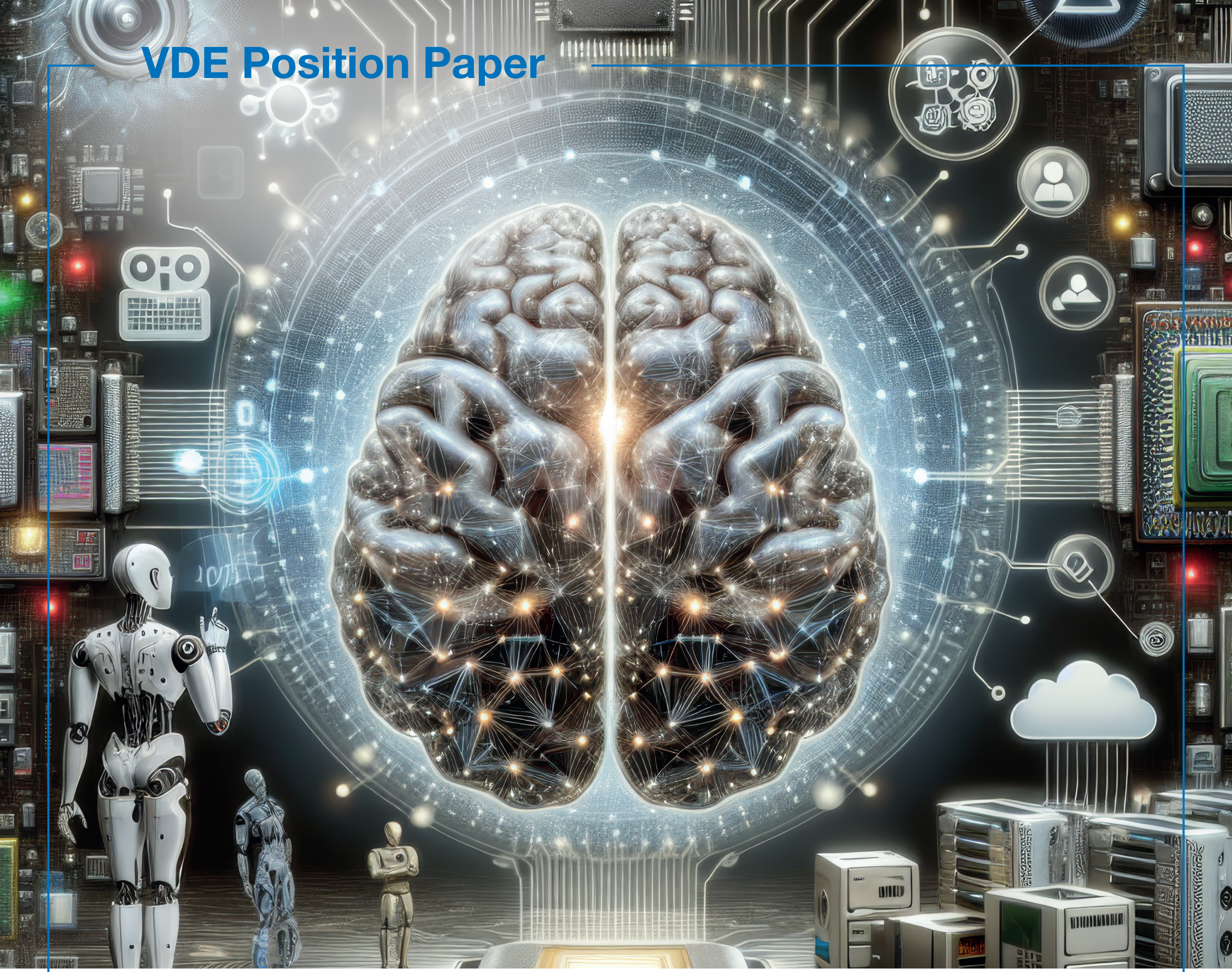
VDE Position Paper



Large Language Models are Transformers in Artificial Intelligence, Industry, Education, and Society

by VDE ITG

VDE

Recommended citation:
VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.:
Large Language Models are Transformers in Artificial Intelligence, Industry, Education, and Society,
VDE Position Paper, Offenbach am Main, November 2024

Editor:
Prof. Dr.-Ing. Rainer Martin, Ruhr-Universität Bochum

Authors and Reviewers:
Prof. Dr.-Ing. Stefan Brüggenwirth, Fraunhofer FHR, Wachtberg
Dr. phil. Aljoscha Burchard, DFKI Berlin
Prof. Dr.-Ing. Tim Fingscheidt, TU Braunschweig
Prof. Dr. rer. nat. Holger Hoos, RWTH Aachen
Dr.-Ing. Klaus Illgner, K|Lens GmbH, Saarbrücken
Dr. rer. nat. Henrik Junklewitz, VDE Verband der Elektrotechnik Elektronik Informationstechnik e.V.
Prof. Dr.-Ing. André Kaup, Friedrich-Alexander-Universität Erlangen-Nürnberg
Dr. phil. Katharina von Knop, VDE Verband der Elektrotechnik Elektronik Informationstechnik e.V.
Dr.-Ing. Joachim Köhler, Fraunhofer IAIS, St. Augustin
Prof. Dr. rer. nat. Gitta Kutyniok, Ludwig-Maximilians-Universität München
Prof. Dr.-Ing. Rainer Martin, Ruhr-Universität Bochum
Prof. Dr.-Ing. Dorothea Kolossa, TU Berlin
Prof. Dr.-Ing. Sebastian Möller, TU Berlin
Dr. rer. nat. Ralf Schlüter and David Thulke, M.Sc., RWTH Aachen
Dr. rer. nat. Vera Schmitt, TU Berlin
Prof. Dr.-Ing. Ingo Siegert, Otto von Guericke Universität, Magdeburg
Dr.-Ing. Volker Ziegler, Nokia, München

Acknowledgement:
The authors would like to thank Luca Becker (RUB), Julia Ringeis (RUB), and Franziska Bienek (ITG) for helping with illustrations and preparation of the manuscript.

Publisher:
VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.
Informationstechnische Gesellschaft im VDE (VDE ITG)
Merianstraße 28
63069 Offenbach am Main
Germany
Phone +49 69 6308-312
itg@vde.com
www.vde.com/itg

Cover image: © Ruhr-Universität Bochum
The cover picture symbolizes the network of information in large language models and the resources required to implement them. It was generated with ChatGPT 4o and the keywords “large language model, brain, neural processor hardware, multiple robots,” and also contains random elements generated by the AI.

Design: Schaper Kommunikation, Bad Nauheim

November 2024

Content

Zusammenfassung

Große Sprachmodelle (engl. large language models, LLMs) haben sich in kürzester Zeit zu einer wesentlichen Grundlage für viele intelligente, informationstechnische Anwendungen entwickelt, deren Potential bei weitem noch nicht ausgeschöpft ist und sich rapide weiterentwickelt. Große Sprachmodelle unterstützen nicht nur typische Sprachanwendungen, wie z.B. automatische Diktier- und Übersetzungssysteme, sondern erlauben auch die Analyse und semantische Interpretation unterschiedlichster Datentypen, u.a. Video, Audio und Radar. Sie sind damit zu einem Treiber, wenn nicht gar zum Inbegriff (z.B. in der Form von ChatGPT) der künstlichen Intelligenz geworden. Mit den großen Sprachmodellen hat sich eine Technologie entwickelt, die nicht nur die Interaktion zwischen Mensch und Maschine in Form von Texten revolutioniert, sondern auch eine Vielzahl von industriellen und medizinischen Anwendungen, wie z.B. die automatische Bildannotation und das Verstehen von visuellen Szenen, neu definiert. Ihre breite Einführung erfordert jedoch die Bewältigung ethischer, technischer und gesellschaftlicher Herausforderungen, um ihre verantwortungsvolle und gerechte Nutzung zu gewährleisten.

In diesem Positionspapier erläutern wir zunächst die technischen Grundlagen und Anwendungen großer Sprachmodelle und befassen uns mit Chancen und Herausforderungen für Industrie, Gesellschaft und Bildung, die in diesem Kontext stehen. Dabei wird auch auf die Bedeutung einer agilen Regulierung, der Ausbildung an Schulen und Universitäten und der nachhaltigen Forschung in diesem Bereich eingegangen. Schließlich werden technische Hürden, gerade auch für kleinere Unternehmen und Forschungseinrichtungen, sowie auch problematische Aspekte dieser Technologie benannt.

Insgesamt bieten große Sprachmodelle wie GPT-4 eine Vielzahl an Möglichkeiten, die derzeit wirtschaftlich weiter erschlossen werden. Es ist jedoch von entscheidender Bedeutung, die damit verbundenen Herausforderungen zu bewältigen, um sicherzustellen, dass ihr Nutzen auf nachhaltige Weise maximiert wird. In Anbetracht ihrer Abhängigkeit von sprachlichen Ressourcen und der damit einhergehenden kulturellen Prägung, müssen deutsche und europäische Firmen und Institutionen in die Lage versetzt werden, hierbei eine führende Rolle zu übernehmen. Daher besteht die Notwendigkeit, die technische Entwicklung und industrielle Verwertung von großen Sprachmodellen in Deutschland, aber auch die Diskussion von Chancen und Risiken in der Öffentlichkeit zu fördern. In diesem Gesprächsdiskurs und der Förderung der Technologie sollte man den wirtschaftlichen Möglichkeiten, dem dynamischen Transfer aus der Grundlagenforschung in die industrielle Anwendung und vor allem auch dem gesellschaftlichen Nutzen große Sichtbarkeit verleihen. Hierbei möchten wir als Verfassende dieses Positionspapiers hervorheben, dass es bei generativer KI nicht nur um die Entwicklung von Algorithmen und Software für bahnbrechende neue Anwendungen geht, sondern dass auch weitere Fortschritte im Bereich der energieeffizienten, digitalen Hardware und des gemeinsamen Entwurfs von Hardware und Software eine wesentliche Rolle spielen. Die enormen Chancen dieser Technologie sind mit neuen Herausforderungen verknüpft, die wir in Verantwortung für den gesellschaftlichen Nutzen und mit Weitblick auf die nationale und europäische wirtschaftliche Entwicklung und unsere technologische Souveränität angehen sollten.

Executive Summary

Large language models (LLMs) have quickly become an essential basis for many intelligent, information technology applications, the potential of which is far from exhausted and is developing rapidly. LLMs not only support typical language processing applications, such as automatic dictation and translation systems, but also allow the analysis and semantic interpretation of a wide variety of data types, including video, audio and radar. They have thus become a driver, if not the epitome (e.g., in the form of ChatGPT) of artificial intelligence. With large-scale language models, technology has emerged to revolutionize not only human-machine interaction in the form of text, but also redefines many industrial and medical applications such as automatic image annotation and visual scene understanding. However, their widespread adoption requires overcoming ethical, technical and societal challenges to ensure their responsible and equitable use.

This position paper begins by explaining the technical foundations and applications of LLMs and looks at the opportunities and challenges for industry, society and education that arise in connection with these models. It also addresses the importance of agile regulation, education at schools and universities, and sustainable research in this area. Finally, technical hurdles, especially for smaller companies and research institutions, as well as problematic aspects of this technology are identified.

