Lernende Systeme

GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

Developing and Applying Large Language Models

White Paper

Löser, A., Tresp, V. et al.
Working Group Technological Enablers and Data Science

Executive Summary

Currently, many of the most advanced generative models, including large language models, are being developed in the United States and China, but they are often not openly accessible. Furthermore, alignment with European values and guidelines is not guaranteed. German companies are therefore dependent on the quality of training data and models provided by these suppliers when utilizing them.

Given the increasing influence and rapid technological development of these AI models, there is a need to address Europe's dependencies on technology and data and to create alternatives to drive innovation and competitiveness in Germany and Europe, ensuring digital sovereignty. This emphasizes the demand and importance of a comprehensive, open, and commercially usable German training dataset, curated according to European values and regulations. Such a dataset could foster the development of various language models in research, business, and civil society, making it easier to transfer them into practical applications.

Application perspectives: Potentials and challenges

Large language models offer tremendous potential for German and European companies by replacing older Natural Language Processing (NLP) technologies and optimizing traditional NLP tasks. Multi-modal models, for example, enable new applications such as searching for product descriptions solely through the input of images. However, a significant value of these models lies in their reusability, as they can be adapted to industry- and company-specific requirements and data without the need to retrain a large, pre-trained AI model from scratch. The use of such models can also take place in areas where little proprietary data was previously available. This opens up opportunities for use cases in companies that have limited or no resources for machine learning and data science.

Many of these use cases, particularly in business areas and services with a supporting function, enable companies to achieve noticeable efficiency improvements, as large language models enhance information access and automate repetitive processes. However, maintaining the confidentiality of data processing is crucial for such use cases to ensure data protection and the protection of sensitive (business) data during processing and data sharing with third parties. Solutions are necessary to address these challenges, as well as legal and planning certainty for stakeholders, especially regarding copyright, licensing, and data protection regulations concerning the training data.

Large language models – overview of applications and task:

- **Application areas:**
  IT industry, sales, marketing, publishing, service sector, knowledge management, healthcare, and many more.

- **Tasks:**
  Verification of information, research support, text analysis, knowledge extraction from documents, document matching and processing, information retrieval, text generation, automatic translation, dialogue systems, chatbots, code generation, and much more.

- **Multi-modal models can fulfill various tasks:**
  Generation of images and videos from text input, creation of 3D shapes from 2D drawings, component creation, robot control, cross-modal search, generation of image captions and descriptions, cross-modal alignment to improve model capabilities, and much more.

The diverse applications of language models highlight the potential and challenges of this AI technology, exemplified by two application areas: business applications and healthcare applications. These areas involve processing sensitive data, including sensitive business data and internal knowledge in the case of business applications, and privacy-sensitive patient data in healthcare applications.

Business applications: Two key areas for future AI-based applications in businesses are digital assistants and document processing. Digital assistants play a crucial role in designing user-friendly business applications. They enhance the user experience through navigation and search assistance, as well as question answering in natural language. Unlike simple chatbots, digital assistants can work across application boundaries, consider context, and continually adjust to users through personalization. In document processing, language technologies and information extraction offer the possibility to ease and automate repetitive tasks. This enables the efficient analysis and linking of complex business documents by recognizing, extracting, and enriching or linking relevant information in the document, whether it's invoices, payment notifications, or orders.

Healthcare: Large language models also open groundbreaking possibilities in healthcare, particularly in outcome prediction and decision support. Precise, robust, and explainable results are crucial in this context. For outcome prediction, language models assist medical personnel in the differential diagnosis by suggesting possible diagnoses and pointing out anomalies based on textual data, vital signs, and laboratory parameters. The combination of language models with medical expertise (e.g. see hybrid AI) and other data sources further enhances predictive power and robustness. In decision support, pre-trained language models serve as the foundation for question-answering and chatbots in the medical field. However, the challenge lies in the transparency of predictions, especially in clinical situations where rapid verifications are necessary.

