bitkom

Large language models

An overview

Publisher
Bitkom e.V.
Albrechtstraße 10
10117 Berlin
P +49 (0)30 27576-0
bitkom@bitkom.org
www.bitkom.org

Contact person
Kai Pascal Beerlink | Policy Officer Artificial Intelligence
P +49 (0)30 27576-242 | k.beerlink@bitkom.org

Bitkom committee responsible
AK Artificial Intelligence

Authors
Lorenz Lehmhaus | Aleph Alpha GmbH
Christopher Kränzler | Lengoo GmbH
Dr. Maria Börner | Westernacher Solutions GmbH

Layout
Anna Stolz | Bitkom e.V.

Cover photo
© Christina Wocintechchat – unsplash.com

This publication represents general non-binding information. The contents reflect the view within Bitkom at the time of publication. Although the information has been prepared with the greatest possible care, there is no claim to factual correctness, completeness and/or up-to-dateness; in particular, this publication cannot take into account the specific circumstances of individual cases. Any use of this information is therefore the responsibility of the reader. Any liability is excluded. All rights, including the reproduction of excerpts, are held by Bitkom or the respective rights holders.

Content

1 What are large language models?

AI language models or Large Language Models (LLMs), an execution of so-called »Foundation Models«, are the latest development in the field of Artificial Intelligence – they have captured the public imagination and garnered a great deal of attention. Their potential can speed up or even completely automate time-consuming human tasks in content creation, revision, transformation and analysis. At the same time, however, the models' capabilities have also raised concerns.

This guide aims to facilitate informed discussions about the consequences of this technological development by explaining the technical basics of large language models. The functionality as well as central aspects of the models are therefore illuminated and general use cases are outlined. These are also exemplified with concrete use cases. Finally, ways in which companies can benefit from this technical development are discussed and some recommendations for the development and use of the models are formulated.

1.1 Basic concept

»GPT« is a well-known name for a number of large-scale language models developed by OpenAI. The most famous GPT model is ChatGPT, which allows users to interact with technology via natural language input. The name »GPT« stands for »generative pre-trained transformer.« This means that these models have been trained to automatically generate text based on experience they have gained beforehand. ChatGPT has garnered a lot of attention since its release in November 2022 and is used by millions of people.

- **Generative** – The »generative« aspect of a large language model is its basic ability to generate text on its own by predicting the best next word in a response. The order of words and phrases in its responses is based on the patterns the model has learned from the training data.

- **Pre-trained** – Large language models are referred to as »pre-trained« because the Deep Learning algorithms that create such models are provided with an enormous amount of training data before a specific task is defined for the neural network. In other words, the learning algorithms look for information and patterns in the training material that will help them generate fluent text. In doing so, they do not try to adapt this information and patterns to specific types of questions or tasks. These models, in turn, can form the basis for explicitly trained AI systems (hence the name »foundation model«) or can be used to perform a variety of tasks via natural language input.

- **Transformer** – Large language models are based on the so-called transformer architecture for neural networks. This architecture has been used by machine translation developers for over five years and has been continuously refined throughout this period. The success of transformer networks in addressing one of the oldest challenges in computer science, machine translation, has helped make them the standard. Their success is based on their ability to efficiently solve sequence-to-sequence problems by converting an input sequence of words in a source language into an appropriate sequence of words in the target language. Compared to older neural network architectures such as feedback neural networks (RNNs) or the convolutional neural networks (CNNs), transformers are simpler in structure. By parallelizing the training process, they can be trained in a much more scalable way. Another significant advantage of the transformer architecture is that they are able to process input simultaneously rather than sequentially. This allows them to take advantage of remote contexts in the input, e.g., the last word in a long sentence that has a bearing on the meaning of the first word.

1.2 Other concepts

Two other important concepts to consider here: **Multimodality** and **new capabilities** that develop large language models.

- **Multimodality** – When a large language model is trained with material containing multiple types (or modalities) of information, such as text, images, audio, video, and even computer code, the resulting model is called multimodal. Such models have the ability to process and understand different types of data and information simultaneously, allowing them to perform a variety of tasks that require knowledge of more than one type of data. For example, a model trained with both text and images can generate descriptive captions for images by combining its understanding of visual content with its knowledge of natural language. For example, a model trained with both text and audio data can be used to transcribe speech because it can associate the sounds of spoken language with the corresponding written words.
- **Emergent capabilities** – What is perhaps an even more remarkable aspect of large language models are new capabilities and behaviors that emerge when the scale of the models exceeds a certain threshold. As these models are trained with massive amounts of textual data, they can learn a great deal about the structure and patterns of human language. Through this analysis, a large language model also builds a semantic mapping of the language, in a sense learning not only the language itself, but also knowledge about the content of the training material. This allows the models to generalize this knowledge and apply it to new tasks and domains for which they have not been specifically trained.