The rise of LLMs like GPT-4 offers exciting opportunities, but also presents a number of challenges. It is vital to address these challenges to ensure their benefits are maximized in a sustainable manner. Given their intrinsic dependence on linguistic (and thus cultural) resources, German and European companies and institutions should contribute to these developments to the fullest extent possible. To this end, we propose that an action plan be drawn up to promote not only the technical development and industrial exploitation of LLMs in Germany, but also the discussion of opportunities and risks among the public. The economic opportunities, the dynamic transfer from basic research to industrial application and, above all, the social benefits should be given high visibility in this dialogue and promotion of this technology. As the authors of this position paper, we would like to emphasize that generative AI is not just about the development of algorithms and software for ground-breaking new applications, but that further advances in the field of energy-efficient, digital hardware and hardware-software co-design methodologies also play a key role. The enormous opportunities offered by this technology are linked to new challenges that we should tackle with a sense of responsibility for the benefits to society and with a vision for national and European economic development and our collective technological sovereignty.

1 Introduction and Motivation

Since their remarkable breakthrough in 2022, large language models (LLMs) have revolutionized many applications in natural language processing and artificial intelligence. As of October 2024, OpenAI's ChatGPT, specifically leveraging the GPT-4 LLM, boasts a user base of 200 million active users globally and 1.5 billion visits per month (Demandsage, 2024). This significant user base reflects the widespread adoption and integration of AI-powered language models (LMs) in various aspects of text-generation and translation, computer code-generation, and business operations.

Powered by huge neural networks and trained on diverse and extensive datasets, these models have demonstrated an unprecedented ability to comprehend and generate human-like text. Their success is built on advancements in deep learning and the increasing availability of computational resources, enabling models like GPT-4 and subsequent iterations to push the boundaries of what applications in many areas of human-machine communication and data interpretation can achieve. These breakthrough achievements have been recently highlighted by the 2024 Nobel Prize in Physics for John Hopfield and Geoffrey Hinton, two pioneers in neural networks and machine learning.

At the heart of most of these LLMs is an architecture known as the *transformer,* an encoder-decoder neural network introduced by Vaswani et al. (2017) and modified into a decoder-only system by Liu et al. (2018). The transformer DNN leverages self-attention mechanisms, allowing it to weigh the importance of different words in a sentence relative to one another and thus to learn relations between words more efficiently than previous architectures. This capability is crucial for capturing context and generating coherent, contextually relevant text. Training these models involves processing massive datasets containing text from diverse sources, enabling them to learn grammar, facts, and even some reasoning abilities. The result is a model that can perform a wide range of language tasks, from translation and summarization to creative writing, and complex question answering. Furthermore, these models may be used in conjunction with other data modalities like video or audio, where they can be used in annotation and semantic interpretation tasks.

Training LLMs requires immense computational resources because their neural networks contain billions of parameters. Training from scratch also requires extensive datasets, necessitates powerful hardware, and requires substantial energy resources. By leveraging widely available text resources, major advances have been made through unsupervised pre-training followed by supervised fine-tuning steps, resulting in the generative *pre-trained transformer* (GPT) as introduced in (Radford et al., 2018). The training of these systems relies on extensive hardware resources including high-performance graphics processing units (GPUs), tensor processing units (TPUs), massive amounts of random access memory (RAM), and large data storage systems. Leading-edge data centers are typically equipped with thousands of interconnected GPUs to manage the computational load.

The business opportunities and implications of LLMs are vast and multifaceted. Companies across various industries are harnessing their capabilities to enhance customer service through sophisticated chatbots and virtual assistants that provide instant, accurate responses. In content creation, LLMs are being used for summarizing the state-of-the-art in science and technology and other fields of interest and for the development of software code, thereby significantly boosting productivity. The healthcare sector benefits from these models through improved diagnostics, personalized medicine, and streamlined administrative tasks. Moreover, in finance, LLMs aid in market analysis, fraud detection, and customer support, offering new avenues for efficiency and service improvement. As highlighted already in (LEAM:AI, 2023), all this justifies substantial and increasing investments into infrastructure, research, and education.

Despite their potential, LLMs present several risks and challenges. One significant concern is the potential for bias in their outputs, as these models can inadvertently learn and propagate biases present in their training data. This can lead to unfair or discriminatory outcomes, particularly in sensitive applications like hiring or lending. Additionally, the widespread use of LLMs raises ethical questions about misinformation and the potential for malicious use, such as generating deepfake content or automated disinformation campaigns. Moreover, the environmental impact of training these large models and deploying them in widely-used applications is a major concern, as they require a substantial amount of computational power, which in turn results in a considerable energy consumption (Desislavov et al., 2023).

All of these aspects will be further explored in the next sections of this position paper. Section 2 and 3 will provide a brief introduction to the background of LLM technology, including hardware and data requirements. **Readers less interested in the technical aspects may skip these sections and move on to Section 4 which outlines applications, services and business potentials.** The necessity and current state of regulatory activities, as well as research gaps, are discussed in detail in Section 5. Section 6 concludes this paper and recommends further actions.

Also, LLMs fine-tuned for specific fields of knowledge will be needed for AI deployment for the enterprise beyond the general-purpose skills of ChatGPT. This applies to e.g., the field of information technology itself where specialized terminology, phrases and acronyms are widespread. AI tools for specialized fields of knowledge typically need less effort in training, provide better accuracy, and require less computation during a session.

2 Technology Basics

2.1 Generative AI

Very broadly speaking, generative models are systems that can be used to create (or generate) data. This data can take many different shapes and forms: For example, as the core focus of this position paper, it can be text, generated in response to a user input, the so-called *prompt or query*. However, and in addition, the idea of generative modeling has been employed for many different types of media, such as for generating images, audio, or video signals.

Generative models have been around for more than half a century, spanning a broad range of rule-based and statistical approaches. These include Noam Chomsky's initial work on language models, cast in the form of sets of rules for the generation of admissible sentences, or hidden Markov models as purely statistical models that – for a long time – undergirded the most successful algorithms in diverse tasks such as decoding mobile communication signals or carrying out automatic speech recognition.

As of today, neural networks have taken generative modeling by storm. Among the first systems of this decade that truly caught the imagination of the public and the scientific world alike, were generative adversarial networks (Goodfellow et al., 2014), which were able to generate naturally looking images by pitting a generator and a discriminator against each other in training, such that the generator model learned to create images (or audio) that were virtually indistinguishable from reality, at least by other neural networks. Since then, generative models have evolved into powerful *foundation models* which are trained on diverse unlabeled data and are capable of supporting a wide variety of tasks. Foundation models encode relationships between data items and concepts in a relatively general way, making them an ideal starting point for developing a broad range of AI applications.

Today, neural generative models are mostly driven by two types of model architectures – the so-called *transformers*, more technical details on which will follow below, and diffusion models, underlying much of current generated art and video and bringing a lot of movement into the field of signal processing as well.

2.2 Large Language Models (LLMs)

In this subsection, we will provide a brief introduction to language models, to transformer-based LLMs and their properties, and eventually to ChatGPT-like applications. Given the vast number of publications and corresponding developments in this field, we aim to highlight the most important concepts only.

What is a language model? *Language models* (LMs) have existed for many decades as fundamental components of speech dialog systems, automatic speech recognition (ASR) such as *Siri* or *Google Home*, spell checking or machine translation tasks. In a nutshell, an LM is a system (be it a set of rules or a Markov chain or a deep neural network, DNN), which takes previous letters, sub-word units, or words, subsumed as "tokens" in the following, as input and predicts probabilities for the next token. One crucial element of these methodologies is the use of *word embeddings* which map words or, more broadly, tokens and their associated context to vectors in a high-dimensional space. By utilizing these vectors, the degree of similarity between tokens can be quantified, and predictions regarding subsequent tokens can be formulated. Thus, in a bird's eye perspective, an LM is an "old-text-in / follow-up-text-out" system. LMs have been successfully employed to reduce the word error rates of ASR systems, as they take care that reasonable output is generated, with words existing in the respective language, and with grammar rules etc. being followed.

How does transformer-based end-to-end automatic speech recognition work? In recent years, *transformers* have become one of the predominant technologies both for end-to-end ASR and machine translation tasks, but also for standalone LMs. As shown in Fig. 1, a transformer-based ASR system typically consists of an encoder with an audio feature sequence as input, and a decoder, jointly called attention-based encoder-decoder (AED) model. The decoder receives the encoder output as conditioning information, i.e., a hidden encoded representation sequence of the same length as the input audio feature sequence, and receives as further input the *tokens*, which have been recognized before (also called *query*). As the decoder predicts the probabilities for the next letter or word, an *argmax* function finds the most likely next token and outputs the respective recognized word or letters.

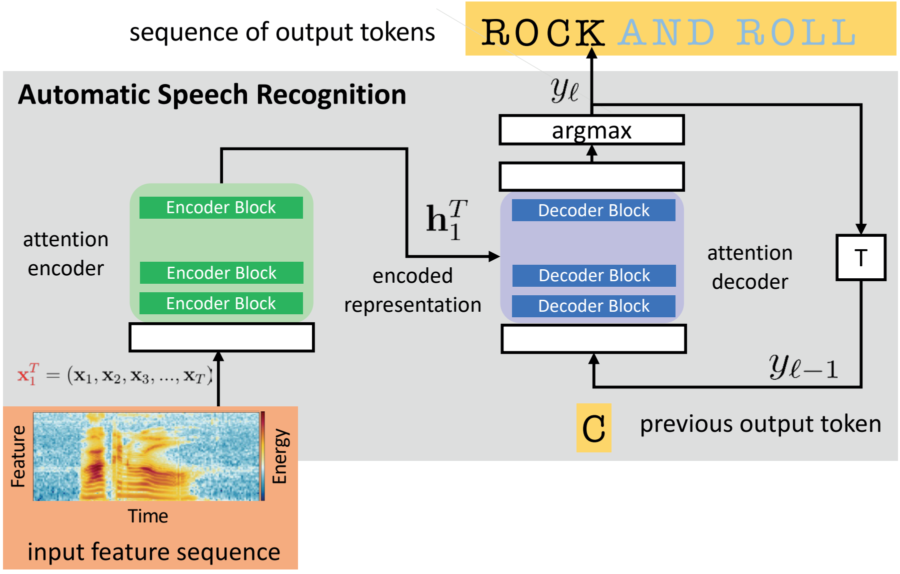


Figure 1: High-level block diagram of an attention-based encoder-decoder (AED) model for end-to-end automatic speech recognition (ASR). While the spoken audio is converted to a sequence of features and encoded, the decoder generates the spoken text in a recursive fashion.

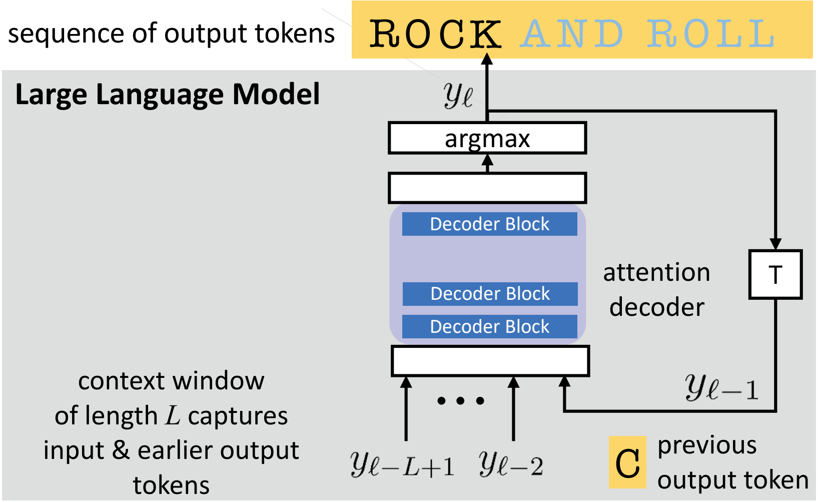


Figure 2: High-level block diagram of an attention-based LLM based on an arrangement of decoder blocks in a recursive generative loop.

