion of science and research in cases of transregional importance. The scheme was created in particular for the financing of large research infrastructures and also offers the federal government, then in cooperation with a host country, the possibility of publicly funding the LEAM project.

In a model of (purely) public funding of the LEAM project, all three legal subjects described above have the strongest impact. In particular:

EU State aid law

In the case of purely or in any case largely public funding, the above explanations of EU state aid law apply. As explained therein, on the basis of the R&D Framework, it is likely that it can be argued that major planned deployments of the AI supercomputer are non-economic activities and therefore the public financing of the planned infrastructure does not constitute aid. This would be advantageous insofar as, on the one hand, a complex notification procedure with the EU Commission would be eliminated and, on the other hand, the restrictions of the AGVO, which generally does not allow full state "cross-financing", would not apply. In a basic public funding scheme, the federal government, together with a seat state, could, in principle, provide funding for the construction of the infrastructure and its operation (including personnel), if it is politically desirable.

However, activities outside this infrastructure that are of an economic nature within the meaning of EU State aid law (e.g. the planned consultancy offer) are not excluded in the first place. The R & D framework also allows such activities, albeit subject to certain requirements. These are summarised below:

The most essential requirement is the introduction of a so-called separation bill. This means: Where the LEAM project carries out both economic and non-economic activities,

public support for non-economic activities is not covered by the prohibition of aid if the non-economic and economic activities and the costs, financing and revenues can be clearly separated so that there is no risk of cross-subsidisation of economic activities. Money that comes from the federal government or a state may not be used to compensate for losses in the economic sector (prohibition of cross-subsidies and requirement of the so-called "separation bill", paragraph 10). 19 of the R & D Framework). Such segregated accounting can be carried out in the financial statements, but it also means that revenues and expenses for non-economic and economic activities must be strictly separated in the accounting and it must always be clear whether an expense (whether it be personnel or infrastructure use) can be attributed to one area or the other. In practice, this requires some administrative work, but it can be solved.

Therefore, the economic sector (e.g. consulting) must not generate a loss. Similarly, the economic sector must not become a major application of the publicly funded infrastructure, pushing the non-economic sector into the background. To this end, the R & D framework also provides for a privilege for research facilities/infrastructures: Where the research facility or research infrastructure is used almost exclusively for a non-economic activity, its support may fall entirely outside the scope of State aid law, provided that the economic use is a purely ancillary activity directly linked to and necessary for the operation of the infrastructure, or that is inseparable from and limited in scope to the main non-economic activity. This is the case where the economic activities use the same inputs (such as materials, equipment, personnel and fixed capital) as the non-economic activities, and where the annual capacity allocated to the economic activity in question does not exceed 20% of the annual capacity of the facility or infrastructure in question (figure 1). 21 of the R & D Framework). For the LEAM AI project, for example, this would mean, very roughly simplified, that 80% of the computing capacity of the AI supercomputer would be allowed to be used for non-economic activities (e.g. to develop large AI foundation models with open access for business and science) and 20% for economic purposes. Details must then be worked out for the actual operation. It appears, however, that even with purely public funding, economic activities are not forbidden to the LEAM project and can be further planned in the business model. In practice, such cases exist, for example, in the case of universities whose infrastructure is publicly funded but which nevertheless have economic offers such as commissioned research or research-related services.

Once a distinction has been made between economic and non-economic activities, it is necessary to ensure that economic activities do not result in cross-subsidisation. This is measured by the so-called private investor test. The economic behaviour of the public body is compared with the hypothetical behaviour of a private investor. In the absence of such an advantage, or at less favourable terms, aid would constitute State aid under Article 107(1) TFEU. Article 107(1) 1 of the Treaty on the Functioning of the European Union. The R & D framework contains detailed rules in this respect (see recitals 19 and 20). Point 26f. R & D framework), which requires pricing to be based either on market prices or on full costing with a profit mark-up. However, the costs that the LEAM project charges to the economy for such economic activities may be significantly lower from an economic point of view than comparable offers from the USA, since even under state aid law, the LEAM project does not have to be guided by the idea of maximising profit in such

economic activities, but rather by the idea of full costs plus a reasonable profit margin. The latter is likely to be significantly lower than offers from US suppliers, which may have a quasi-monopoly position.

In sum, this means that public funding of the LEAM project and its infrastructure is possible under state aid law, especially if the infrastructure created is used predominantly for non-economic purposes and is made available to the scientific community or cooperating companies. However, a certain capacity of the supercomputer and staff can also be made available for economic activities, which are usually non-loss-making and must be offered at a market price. A segregation account must make it clear in the accounting and financial statements that there is no cross-subsidisation of economic activities.

Public procurement law

Procurement law is fully relevant in the case of purely public funding.

If the LEAM project is fully publicly funded according to the first eligible model, then the legal entity to be created for the LEAM project is a public contracting entity within the meaning of §§ 98, 99 GWB. In this case, the conditions set out in Chapter 11.4 are those of § 99 (2) lit. (a) GWB is satisfied, since in this case it is a legal entity established for the specific purpose of carrying out non-commercial tasks in the general interest and financed at least predominantly - here in full - by local authorities.

In this respect, the characteristic of non-commercial tasks is met even if consultancy or the rental of computing time are offered in addition to the core task of AI research and development under the LEAM project, as it is important to have an overall view of the issue of non-commerciality, taking into account factors such as the decoupling of competition and market mechanisms and the existence of a profit-making intention. In this respect, ancillary activities such as consulting and computer time leasing fall behind the largely non-profit-oriented core activity.

Public service and remuneration law

In the purely public financing model, Art. 33 para. 2 GG and the limitations of public remuneration law.

The application of the constitutional principle of selecting the best leads to a certain amount of additional effort to create a legally secure application process. For example, for posts to be filled, a profile of requirements must be defined and questions addressed during the interview, and documentation requirements apply. In addition, vacancy announcements must be limited by a time limit.

Depending on the structure of the project, TVöD or TVL will also be used, with the result that they will be able to pay only comparatively unattractively. Employees of purely or predominantly state-funded companies, even if they have a private-law form, are subject to the prohibition on improvement. This must be observed in the case of institutional support (grants to cover all or an undetermined portion of the recipient's expenditure) and in the case of project support where the recipient's total expenditure is predominantly financed by public grants (see section 8 of the 2021 Budget Act). Although the Federal Minister of Finance may grant exemptions on compelling grounds, this is done in a restrictive manner.

However, there is one common point for EU state aid law and public service and remuneration law: Even with purely or predominantly public funding of the LEAM project, it is possible that, in addition to the support structure of the LEAM project, a service company will be established, focusing on economic activities and leasing the infrastructure of the LKS for their provision (taking into account the capacity constraints described above). The separation of economic activities within the meaning of EU State aid law into an economic entity is, in principle, permissible under state aid law, subject to compliance with the rules prohibiting cross-subsidisation and capacity restrictions for the economic use of supported infrastructure.

It has been proven in the German science landscape and offers several advantages:

- The segregation accounting that the assisted entity must maintain will be significantly simplified if all economic activities are segregated into a specialised LLC.
- The economic entity legally separate from the supported structure (e.g. a separate service company) enjoys greater discretion in terms of remuneration and pricing. In particular, it may, for example, employ and remunerate employees of the assisted non-economic entity as a secondary activity in order to provide an attractive remuneration package within the limits of what is legally permissible. Moreover, since the employees then have to make their own efforts with the secondary activity to achieve this additional compensation, acceptance of such solutions can usually also be achieved in external representation and policy. This approach has been proven, in particular, in the field of science. Legal parameters need to be elaborated in detail.
- In implementation, the infrastructure/capacity of the economic unit/commercial entity would be leased by the supported entity, which is in principle permissible under state aid law and allows for a clear separation between economic and non-economic sectors.

In the context of the concrete design, it may be useful to examine the separation of non-economic and economic activities of the LEAM project into two legally separate structures.

SPOTLIGHT

Ubermetrics Technologies GmbH

At various points in the study, we introduce selected German startups and companies that are already using AI foundation models. They report on their use cases, challenges and the advantages of European models.



Patrick Bunk, founder and CEO of Ubermetrics.

The **Ubermetrics** Content Intelligence platform is based on artificial intelligence, specifically designed and trained for the needs of communicators, that transforms publicly available content and data into meaningful insights for effective communication strategies. Processing over 50,000 articles per minute and content from more than 460 million sources, Ubermetrics is the leading content intelligence platform for marketing and PR experts.

Where do you use foundation models? What is your use case?

Ubermetrics extracts texts from millions of Internet sources and evaluates them using human experts and automated AI procedures for customers. So far, we have limited the use of foundation models in some parts of the results presentation. For example, we use a RoBERTa model to automate entities (e.g. People, places, products, etc.) in texts to make them usable for further analysis (so-called Named Entity Recognition / Linking). However, we are only beginning to exploit the possibilities of the models for our products.