  Modern large language models can be made to perform a variety of tasks beyond traditional natural language processing, such as mathematical computation, music composition, and even computer programming.

For example, when ChatGPT was asked to calculate the area of a triangle with a height of 2.5 meters and a base of 3.7 meters based on GPT 3.5 and explain its approach, it gave the following answer:

> To calculate the area of a triangle, you must use the formula:
> Area = (Base x Height) / 2
> In this case, the height of the triangle is given as 2.5 meters and the width of the triangle (which is also the base) is given as 3.7 meters. So we can replace these values in the formula to get:
> Area = (3.7 meters x 2.5 meters) / 2
> Area = 4.625 square meters
> Thus, the area of the triangle is 4.625 square meters.
> The reason we divide by 2 in the formula is that a triangle is half of a rectangle. If we multiply the base and the height of a triangle, we get the area of the rectangle that contains the triangle. But since we only want the area of the triangle, we divide by 2 to get half the area of the rectangle. Therefore, the formula for the area of a triangle is (base x height) / 2.

Thus, an emergent capability is a task that randomly and therefore unreliably produces results in successively larger iterations of a model but then suddenly produces significantly improved results after the model reaches a certain size. Based on the poor performance of such a task in successively larger models, one would not expect a sudden improvement in performance on this task when the model reaches a certain level. [2]

Overall, new capabilities of large language models are a testament to the power of deep learning and the ability of these models to learn about language and build a kind of world knowledge from massive amounts of data. As these models evolve and improve, we are likely to see more examples of emergent capabilities in the future.

2 Functionality

»Generative pre-trained transformers« can be likened to a large, multi-dimensional fishing net in which each node is a word and each rope connecting two points is the probability that those two words will appear next to each other. An LLM simply predicts the most likely next word based on the previous words in a sentence. Training material is collected from all over the Internet – human-generated content – and the models learn the probabilities of words appearing next to each other (also called weightings). The models become exceptionally good at producing word orders that sound natural to humans, and this creates an aura of intelligence. Using this analogy, it is clear that these models do not really understand what they are doing. One might even go so far as to say that earlier machine learning-based approaches, trained with very task-specific data to find patterns and make predictions about individual events, may have more »intelligence« than these new systems.

> An LLM simply predicts the most likely next word based on the previous words in a sentence.

The important difference is that LLMs mimic human language very well and therefore appear intelligent to us because we equate language use with intelligence. However, it would be more accurate to represent LLMs as a kind of language layer that allows us to interact with data through natural language. And here's where it gets interesting: we can not only control these models using so-called »prompt engineering,« but we can also use it to give the models context.

2.1 Process steps

The process by which an LLM processes an input or »prompt« can be divided into the following steps:

Tokenization of the input

The model examines the sentence and divides it into small pieces, called tokens. Each token represents a word or part of a word. For example, the sentence »How does a large language model work?« is divided into tokens such as »How«, »does«, »a«, »large«, »language«, »model«, »work« and »?«.

Optimization of the input for the LLM

The model analyzes the prompt and uses information from the training material and the position of words in the sentence to understand their meaning. This information is stored as numbers. Combining the numbers for all the words creates a summary of the sentence. The model uses this summary to organize the sentence in an appropriate form for further processing.

Generate a response

The goal of the LLM is to generate a response to the input (prompt). The model looks at the prompt and tries to predict which word or phrase will fit best next in the response. In doing so, it uses a series of transformation layers that help it focus on different parts of the input and make connections between them. In the end, the model generates a list of possible words to use next in the response. It selects the most likely word and adds it to the answer. This process is repeated until the answer is finished, such as when a maximum length is reached or a special »end-of-sequence« word is generated.

Make the answer readable in natural language

The model transforms the selected words or phrases into a text that people can read and understand.

Response output

The model provides the generated text in response to the prompt.

2.2 The correctness of the answer

Probability,
not truth models

Language models always try to give an answer – even if they have only limited information about a certain issue. Because of this, it is a constant challenge for users to be able to comprehend the derivation of a response. There are already technical possibilities to increase the correctness and comprehensibility by a continuous, critical evaluation by the actual user and by a consideration of the respective use case.

Human feedback

The potential of using human feedback to refine these systems has already been demonstrated by OpenAI through their investment in human assessment of GPT-3 responses. The feedback collected led to the significant improvements in the GPT-3.5 release that allowed the ChatGPT application to attract so much attention in its initial release. The quality of these human ratings has a significant impact on the response quality of the model. Expert users, sometimes referred to as Data Curators, Citizen Data Scientists or Humans-in-the-Loop, will in the future make an important contribution to the successful deployment and continuous improvement of artificially intelligent systems, especially in the workplace.