This recognized output is then fed back ("T" block in Fig. 1) into the decoder at the next decoder timestep ℓ. It is important to understand that the encoder timesteps t are regular, for example one feature vector per 10 ms, while for a given audio input and hidden sequence, in general the autoregressive call of the decoder could run forever. This asynchronicity of encoder timestep t and decoder timestep ℓ is solved in practice by terminating decoding, once an <EOS> token has been decoded, marking the *end of sentence*.

The example in Fig. 1 displays the moment, when the entire input audio feature sequence is available, accordingly also the entire hidden representation sequence, but the decoder has only been called four times (decoder timestep $\ell = 4$), with output tokens (here: letters) R, O, C, and K. We still see the past output token C being input to the decoder.

LMs help speech recognition. Note that even an end-to-end ASR system is in practice equipped with an extra LM, which would have to be attached to Fig. 1 on the right hand side. This LM also takes the previous output token as input, and outputs a probability distribution for the next token. The LM's output probabilities would then have to be fused with the ASR decoder output probabilities, before the argmax is taken. We note that from the input/output relations point-of-view, an LM is very similar to an AED decoder, except that the LM model has no extra conditioning input from the left. This is of high advantage for an LM, as it does not need paired audio/text training data (available only in limited amounts), but it can be trained on text only, which is available in abundance on the Internet. The reason is that both its input and output are letters or words.

What does an LLM look like? Figure 2 shows such an LM with only text input and predicted text output. It consists of a serial call of a number of topologically equal decoder blocks, which are similar to the AED encoder blocks in Fig. 1, as there is no second type of input. This type of network block is called *self-attention* block, whereas the AED decoder in Fig. 1 consists of *cross-attention* blocks. A typical AED decoder consists of 6 decoder blocks, whereas an LM can consist of 12 to 100 and more decoder blocks, which then justifies the term large language model, resulting in network sizes of 100 million parameters or even a million times more.

Do transformer-based LMs scale? Early LMs were just a set of (handmade) rules, data-driven Markov chains, or, more recently and still today, recurrent neural networks. The attention-based transformer displayed in Fig. 2 is a powerful technology for language modeling that is available in very large network sizes (100 billion parameters and more), whereby any shrinking down of the network size also tends to incur a certain loss in performance. Going the other way and increasing the network size more and more, so far, transformers have shown to get better and better, indicating an effective scaling behavior. The expansion of training data and the enhancement of computational capabilities have so far consistently led to ever-increasing transformer performance.

Nonetheless, while training time increases with the number of decoder blocks, it increases quadratically with the signal context length in the attention blocks. This has been recently relaxed by selective state-space models, which are recurrent in nature, leading to the so-called MAMBA model (Gu and Dao, 2023) as one interesting example. Selective state-space models like MAMBA replicate the success of the attention model by selectively using – or not using – parts of the input sequence through a selective scan algorithm. In effect, this realizes a time-variant state-space model, which is nonetheless well-adapted to GPU computation and efficient in learning, circumventing the usual instabilities and vanishing gradients that were previously observed in recurrent models. While any predictions on the evolution of this dynamic field of work are necessarily speculative to an extent, the MAMBA model and the idea of selective state-space models appears to be highly attractive, as it offers a solution to the extreme growth of inference time in long-context models, while remaining easily parallelizable and stable during training time.

Additional work is ongoing on addressing the shortcomings of the long-dominant and well-known long short-term memory (LSTM) models. While LSTMs (Hochreiter and Schmidhuber, 1997) were highly successful in learning sequential data and were indeed long counted among the prevalent architectures in machine learning, especially so for language-related tasks, they have been outperformed by the recent transformer architectures. In an effort to reverse this trend, the so-called *xLSTM*-model (Beck et al., 2024) was recently introduced. It incorporates exponential gating and a new memory architecture allowing for parallelization. With these updates, the xLSTM, used as an LLM, has also recently achieved results that are comparable to both transformer models and state-space models when trained at scale, i.e., using billions of parameters and large training corpora.

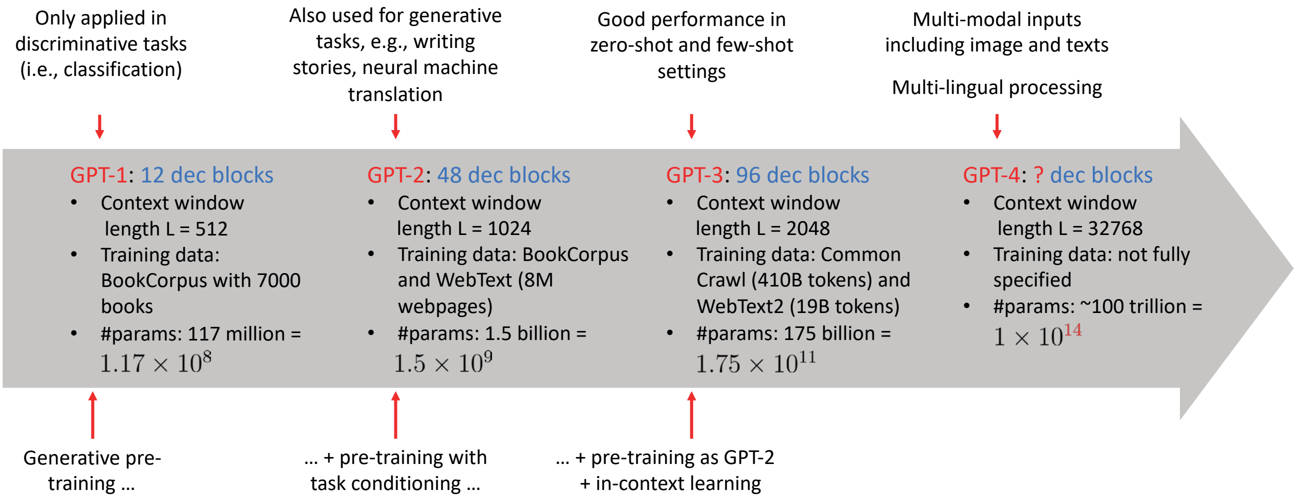

Only applied in discriminative tasks (i.e., classification)
Also used for generative tasks, e.g., writing stories, neural machine translation
Good performance in zero-shot and few-shot settings
Multi-modal inputs including image and texts
Multi-lingual processing
GPT-1: 12 dec blocks
Context window length L = 512
Training data: BookCorpus with 7000 books
#params: 117 million = 1.17×10^8
GPT-2: 48 dec blocks
Context window length L = 1024
Training data: BookCorpus and WebText (8M webpages)
#params: 1.5 billion = 1.5×10^9
GPT-3: 96 dec blocks
Context window length L = 2048
Training data: Common Crawl (410B tokens) and WebText2 (19B tokens)
#params: 175 billion = 1.75×10^{11}
GPT-4: ? dec blocks
Context window length L = 32768
Training data: not fully specified
#params: ~100 trillion = 1×10^{14}
Generative pre-training …
… + pre-training with task conditioning …
… + pre-training as GPT-2 + in-context learning

Figure 3: Evolution of the generative pre-trained transformer (GPT) LLM (Radford et al., 2018; Radford et al., 2019; Brown et al., 2020; OpenAI, 2024).

Evolution of the generative pre-trained transformer (GPT) LLMs: The initial versions of transformer-based LLMs were referred to as generative *pre-trained transformers* (GPTs) since they were only trained to predict next tokens on large text databases, without any additional human supervision, i.e., in a *self-supervised* fashion.

Subsequent experiments, however, showed that pre-trained models can be further improved via a variety of fine-tuning methods. These include supervised learning approaches based on one or multiple domain-specific data sets but also on data sets that contain typical user instructions ("Translate …", "Explain …", "Summarize …") and corresponding answers. While these purely data-driven approaches enhance the relevance of the LLM's output and may relieve the user of complicated *prompt engineering* tasks to some extent, LLMs could also be further optimized by fine-tuning on human labels. Specifically, the technique of reinforcement learning with human feedback (RLHF), where humans are asked to rank different model outputs and this ranking is used to improve generation, led to significant improvements in the model output.

In contrast, the LLM network topology remained largely unchanged: Figure 3 illustrates the relatively minor topological evolution that took place over the recent years: The number of decoder blocks increased steadily from 12 to 96 and beyond, and the input context window grew steadily, from 512 tokens for GPT-1 to 32768 tokens for GPT-4. This input context is the maximum prompt or query text size that can be processed at the LLM input. It can also be seen how the amount of training material grew from one generation of GPT to the next. Most impressive, however, is the model size: While GPT-1 consisted of 117 million network parameters, GPT-4 already is one million times larger as it features about 100 trillion parameters.

In-context learning: GPT-3 introduced in-context learning, which is *not a training strategy*, but a prompt engineering strategy. Acknowledging that a so-called zero-shot query is still somehow difficult ("Translate English to French: cheese =>"), a one-shot in-context learning approach results in a performance improvement of the system ("Translate English to French: love => amour; cheese =>"). This resembles the fact that for humans, a task also appears to be simpler if it was specified by examples. For the LLM, it can be observed that the larger the model, the more effective a K-shot query becomes with K>0.

Both scientifically and politically, however, the most interesting and yet still partially unanswered question is, how the *training* of LLMs needs to be performed, both to achieve maximal performance on specific tasks, and to ensure safety (e.g. avoidance of harmful outputs) and alignment (correspondence of model outputs with human values and preferences). To address these issues, we need to look at two aspects: task-conditioning and alignment to human feedback, as outlined in the following.

Task conditioning: When an LLM is pre-trained, this is done by letting the model predict next tokens (or, for masked LLMs, missing tokens) in a given sequence of text tokens. This training employs extremely large corpora, but without any additional knowledge or structure being added. It already results in surprisingly successful models, but in models that are merely trained to predict well what a human author would write next. It is not, however, yet optimized to produce the best output, giving the highest-quality answer to all queries and in any application domain.

Thus, once a basic, predictive model has been trained, the next step of task conditioning adapts the parameters in such a way that the model achieves better performance on a set of language processing tasks. Examples of these tasks include natural language inference and question answering. The former task requires the model to determine whether a second statement is logically inferred from a first one, whether it is neutral to it, or even contradictory. In contrast, question answering involves a set of questions and correct answers provided as training material. While this step of supervised task conditioning is indeed helpful to achieve better performance on the trained tasks, it turns out that reinforcement learning from human feedback is far more effective yet in improving model quality for open-world tasks, especially when diverse human-written prompts are involved.

Alignment via reinforcement learning from human feedback (RLHF): All LLM architectures we have considered here are *probabilistic,* i.e., their output is the result of sampling the probability distribution of the next word. This can be conceptualized as a two-step process:

(1) The transformer finds the set of N possible continuations of a sentence, and the associated probability distribution. As one example, imagine that the model is tasked with continuing the sentence: "For transformers, the greatest…" The transformer will then derive the next N most likely words and their probabilities (given in parentheses), which might be (for N = 3): challenge (47 %), advantage (34 %), problem (19 %).

(2) Then, the model will randomly draw the next word from this intermediate distribution. So, the sentence would most likely continue with the word “challenge,” but there is also a 34% chance that it continues with “advantage,” likely giving the entire continuation (which will then of course build on the previous text “For transformers, the greatest advantage”) a much more positive turn.