What impact do AI foundation models have on your business model or your projects?

foundation models are of paramount importance to the competitiveness and success of Ubermetrics, or UNICEPTA, respectively. Until now, we have had to implement or research a new AI process for each aspect of our production chain and, in some cases, for each language (e.g. Sentiment analysis). These could in large part be replaced by methods based on foundation models – while improving the safety of hits. This affects both the collection of text and the processing (e.g. Speech recognition) and evaluation. foundation models now allow us a comprehensive basis on which to build the development of better processes and the development of new products. For us, therefore, on the one hand, the effort in production is reduced and, on the other hand, we are able to offer better or completely new products and features. E.g. we could partially automate summaries across several articles and article abstracts to assist our knowledge workers with less demanding tasks. We estimate the economic value of the implementation at 60% of our turnover.

What difficulties and problems do you see that only the U.S. and China are currently providing AI foundation models on a comprehensive basis?

The biggest difficulties and problems lie in the loss of digital sovereignty and weaker support for relevant European languages. If we were forced to use non-European models via APIs, we would have to disclose relevant and sometimes sensitive business data. This is hardly compatible with European data protection standards and is a major reason for our sporadic use of such models so far.

How would European models - open source, covering all European languages, with high standards of data protection and minimal bias - help you?

From European open-source models, we expect, in addition to better language coverage, including through European data protection standards, greater acceptance by our customers. At the same time, we expect better planning in terms of further developments.

The private financing model

In principle, the model of private financing or predominantly private control of the chosen structure contrasts with purely public capital raising.

Assuming that no state resources are involved, but that the model is purely privately financed and operated, the above-mentioned restrictions do not apply. While the LEAM project in this case must ultimately be "self-supporting" (if the non-profit form is not chosen) and even generate profits in order to survive economically, there are no state aid, award or tariff restrictions in this case. The LEAM project can (and must) then operate "free on the market". Contracts can be awarded quickly, market-standard salaries paid. This is true even if the choice falls on a non-profit corporation for tax reasons (with certain limitations of non-profit law, which cannot be presented conclusively here).

EU State aid law

Such a privately structured or mostly privately funded institution can then in general apply for funding for scientific infrastructure and projects, for example through project support from the DFG, the Federal Ministry of Research and Education or the EU - just like any other public or private institution. However, this project support is not only awarded on a competitive basis for a limited period of time (i.e. the funds are limited in time and are not guaranteed to be received due to the competitive nature of public tenders), but it is also associated with public service obligations, where appropriate. For example, these may include restrictions on remuneration (e.g. the prohibition of improvement, see above) or public tendering procedures when funding is awarded to companies through the supported infrastructure (e.g. the LEAM project, which wants to buy computing capacity, for example). This depends on the specific pot to be used and the intensity of the subsidy.

However, it is not excluded that the federal government or a state may also provide funding assistance if the underlying entity is to be privately financed by its nature and operated in the shareholder structure. Germany has numerous support instruments for start-ups and technology-related companies, but not all of them can be presented in their range here. The well-known high-tech startup fund is also a part of this, as are the credit

programs of the federal state-run promotional bank KfW or the promotional banks at the state level. Notwithstanding the fact that this can also lead to state aid examinations (immediately), such funding can often only be a limited amount of "stimulus financing" in addition to substantial private funding (see also the comments on the AGVO and the Public-Private Partnership).

In order to avoid a lengthy notification procedure, such financing, if it is intended to be of a significant size, will probably have to be assessed as "aid" to a company, in particular by reference to the rules of the AGVO. Aid granted on the basis of the AGVO must meet certain transparency requirements and have an incentive effect. In addition, information on aid must be published on an "aid website" of the Member State or its operating body and communicated to the Commission. The Member States have an obligation to monitor. In particular, the following categories of the AGVO could be relevant to the promotion of the LKS:

Art. 17 AGVO: Investment aid for SMEs

Up to 20% of the investment costs of small and 10% of the investment costs of medium-sized enterprises for intangible and tangible capital goods. This includes, inter alia, the construction of a new production facility and the cover of the labour costs of the jobs created by the investment project.

Art. 18 AGVO: Investment aid for SMEs for the use of consultancy services

Consultancy services by external consultants may be subsidised up to 50% of the cost, insofar as they are not ordinary advertising costs and ongoing services.

Art. 26 AGVO: Investment aid for research infrastructures

Construction and expansion of research infrastructures can be supported as preferential aid under the rules of the AGVO up to 50% of the cost. The price for the operation and use of the infrastructure so supported must be in line with the market price. The infrastructure must also be open to multiple users and access must be provided on transparent and non-discriminatory terms. Companies that have funded at least 10% of the investment costs of the infrastructure can receive preferential access. If a research infrastructure carries out both economic and non-economic activities, a segregated billing is mandatory.

Art. 28 AGVO: Innovation aid for SMEs

Up to 50% of the costs incurred by small and medium-sized enterprises, including the posting of highly qualified personnel for research and knowledge dissemination, or for innovation advisory or support services, may be financed as AGVO grants. Up to an amount of €200,000 over three years, up to 100% of the cost of innovation advisory services and innovation support services may also be funded.

Public procurement law

In the case of purely private financing, the LEAM project is not a public contracting authority and therefore the procurement law does not apply to the procurement of services. This leads to significantly freer procurement processes.

Where appropriate, in the context of project financing requested by a private entity, the eligibility conditions may include rules on the procurement of infrastructure to be procured. This may or may not be the case.

Public Service Law

Civil service law does not apply to recruitment in this case. Since the prohibition on improvement only applies to predominantly public funding, it does not apply to (predominantly) private funding. In terms of employees, purely private financing is generally much more free and higher remuneration is more attractive for employees.

The Public-Private Partnership Model

The mixed form of the previous models is the concept of public-private partnership, i.e. financial and/or (societal) legal cooperation between the private and the state. In practice, there are a variety of possible aspects to this.

Before explaining the applicability and significance of state aid, public procurement and public service law to the public-private partnership model, it is first necessary to explain what it is, what forms it can take, and to what extent the state can participate in it.

Definition

A so-called public-private partnership (hereinafter also "PPP"), in German public-private partnership, or PPP for short, is a mixture of purely public funding and structure and an entirely private arrangement.

A PPP is a partnership between the public sector and private companies for the purpose of carrying out a public project or providing services to fulfil a public service mission.

Consideration of the legal possibilities, requirements and relevant regulations depends on the specific structure of the respective cooperation. For PPPs, there are only loopholes in various laws, but no comprehensive legal requirements.

Design models

There are many different organisational models for the practical design of PPPs. Again, these are not defined by law, but have emerged in practice. The categorisation of organizational models is often inconsistent.

The number, names and exact description of the forms of organization vary depending on the consideration. However, roughly speaking, the following categorisation can be made:

Concession model

In the concession model, the public authority entrusts a public task to a private company. In doing so, the private operator operates the institution concerned in his own name, on his own account and at his own economic risk. Only he acts in this respect towards third parties. It generates revenue by charging third parties for the use of a thing or service. In the case of LEAM, it would be possible to rent computing time and consultancy in this respect. In some cases, the concession model also provides for a concession fee to be paid by the public as the concessionaire to the private party. However, in this model, it should be noted that such a concession award requires a pan-European tender procedure.

However, the problem with a concession model is that the operation of the AI HPC must logically be a (compulsory) public task for the state (as opposed to government funding of desirable activities, for example in science). This seems difficult to imagine at the moment,

but it also needs to be discussed with political leaders whether, if necessary, they wish to establish such a situation by regulation or law. Until then, there have been doubts about the applicability of the concession model.

Leasing model

In the leasing model, the public authorities contract a private individual or a leasing company to design, build and finance a project against payment of an agreed lease rate. The private person becomes the owner or owner of the property, but grants the public authority the right of use for the agreed lease term. For this purpose, an act of dedication may be carried out, by which the object becomes a public good.

Operator model

In the operator model, a private organisation carries out contractual parts of the tasks of planning, development, operation and financing and receives an agreed remuneration for them. It is both possible that the private party acts only in an internal relationship with the administrative entity rather than as a technical enforcement agent, and that, more closely to the concession model, it also acts vis-à-vis third parties and collects a fee from them.

Management model

In the management model, a private organisation operates a public entity named and on behalf of the public authorities at their own risk and at their own expense. Thus, he only acts in foreign names vis-à-vis third parties.

The abovementioned models can be summarized under the general term "contract models". They have in common the fact that public and private authorities work together on a contractual basis, but they do not create a common legal entity.

Participation or cooperation model

It is different with the so-called participation or cooperation model. In this way, the public sector and private actors form a joint company. Both public bodies (e.g. the federal government and/or the states) and private companies participate as shareholders.

As a result of the creation of a new company as a separate legal entity to be distinguished from its partners, this model is also referred to as an "institutional PPP" and a "mixed economy enterprise" because of the joint ownership of public and private entities.

A partnership contract is concluded. In addition to the legal form of the company (see Chapter 12), it first determines the share of the shares held by the respective shareholders. The share of the public sector can vary in size. It may be a majority shareholder, as is often the case in practice, but it is also possible to have a shareholding of less than 50%.

The company contract contains further rules on the management and representation of the company, on the rights and obligations of the members, on the distribution of profits and losses, and on the appointment of members of the company's organs. The advantages and disadvantages of different types of companies (see Chapter 12) include improved control and control by the public authorities, as well as opportunities for corporate influence (see Chapter 12).