Language models are not truth machines. Context, comprehensibility and explainability can help here.

Contextualization & Finetuning

The second important lever is the ability to provide the language model with as much relevant context as possible. With so-called prompt engineering, one can increase the quality of the response via optimized prompts in which one defines the context in which the model should act. Although the amount of context that can be provided is limited, the potential is already clearly evident. The next logical step of this concept would be to include the context already in the training of the model. Here, a significantly larger amount of context can be used. Today, it is already possible to train customized models with data that have their industry or organization as content.

3 Application areas

Predicting the next word in a fluent natural language output may not sound very impressive until you consider the myriad real-world use cases. The number of use cases is basically only limited by how creative we are in producing prompts. Below are potential use cases for the technology. Towards the end of the publication, there are also specific examples of how the models could be used, with particular emphasis on sustainable value for business and return on investment (ROI).

> The number of use cases is basically only limited by how creative we are with prompts

Text generation

A large language model can be used with appropriate inputs to create texts such as articles, product descriptions, e-mails, and social media posts, significantly speeding up the process of text production. Some applications can even be fully automated, while for others these texts are used as a draft and can be further edited by a copywriter as needed. For example, a language model can generate entire product descriptions from a dataset of product names, keywords, and specifications, greatly simplifying the creation of a catalog.

Text Revision

A large language model can help improve texts by correcting grammar, punctuation, and spelling and making them more understandable to a target audience. It can even take into account specific style specifications of a company.

An example would be that a language model can rewrite a complex text into a simpler version by using simpler words and sentence structures. This facilitates effective communication with a wider audience by adapting the content to the needs of the readers.

Text summary

Large language models can collapse long texts into shorter ones, while still preserving important information and context. This allows employees to save time and focus on the information that is relevant to them.

Knowledge Management

Large language models can help to improve the organizing and structure of information in companies. In doing so, they can categorize, summarize and link similar data or concepts. Often, companies have many different sources of information, such as knowledge bases, document archives, shared drives, or Slack channels. The problem with this is often that relevant information has to be searched for laboriously. A large language model, through its ability to categorize and summarize information, could provide every employee with important information each morning, such as a summary of the latest meeting notes, links to relevant documents, or an overview of important white papers.

These models can also answer direct questions by analyzing the query and responding to it with relevant information from the training data or from a referenced dataset.

Customer Service

AI language models can be used to create chatbots and virtual assistants that can handle customer queries and provide support around the clock. If support chatbots are built based on a knowledge base, training data is needed to map customer requests to specific information in the knowledge base. In that case, the training data are potential queries, which can be quickly and easily extended by AI language models. Building a chatbot can be done more efficiently and quickly using LLMs, and the application can subsequently reduce the workload of human customer service agents and improve response time.

Machine translation

Large language models have been setting new standards for innovation for machine translation providers for years. They have already reached a maturity in this area that predestines them for lighthouse projects in any company. Not only do they demonstrate the enormous potential inherent in the customization capabilities of such models, but they offer a quick payback use case that subsequently offers high budget savings and a significant stream of company-specific language data. Companies that operate internationally can benefit from using large-scale, generative AI language models to automate content translation. This enables them to translate their content faster and more efficiently into different languages and expand their reach on a global scale.

Automated customer care

By using large-scale, generative AI language models, companies can automate their customer support. Chatbots and virtual assistants can respond quickly and efficiently to customer inquiries and help them when needed.

Personalized product suggestions

Using large-scale, generative AI language models, companies can create personalized product suggestions for their customers. These suggestions are based on customer behavior and can help increase customer satisfaction and sale.

Automated content creation

Using large-scale, generative AI language models, companies can have content created automatically. This can help reduce the time and cost of content creation while improving content quality.

Market research and data analysis

Companies can use large, generative AI language models to conduct market research and data analysis. By analyzing data , they can identify trends and patterns that can help them improve their products and services and better understand their customers.

And beyond that ...

There are countless ways in which Large Language Models can be used. For example, they can be used to categorize data or to create special training materials. It is also possible to write computer code based on simple descriptions. As each company gains more experience with LLMs, more use cases will emerge that are specific to that company's needs.

4 How companies can benefit

Large language models can be used for many different purposes. Their use offers companies several advantages, which are explained in more detail below.

Improved customer retention and loyalty

Generative AI language models can help companies better engage with their customers by enabling more natural language interactions, more personalized messaging, and better recommendations. This can help them engage customers and improve their loyalty in the long term.

Increased efficiency and productivity:

Using generative AI language models, large enterprises can automate and streamline many business processes, such as customer support, content creation, and marketing campaigns. This can help increase productivity, reduce costs, and free up valuable resources.