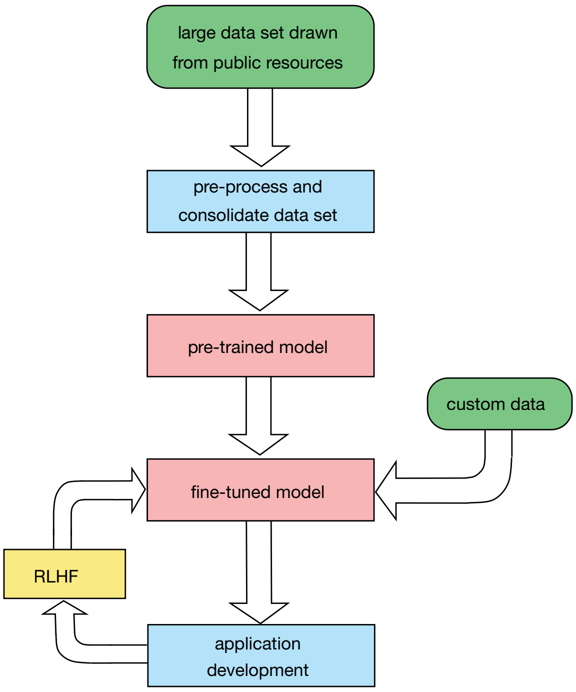

large data set drawn from public resources
pre-process and consolidate data set
pre-trained model
custom data
fine-tuned model
RLHF
application development

Figure 4: Training of fine-tuned LLMs using custom data and reinforcement learning from human feedback (RLHF).

In this way, as the generation of the next word is a process guided both by knowledge (in the computation of the next possible words) and by chance (in the random draw of the next concrete word), it is also possible to draw a number of possible answers to any question. This is where RLHF comes in: The model under training will be tasked to generate multiple possible continuations to an input. Human raters will then be asked to re-rank these, i.e., to order them by preference. Finally, the model will be re-trained to make the preferred outputs more likely than the less highly ranked ones, as depicted in Figure. 4.

This step of reinforcement learning has proven vital to the performance of LLMs, and it also gives designers and developers a handle in the training process, with which to adjust the alignment of the model to human preference and value judgements. This is most important, of course, to ensure safety of models, and has also been used, with typically great but also varying success, to prevent models from bias and prejudice, as may result from imbalances or biases implicitly encoded in the training data.

What is ChatGPT? The technologies of GPTs with RLHF forms the core of ChatGPT and of many related models, such as Anthropic’s Constitutional AI model or X.AI’s Grok. The fine-tuning stage is vital to ensure alignment with human value functions, and much work is being expended on making resulting models impervious to adversarial attacks or other attempts to elicit unwanted and potentially harmful responses. The scientific community in the field identified major values an LLM in direct dialogue with human users should follow. Among these are that the LLM shall be (1) truthful, (2) unbiased, (3) not toxic, and (4) not harmful. As mere LLMs just after training do not meet these standards well enough, OpenAI as the creator of ChatGPT invested a lot of work to derive ChatGPT from the more generic GPT-x family of models.

Based on prior academic work, LLMs such as Google’s Gemini or GPT-4o can also process multimodal data as they now seamlessly allow interfacing also to audio, image, and video. This has been achieved with single end-to-end models, instead of a former modular approach with user-perceivable delays during its use.

2.3 LMs for Annotating Multimodal Data and Scene Understanding

While LLMs have been originally designed for processing spoken or written language as explained above, their underlying architecture, specifically the transformer-based models, can be adapted and extended to also process other types of information effectively. The technical principle behind this adaptation lies in the ability of transformer-based architectures to learn complex patterns and relationships within sequential data, which can be applied not only to linguistic data but also to a wide variety of modalities. Of course, the selection of modalities depends on the desired application and the available sensors. In the context of combining text and image or video data, for instance, LLMs can be leveraged along several pathways:

Pre-training with multimodal data: To enable LLMs to understand visual information, models can be pre-trained on large datasets containing both textual and visual information. This pre-training process involves presenting the model with pairs of images and corresponding textual descriptions or labels. By processing these multimodal inputs together, the model learns to associate visual features with linguistic representations, effectively capturing the relationship between images and their textual descriptions.

Fine-tuning on specific visual tasks: After pre-training on multimodal data, LLMs can be fine-tuned on specific visual tasks, such as image classification, object detection, image captioning, or video understanding. During fine-tuning, the parameters of the model are adjusted using labeled visual data to optimize performance on the target task. This process allows LLMs to adapt their learned representations to better suit the nuances of visual processing tasks.

Integration of visual attention mechanisms: Attention mechanisms, which are central to transformer architectures, play a crucial role in processing both linguistic and visual information. These mechanisms allow the model to focus on relevant parts of the input data while generating outputs. In the context of visual processing, attention mechanisms enable the model to selectively attend to different regions of an image or video, allowing for more precise analysis and understanding.

Cross-modal interaction: LLMs facilitate cross-modal interaction, enabling information exchange between linguistic and visual modalities. This interaction allows the model to leverage the semantic context provided by textual descriptions to better understand the visual content, and vice versa. For example, when generating textual descriptions for images, the model can simultaneously incorporate visual features to ensure that the generated captions are more accurate.

Employing these technical principles, LLMs can effectively process visual information, enabling tasks such as image understanding, automatic image annotation, and video processing.

In *image understanding,* LLMs can analyze and interpret images or videos by generating descriptions, identifying objects, and inferring relationships between elements within the scene. This capability opens up new applications across industries. For instance, in e-commerce, LLMs can automatically tag products within images, enabling more efficient cataloging and search functionalities. In healthcare, they can assist in medical imaging analysis, aiding in the diagnosis of diseases or abnormalities. Moreover, in security and surveillance, LLMs can help in identifying suspicious activities or objects.

Automatic image annotation is another area where LLMs improve performance. By leveraging their understanding of language and context, they can generate descriptive captions or labels for images, significantly reducing the need for manual annotation efforts. This has profound implications for content management systems, social media platforms, and digital archives, where vast amounts of images require organization and categorization.

In *video processing*, LLMs enable advanced capabilities such as automatic video summarization and content recommendation. They can automatically parse through hours of video footage, extracting key information, and providing concise summaries. This is valuable in fields like media and entertainment, where content creators can utilize LLMs to streamline the editing process or personalize content recommendations for users.

Finally, we note that LLMs also have the potential to explain output data from a variety of other modalities, such as proprioceptive sensors in robotics, radar sensors, or other difficult to interpret sensors by creating a description of the multimodal data and eventually the underlying scene. The prompt that is typically used to interact with the LLM is then replaced by sensor data and the model needs to be re-trained. Since the amount of data available for less common types of sensor output is usually not comparable to the size of text or image-based training data sets, the use of a smaller language model, i.e., a reduced LLM, may be a more viable option. Such a procedure would also support the explainability of neural network sensor processing, especially if a human interpretable output can be produced by the LLM.

How sensor data is converted and fed into the LLM is an open problem. In the case of radar signals, one option might be the approach of a talking radar (Visnevski et al., 2007) converting pulses to letters and thus building words and phrases. A potential application might be a warning interface in a car giving a detailed description of the current situation around the car, e.g. “Attention, animals 100m ahead on the left side of the road between the trees.” A safety-critical aspect of this application scenario is the highly reliable identification of objects and scene understanding from the sensor data.

Despite their immense potential, the integration of LLMs in novel (and possibly non-linguistic) applications poses several challenges. Firstly, ensuring the ethical and responsible use of these models is paramount, particularly in sensitive domains like healthcare, cybersecurity or surveillance. Issues surrounding bias, privacy, and security need to be addressed. Additionally, LLMs require substantial computational resources and data to train effectively, limiting their accessibility to smaller organizations or educational institutions with fewer resources. Nevertheless, the range of (multimodal) applications of LLMs and generative AI has not yet been fully explored. The examples of use in vision and radar processing discussed above merely scratch the surface of the diverse spectrum of potential applications.

3 Data and Hardware Requirements

3.1 Training Data

For the training process of powerful pre-trained LLMs huge amounts of textual data resources are required. Training data for state-of-the-art LLMs (e.g. Mistral, OpenGPT-X, Poro, Llama, etc.) consists of up to 15 trillion tokens to train models of increasing size, ranging from 7 billion, 24 billion, 70 billion to more than 400 billion parameters. These tokens are pragmatic subword units which are created automatically by a tokenizer module. The selection of training data is a crucial process to achieve ethical and non-biased LLMs. The amount of training data should be huge (measured in terms of words or tokens) as more data typically leads to better model performance. It should cover different domains (such as medicine, finance, or the legal domain) to be able to support tasks in these domains. The data should comprise different types of textual content (such as dialogue, speeches, books, news articles) to learn about different types of queries and responses corresponding to styles of user interactions. Further, the data should have high quality (including aspects such as noisy content, toxic or offensive content, and duplicate content). LLM training data is derived from large text corpora, predominantly from internet crawls of web pages and publically available text resources.

In the OpenGPT-X project (funded by the German Federal Ministry for Economic Affairs and Climate Action, BMWK), for instance, text corpora are processed with a data pipeline which transforms raw data into high-quality training data, addressing the requirements mentioned above. The data comprises i) web data (originating from CommonCrawl crawled web pages) and ii) publically available curated data sets. As a first step, the data is converted into a common format, converting CommonCrawl WARC/WET files and curated datasets in a variety of different formats (e.g., CSV, plain text, SQL dumps, XML) and into JSONL files, which can be used for the training process. One line in a JSONL file corresponds to one document, stored together with document metadata.

Overall, OpenGPT-X uses more than 2.5 trillion tokens for training LLMs, consisting of roughly 80% web data and 20% curated data. The web data is processed with the *Ungoliant* pipeline (Abadji et al., 2021), which detects the content language and computes quality warnings (e.g., for harmful content). Subsequently, only documents in languages of interest are kept and documents with a quality warning are filtered out. The remaining data set is then de-duplicated, using MinHash/LSH with a Jaccard (similarity) index of 0.7 to filter out nearly identical documents. The amount of raw data is reduced by filtering by about 17% and by de-duplication by about 25%. Curated datasets are processed in a similar way: The content language is detected and documents are enriched with metadata corresponding to quality metadata. Low quality documents are filtered out. The list of curated datasets used for training of the OpenGPT-X models includes Wikimedia datasets (Wikipedia, Wikisource, Wikibooks, Wikinews, Wikivoyage), scientific articles (ArXiv, Pes2o, dissertations of the German national library), patents (from the European, World, and US patent offices), legal documents (e.g., OpenLegaldata, court decisions and laws), publications of German institutions (e.g., plenary protocols of the German Bundestag), and source code (e.g., StarCoder).

OpenGPT-X has scaled up processing from initially two languages (EN, DE) to five languages (EN, DE, FR, IT, ES), to the 24 official EU languages. Wide coverage of languages ensures a fully multilingual model for European societies. The representation of under-resourced languages, e.g., Irish/Gaelic or Maltese, and dialects is an open research issue which is addressed in several European initiatives (e.g., European Language Grid).

For multimodal foundation models additional data sets are required, e.g., huge speech corpora, video/image, or radar signal corpora. Audio, video and image data typically also includes a textual component in the form of audio transcripts or image captions. Many LLMs are also trained with code in a variety of programming languages to support the usage of LLMs for programming co-pilots and code generation. However, because of their limited reasoning capabilities, the use of LLMs for these purposes remains problematic, with known shortcomings in terms of the correctness, safety and efficiency of the programs thus obtained.