In view of LEAM as a long-term project, the creation of a joint company appears to be the most appropriate form of PPP. Therefore, the following comments focus on the legal implications of such an arrangement.

Prerequisites for public-private partnerships

In principle, the State has a wide leeway in the performance of its tasks. This includes the question of whether it acts under private law and the extent to which it can use private actors to carry out public tasks. Participation in public-private partnerships is therefore also possible, apart from exceptions that are not relevant to LEAM.

The LEAM project can also comply with other budgetary rules governing public participation in companies, such as limiting the obligation to deposit to a certain amount and ensuring appropriate government influence in the company's supervisory bodies.

EU State aid law

There are no specific legal issues for public-private partnerships under EU state aid law. The State participation and the financing of the project from State resources will in principle continue to apply state aid law, but the financing of non-economic activities will be subject to the same criteria as purely public financing.

The principles set out in Chapter 11.4 apply to the distinction between economic and non-economic activities, which is crucial to the question of whether there is State aid, since it does not depend on the legal form and nature of the shareholders involved, but rather on the nature of the activities involved. The R & D framework is therefore also relevant in this respect. Its requirements, i.e. the requirement to introduce a segregated invoice and the prohibition of cross-subsidisation of economic activities should be transferred to PPPs. However, in the case of private financing, it is possible to tailor the structure to enable economic activities to be carried out to a greater extent, in particular if private investors agree to compensate for any losses of the economic activities themselves (so that there is no cross-subsidisation of state resources). Details must then be evaluated in the concrete design.

Insofar as the concept of aid is concerned, the PPP financing model may also be exempted. In this respect, insb. the AGVO aid listed above. In the case of State aid which is not exempted, a notification procedure must be carried out.

Public procurement law

In order to determine the extent to which public procurement law applies to PPPs, there are a number of issues to be addressed: A distinction should be made between, on the one hand, cooperation which falls within one of the contractual models described above (concession, leasing, operator or management model) and, on the other hand, PPPs under the ownership model which form a joint venture. These, in turn, must be differentiated according to whether they concern the applicability of public procurement law to the formation of a company or to the procurement of services by the existing company.

Cooperation under one of the contract models cannot at this stage be fully assessed in the context of public procurement due to the variety of possible models. However, since they have in common the fact that the state here delegates all or some of the project management to private parties, the project contract is often a public contract within the meaning of Section 103(1) of the Code of Conduct. 1 GWB or a concession under § 105 Abs. 1 GWB act. Therefore, the conclusion of a project contract with a private cooperation partner usually requires a formal award procedure, given the high volume of projects with pan-European publicity.

For the creation of a joint venture (participation model), the legal situation is as follows: The establishment of the mixed economy company itself is, in principle, not a procurement operation and is therefore not subject to public procurement law. However, the situation is different if the act of incorporation is accompanied by a direct transfer of tasks from the public authority to the company. In this case, the private undertaking involved becomes not only a shareholder but also a service provider to the State. In the context of an overall examination to be carried out, this overall design is considered as a subject-matter process in the form of a public contract within the meaning of Article

103(1) of the Treaty. 1 GWB or as a concession under § 105 Abs. 1 GWB, since the establishment of a company and the performance of tasks form an indivisible whole. The same applies when a mixed-economy enterprise is set up by the public and private entities to carry out their own tasks in an appropriately defined area and the services are subsequently provided by the members.

If contracts required for the company are subsequently awarded to third parties by the company at a downstream stage, which is to be distinguished from the formation of the company, the company is usually itself a contracting authority under § 99 (2) lit. (a) To be classified as GWB. If it procures services on the market, such a procurement act is generally subject to the requirements of public procurement law. The conditions for this requirement, as set out in Chapter 11.4 above, are met. In particular, the following are given: As explained in the comments on public procurement in the case of purely public funding, the overall overview of the activities planned for the LEAM project is based on the assumption that the tasks of general interest are of a non-commercial nature. There are no differences in this respect. The further requirement of § 99 (2) lit. a) GWB that the company is predominantly funded by local authorities - presumably the federal government and the country of location - is present when, as expected, the funding is more than 50% government-owned.

Public service and remuneration law

Even in public-private partnerships (i.e. those with more than 50% state ownership) where the public-private partnership is dominated, the vacancy requirements under Art. 33 para. 2 GG. It is clear from the Federal Labour Court's previous case law on privately owned public enterprises that the rule will also apply to publicly-controlled mixed economy enterprises.

Restraints on remuneration law only apply if the amount of aid makes it possible to apply the prohibition on improvement.

11.6 Final overview

The following table summarizes the legal advantages and disadvantages of each model in relation to the three basic types of financing:

	Publicly funded	Privately funded	Public-Private-Partnership
EU State aid law	<ul style="list-style-type: none"> - fully applicable - Separation invoice - Restrictions on economic activities - R&D framework 2022 with many privileges 	<ul style="list-style-type: none"> - Not applicable or limited - Possible aid in the light of AGVO 	<ul style="list-style-type: none"> - possible restriction of economic activity and tax separation; R & D framework - In addition, any aid envisaged under AGVO
Public procurement law	<ul style="list-style-type: none"> - fully applicable - public tenders necessary 	<ul style="list-style-type: none"> - In principle not applicable 	<ul style="list-style-type: none"> - applicable to the act of incorporation if it is simultaneous with the transfer of tasks to the private partner and in the case of a concession - Generally applicable to the procurement of services by the mixed economy company
Public service and remuneration law	<ul style="list-style-type: none"> - fully applicable - Prohibition of repairs intervenes 	<ul style="list-style-type: none"> - free in the selection of employees and remuneration (e.g. tax restrictions in the case of charitable activities) - Market remuneration is possible 	<ul style="list-style-type: none"> - Employee profitability is subject to best-selection requirements if government ownership exceeds 50% - Prohibition of repairs possible depending on the amount of aid

Table 21: Overview of the advantages and disadvantages of the three funding models in terms of the regulatory framework

Corporate structure of LEAM

.....●



12. Corporate structure of LEAM

The corporate structure of LEAM is highly dependent on the type of financing of the project. Different types of company are available depending on the financing structure. In the context of funding, these should be more thoroughly evaluated and evaluated. In general, due to the structure of the project, the creation of a project development company is conceivable.

In principle, different types of company can be distinguished. These are to be briefly highlighted on the basis of the form of financing.

12.1 Public funding

If purely public financing is chosen, several forms of private law are in principle considered in addition to the legal forms of public law (especially the public law institution). Also in German funding and scientific practice, there are institutions that are organized as e.V. (for example, the Max-Planck-Gesellschaft or Fraunhofer) or also as GmbHs. The GmbH in particular - possibly also in the form of the non-profit GmbH with the special tax impact - can in practice play a role in the performance of public authorities' tasks today. The federal government or a federal state may have a shareholding under German financial law in the amount of the investment (i.e. if, for example, the federal government bears 90% of the investment, it would also hold 90% of the shares in the GmbH). This also follows from the logic that EU state aid law and public procurement law, with their possibilities and limitations, are not based on legal form but rather on the activity actually carried out. GmbH law is also flexible enough to include special organisational structures and at the same time allow for a certain degree of flexibility on the part of the parties involved, in particular the management. In the case of a GmbH construction, it may also be possible to combine a more non-economic carrier company for the infrastructure (e.g. a gGmbH) with an economically oriented company.

The joint-stock company and the cooperative, on the other hand, do not appear to make sense in this model, as they do not appear to be representative from a budgetary point of view for the federal government and a participating federal state, and can also cause problems in terms of legal flexibility.

12.2 Private financing

In the context of private financing, in principle all the forms of company described above are conceivable.

The legal form of the AG is flexible with regard to the transferability of shares and the raising of capital, including in the short term. Strict legal requirements, however, reduce the scope for entrepreneurial design. In addition, an AG makes more sense for a capital market-oriented company. In comparison, transferability and short-term capital raising are more difficult for a GmbH. Unlike the management board of an AG, the management of a GmbH is bound by the instructions of the shareholders. Formal requirements are less stringent. The democratic structures and flexibility of the cooperative can be seen as an advantage or a disadvantage, depending on the point of view. The disadvantage of a foundation is the loss of access to the inflow of assets. The operation of an economic enterprise with a profit-making intent as a registered association is thus not possible.

12.3 Public-Private-Partnership

Again, there are fundamentally different forms of society conceivable.

For PPPs, the legal form of the GmbH is often chosen. These are supported by the fact that, by virtue of their freedom of constitution and the obligation to give instructions to the management, they are well placed to give adequate consideration to the budgetary obligations of the public authorities to influence and control the management. In the form of the GmbH, the public authority also protects against unacceptable liability risks.

The statutory definition of the public task in a mixed economy enterprise also has the important importance of addressing the risk of a third party being challenged for breach of loyalty obligations or the pursuit of special advantages of shareholders' decisions oriented towards the common good.

The advantage of the GmbH over the AG is that, in the context of public-private partnerships, it allows more control over the shareholders. In this respect, GmbH law is more open in that it largely leaves the regulation of the legal relations in the form of a company contract to the shareholders.