In the healthcare sector, it is evident that applications of language models have primarily focused on English-language medical and clinical texts due to the availability of relevant datasets, including publications from the PubMed database or clinical patient letters and data from the MIMIC database (Medical Information Mart for Intensive Care). This underscores the need to develop domain-specific language models for the German language, as pre-training on medical texts in comparison to domain-independent models yields better results.

Levels of digital sovereignty regarding large language models

Large language models are at the core of diverse applications. At the same time, these AI models must be (further) developed in line with European values. They should provide legal certainty for researchers, developers, and users to ease implementation in Germany and Europe. With such models, it is ensured that in sensitive areas such as healthcare, disaster management, and crucial industries, access to such AI models can be achieved without creating economic or technological dependencies. This enhances digital sovereignty and, thus, the ability to shape digital transformation in accordance with European values for fair competition by creating conditions for the sovereign realization of the economic and societal potential of large language models according to European standards. In addition to the technology- and data-related levels of digital sovereignty, other relevant levels play a central role in the context of large language models, too.

- **European values and legal system:** In the European context, language models raise both legal and ethical questions, including issues related to data protection, copyright, and the spread of misinformation. The European Union's AI Act aims to provide a legal framework for AI and ensure compliance with European values regarding large language models. Companies need clear guidelines for the use of generative AI models, including transparent information about training data and compliance with European law. Therefore, developing German-language and European language models is essential to ensure ethical standards in the selection of training data, avoid bias, and protect privacy.

- **Data – an essential requirement:** Access to extensive and high-quality datasets is crucial for the development of powerful language models. While custom models tailored to specific requirements with limited data are possible, transparency and legal certainty in training datasets are often not guaranteed. This poses challenges for researchers and companies when deploying and implementing language models in Germany and Europe. The relatively low percentage of German-language data in current multi-lingual models is particularly problematic, leading to inaccuracies.

- **Graphics processors:** Graphics processors are crucial as AI accelerators, primarily used for computing AI models, especially for transformer architectures in machine learning. Simultaneously, efforts are underway to develop chips optimized for transformers and chip technologies capable of handling large AI algorithms on small devices. The competition for the development of advanced language and multimodal models is driving increased demand for graphics processors.

- **Computing Infrastructure:** The computational requirements, model parameters, and data volumes for large AI models have significantly increased in recent years, a trend likely to intensify with future multimodal models. Initiatives in Europe and Germany aim to strengthen digital sovereignty in the field of computing infrastructure through high-performance computers. Operating commercial data centers is significant to continuously use language models in commercial operations and create alternatives to non-European offerings. Currently, Germany operates 36 of the world's top 500 most powerful computers, three of which are a part of the Gauss Centre for Supercomputing.

- **Cloud-based and locally executable models:** Access to AI models impacts digital sovereignty on three levels: training a proprietary model, using APIs from major AI companies, and customizing open-source models. The first approach requires considerable expertise and resources. The second option carries the risk of AI research being dominated by a few large models. European start-ups like Aleph Alpha could provide alternatives. Finally, locally executable models based on open source can reduce dependencies and meet specific requirements.

- **Talents:** Successful language model development requires data, computing power, suitable algorithms and a well-connected and coordinated AI community with relevant expertise. Two communities are key: a technical one focusing on natural language processing and machine learning, and another consisting of businesses, consulting firms, and universities focusing on adapting large

language models. Developing language models requires specialized talents with master's or doctoral degrees in natural language processing, data engineering, or machine learning, as well as complex skills, including domain and customer understanding, transferring domain-specific circumstances into machine learning procedures, and estimating feasibility and costs. While the talent pool in Germany and Europe is comparatively large, many top talents leave Germany for leading AI locations, making talent shortages a challenge, particularly for AI start-ups.

Conclusion and options for action

Given the enormous potential of large language models, it is crucial to gain clarity on what digital sovereignty entails, especially concerning this key technology, and to identify the levels and components contributing to realizing digital sovereignty in this overall context.