3.2 Hardware and System Concepts

The training of LLMs and corresponding large-scale inference requires a huge amount of computational resources, which nowadays are provided by dedicated multi-core systems, utilizing Graphics Processing Units (GPUs) with software environments such as CUDA to allow for highly parallelized processing with large throughput and high memory bandwidth. Especially for the training of LLMs from scratch, huge and specialized high-performance computing (HPC) centers are required, typically using 1000 and more GPUs. The latest advancements in GPU technology have significantly contributed to the feasibility of training LLMs. For instance, NVIDIA's A100 and H100 Tensor Core GPUs are designed specifically for AI and deep learning tasks. NVIDIA's Hopper architecture, featured in the H100 GPU, incorporates innovations like the Transformer Engine, which is optimized for the computations used in LLMs. Currently only a few of these HPC systems are available in Germany, like the JUWELS HPC system at Forschungszentrum Jülich including the JUWELS booster module, which contains 3744 A100 GPUs. This HPC system has been used to train the OPENGPT-X models. For successful LLM training, the parallelization of GPUs and the fast transfer of data and model parameters are important and therefore InfiniBand networking is used to guarantee low latency data transmission across GPUs. Beside Forschungszentrum Jülich, the Technical University of Dresden and the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, for instance, operate scalable HPC facilities for model training. In other European countries, EuroHPC centers, like Lumi, Leonardo, MareNostrum, and Melunxina offer large GPU capacities for the pre-training of LLMs. Obviously, these and other AI-related developments will consume not only hardware resources but also huge amounts of electric power. For instance, OpenAI recently announced plans for data centers consuming as much as 5 GW of electric power. Providing this amount of power in a carbon-free fashion in the near future will be a tremendous challenge, and will require detailed considerations of renewable as well as alternative power resources.

For fine-tuning and model inference (operation of LLMs for usage), different HPC set-ups are required. Fine-tuning of pre-trained models requires significantly less resources and can thus also be achieved outside large and centralized data centers. Here, the main challenge resides in the acquisition of sufficient amounts of task-specific data. Model inference often requires the deployment within an embedded system architecture. With the help of neural co-processing cores, small-to-midsize LLMs may be evaluated on embedded platforms. In general, AI-accelerators are currently moving also into (rechargeable) battery-operated devices, making advanced inference available in mobile devices such as smartphones and hearing aids. Nevertheless, energy-efficiency of such methods remains a pressing issue.

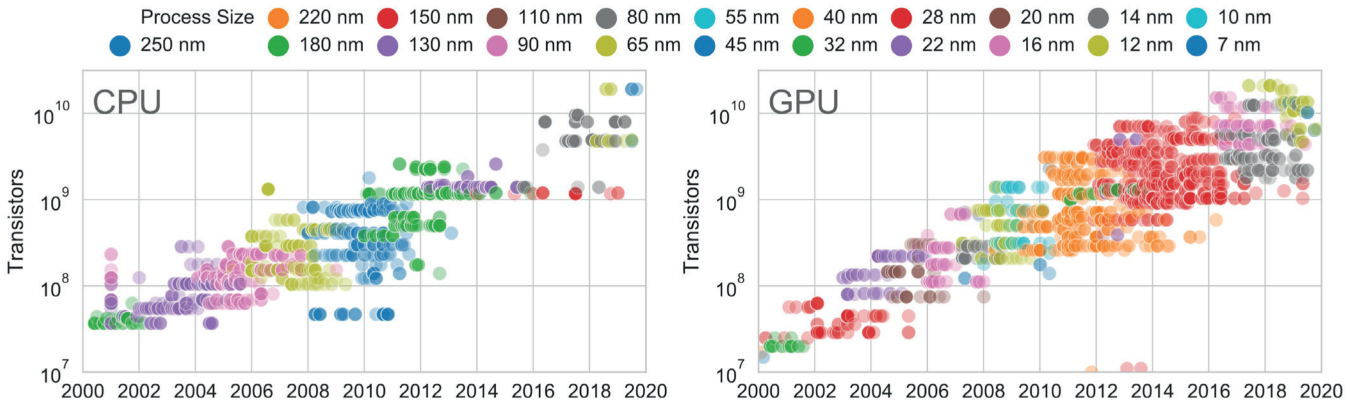
Although today's digital technology is very much optimized for low energy consumption, there is still a physical limitation as transistors generate heat with each read or write step. As large temperatures will degrade the efficiency of the device, or even damage it, heat management is an integral part of the system design at all scales, from individual dies to large GPU clusters and computing centers. Down-scaling the feature size to sub-nm regions leads to less energy storage and conduction on the gates of the semiconductor device, but also degrades the stability of binary data. To counteract this effect and to improve the transistor's physical properties, engineers have devised specific architectures for the core component, the field-effect transistor (FET). A major improvement has been the *fin field-effect transistor* (FinFET) using a multigate structure. It was transferred from research into the fabrication lines already in 2010 for a feature size down to 14nm. Using this approach, the most advanced, high-volume devices fabricated by TSMC in Taiwan have reached 3nm.

Beside down-scaling the feature size and increasing the number of transistors per chip area in accordance with Moore's law (see Figure 5), there are several new approaches to overcome the limitations of these traditional chip design methods. A selection of novel research fields which have reached different levels of technological readiness (TRL) and possess the potential to open new routes in data processing in the context of LLMs are summarized as follows:

- multilevel data storing in a single memory device as in phase-change-materials
- mixed-signal data processing circuits based on the state-of-the-art technology in CMOS integration
- integration of electrical-optical signal processing devices within system-on-chip (SoC) architectures

- partial analog signal processing with customized analog circuit devices
- neuromorphic computing approaches including novel information processing paradigms inspired by signal processing in biological systems
- quantum computing approaches on different hardware platforms, e.g., using spintronics.

All of these technologies are still under research and therefore have very different levels of maturity. One of the most prominent is the very active research field of neuromorphic computing, where various implementations in CMOS technology have been successfully demonstrated. Even if most of these implementation and integration examples are not available in the most advanced semiconductor process nodes, their functionality is scalable under defined conditions. The principle is based on the fact that with the recently introduced memristor device a new kind of information processing and data memory cell can be merged in one nm-scaleable device. One of the promising types of memristors is the resistive RAM called RRAM, used primarily for in-memory-computing. These novel electronic devices play a similar role for memory, learning and inference as neurons and synapses in biological neural systems (Schuman et al., 2022). As in nature, neural processing may be implemented in the form of spiking neural networks (SNN) and continuous reinforcement learning, whereby the latter aims to establish the defining weights for the information pathways in the network. Regarding the design of such systems there are no boundaries yet defined and therefore they are a perfect platform for fundamental research. However, the different research results and activities are in most cases not sufficiently mature for the integration in industry processes.



(a) Transistor count of CPUs.

(b) Transistor count of GPUs.

Figure 5: Summarizing CPU and GPU design trends with product data a) CPU and b) GPU transistor count vs. the years in development (Sun et al., 2020)

The motivation for industry applications is derived from the potential applications of these novel systems in consumer electronics or, more generally, in marketable products. The everlasting question is, how to transfer research into the market. Here, the standardization of engineering parameters might open a valuable pathway. The promising aspects of neuromorphic computing is to achieve a higher integration and more energy-efficient data processing and storage with a computing performance similar to the human brain. This could be especially attractive for adaptive on-the-fly training of neural networks, and eventually LLMs. As this development still requires major research and development efforts, the Information Technology Society (ITG) in the Association for Electrical, Electronic & Information Technologies (VDE) is currently working on a harmonization of common terminology and a layer model from defined application requirements, algorithms, over hardware systems, circuit architectures, chip design, devices and material choices. This document is titled “Neuromorphic computing based on new devices – a layer model for developing AI hardware”, and is referenced as VDE SPEC 90033 V1.0 (en). The intention of this document is to advance this technology and to specify uniform definitions in this field of research and development, as well as developing a coordinated joint understanding of its pros and cons.

The second field that deserves a closer look is *quantum computing*. Quantum natural language processing (QNLP) is an emerging field that applies quantum computing principles to enhance natural language processing. Companies like IBM, Microsoft, and Quantinuum are developing QNLP architectures to process language in quantum circuits (Lorenz et al., 2023). Quantum computers offer theoretical advantages in certain types of computations, particularly those that require high parallelism or the ability to exploit quantum entanglement and superposition. Quantum computers have the potential to solve problems significantly faster than would be possible on classical computers, especially for problems that would take exponentially longer on classical systems. This could be particularly useful for complex NLP tasks such as semantic analysis or machine translation. While quantum systems promise faster solutions for complex NLP tasks like semantic analysis, they currently face challenges, such as the need for extensive cooling and scaling for large-scale applications. Quantum algorithms specific to QNLP are still in the research and development phase and need further optimization and adaptation to provide realistic advantages over classical systems.

4 Applications, Services, and Business Potential

4.1 Applications

LLMs are an enabling technology for a variety of applications. In what follows, we will briefly summarize the most prominent areas.

Machine translation: As LLMs have emerged from natural language processing (NLP) tasks, it is obvious that natural language applications are amongst the most important commercial use cases. One first such task is machine translation. Although the majority of large-scale systems still rely on traditional task-specific encoder-decoder models, some commercial products start to utilize LLMs for text-to-text translation, and achieve an acceptable quality for many tasks, such as book translations, translations of social media texts, and websites (Gao et al., 2024). The quality is comparatively lower for speech-to-speech translation tools, but first commercial products already enter the market, e.g., on mobile phones, and also mixed forms (e.g., speech-to-text translation in live dubbing) may result in valuable applications. Translation usually comes to its limits when low-resource languages are addressed, and even for the German language there are far less models available than for widespread languages such as English. Providing LLMs for a large set of (low-resource) languages is a primary task for language equality in Europe, and should be supported on a national level via funded projects to set up competitive open-source models (Robinson et al., 2023). This also relates to language diversity within a language, such as easy language ("leichte Sprache") which is legally required to improve accessibility, but where comprehensive resources are more-or-less completely missing (Freyer et al., 2024). Another example is speech-to-sign-language translation. As tools for such languages address a smaller portion of the population, they will be commercially less viable and thus require public funding.

Customer chat tools: A second popular natural language task for LLMs is customer chat. Chatbots offering customer care services are becoming increasingly prevalent across a diverse range of domains, and the larger the domain is, the more interesting it gets to use LLMs. LLMs may be used in the front end (e.g., for natural language understanding, domain classification, call redirection), in dialog management (selecting the next turn), in response generation, as well as for backend tools such as question answering and summarization for call-center agents. In such situations, LLMs may improve performance of the individual tasks, and may cater for a larger variety of users, including different languages. In turn, LLMs might be less appropriate when information for processing tools with defined and limited functionalities is to be collected; in such use cases, well-structured dialogues following fixed grammars may provide better results (Waisberg et al., 2024).

Computer programming assistance: Language-related applications which go beyond natural language include LLM-based programming tools. Such tools may facilitate the programming process by co-programming, i.e., an LLM-based wizard providing a raw code with basic functionality, which then can be improved by a human programmer. They may also be used for source code conversion, i.e., from one programming language to another one, or for data visualization and understanding. While such tools have already increased the productivity of software developers significantly, it is important to note that the human programmer is still responsible for the correctness and consequences arising from the code; thus, appropriate validation steps need to be foreseen in the code generation process (Jury et al., 2024; Gabbay and Cohen, 2024).

Knowledge representations: As LLMs so far are LMs learning the structure of text and the implicit connections, there is actually still a gap of explicit knowledge representation (d'Amato et al., 2023). But efforts are made to combine LLMs with knowledge graphs to enable knowledge management, contextual analysis, and collaborative information sharing (Pan et al., 2024). This combination would help to convert unstructured data to structured graphs and thereby gain a deeper understanding of the initial data, also in terms of identifying flaws or open issues.