A GmbH also allows a PPP to allocate business shares between the federal, state and private sector in an appropriate way, while at the same time ensuring certain freedom and flexibility for the management of the GmbH to implement the objectives of the LEAM project.

On the other hand, it is not excluded that mixed economy enterprises may be set up in a public legal form. However, this requires increased requirements, because the constitutional requirement of legitimation applies extensively to public legal forms.



Scenario for a LEAM AI Service Centre

.....●



13. Scenario for a LEAM AI Service Centre

On the basis of the foregoing, various scenarios for the design and implementation of the LKS can be derived. A possible concept is developed and explained below. It should be borne in mind that there may still be design options and alternatives in all dimensions. These will then have to be specified in more detail in the further course of design and negotiation and brought to a decision.

The governance structure of the LKS is shown schematically in Figure 27 and consists of the following components:

LEAM Institute

The LEAM Institute is responsible for:

- Overall coordination of LEAM activities and the LKS
- Managing the development of AI foundation models
- Providing foundation models as open source
- Development and implementation of the "Training-as-a-Service" service

Funding is provided through public funds, taking into account the compatibility of the aid with public service pay structures (TVöD). The lease of infrastructure (housing, office, possibly computing capacity) or the purchase of an AI supercomputer must be publicly tendered for. In order to comply with EU State aid rules, 80% of the supercomputer's capacity is used for non-economic purposes - i.e. research and development by business and science in the open source process - and 20% for economic purposes.

LEAM service company

The LEAM service company undertakes the following tasks:

- Offering a model tuning service with a focus on business clients
- Provision of capabilities for the operation of AI applications (Inference)
- Provision of GPU computing capabilities
- Consulting services

These services are usually offered commercially.

Possible sources of funding for the service company include:

- Financial investors
- Companies or joint ventures
- A Public-Private Partnership (PPP) involving the federal/national government

The possibility of further public support (e.g. start-up funding) needs to be explored in the future and the legal form and commercial orientation (for-profit/non-profit) should be defined.

Since the company is not subject to public service and remuneration law, it can pay market wages. This increases the attractiveness of highly skilled talent.

Infrastructure

Two alternative approaches can be used to provide the necessary supercomputing infrastructure (see Chapter 9.2):

- Purchase of an AI supercomputer
- Purchasing GPU RZ performance from an external provider that provides an AI supercomputer

Housing

The necessary housing infrastructure is rented from an external service provider (see Chapter 9.2).

Support projects

An initial foundation model will be created as part of a publicly-funded anchor project involving business and academia working together. The design of this project can be guided by the existing OpenGPT-X project and take over and develop the results of OpenGPT-X.

The objectives of this anchor project are to:

- Building and maintaining a data body for language models
- Implement and test the necessary training-as-a-service processes
- Providing the developed foundation model as an open source for business

AI Compute Voucher

As part of the LEAM initiative, start-ups will be given the opportunity to participate in the development of foundation models as well as - above all - to use the provided foundation models and develop their own AI applications and business models on this basis.

To support this objective, the public authorities provide so-called AI-compute vouchers. This allows (even smaller) development projects for start-ups to be requested. The exact structure and scope of this programme will have to be defined in the future.

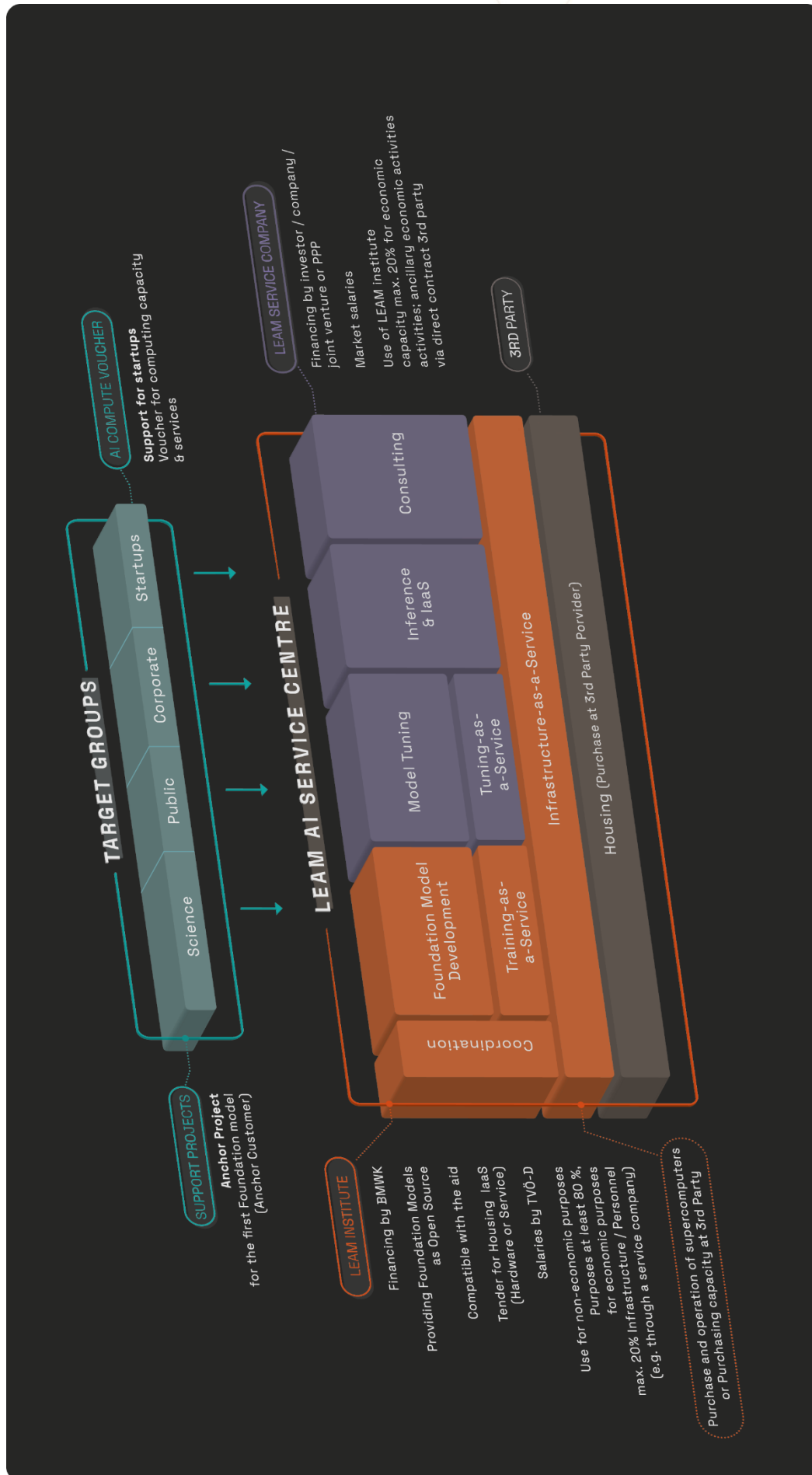


Fig. 27: Scenario for the structure of the LEAM KI Service Centre

Conclusion

.....●



14. Conclusion

The surveys and interviews with experts from business and academia give a clear picture: AI foundation models represent the next development in the Artificial Intelligence success story. The currently popular language models are only the first step. In the next few years, even more powerful models trained on even more diverse data will continue to revolutionize the market.

The German economy has recognized this trend and is already evaluating ways to effectively use the AI foundation models in internal processes and as products. At present, however, it is largely dependent on proprietary, American foundation models. This presents companies with major challenges in terms of privacy, quality and access to the models. A decisive competitive disadvantage over foreign competition is an imminent threat. foundation models compliant with European standards, trained with high-quality and diverse data and available open-source, would address these challenges and help the German economy to benefit widely from AI foundation models.

Three key challenges have been identified by both business and academia:

- (1) There is a need for a large number of experts on the subject of AI foundation models.
- (2) These experts need access to high-quality data of various types, as well as
- (3) Access to a sufficiently powerful infrastructure optimized for current AI technology and AI development processes.

These challenges need to be solved in a joint effort by business, academia and the government.

14.1 Feasibility assessment

This study has explored the need, opportunities and demand for AI foundation models in Germany. In particular, the feasibility of setting up and operating a dedicated AI data center was highlighted. For this purpose, this study looked at the areas of software, hardware, building infrastructure, personnel, organizational structure and finance. A final assessment of the feasibility in these areas is to be made here.

Software

The software frameworks and tools needed to train and develop the AI foundation models do not present a central challenge. The necessary technologies are already in place and mostly available as open-source software.

Hardware

There is currently no dedicated AI data center in Europe that is sufficient to develop internationally competitive foundation models. Around 4,500 GPUs need to be installed in the data center to keep up with the current state of the American hyperscaler. Manufacturers and collocation providers have signalled that building such a cluster is possible in the near future in terms of hardware.

Construction infrastructure

Due to the cost and expected construction time, construction of the program's own building infrastructure should be avoided. However, several collocation providers have already signaled that they would be able to operate in their infrastructure. This possibility should therefore be used.

Staff

A team of highly specialized experts is needed to operate an AI data center. Building the team to build and operate the services is a challenge in the short term. A great opportunity to retain talented scientists in the medium and long term.