To advance and promote the development of large language models in line with digital sovereignty in Germany and Europe, various measures need to be addressed:

Table 1: Levels of Digital Sovereignty (DS) in Large Language Models – Summary

Levels of Digital Sovereignty (DS)	Elaborations
European values and legal system	• The AI Act can become an instrument to enforce European values in large language models. **(+)** • Currently, many well-known models do not adhere to criteria (or uncertainties persist). **(-)**
Data	• Several larger German-language text corpora already exist. **(+)** • A comprehensive (10 to 15 terabytes), widely available, and curated German-language text dataset in line with European values and regulations is needed. **(-)** • A relatively small proportion of German text in existing well-known models can affect the quality of model outputs in the German language. **(-)** • Copyright and licensing challenges in creating extensive, broadly usable corpora for model training. **(-)**
(Graphics) Processors	• The EU holds only a 10 percent market share in the chip market. **(-)** • Chip manufacturers depend on complex production machinery, where European producers are market leaders. **(+)** • In general, the European Chips Act and the establishing of production facilities sites counteract dependencies. **(+)** • There remains a general dependency, especially regarding the best and most powerful GPUs (see NVIDIA). **(-)**
Computing infrastructure	• GU EuroHPC and GSC initiatives contribute to DS, especially but not exclusively in research. **(+)** • Some private initiatives contribute to DS. **(+)** • 25 percent of companies intend to invest in or expand their own resources. **(+)** • However, infrastructure needs to grow in accordance with increasing requirements, and more European commercial solutions are necessary. **(-)** • However, 74 percent of companies are dependent on external resources and often on non-European cloud providers. **(-)**
Models	• Open Source with local language models (local LLM) can contribute to DS. **(+)** • European start-ups are an opportunity for more DS. **(+)** • But most models are created in the USA and China. Large, new, and intricately trainable models will continue to be developed by major tech companies and institutions. **(-)**
Talents	• In comparison to other countries, Germany has a good overall situation. **(+)** • 35 percent of positions at AI start-ups remain unfilled. **(-)** • Emigration of top talents. **(-)** • Talents at the intersection of natural language processing and machine learning are not sufficiently available. **(-)**

Source: Own compilation.

In order to advance and promote the development of large language models in the sense of digital sovereignty in Germany and Europe, various measures need to be taken:

- **Make extensive and curated training datasets available as open source:** provide comprehensive and curated training datasets, aligned with German and European values, as open source. This will support the development of AI models and foster collaborative cooperation within the AI ecosystem.

- **Ensure suffcient access to AI Chips and computing infrastructure:** guarantee the adequate availability of AI accelerators and drive the expansion of computing infrastructure in Germany and Europe as needed.

- **Develop language models in alignment with European values:** develop open and proprietary language models in line with European values to establish a reliable foundation.

- **Promote community building and development:** foster community building and development to effectively integrate data collection, computing infrastructure, and AI expertise.

- **Promote and build up talent:** encourage and build talent through internships, research projects, and collaborative initiatives in both research and industry.

Imprint
Editor: Plattform Lernende Systeme – Germany's Platform for Artificial Intelligence | Managing Office | c/o acatech | Karolinenplatz 4 | D-80333 Munich | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de | Follow us on X: @Lernende Systeme | LinkedIn: de.linkedin.com/company/plattform-lernende-systeme | Mastodon: social.bund.de/@LernendeSysteme | Status: December 2023 | Photo credit: nuttapong punna/iStock/Title | Translated using DeepL.com

This executive summary is based on the white paper: *Große Sprachmodelle entwickeln und anwenden. Ansätze für ein souveränes Vorgehen*. Munich, 2023. The authors are members of the working group Technological Enablers and Data Science of Plattform Lernende Systeme. The original version of this publication is available at: *https://doi.org/10.48669/pls_2023-6*

SPONSORED BY THE

Federal Ministry of Education and Research

acatech
NATIONAL ACADEMY OF SCIENCE AND ENGINEERING