Improved decision making

Companies can use the insights and analytics provided by generative AI language models to make better decisions based on data-driven insights. This can help them stay ahead of their competitors and identify new opportunities for growth and expansion.

Competitive advantage

Generative AI language models can help companies differentiate themselves from their competitors by offering unique, personalized experiences to their customers. By leveraging the latest AI technologies, they can stay ahead of the curve and establish themselves as leaders in their respective industries. If proprietary company data is included in the training of the models, this can also create a long-term competitive advantage..

Innovation and agility

By leveraging generative AI language models, companies can foster a culture of innovation and agility that allows them to adapt to changing market conditions and remain relevant in an ever-evolving business landscape. This can help them stay ahead of their competitors and drive long-term growth and success.

5 Recommendations

Well-positioned companies in various industries have been benefiting for years from the increase in productivity and competitive advantage from digitization and computerization of tasks. They understand that business is not lost to technological progress, but to competitors who apply technology better than they do.

This is not a time for moratoriums and bans, but for applying new technologies, using them wisely and quickly, and learning their strengths and weaknesses through real-world applications. As with any significant business decision, it is incumbent on companies that see great potential in LLMs to take a balanced and responsible approach that maximizes benefits and minimizes risks. These strategies include:

> Well-positioned companies understand that business is not lost to technological advances, but to competitors who apply technology better than they do.

Responsible development and use of AI

- Assume responsibility: Companies should prioritize ethical considerations when developing or using large language models. This includes a risk-based approach to the models, guided by the requirements of the upcoming AI Act.
- Leverage context: Application-specific custom models that use an LLM as a foundation can add context to increase response accuracy and improve traceability.
- Ensure qualified feedback: Companies should ensure that only qualified users contribute to the training of the models, as the quality of human feedback contributes significantly to the quality of the language model.
- Implementation of data protection measures: Since generative AI language models process and analyze large amounts of data, companies must ensure that data protection measures are implemented. This includes, among other things, the protection of customer data and compliance with applicable data protection laws.
- Employee training and awareness: Companies need to ensure that their employees are well trained and have a basic understanding of generative AI language models. This is important to ensure that the technology is used correctly and that employees understand the risks and potentials.
- Regularly review and update models: Generative AI language models can become outdated or inaccurate over time. Organizations need to ensure that their models are regularly reviewed and updated to ensure accuracy and relevance of results.
- Implementing controls: Organizations need to implement controls to ensure that their generative AI language models do not produce unanticipated results or have undesirable effects. This may include implementing tests and checks to ensure that the results are consistent and correct.

- Collaborate with people with expertise and stakeholders: Companies should ensure that they engage individuals with expertise and stakeholders to ensure that their generative AI language models are used ethically and responsibly. This may include collaboration with experts in the field of artificial intelligence, privacy and ethics committees, and other stakeholders.

Take control of AI in your own business environment

- Companies should use customized models based on LLMs. A company has full control over these models, and using the companies' own data for their own employees ensures that it is accurate, traceable and free of bias.

- Companies should invest in improving the accuracy, efficiency, and security of large language models. This could include working with the AI research community, academia, and other industry partners to actively identify and mitigate biases in large language models.

- Companies should collect proprietary data generated when employees complete tasks and use it for future contextualization. Proprietary data can play a central role in creating a competitive advantage here.

- Organizations should provide guidance and training for users that explains how to use large language models effectively and responsibly and includes understanding their limitations and potential biases.

- Companies should establish an ecosystem with their own models that are used and trained by professionals to realize the full potential of the technology with full control.

Applying these strategies can help organizations responsibly develop, continually improve, and leverage large language models.

References

[1] Vaswani, A, Shazeer, N, et al. ↗ »Attention Is All You Need«, December 6, 2017.

[2] Wei, Jason, et al. ↗ »Emerging Abilities of Large Language Models«, August 2022.

Bitkom represents more than 2,200 companies from the digital economy. They generate an annual turnover of 200 billion euros in Germany and employ more than 2 million people. Among the members are 1,000 small and medium-sized businesses, over 500 start-ups and almost all global players. These companies provide services in software, IT, telecommunications or the internet, produce hardware and consumer electronics, work in digital media, create content, operate platforms or are in other ways affiliated with the digital economy. 82 percent of the members' headquarters are in Germany, 8 percent in the rest of the EU and 7 percent in the US. 3 percent are from other regions of the world. Bitkom promotes and drives the digital transformation of the German economy and advocates for citizens to participate in and benefit from digitalisation. At the heart of Bitkom's concerns are ensuring a strong European digital policy and a fully integrated digital single market, as well as making Germany a key driver of digital change in Europe and the world.

Bitkom e.V.

Albrechtstraße 10
10117 Berlin
P +49 (0)30 27576-0
bitkom@bitkom.org

bitkom.org