Organizational structure

In order to operate an AI data center, the authors propose setting up their own organization, the LEAM AI Service Center. This LKS will support the construction of the infrastructure and provide specialized services to the identified audiences. The focus is to provide industry with AI foundation models for application development.

Financing

The authors estimate a need of around 380 million euros over four years to build and operate an AI data center. A model for funding has been developed that takes into account both public and private resources.

In conclusion, the development of LEAM infrastructure and services is a crucial economic factor for the German economy. Implementation can be achieved with the participation of public authorities, research and academia, as well as business.

14.2 Forecast

The issue should now be further promoted in close dialogue with politics, business, potential locations, suppliers and users. Above all, it is important to further clarify the financing possibilities.

Recent developments such as the hype surrounding ChatGPT show how dynamic the research and application of foundation models is evolving. The risk of monopoly structures forming in the US – similar to Google in search engines – is growing steadily. A timely implementation of the LEAM infrastructure concept is needed to prevent Germany from falling further behind. The momentum gained in the preparation of this feasibility study – through close collaboration among research, academia, business and start-ups – should be used to advance the objective.

One possibility is the prompt establishment of a project development organization to work out the foundations for building this strategic AI lighthouse project.

Directories & Methodology LEAM Feasibility Study



I. Source Directory

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, 265-283. <https://dl.acm.org/doi/10.5555/3026877.3026899>
- Abseil Python Common Libraries. (2022). [Python]. Abseil. <https://github.com/abseil/abseil-py> (Original work published 2017)
- Agarwal, O., Ge, H., Shakeri, S., & Al-Rfou, R. (2021). *Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training* (arXiv:2010.12688). arXiv. <https://doi.org/10.48550/arXiv.2010.12688>
- AI accelerator. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=AI_accelerator&oldid=1123373022
- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in Neural Information Processing Systems*. <https://openreview.net/forum?id=163556id=EbMuimAbPbs>
- AMD. (o.D.). *AMD ROCm™ Open Ecosystem*. Retrieved 28 November 2022, from <https://www.amd.com/en/graphics/servers-solutions-rocm>
- AMD. (2021, August 11). *New AMD Instinct™ MI200 Series Accelerators Bring Leadership HPC and AI Performance to Power Exascale Systems and More*. <https://www.amd.com/en/press-releases/2021-11-08-new-amd-instinct-mi200-series-accelerators-bring-leadership-hpc-and-ai>
- An updated set of basic linear algebra subprograms (BLAS). (2002). *ACM Transactions on Mathematical Software*, 28(2), 135-151. <https://doi.org/10.1145/567806.567807>
- An, W., Guo, Y., Bian, Y., Ma, H., Yang, J., Li, C., & Huang, J. (2022). MoDNA: Motif-oriented pre-training for DNA language model. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1-5. <https://doi.org/10.1145/3535508.3545512>
- Apache Flink. (2022). [Java]. The Apache Software Foundation. <https://github.com/apache/flink> (Original work published 2014)
- Apache Hadoop. (2022). [Java]. The Apache Software Foundation. <https://github.com/apache/hadoop> (Original work published 2014)
- Apache Spark. (2022). [Scala]. The Apache Software Foundation. <https://github.com/apache/spark> (Original work published 2014)
- Arakelyan, G., Soghomonyan, G., & The Aim team. (2020). *Aim* (3.9.3) [TypeScript]. <https://doi.org/10.5281/zenodo.6536395>
- ASHRAE. (2022). *Data Center Power Equipment Thermal Guidelines and Best Practices*. ASHRAE Technical Committee (TC) 9.9 Mission Critical Facilities, Data Centers, Technology Spaces, and Electronic Equipment.
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), 871-876. <https://doi.org/10.1126/science.abj8754>
- Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 12449-12460.
- Bannour, N., Ghannay, S., Névél, A., & Ligozat, A.-L. (2021). *Evaluating the carbon footprint of NLP methods: A survey and analysis of existing tools*. 11-21. <https://doi.org/10.18653/v1/2021.sustainlp-1.2>
- Basic Linear Algebra Subprograms. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Basic_Linear_Algebra_Subprograms&oldid=1120747813
- Benaic, N., & Hogarth, I. (2022). *State of AI 2022*. <https://www.stateof.ai>

- Bitkom e.V. (2022a, January 3). *The IT skill gap is widening: 96,000 open jobs* / Bitkom e.V. <https://www.bitkom.org/Presse/Presseinformation/IT-Fachkraefte-luecke-will-grow>
- Bitkom e.V. (2022b). *Bitkom Positions paper: Prospects for a sustainable data center economy by 2030*. https://www.bitkom.org/sites/main/files/2022-10/2210-Position_Paper-Sustainable-Data_Centers.pdf
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. <https://doi.org/10.48550/arXiv.2108.07258>
- Boroditsky, L. (2012, March 15). *Linguistics: How language shapes thinking*. Spektrum.de. <https://www.spektrum.de/news/linguistik-wie-die-sprache-das-denken-formt/1145804>
- Branch, H. J., Cefalu, J. R., McHugh, J., Hujer, L., Bahl, A., Iglesias, D. del C., Heichman, R., & Darwishi, R. (2022). *Evaluating the Susceptibility of Pre-Trained Language Models via Handcrafted Adversarial Examples* (arXiv:2209.02128). arXiv. <https://doi.org/10.48550/arXiv.2209.02128>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- Brundage, M., Mayer, K., Eloundou, T., Agarwal, S., Adler, S., Krueger, G., Leike, J., & Mishkin, P. (2022). Lessons learned on language model safety and misuse. *OpenAI*. <https://openai.com/blog/language-model-safety-and-misuse/>
- Federal Ministry of Digital and Transport. (2022). *Digitalstrategie Deutschland*. <https://digitalstrategie-deutschland.de/medien/>
- Campa, C., Kawalek, C., Vo, H., & Bessoudo, J. (2020, May 14). *Defining AI Innovation with NVIDIA DGX A100*. NVIDIA Technical Blog. <https://developer.nvidia.com/blog/defining-ai-innovation-with-dgx-a100/>
- Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., & Erlingsson, U. (2021). Extracting training data from large language models. *30th USENIX Security Symposium (USENIX Security 21)*, 2633-2650. <https://www.usenix.org/system/files/sec21-carlini-extracting.pdf>
- Carpintero, A. G. (2021, November 21). *MLOps with Docker and Jenkins: Automating Machine Learning Pipelines*. Medium. <https://towardsdatascience.com/translation.php?a=17&l=16&l=17>
- Cen, S., & Shah, D. (2021). Regulating algorithmic filtering on social media. *Advances in Neural Information Processing Systems*, 34, 6997-7011. <https://proceedings.neurips.cc/paper/2021/hash/38b4f06e27fd4f6fdccceabc6f5c068ea-Abstract.html>
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (Vol. 34, pp. 15084-15097). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf>
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., & others. (2021). Evaluating large language models trained on code. *ArXiv Preprint ArXiv:2107.03374*.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... (2022). *PaLM: Scaling Language Modeling with Pathways* (arXiv:2204.02311). arXiv. <https://doi.org/10.48550/arXiv.2204.02311>
- Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G. M., Sorger, P. K., & AlQuraishi, M. (2022). Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11), Article 11. <https://doi.org/10.1038/s41587-022-01432-w>
- Climate Neutral Data Centre Pact – The Green Deal need Green Infrastructure*. (o.D.). Retrieved 19 December 2022, from <https://www.climate-neutral-data-centre.net/>
- CodeCarbon. (2020). *CodeCarbon—CodeCarbon 2.0.0 documentation*. <https://mlco2.github.io/codecarbon/>
- Cremers, A. B., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., Julia Rosenzweig, J., Rostalski, F., Volmer, J., & Voosholz, J. (2019). Trusted use of Artificial Intelligence. Areas of action