Social media analysis and the generation and detection of fake news: The application of LLMs in social media spans various functionalities, significantly enhancing user experiences and content management. LLMs are pivotal in facilitating co-writing skills. Additionally, LLMs empower efficient search functionalities for users that do not know or have difficulties in defining exact and suitable search terms. LLMs' translation capabilities further break down language barriers, fostering global connectivity and communication. Sentiment analysis is another vital aspect wherein LLMs discern and analyze user emotions expressed in social media content, offering valuable insights into public opinion trends (Ampel et al., 2024). Furthermore, as LLMs enable text-style conversion while preserving information and context, they can be utilized as an anonymization method to ensure user privacy and security, safeguarding personal data in the digital realm (Sinha et al., 2024). However, the widespread integration of LLMs in social media also raises consequential considerations. Currently, there is a contentious discourse surrounding the role LLMs could play in the generation of fake news. However, LLMs can equally be utilized in the detection of fake news, aiding in the preservation of digital integrity and the dissemination of accurate information.

Public administration: Another use case of LLMs can be found in politics and public administration. Here, LLMs can leverage their previously described strengths to provide customized chat services and elucidate processes in various languages including simplified language, as well as perform text-related tasks such as summarizations, key wording, cross-referencing (Cheong et al., 2024; Musumeci et al., 2024).

4.2 Performance Evaluation and Legal Aspects of Deployment

Whereas most current assessments of LLMs focus on the performance of the models to achieve certain tasks (Yang et al., 2024), less effort has been spent on the subjective evaluation of LLM-based tools, including the perceived quality of LLM-generated texts. Performance indicators include the accuracy of LLMs to answer questions, extract information, identify disinformation or sentiments, and corresponding task-based metrics serve the comparative assessment of different tools fulfilling such tasks (LLM-based or not). In turn, perceived quality requires users to rate the quality of generated text, or of interactions held with LLM-based chatbots. Such subjective evaluations should cover both, the content of the generated output such as correctness and avoiding hallucinations, and the form such as grammaticality, coherence, transparency, or complexity. Corresponding methods are frequently developed ad-hoc, and standards to make such evaluations comprehensive and comparable still need to be developed.

Thus, LLMs contribute to barrier-free access, facilitating inclusive communication and information dissemination. However, their widespread adoption mandates robust regulation on open-data usage to mitigate potential risks and ensure ethical deployment, including intellectual property (IP) remuneration and licensing models, and responsibility for correctness and liability. This applies especially to the areas of data sovereignty and sensitive data, highlighting the necessity for an essential aspect: the establishment and operation of extensive European hosted server capacity, combined with the development of skills within companies, enabling them to train or fine-tune decentralized models for privacy-sensitive or IP-sensitive models.

4.3 LLMs in Education

As mentioned before, comprehensive user education programs become imperative to promote digital literacy and responsible engagement with LLM-powered features. It is short-sighted to offer user training solely within the professional context; it is crucial to embed this knowledge as early as possible in school education, both as tools and subjects of study. This is usually denoted by the term GenAI readiness (Dettmer et al., 2024). An important skill in utilizing LLMs pertains to prompt engineering, i.e., the successive refinements of the instructions given to an LLM in order to provide the desired output. The output of an LLM may vary significantly with each (additional) prompt and may include incorrect responses (also known as "hallucinations") or biases. Therefore, the user must be educated to craft appropriate prompt sequences and empowered to assess the validity of the answers given by an LLM.

In an educational context, LLMs can then be used to facilitate text generation, quiz/examination creation, educational dialogues, and even hypothesis testing, argument generation and texting. However, their integration into the educational setting implies significant consequences and requirements. Firstly, the type of examination has to change from knowledge repetition and checking the final results, moving towards accompanying and querying the learning process and explaining correlations. Further, emphasizing examinations that highlight original and personal work, fostering critical thinking and creativity is essential. Additionally, educator training must incorporate fundamentals of LLMs, ensuring teachers are equipped to leverage these tools effectively (Adarkwah, 2024). Moreover, equitable access to LLM services for all educators and students is paramount. Also, education on intellectual property rights (IPRs) becomes imperative in light of the prolific use of LLMs. Thus, embracing LLMs in education requires a comprehensive approach that addresses both pedagogical and ethical considerations.

5 Seizing Opportunities, Addressing Challenges, and Mitigating Risks

The recent developments of AI based on foundation models have been surprising even for experts that have long been in the field. Although the developments are somewhat more predictable and interpretable for experts, it is not surprising that regulatory bodies face many challenges in keeping pace with these developments and creating reliable guardrails for the technology while supporting innovation.

While consultants and economists have tried to foresee the opportunities of LLMs and even to estimate their economic potential, it will be difficult to provide reliable figures. It may even be the case that developments will follow Amara's Law, so that we tend to overestimate the speed of developments in the short term and underestimate the long-term impact today. Indeed, after a substantial hype surrounding generative models, some disillusion is also setting in: generative models are becoming ever more expensive to train, are exceedingly data hungry – leading to law-suits and debates around fair use of human-written text and human art without licensing – and they may not always lead to the strong business models that were originally envisioned. Nevertheless, LLMs have triggered a large number of new ideas, developments and applications, some of which are outlined in Section 4. Sustained investments are now necessary to harness their added value, to create profitable business models, and to further develop LLMs towards intelligent agents in a variety of contexts.

While there are encouraging examples of how large-scale computational resources and shared data repositories can be developed in Germany (e.g., project nxtAIM, creating foundation models for autonomous driving), small and medium enterprises (SMEs) still face substantial challenges when adopting LLMs and tailoring them to their business needs. Addressing these challenges requires strategic planning, and a careful management of investments in skills and infrastructure. On the one hand, the costs associated with acquiring high-performance GPUs, data storage, and the energy consumption needed for training and maintaining large models present a significant financial burden. On the other hand, dependencies on third-party providers for LLM solutions limit flexibility and increase long-term costs. While the cost of training an LLM from scratch may be prohibitive, added value can also be developed by fine-tuning existing models, a route that is more viable for SMEs and academic institutions. Furthermore, the use of LLMs may raise significant privacy and security concerns, particularly in highly regulated industries such as finance, healthcare, and legal services. Moreover, SMEs must protect their own sensitive information from breaches. Compliance with industry-specific regulations, standards, and best practices can add an extra layer of complexity to the deployment and use of LLMs.

It is important to note that this technological window of opportunity coincides with a significant loss in workforce due to retirement of the "baby boomers". This has already begun to impact Germany. It is absolutely clear that we will be more dependent on technology in the future to ensure the functioning of public administration, mobility and logistics, the medical system, production, commerce, farming, and many other areas. LLMs require specialized knowledge to deploy and fine-tune effectively. Especially SMEs may lack the in-house expertise needed to manage these sophisticated technologies and hiring or training staff with the requisite skills in machine learning, data science, and software engineering is challenging and costly. This expertise gap can hinder the effective implementation and optimization of LLMs. Nevertheless, it would be unwise to not attempt to establish technological sovereignty in this field.

5.1 Scientific Challenges and Research Gaps

LLMs also pose significant scientific challenges for research and development of products under regulatory demands, especially in the wider socio-technological contexts and regarding AI trustworthiness (European Commission, 2019). The more technically focused scientific challenges, as described in previous sections, are largely due to the frontier nature of LLMs and foundation models. These encompass questions of model robustness and reliability, limits in computing resources, energy efficiency of model training and inference, availability of data, developing multimodal capabilities, or the reproducibility of scientific results (Maslej et al., 2024; OECD, 2023). However, even of more public significance are connected challenges of broader socio-technological nature, amplified by the fast technological advancement of AI models (Dahlin, 2021). These challenges and related opportunities are the subject of an ongoing and important societal debate, see (Simon et al., 2024) for a recent discussion, and they are reflected in major research gaps as follows:

Assessing trustworthiness of AI systems, especially concerning factual accuracy, bias, and transparency is essential. Ensuring that the outputs of AI systems, and in particular of LLMs, are reliable and grounded in factual information is a major challenge. Even though some LLMs are able to reach impressive performance on benchmarks, they struggle with complex reasoning, providing consistent factual answers, and explaining their decisions, which affects their overall trustworthiness and quality (Maslej et al., 2024). Moreover, the inability to properly manage misinformation in the use of LLMs and its potential impact on democratic processes (e.g., elections) is a major concern that requires further in-depth research attention. Additionally, there is a lack of efficient and effective international standards on AI trustworthiness and AI quality, particularly regarding privacy, accountability and ethical use of AI, underscoring the need for deeper research into AI policy and governance structures. To develop effective and efficient international standards, independent tests encompassing real live environments would be useful. This research challenge is not just technical, but also involves legal, regulatory and societal dimensions (Bourg et al., 2024; Bareis, 2024; Orwat et al., 2022; Folberth, 2022).

Ensuring bias and fairness across diverse datasets and applications is an unresolved challenge. There is a need for more advanced methodologies to detect and correct biases in both the data and the model training process. Generative models amplify existing biases present in the training data and may perpetuate harmful stereotypes or influence decision-making processes in critical areas like healthcare, finance and human resources. Finding solutions addressing this gap is critical, ensuring trust and quality in the ability of AI to operate without perpetuating social inequalities. The results would be very useful for many applications including the finance and insurance sector (Agu et al., 2024).

Addressing data quality and transparency, especially concerning insufficient transparency in how the datasets are sourced. Non-transparent data collection may lead to contamination with unreliable or biased data and potential copyright issues (Hardinges, 2024). Furthermore, the performance of generative models may degrade over time, especially when training on their own outputs, a phenomenon known as “model collapse” (Alemohammad, 2023). Addressing these challenges requires transparent data, effective data governance, transparent data quality processes, improved curation of data, and transparent documentation of training datasets.

Explainable AI (XAI) for ensuring trustworthiness and quality, particularly in high-risk and highly regulated areas like healthcare or legal systems. Despite some progress in explainability and interpretability techniques, such as Saliency Maps, LIME, and SHAP, there remains a significant research gap on making complex models more interpretable and explainable to both developers and end-users (Mane, 2024; Salih et al., 2024), especially in the context of large generative models, see for example (Saranya and Subashini, 2023; Zhao et al., 2024).

Ensuring security and adversarial robustness in unpredictable or adverse environments and in the presence of unexpected or malicious inputs or adversarial attacks is particularly crucial when AI systems operate in real-world applications where the data they encounter may not match their training sets, where cybersecurity risks are high, and where misuse of AI systems such as LLMs may have a large impact (Maslej et al., 2024). Resources such as the ENISA multilayer framework for good cybersecurity practices in AI[1] or the NIST Taxonomy and Terminology of Attacks and Mitigations[2] are already well elaborated frameworks which need further developments and more implementation practice. Addressing the additional AI-related cybersecurity risks and vulnerabilities in AI systems is essential for the security and safety of AI applications.

Assessing the safety of AI systems, particularly related to how AI models can be systematically tested for robustness and safety in real-world applications. Stronger benchmarks and frameworks to assess AI performance, particularly in dynamic, high-stakes environments like healthcare and autonomous systems, are being needed (Mirzarazi et al., 2024; Alenjareghi et al., 2024).

[1] See Multilayer Framework for Good Cybersecurity Practices for AI – ENISA

[2] See AI 100-2 E2023, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations | CSRC

Automatic, continuous and exhaustive AI software testing currently relies largely on snapshots whose methodologies still involve substantial manual testing and is limited by a lack of tools and methods for proper verification. To ensure quality of testing, continuous, autonomous, independent and verified testing results are essential. AI system provider, deployer and user that use products which fall under the high risk class according to the EU AI Act would greatly benefit from permanent, automated independent software testing.

Overall, these challenges are not unlike those encountered in AI research in general, but are exacerbated by the complexity of LLMs and the limited transparent access to typical foundation models such as GPT-4, see (Maslej et al., 2024) for a general overview.

5.2 Risk Mitigation through Agile Regulation

Addressing these challenges and risks increasingly becomes the subject of regulatory demands and the need for compliance testing and product certification schemes. Still, as with any other technology, we advise not to take current limitations of systems in one or several of these points as "show-stoppers", but rather as demands that can be met by further research, guard-railed by societal monitoring and regulation. The final version of the EU AI Act makes it clear that AI systems – including LLMs – are mostly exempt from regulation when only developed for purposes of research, thus ensuring that future research can rise to the challenges. This will facilitate the testing of potential regulations in actual operational contexts, allowing for an assessment of their suitability in achieving the defined goals before their issuance[3].

European AI Regulation largely aims at mitigation of risk associated with those challenges that are deemed relevant from a societal point of view. Within the EU, good examples with relevance for LLMs are regulations such as the General Data Protection Regulation (GDPR), the AI Act, and the Digital Services Act (DSA). For instance, in the AI Act, considered risks are those to the fundamental rights, health and safety of human beings, which, by and large, connects to the scientific challenges listed above.

The risk-based approach in the AI Act (European Parliament & Council, 2024) defines different sets of requirements and obligations, which providers and deployers of AI systems need to follow, with most requirements focused on high-risk AI systems. There are specific definitions[4] on which systems would be considered high-risk, depending on their field of application (e.g., in biometrics or critical infrastructure) and the specifics of their use, e.g., if the AI system is part of a safety component of a product. Additional guidance is provided for the self-conducted risk assessment of AI systems by providers and deployers, and beyond the clearly defined classes of high risk AI systems, the categorization of risks can still be interpreted flexibly, depending on the specific needs of the providers or deployers. Highly relevant for LLMs, there are additional specific requirements for general purpose AI (GPAI) models[5], and even more strict requirements for the largest GPAI models identified as of "systemic risk". LLMs and, more generally, large generative AI models (LGAIMs) fall under the category of GPAI models, with the largest current LLMs being already considered of systemic risk[6]. These requirements always need to be fulfilled, regardless of the risk class category of the overall AI system. The regulation of GPAI models has been added to the AI Act as a direct response to the GPT models taking the discourse of LLMs to the forefront in society.

Requirements and obligations in the AI Act directly aim at a wide range of risks connected to the discussed scientific challenges and can largely be considered an implementation of the concepts of trustworthy AI into requirements, ranging from transparency and human oversight to the completeness of training data, robustness and security of AI systems. Since the AI Act follows the new legislative framework (NLT) on product regulation[7], by design, it only sets down requirements around high-level concepts and explicitly aims at developing more detailed technical guidance in standardization together with the European standardization bodies, most notably CEN-CENELEC. Currently, the joint tech-

[3] For a recent example, see the foreseen use of regulatory sandboxes in the AI Act, in Chapter VI, Art.57, AI Act.

[4] See Chapter 3, Article 6, AI Act and Annex III AI Act for the definitions.

[5] Defined in Chapter 2, Article 3 AI Act as: "general-purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market"

[6] See for example https://epochai.org/blog/tracking-large-scale-ai-models

[7] See https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en

nical committee 21 (JTC21)[8] is developing a set of standards for the high-risk requirements that are planned to be harmonized throughout the EU, which will come with a presumption of compliance to the AI Act. VDE and DKE (German Commission for Electrotechnical, Electronic, and Information Technologies of DIN and VDE) play a leading role in these efforts. In addition, the European Commission has recently started a process to develop a Code of Practice[9] as guidance for the specific obligations to GPAI models, including LLMs, which involves a process including a large list of stakeholders from academia, industry and civil society.

A shortcoming of this approach to legislation is the currently limited transferability of the obligations themselves to concrete technical parameters without the guidance of additional standards or other frameworks, with harmonized standards not foreseen to be published before the end of 2025. For example, the transparency obligation outlined for high-risk AI systems in Chapter 2 of the AI Act considered separately, falls short in providing an adequate level of detail for practitioners to implement this obligation adequately, such that actual transparency can be provided to users with different levels of expertise, domain knowledge and backgrounds (Helberger and Diakopoulos, 2023). Another example would be the obligation of error-free and unbiased datasets for training high-risk AI systems, which cannot be achieved without detailing how to overcome the representative sample and bias tradeoff. Here, the challenge is to balance the need for a representative sample in training data (to capture diverse real-world scenarios) with the risk of introducing biases, as achieving full representativeness can inadvertently amplify certain biases or introduce new ones. Overall, because of the NLT approach, the obligations alone, as defined in the AI Act, lack concreteness, remain opaque and are not enough to guide developers in practical matters, especially with regard to GPAI models. Also, some AI process and lifecycle stages as well as some sectoral applications are not covered in any detail in the AI Act alone and additionally fall under different regulatory measures[10]. A second regulatory example for the challenges of policy making to keep pace with developments in LLMs is the DSA, which was crafted when LGAIMs were not present in the public discourse (Hacker et al., 2023). The new regulations for intermediaries (e.g., major social media platforms) set out the procedures for dealing with harmful content and disinformation shared on their platforms. The DSA, while comprehensive in regulating online platforms and content moderation, falls short in addressing the unique challenges posed by LGAIMs in disinformation generation. The DSA primarily targets platforms but does not directly regulate the AI models themselves, leaving gaps in managing the source of disinformation. LGAIMs can create content at massive scales and high sophistication, overwhelming current moderation systems. Furthermore, the DSA's focus on reactive measures and algorithmic transparency is insufficient for proactively combating AI-generated disinformation. Coordination with AI-specific regulations like the AI Act is necessary to effectively address these emerging challenges.

[8] See https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/

[9] See https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice

[10] For example regarding the question of how to translate a research model later into a product, for medical products containing AI, which also fall under the medical devices regulation, or with regards to cybersecurity, where a high risk AI system may need to follow the Cyber Resilience Act as well.

6 Recommended Actions

AI technologies already permeate all areas of life and all sectors of the economy. After the ChatGPT moment in November 2022, AI has been leveraged across our society and many business areas. AI, especially generative AI (GenAI) based on large foundation models, will continue to have a profound impact on our society and economic ecosystems across all domains. Several professions and job profiles will change dramatically, and we are still at the beginning of this new technological era. It is essential that we dedicate our full attention and focus to AI in order to contribute its continued development and integration into industry and academia, and to assume a leading role in this field. It also implies that political, technical and administrative organizations, users and the entire society should be aware of the forthcoming opportunities and changes driven by AI technologies. The following areas should be addressed to ensure that we achieve a maximum impact and high degree of technology sovereignty in this field:

- Resources for Research: AI and particularly future generations of LLMs require additional research efforts, concerning both theoretical and practical aspects. The list of research topics is extensive, including factual correctness, reasoning abilities, removing bias, following ethical values, energy efficiency, memory consumption both in fine-tuning and inference, alternative model architectures beyond the transformer, data-efficient processing to avoid intellectual property issues and toxicity, multimodal approaches to combine different input channels (speech, audio, images, video, text, etc.), and multilingual approaches. These research activities are directed toward novel model architectures and even stronger large foundation models. To ensure the ability to perform significant research work, it is necessary to establish the skills, infrastructure and research networks to train large foundation models from scratch and to democratize fine-tuning and inference towards smaller organizations and platforms. Additionally, many other research fields (e.g., social sciences, chemistry and biological sciences) will undergo a radical transformation as a result of the utilization of AI. For all upcoming research on AI or applying AI in research, substantial resources and skilled researchers are needed. This requires sustained public funding on both the German and European level and improved access to venture capital for research-intensive startup companies.

- Transfer of AI technologies: Germany is a world leader in many areas of science and technology, and its industry and research institutions are well positioned to capitalize on the current momentum of AI developments. The full benefit of AI can be achieved through a trustful and broad transfer of AI research and innovative technologies into the industrial context and real-world applications. Therefore, existing solutions have to be enriched with AI functionalities, like chatbots, retrieval-augmented generation (RAG) systems, machine translation, and upcoming agentic AI architectures. The adaptation (e.g., fine-tuning) and tailoring of large foundation models require sufficient access to scalable HPC infrastructure. For the successful transfer of AI technologies, access to AI models and the availability of skilled experts in this area are essential. Here we need strong transfer mechanisms on both the national and European level, including the definition of important projects of common European interest (IPCEI) in areas with advanced levels of technological readiness.

- Trustworthy AI, Ethical and Social Aspects: The enormous power of AI needs regulation to ensure the fair and ethical usage of this new technology. The European AI Act and standardization efforts (including activities of VDE, see Section 5) are important first steps to ensure the trustworthiness of AI in society. Developing techniques to identify and mitigate bias, create safe and secure models, ensuring transparency and accountability in AI systems, and promoting sustainable AI practices are crucial steps. Regulatory frameworks and industry standards must evolve rapidly to keep pace with these advancements, ensuring that the deployment of LLMs benefits society while minimizing risks. Here we are also at the beginning of a process that needs more effort and networked activities, for instance, in providing more concrete guidance in translating the vaguely defined obligations into technical parameters. VDE and other partners are already well positioned to push these activities further. Other societal and economic aspects have to be considered as well, like energy efficiency of large foundation models, or the transparency and acceptance of AI in our society.

In summary, AI requires a collaborative effort between research institutions, industry (SMEs and large companies alike) as well as the public sector, and government. The success of new AI technologies, including LLMs, will only be achieved if society is willing to support, use, finance and secure this new generation of technology. It offers an enormous opportunity to address upcoming societal challenges. It is therefore essential that we utilize it responsibly. Technology will be the basis of our society and economy more than ever.

7 References

Abadji, J., Ortiz Suárez, P. J., Romary, L., & Sagot, B. (2021). Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9), Limerick, 12 July 2021 (Online Event).* https://doi.org/10.14618/IDS-PUB-10468

Adarkwah, M. A. (2024). GenAI-Infused Adult Learning in the Digital Era: A Conceptual Framework for Higher Education. *Adult Learning*, Article 10451595241271161. Advance online publication. https://doi.org/10.1177/10451595241271161

Agu, E. E., Abhulimen, A. O., Obiki-Osafiele, A. N., Osundare, O. S., Adeniran, I. A., & Efunniyi, C. P. (2024). Discussing ethical considerations and solutions for ensuring fairness in AI-driven financial services. *International Journal of Frontline Research in Multidisciplinary Studies, 3*(2), 1–9. https://doi.org/10.56355/ijfrms.2024.3.2.0024

Alemohammad, S., Humayun, A. I., Agarwal, S., Collomosse, J., & Baraniuk, R. (2024, August 29). Self-Improving Diffusion Models with Synthetic Data. https://doi.org/10.48550/arXiv.2408.16333

Alenjareghi, M. J., Keivanpour, S., Chinniah, Y. A., Jocelyn, S., & Oulmane, A. (2024). Safe human-robot collaboration: a systematic review of risk assessment methods with AI integration and standardization considerations. *The International Journal of Advanced Manufacturing Technology, 133*(9-10), 4077–4110. https://doi.org/10.1007/s00170-024-13948-3

Ampel, B., Yang, C.-H., Hu, J., & Chen, H. (2024). Large Language Models for Conducting Advanced Text Analytics Information Systems Research. *ACM Transactions on Management Information Systems*, Article 3682069. Advance online publication. https://doi.org/10.1145/3682069

Bareis, J. (2024). The trustification of AI. Disclosing the bridging pillars that tie trust and AI together. *Big Data & Society, 11(*2), Article 20539517241249430. https://doi.org/10.1177/20539517241249430

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). xLSTM: Extended Long Short-Term Memory. https://doi.org/10.48550/arXiv.2405.04517

Bourg, C., Kriegsman, S., Lindsay, N., Sardis, H., Stalberg, E., & Altman, M. (2024). Generative AI for Trustworthy, Open, and Equitable Scholarship. *An MIT Exploration of Generative AI.* Advance online publication. https://doi.org/10.21428/e4baedd9.567bfd15

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. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cheong, I., Xia, K., Feng, K. J. K., Chen, Q. Z., & Zhang, A. X. (2024). (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2454–2469). ACM. https://doi.org/10.1145/3630106.3659048

Dahlin, E. (2021). Mind the gap! On the future of AI research. *Humanities and Social Sciences Communications, 8*(1). https://doi.org/10.1057/s41599-021-00750-9

d'Amato, C., Mahon, L., Monnin, P., & Stamou, G. (2023). Machine Learning and Knowledge Graphs: Existing Gaps and Future Research Challenges. *Transactions on Graph Data and Knowledge (TGDK), 1(*1), 8:1-8:35. https://doi.org/10.4230/TGDK.1.1.8

Demandsage (2024), https://www.demandsage.com/chatgpt-statistics/, last accessed on October 16, 2024.