- from a philosophical, ethical, legal and technological point of view as a basis for an Artificial Intelligence certification. *Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)*. http://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf
- CSTB Releases Report *Fostering Responsible Computing Research: Foundations and Practices* » CCC Blog. (2022, May 16). <https://cccblog.org/2022/05/16/cstb-releases-report-fostering-computing-research-foundations-and-practices/>
- Dask. (2022). [Python]. dask. <https://github.com/dask/dask> (Original work published 2015)
- Data protection in the EU*. (o.D.). [Text]. European Commission - European Commission. Retrieved 28 November 2022, from https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en
- Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, 137–150.
- Deep Lake. (2022). [Python]. Activeloop. <https://github.com/activeloopai/deeplake> (Original work published 2019)
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. <https://doi.org/10.48550/arXiv.1810.04805>
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. <https://doi.org/10.18653/v1/2021.emnlp-main.98>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2022, March 23). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. International Conference on Learning Representations. <https://openreview.net/forum?id=163556id=YicbFdNTTy>
- Dürr, B. (2018). *IT rooms and data centers plan and operate: Manual of Structural Measures and Technical Building Equipment* (Vol. 2). Publishing house Bau+Technik.
- ELE Consortium. (2022). *Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap*. <https://european-language-equality.eu/agenda/>
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2022). Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1), 4839–4886.
- FFCV. (2022). [Python]. FFCV. <https://github.com/libffcv/ffcv> (Original work published 2021)
- Frostig, R., Johnson, M., & Leary, C. (2018). *Compiling machine learning programs via high-level tracing*. <https://mlsys.org/Conferences/doc/2018/146.pdf>
- Gehlhaus, D., & Koslosky, L. (2022). Training Tomorrow's AI Workforce. *Center for Security and Emerging Technology*. <https://cset.georgetown.edu/publication/training-tomorrows-ai-workforce/>
- Gehlhaus, D., Koslosky, L., Goode, K., & Perkins, C. (2021). U.S. AI Workforce: Policy Recommendations. *Center for Security and Emerging Technology*. <https://cset.georgetown.edu/publication/u-s-ai-workforce-policy-recommendations/>
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models* (arXiv:2009.11462). arXiv. <https://doi.org/10.48550/arXiv.2009.11462>
- Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P.-S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., ... Irving, G. (2022). *Improving alignment of dialogue agents via targeted human judgments* (arXiv:2209.14375). arXiv. <https://doi.org/10.48550/arXiv.2209.14375>
- GlusterFS. (2022). [C]. Gluster.org. <https://github.com/gluster/glusterfs> (Original work published 2011)
- Gopani, A. (2021, July 16). *JAX Vs TensorFlow Vs PyTorch: A Comparative Analysis*. Analytics India Magazine. <https://analyticsindiamag.com/jax-vs-tensorflow-vs-pytorch-a-comparative-analysis/>
- Hansell, S. (2002, April 8). Google's Toughest Search Is for a Business Model. *The New York Times*. <https://www.nytimes.com/2002/04/08/business/google-s-toughest-search-is-for-a-business-model.html>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J.

- F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), Article 7825. <https://doi.org/10.1038/s41586-020-2649-2>
- Hensel, M., & Ostler, U. (2020, November 22). The most popular providers of technology for high-performance computing. *Datacenter Insider*. <https://www.datacenter-insider.de/the-most-popular-vontechnik-provider-for-high-performance-computing-a-980532/>
- Hickmann, H., & Koneberg, F. (2022). *The occupations with the greatest shortage of skilled workers at present*. 67. <https://www.iwkoeln.de/studies/helen-hickmann-filiz-koneberg-the-professions-with-the-currently-growth-of-the-hard-crafts.html>
- Hintemann, Dr. R., Hinterholzer, S., Graß, M., & Grothey, T. (2022). *Bitkom study: Datacenters in Germany 2021 - Current market developments*. Bordstep Institute. <https://www.bitkom.org/sites/main/files/2022-02/10.02.22-study-datacenters.pdf>
- Hintemann, Dr. R., Hinterholzer, S., & Grothey, T. (2021). *Challenges and opportunities due to the boom in the construction of new data centers*. Hessian State Chancellery, Minister for Digital Strategy and Development. https://digitales.hessen.de/sites/digitales.hessen.de/files/2022-05/rechenzentrumsmarkt_hessen.pdf
- Hintemann, R. (2020). *Data centers 2018. Efficiency gains are not enough: Data center energy consumption continues to rise significantly - Cloud computing boosts growth*. <https://doi.org/10.13140/RG.2.2.26033.40800>
- Hintemann, R., & Clausen, J. (2018). *Importance of digital infrastructures in Germany*.
- Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network* (arXiv:1503.02531). arXiv. <https://doi.org/10.48550/arXiv.1503.02531>
- Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022). *Imagen Video: High Definition Video Generation with Diffusion Models* (arXiv:2210.02303). arXiv. <https://doi.org/10.48550/arXiv.2210.02303>
- Hugging Face. (o.D.). *Model Parallelism*. Retrieved 28 November 2022, from <https://huggingface.co/docs/transformers/v4.15.0/parallelism>
- Hydra. (2022). [Python]. Meta Research. <https://github.com/facebookresearch/hydra> (Original work published 2019)
- Intel. (2022, June 29). *Second-Gen Habana Gaudi2 Outperforms Nvidia A100*. Intel. <https://www.intel.com/content/www/us/en/newsroom/news/second-gen-habana-gaudi2-outperforms-nvidia-a100.html>
- ISO - ISO/IEC 27001 and related standards—Information security management. (o.D.). ISO. Retrieved 28 November 2022, from <https://www.iso.org/isoiec-27001-information-security.html>
- ISO 27017 and ISO 27018 Certification | DEKRA. (o.D.). Retrieved 6 December 2022, from <https://www.dekra.com/en/iso-27017-and-iso-27018-certification/>
- Izcard, G., & Grave, E. (2021). *Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering* (arXiv:2007.01282). arXiv. <https://doi.org/10.48550/arXiv.2007.01282>
- Jawahar, R. (2021, October 14). *Teaching AI to perceive the world through your eyes*. Meta AI. <https://ai.facebook.com/blog/teaching-ai-to-perceive-the-world-through-your-eyes/>
- Joseph, E., Riddle, M., Sorensen, T., & Conway, S. (2022). *The Economic and Societal Benefits of Linux Supercomputers*. <https://davidbader.net/publication/2022-hyperionresearch/>
- Jülich Research Center. (2022, June 15). *The first European exascale supercomputer arrives in Jülich*. https://www.fz-juelich.de/en/news/news/press_releases/2022/exascale-location-decision
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), Article 7873. <https://doi.org/10.1038/s41586-021-03819-2>
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kale, M., Siddhant, A., Al-Rfou, R., Xue, L., Constant, N., & Johnson, M. (2021). nmT5 - Is parallel data still relevant for pre-training massively multilingual language models? *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 683–691. <https://doi.org/10.18653/v1/2021.acl-short.87>

- [illegible]

- Mudigere, D., Hao, Y., Huang, J., Jia, Z., Tulloch, A., Sridharan, S., Liu, X., Ozdal, M., Nie, J., Park, J., Luo, L., Yang, J. A., Gao, L., Ivchenko, D., Basant, A., Hu, Y., Yang, J., Ardestani, E. K., Wang, X., ... Rao, V. (2022). *Software-Hardware Co-design for Fast and Scalable Training of Deep Learning Recommendation Models* (arXiv:2104.05158). arXiv. <https://doi.org/10.48550/arXiv.2104.05158>
- Mujkanovic, N., Sivalingam, K., & Lazzaro, A. (2020). *Optimising AI Training Deployments using Graph Compilers and Containers* (arXiv:2008.11675). arXiv. <https://doi.org/10.48550/arXiv.2008.11675>
- Nagrani, A., Seo, P. H., Seybold, B., Hauth, A., Manen, S., Sun, C., & Schmid, C. (2022). Learning Audio-Video Modalities from Image Captions. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision - ECCV 2022* (pp. 407–426). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19781-9_24
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2021). *WebGPT: Browser-assisted question-answering with human feedback* (arXiv:2112.09332). arXiv. <https://doi.org/10.48550/arXiv.2112.09332>
- Narayanan, D., Shoneyi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V. A., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A., & Zaharia, M. (2021). *Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM* (arXiv:2104.04473). arXiv. <https://doi.org/10.48550/arXiv.2104.04473>
- Nguyen, T. T., Trahay, F., Domke, J., Drozd, A., Vatai, E., Liao, J., Wahib, M., & Gerofi, B. (2022). Why Globally Re-shuffle? Revisiting Data Shuffling in Large Scale Deep Learning. *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 1085-1096. <https://doi.org/10.1109/IPDPS53621.2022.00109>
- Nichol, A. (2022, June 28). DALL·E 2 Pre-Training Mitigations. *OpenAI*. <https://openai.com/blog/dall-e-2-pre-training-mitigations/>
- NVIDIA. (o.D.). *NCCL and MPI - NCCL 2.15.5 documentation*. Retrieved 28 November 2022, from <https://docs.nvidia.com/deeplearning/nccl/user-guide/docs/mpi.html>
- NVIDIA Developer. (2013, July 2). *CUDA Toolkit—Free Tools and Training*. NVIDIA Developer. <https://developer.nvidia.com/cuda-toolkit>
- Ofeidis, I., Kiedanski, D., & Tassioulas, L. (2022). *An Overview of the Data-Loader Landscape: Comparative Performance Analysis* (arXiv:2209.13705). arXiv. <https://doi.org/10.48550/arXiv.2209.13705>
- Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19, 1750-1758. <https://doi.org/10.1016/j.csbj.2021.03.022>
- Open MPI. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Open_MPI&oldid=1120683830
- OpenAI. (2022a, July 18). Reducing Bias and Improving Safety in DALL·E 2. *OpenAI*. <https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2/>
- OpenAI. (2022b, November 30). *ChatGPT: Optimizing Language Models for Dialogue*. OpenAI. <https://openai.com/blog/chatgpt/>
- Paaß, G., & Giesselbach, S. (2023). *Foundation Models for Natural Language Processing*. Springer Cham. <https://link.springer.com/book/9783031231896>
- Papers with Code - Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. (o.D.). Retrieved 10 January 2023, from <https://paperswithcode.com/paper/exploring-the-limits-of-transfer-learning>
- Parliament, E. (2018). *European Parliament resolution of 11 September 2018 on language equality in the digital age (2018/2028 (INI))*. http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (arXiv:1912.01703). arXiv. <https://doi.org/10.48550/arXiv.1912.01703>
- Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer*, 55(7), 18–28. <https://doi.org/10.1109/MC.2022.3148714>
- Philippe Lorenz & Kate Saslow. (2019). *Demystifying AI & AI Companies. What foreign policy makers need to know about the global AI industry*. <https://www.stiftung-nv.de/en/publication/demystifying-ai-ai-companies-what-foreign-policy-makers-need-know-about-global-ai>

- Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9), Article 9. <https://doi.org/10.1038/s41562-022-01394-8>
- Poretschkin, M. (2022). *CERTIFIED AI / Ensure quality. Shaping progress*. ZERTIFIZIERTE KI. <https://www.zertifizierte-ki.de/>
- Poretschkin, M., Mock, M., & Wrobel, S. (2021). To systematically assess the trustworthiness of AI systems. *Regulation for Algorithms and Artificial Intelligence*, 175-202. <https://doi.org/10.5771/9783748927990>
- Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI* (arXiv:2204.01075). arXiv. <https://doi.org/10.48550/arXiv.2204.01075>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748-8763. <https://proceedings.mlr.press/v139/radford21a.html>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents* (arXiv:2204.06125). arXiv. <https://doi.org/10.48550/arXiv.2204.06125>
- Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., & de Freitas, N. (2022). *A Generalist Agent* (arXiv:2205.06175). arXiv. <https://doi.org/10.48550/arXiv.2205.06175>
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2022). *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech* (arXiv:2006.04558). arXiv. <https://doi.org/10.48550/arXiv.2006.04558>
- Research and Markets Ltd. (2021, June). *Europe Data Center Colocation Market: Market Size, Forecast, Insights, and Competitive Landscape*. <https://www.researchandmarkets.com/reports/5511065/europe-data-center-colocation-market-market>
- Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., & Kepner, J. (2022). *AI and ML Accelerator Survey and Trends* (arXiv:2210.04055). arXiv. <https://doi.org/10.48550/arXiv.2210.04055>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models* (arXiv:2112.10752). arXiv. <https://doi.org/10.48550/arXiv.2112.10752>
- Schödlwell, B., Zarnekow, D. R., Liu, R., Gröger, J., & Wilkens, M. (2018). *Indicators and indicators for assessing the resource efficiency of data centers and testing their practical applicability*.
- Schreiner, M. (2022, January 29). *Meta's AI chief: Three major challenges of artificial intelligence*. THE DECODER. <https://the-decoder.com/metasis-ai-chief-three-major-challenges-of-artificial-intelligence/>
- Schuhmann, C. (2021, August 8). *LAION-400-MILLION OPEN DATASET / LAION*. LAION. <https://laion.ai/blog/laion-400-open-dataset>
- Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning* (arXiv:2202.05924). arXiv. <https://doi.org/10.48550/arXiv.2202.05924>
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). *Retrieval Augmentation Reduces Hallucination in Conversation* (arXiv:2104.07567). arXiv. <https://doi.org/10.48550/arXiv.2104.07567>
- Simons, G. J., & Frese, A. (2021). *Future regional - digital: The Rhineland Revier; Feasibility study of data infrastructures in the Rhineland Revier*. The Ministry of Economy, Innovation, Digitalisation and Energy of the Land of North Rhine-Westphalia. http://www.economy.nrw/sites/default/files/documents/feasibility_study_datainfrastructures_lang_en.pdf
- Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., & Taigman, Y. (2022). *Make-A-Video: Text-to-Video Generation without Text-Video Data* (arXiv:2209.14792). arXiv. <https://doi.org/10.48550/arXiv.2209.14792>
- Sohofi, A., Yu, T., Aribal, A., Loetzsch, W., Team, S. D., & Wollmann, T. (2022). *Squirrel* [Python]. <https://github.com/merantix-momentum/squirrel-core> (Original work published 2022)
- Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., LeCun, Y., Müller, K.-R., Pereira, F., Rasmussen, C. E., Rä G., tsch, Schölkopf, B., Smola, A., Vincent, P., Weston, J., & Williamson, R.

- [illegible]

000

00000000000000000000000000title=Selene_(supercomputer)&oldid=1109224992

- Wu, C., Jiang, J., Ji, L., Wang, F., Jiang, D., & Duan, N. (2022). NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision - ECCV 2022* (pp. 720–736). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19787-1_41
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). *mT5: A massively multilingual pre-trained text-to-text transformer* (arXiv:2010.11934). arXiv. <https://doi.org/10.48550/arXiv.2010.11934>
- Yazdani-Jahromi, M., Yousefi, N., Tayebi, A., Kolanthai, E., Neal, C. J., Seal, S., & Garibay, O. O. (2022). AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. *Briefings in Bioinformatics*, 23(4). <https://doi.org/10.1093/bib/bbac272>
- Yin, P., Neubig, G., Yih, W., & Riedel, S. (2020). TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8413-8426. <https://doi.org/10.18653/v1/2020.acl-main.745>
- Yoo, A. B., Jette, M. A., & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. In D. Feitelson, L. Rudolph, & U. Schwiegelshohn (Eds.), *Job Scheduling Strategies for Parallel Processing* (pp. 44–60). Springer. https://doi.org/10.1007/10968987_3
- Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: Story Writing With Large Language Models. *27th International Conference on Intelligent User Interfaces*, 841-852. <https://doi.org/10.1145/3490099.3511105>
- Yuan, B., He, Y., Davis, J. Q., Zhang, T., Dao, T., Chen, B., Liang, P., Re, C., & Zhang, C. (2022). *Decentralized Training of Foundation Models in Heterogeneous Environments* (arXiv:2206.01288). arXiv. <https://doi.org/10.48550/arXiv.2206.01288>
- Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., & Choi, Y. (2022). *MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound* (arXiv:2201.02639). arXiv. <https://doi.org/10.48550/arXiv.2201.02639>
- Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Sellitto, M., Sakhaee, E., Shoham, Y., Clark, J., & Perrault, R. (2022). *Artificial Intelligence Index Report 2022*. Stanford Institute for Human-Centered AI, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf
- Zhang, H., Gong, Y., Shen, Y., Li, W., Lv, J., Duan, N., & Chen, W. (2021). Poolingformer: Long document modeling with pooling attention. *International Conference on Machine Learning*, 12437-12446. <https://arxiv.org/abs/2105.04371>
- Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., & Wu, Y. (2020). *Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition* (arXiv:2010.10504). arXiv. <https://doi.org/10.48550/arXiv.2010.10504>
- Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). *Automatic Chain of Thought Prompting in Large Language Models* (arXiv:2210.03493). arXiv. <https://doi.org/10.48550/arXiv.2210.03493>

II. Picture Directory

Education number	Name	Page
Fig. 1	Number of parameters of large AI language models since GPT-3 (open source models marked in red). Source: state of ai Report 2022 (Benaic & Hogarth, 2022)	19
Fig. 2	Clarification of self-attention in the phrase "The bank lends money". The token bank (below) has a high correlation to the token money (above), with the correlation to the other tokens being lower.	24
Fig. 3	Layers with self-attention blocks (blue) that compute context-sensitive embedding vectors (violet) of input tokens (green) are central to the foundation models. The logistical layer L predicts the probability of the output tokens. During training, the parameters are optimized so that the probabilities of the correct missing or next token (yellow) are as high as possible.	27
Fig. 4	A foundation model can take into account information from different data sources of different modalities. This one model can then solve a variety of downstream tasks (Bommasani et al., 2021).	29
Fig. 5	The accuracy of GPT-3's "few-shot" learning is enhanced by expanding the model size and the number of examples presented (Brown et al., 2020).	31
Fig. 6	Answer to a question in the Natural Question Benchmark found by a foundation model using Retrieval. Current models achieve an accuracy (F1) of 80% (Zhanag et al., 2021).	35
Fig. 7	Summary of an 800-word text by the BRIO model compared to the summary produced by experts (Liu et al., 2022).	36
Fig. 8	Images generated on different texts from DALL-E 2 (Ramesh et al., 2022)	40
Fig. 9	Videos produced by CogVideo on different texts (Hong et al., 2022)	41
Fig. 10	The Gato model generates new actions (dark blue) from the current states (light blue). The environment produces new states, etc. The model can process text, readings, images, etc.	44
Fig. 11	Around 2016, a new trend to very large models trained by major Internet companies emerged (red). These were able to finance the necessary investments. The bottom blue line illustrates the computational effort of the other models, e.g. of universities (Sevilla et al., 2022).	55
Fig. 12	Training performance (1 ExaFLOPs = 10^{18} FLOPs) of different foundation models compared to year of release by country	56

Fig. 13	Training performance (1018 FLOPs = 1 ExaFLOPs) compared to year of publication by country. Figure 1.12 (internal): Most of the foundation models were developed in the United States. From Germany and other EU countries, smaller foundation models with a lower training effort have tended to be released. The BLOOM model is an exception and was trained by a scientific collective of over 250 institutions on the Jean Zay supercomputer in France.	57
Fig. 14	Results of the survey of SMEs on their use of foundation models	69
Fig. 15	Results of the survey of SMEs on barriers to the use of foundation models	70
Fig. 16	Results of the SME survey on the importance of different aspects of foundation model development	72
Fig. 17	Results of the SME survey on the relevance of different types of AI models in foundation model development (responses with a response rate of less than 20% were omitted. For full answers, see Annex A.2.)	73
Fig. 18	Simplified presentation of the hardware and software infrastructure of HPCs	110
Fig. 19	The architecture of the training & application layer in detail	112
Fig. 20	The architecture of the system and data storage & loading layers in detail	120
Fig. 21	The architecture of the Framework & Service Layer in detail	122
Fig. 22	MLPerf hardware: accelerators (Zhang et al., 2022, p.18)	138
Fig. 23	Example invoice Building and operating your own HPC data center	160
Fig. 24	Delphi query: How do you assess the following risks for the development of the data center market in Germany? (Hintemann et al., 2022, S. 37)	162
Fig. 25	Organizational units of the LEAM-KI Service Centre	169
Fig. 26	The LEAM Board as the central governance unit of the LKS	181
Fig. 27	Scenario for the structure of the LEAM KI Service Centre	225

III. Table List

Table number	Name	Page
Table 1	A selection of possible applications based on language models	33
Table 2	A selection of possible applications based on multimodal models	39
Table 3	Overview of the most important information on the Economic Survey	67
Table 4	Business experts interviewed	68
Table 5	Thousands of GPUs are needed for training, tuning and inference in the operation of LEAM	111
Table 6	Examples of data centers	128
Table 7	Data center sizes	129
Table 8	Overview of cooling facilities in data centers	132
Table 9	Compute requirements for calculating a foundation model	137
Table 10	HPC locations in Germany	147
Table 11	Availability classes (VK1 - VK4)	152
Table 12	Protection classes according to DIN EN 50600-1	154
Table 13	Overview of OE Housing	170
Table 14	Overview of the Training-as-a-Service	174
Table 15	Overview of the costs of the LEAM KI Service Centre	185
Table 16	Total cost structure of the LEAM KI service centre with a four-year depreciation period	186
Table 17	Cost of the LEAM-KI service center when purchasing the GPU-RZ power	188
Table 18	Overview of revenues from the model training	190
Table 19	Overview of the revenue generated by the consultancy	191
Table 20	Comparison of the three financing scenarios for the LKS	193
Table 21	Overview of the advantages and disadvantages of the three funding models in terms of the regulatory framework	218
Table 22	Overview of the three online surveys	247

IV. Abbreviations list

Abb	Figure
Abs	Paragraph
AEUV	Treaty on the Functioning of the European Union
AI	Artificial Intelligence - Artificial Intelligence
API	Application Programming Interface
Art	Article
B2B	Business to Business
BERT	Bidirectional Encoder Representations from Transformers
BMWK	Federal Ministry of Economic Affairs and Climate Change
bspw	for example
bzgl	concerning
bzw	respectively
ca	circa
DMZ	Demilitarized Zone
DNA	Deoxyribonucleic acid
EU	European Union
F&E	Research and development
FLOP	or FLOPs - Floating Point Operations
FTE	Full-Time-Equivalents
GG	Basic Law
ggf	Where applicable
GPT	Generative Pre-trained Transformer
GWB	Anti-competition law
HPC	High Performance Computing
IaaS	Infrastructure-as-a-Service
IT	Information Technologies

KI	Artificial intelligence
KMU	small and medium-sized enterprises
LEAM	Large European Artificial Intelligence Models
LKS	LEAM AI Service Centre
Mio	Millions
ML	Machine Learning
Mrd	Milliarden
NLP	Natural Language Processing
NLU	Natural Language Understanding
o.D.	without date
OE	Organizational unit
OSS	Open Source Software
Pers	Person
PPP	Public-Private-Partnership
PUE	Power usage effectiveness
RZ	Data Center
S	Page
Seq2Seq	Sequence to Sequence
sog	the so-called
TVöD	Public Service Collective Agreement
u.a.	among others
USV	Uninterruptible power supply
usw	And so on.
vgl	comparisons
VPN	Virtual Private Network
z.B.	For example:
Point	Point

V. Feasibility study methodology

The methodologies used to conduct the feasibility study are described below. They can be divided into three main areas: Literature and Internet research, the collection of primary data in qualitative and quantitative form and their analysis.

In order to make it easier for the reader(s) to understand, the different steps of the methodological approach can be outlined as follows:

1. Literature and Internet research on AI foundation models, software requirements, high-performance data centers, and organizational and financial frameworks
2. Identification of relevant target audiences for the quantitative online survey:
 - a. Small and medium-sized AI companies and AI initiatives
 - b. Large companies with AI departments
 - c. AI researchers
3. Create, send and evaluate an online survey per target group
4. Identification of relevant stakeholder groups for qualitative interviews:
 - a. Large companies with AI departments
 - b. Data center operators and hardware experts
 - c. AI researchers
 - d. Other (e.g. AI initiatives, representatives of the Länder)
5. Creation of several guides/catalogues of questions based on the target audiences
6. Conducting and evaluating the guide-based interviews
7. Assessment of feasibility on the basis of all available information

Internet and literature research

The authors of each chapter used Internet and literature research as a starting point for the study. In many cases, it provides an appropriate introduction to the subject. The findings were also fed into the online survey as well as the interview guides.

As part of the research, relevant sources from the literature were used to gain basic knowledge of AI foundation models, their properties and peculiarities, and currently existing model types. In addition, the current state of development and use of AI foundation models in international comparison was researched.

In the area of software requirements, recent studies and the author's expertise in foundation model development were used to provide a detailed overview of the required software and personnel. In addition, existing literature was used to gain an overview of existing data centers and the challenges for AI training. Finally, the research also provided an introduction to possible governance, company form, and funding options.

Data collection

The data collection consisted of two aspects: quantitative online surveys and qualitative expert interviews. The quantitative data were collected in the form of a web-based online survey and sent to three groups of stakeholders:

- (1) Small and medium-sized AI companies and AI start-ups,
- (2) large enterprises and
- (3) AI Scientists: Inside.

The objective of the qualitative data collection was to gain a general view of the economic needs of AI foundation models and their scientific and technical prerequisites. The qualitative, guide-based interviews were conducted with selected experts from the fields of science, business, politics and law. The interviews were able to go deeper on findings from the quantitative survey as well as emerging topics.

Online surveys

The online survey was largely derived from Internet and literature research and is used to quantify key factors such as the relevance, use and development of various AI foundation models. The interview took place from 31 October to 23 November 2022.

The three catalogues of questions were prepared by the project partners of the feasibility study under the coordination of the KI Bundesverband. They consist of 20 to 46 questions (see Table 22), mostly in a multiple-choice format. Therefore, when evaluating their results, it is necessary to take into account that most of the time there were several options for each question. In addition to the multiple-choice questions, demographic information was requested and, at the end, an open-ended comment box allowed the opportunity to make further comments on the requirements of AI foundation models or the feasibility study. Details of the surveys can be found in Annex A.1.

In order to obtain a high quantitative number of cases, the three surveys were sent to a wide range of contacts. There were a total of 71 responses from small and medium-sized AI companies and 21 responses from AI scientists. Due to the low response rate among large companies, selected respondents were also asked for interviews in the online survey. Thus, the focus of this evaluation is on the interviews.

Survey	Number Questions	Number Recipient	Number Answers	Response-rate
Small and medium AI companies	44*	398	71	17,8 %
large companies	20	94	10	10,6 %
Science	23	74	21	28,4 %

*Questions were only partially displayed depending on the answer

Table 22: Overview of the three online surveys

Expert interviews

The interviews were based on guidelines that provided for open answers to the questions. There were individual guides for each target group with a focus on their expertise. The interviews were conducted by videoconference. The period of the survey was from 25 October 2022 to 12 January 2023.

In the selection of interviewers, a first step was to identify and recruit relevant stakeholder groups. As a result, the following groups or institutions have been identified:

- Large enterprises
- AI researchers
- Data center and hardware expert:
- Other experts (e.g. regional representatives, AI initiatives, etc.)

A total of 71 interviewees were interviewed. The full list of interviewers can be found in Annex B.1.

The analysis of the different data was undertaken by the project partners responsible and linked to the results of literature and internet research.

Analysis of online surveys

The results of the online survey of AI companies/initiatives, on the one hand, and AI researchers, on the other, were collected, evaluated and graphically edited to evaluate the quantitative data collection. The open-ended questions were searched for quantifiability on specific keywords. A graphical presentation of the multiple-choice questions is provided in Appendix A.2. Due to the low number of participants, there was no evaluation of the survey of large companies.

Evaluation of the expert interviews

In order to evaluate the qualitative data in the form of the guidebook-based interviews, the interviews were transcribed and compiled into result protocols. In the case of AI researchers, the results were compiled into an interview matrix for internal evaluation. The business interviewees were asked to release the results protocols for printing. The results protocols released by the end of the editorial period are set out in Annex B.3.

CONTRIBUTORS



FEASIBILITY STUDY

for the establishment and operation of a
dedicated AI high-performance computing centre
for training large AI models in Germany.