Desislavov, R., Martínez-Plumed, F., & Hernández-Orallo, J. (2023). Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems, 38*, 100857. https://doi.org/10.1016/j.suscom.2023.100857

Dettmer, S., Eisenbardt, M., Eisenbardt, T., Gafni, R., Gal, E. Kurtz, G., Leiba, M., Mullins, R., & Siegert I. (2024). Students' readiness to adopt GenAI in their learning and ethical considerations – An international comparative study. in *Refereed Extended Abstract Proceedings - KM Conference.*

European Commission: Directorate-General for Communications Networks, Content and Technology (2019). Ethics guidelines for trustworthy AI. Publications Office. https://data.europa.eu/doi/10.2759/346720

European Parliament and Council (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. REGULATION (EU) 2024/1689.

Folberth, A., Jahnel, J., Bareis, J., Orwat, C., & Wadepuhl, C. (2022). Tackling problems, harvesting benefits: A systematic review of the regulatory debate around AI. *KIT Scientific Working Papers, 197,* Karlsruher Institut für Technologie (KIT).

Freyer, N., Kempt, H., & Klöser, L. (2024). Easy-read and large language models: on the ethical dimensions of LLM-based text simplification. Ethics and Information Technology, 26(3), Article 50. https://doi.org/10.1007/s10676-024-09792-4

Gabbay, H., & Cohen, A. (2024). Combining LLM-Generated and Test-Based Feedback in a MOOC for Programming. In D. Joyner, M. K. Kim, X. Wang, & M. Xia (Eds.), *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 177–187). ACM. https://doi.org/10.1145/3657604.3662040

Gao, D., Chen, K., Chen, B., Dai, H., Jin, L., Jiang, W., Ning, W., Yu, S., Xuan, Q., Cai, X., Yang, L., & Wang, Z. (2024). LLMs-based machine translation for E-commerce. *Expert Systems with Applications, 258*, 125087. https://doi.org/10.1016/j.eswa.2024.125087

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks, https://arxiv.org/abs/1406.2661

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*. https://doi.org/10.48550/arXiv.2312.00752

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. In *2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1112–1123). ACM. https://doi.org/10.1145/3593013.3594067

Hardinges, J., Simperl, E., & Shadbolt, N. (2023). Special Issue 5: Grappling With the Generative AI Revolution. *Harvard Data Science Review.* Advance online publication. https://doi.org/10.1162/99608f92.a50ec6e6

Helberger, N., & Diakopoulos, N. (2023). The European AI Act and How It Matters for Research into AI in Media and Journalism. *Digital Journalism, 11*(9), 1751–1760. https://doi.org/10.1080/21670811.2022.2082505

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jury, B., Lorusso, A., Leinonen, J., Denny, P., & Luxton-Reilly, A. (2024). Evaluating LLM-generated Worked Examples in an Introductory Programming Course. In N. Herbert & C. Seton (Eds.), *Proceedings of the 26th Australasian Computing Education Conference* (pp. 77–86). ACM. https://doi.org/10.1145/3636243.3636252

LEAM:AI (2023). Large AI Models for Germany. German AI Association. *Feasibility study on behalf of the Federal Ministry for Economic Affairs and Climate Action (BMWK),* https://leam.ai/feasibility-study-leam-2023/, last accessed on October 16, 2024.

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating Wikipedia by Summarizing Long Sequences. https://doi.org/10.48550/arXiv.1801.10198

Lorenz, R., Pearson, A., Meichanetzidis, K., Kartsaklis, D., & Coecke, B. (2023). QNLP in Practice: Running Compositional Models of Meaning on a Quantum Computer. *Journal of Artificial Intelligence Research, 76*, 1305–1342. https://doi.org/10.1613/jair.1.14329

Mane, D., Magar, A., Khode, O., Koli, S., Bhat, K., & Korade, P. (2024). Unlocking Machine Learning Model Decisions: A Comparative Analysis of LIME and SHAP for Enhanced Interpretability. *Journal of Electrical Systems, 20*(2s), 598–613. https://doi.org/10.52783/jes.1480

Mirzarazi, F., Danishvar, S., & Mousavi, A. (2024). The Safety Risks of AI-Driven Solutions in Autonomous Road Vehicles. *World Electric Vehicle Journal, 15*(10), 438. https://doi.org/10.3390/wevj15100438

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). The AI Index 2024 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.

Musumeci, E., Brienza, M., Suriani, V., Nardi, D., & Bloisi, D. D. (2024). LLM Based Multi-agent Generation of Semi-structured Documents from Semantic Templates in the Public Administration Domain. In H. Degen & S. Ntoa (Eds.), *Lecture Notes in Computer Science. Artificial Intelligence in HCI* (Vol. 14736, pp. 98–117). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60615-1_7

OECD (2023). Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research. OECD Publishing, Paris.

OpenAI, Achiam, J., Adler, S., Agarwal, S [Sandhini], Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., . . . Zoph, B. (2023). *GPT-4 Technical Report.* https://doi.org/10.48550/arXiv.2303.08774

Orwat, C., Bareis, J., Folberth, A., Jahnel, J., & Wadephul, C. (2022). Risikoregulierung von künstlicher Intelligenz und automatisierten Entscheidungen. In T. Hoeren & Pinelli Stefan (Eds.), *Schriftenreihe Information und Recht: Band 87. Künstliche Intelligenz: Ethik und Recht.* C.H. Beck.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering, 36*(7), 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *Technical report*, OpenAI.

Robinson, N., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), *Proceedings of the Eighth Conference on Machine Translation* (pp. 392–418). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.40

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Advanced Intelligent Systems*, Article 2400304. Advance online publication. https://doi.org/10.1002/aisy.202400304

Saranya, A., & Subashini, R. (2023). A Systematic Review of Explainable Artificial Intelligence Models and Applications: Recent developments and future trends. *Decision Analytics Journal, 7*, 100230. https://doi.org/10.1016/j.dajour.2023.100230

Schuman, C. D., Kulkarni, S. R., Parsa, M., Mitchell, J. P., Date, P., & Kay, B. (2022). Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science* 2, 10–19. https://doi.org/10.1038/s43588-021-00184-y

Simon, J., Spiecker gen. Döhmann, I., & von Luxburg, U. (2024). Generative KI – jenseits von Euphorie und einfachen Lösungen. *Diskussion Nr. 34,* Nationale Akademie der Wissenschaften Leopoldina.

Sinha, Y., Raivakhovskyi, M., Schubert, M., & Siegert, I. (2024). Safeguarding Speech Content Style: Enhancing Privacy Beyond Speaker Identity. In *4th Symposium on Security and Privacy in Speech Communication* (pp. 92–101). ISCA. https://doi.org/10.21437/SPSC.2024-16

Sun, Y., Agostini, N. B., Dong, S., & Kaeli, D. (2020). Summarizing CPU and GPU Design Trends with Product Data, https://arxiv.org/abs/1911.11313

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)* Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Visnevski, N., Krishnamurthy, V., Wang, A., & Haykin, S. (2007). Syntactic Modeling and Signal Processing of Multifunction Radars: A Stochastic Context-Free Grammar Approach. *Proceedings of the IEEE, 95*(5), 1000–1025. https://doi.org/10.1109/JPROC.2007.893252

Waisberg, E., Ong, J., Masalkhi, M., & Lee, A. G. (2024). Large language model (LLM)-driven chatbots for neuro-ophthalmic medical education. *Eye, 38*(4), 639–641. https://doi.org/10.1038/s41433-023-02759-7

Yang, D., Chen, F., & Fang, H. (2024). Behavior Alignment: A New Perspective of Evaluating LLM-based Conversational Recommendation Systems. In G. Hui Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, & Y. Zhang (Eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2286–2290). ACM. https://doi.org/10.1145/3626772.3657924

Zhao, H., Chen, H [Hanjie], Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for Large Language Models: A Survey. *ACM Transactions on Intelligent Systems and Technology, 15*(2), 1–38. https://doi.org/10.1145/3639372

Note: Language and style of this document were refined in parts with the help of AI tools based on large language models.

Abbreviations

AED	attention-based encoder decoder
AI	artificial intelligence
ASR	automatic speech recognition
DKE	Deutsche Kommission Elektrotechnik, Elektronik, Informationstechnik
DNN	deep neural network
DSA	Digital Services Act
GenAI	generative AI
GPU	graphics processing unit
HPC	high-performance computing
IPR	intellectual property right
LGAIM	large generative AI model
LLM	large language model
LM	language model
LSTM	long short-term memory
RLHF	reinforcement learning from human feedback
SME	Small and medium enterprise
TPU	tensor processing unit
XAI	explainable artificial intelligence

About VDE

VDE, one of the largest technology organizations in Europe, has been regarded as a synonym for innovation and technological progress for more than 130 years. VDE is the only organization in the world that combines science, standardization, testing, certification, and application consulting under one umbrella. The VDE mark has been synonymous with the highest safety standards and consumer protection for more than 100 years.

Our passion is the advancement of technology, the next generation of engineers and technologists, and lifelong learning and career development “on the job”. Within the VDE network more than 2,000 employees at over 60 locations worldwide, more than 100,000 honorary experts, and around 1,500 companies are dedicated to ensuring a future worth living: networked, digital, electrical.

Shaping the e-dialistic future.

The VDE (VDE Association for Electrical, Electronic & Information Technologies) is headquartered in Frankfurt am Main. For more information, visit www.vde.com

About the Information Technology Society within VDE (VDE ITG)

The Information Technology Society within VDE (VDE ITG) is the national association of all people working in the field of information technology in business, administration, teaching and research and science. Its objectives are to promote the scientific and technical development and evaluation of information technology in theory and practice. Founded in 1954 as the Nachrichtentechnische Gesellschaft, it is the oldest professional association in the VDE. Its nine technical divisions, to which more than 80 technical committees are assigned, represent the entire spectrum of information technology. About 10,000 VDE members have assigned themselves to the ITG and more than 1,000 experts work voluntarily in the committees.

For more information, visit www.vde.com/itg

VDE Verband der Elektrotechnik Elektronik
Informationstechnik e.V.

Merianstraße 28
63069 Offenbach am Main
Germany

Phone +49 69 6308-0
service@vde.com
www.vde.com