Generative AI: from technology to implementation

GenAI guide - beyond the hype

Dr. Reinhold-Julius Heitmann
Denis Hohan
Tatsiana Bychkouskaya
Jonas Heger
Theresa Sporn

July 2023
Blogpost

Public

1. Renaissance of the AI hype

Artificial intelligence (AI) has been around since the middle of the last century. However, with the introduction of Generative AI (GenAI), characterized by tools such as ChatGPT, it is currently not only experiencing a boom among the masses, but is also opening the door for companies to exploit its potential in handling multi-dimensional data such as images, video and natural language to scale sales and reduce costs.

1.1. Not new, but different

A generic AI model essentially works like a mathematical function: it receives an input, performs a transformation and delivers an output. For example, is a conceivable model that calculates the risk factor for heart disease (output) based on three input data: Heart rate, blood pressure and average daily step count (input).

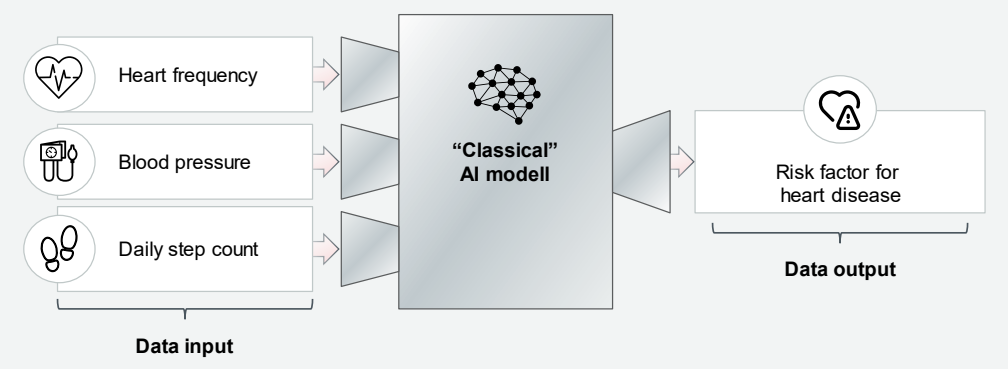

Figure 1 Example of a classic, non-generative AI model that can be realized via machine learning (see below: '1.2. Increasing market momentum')

Generative AI, which is trained using extensive data sets (input) and a complex architecture (transformation), has the ability to create convincingly human-like digital content as output. This content is often innovative in terms of structure and content: GPT-4 from OpenAI, for example, accepts text input and produces text outputs with human speech patterns and semantics. GenAI can generate outputs in the form of text, images or sound, as well as combinations of these in the form of videos and texts set to music. Text generation is covered by so-called Large Language Models (LLM), such as GPT (Generative Pre-trained Transformer from Open AI), LLaMA (Large Language Model Meta AI from Meta) or PaLM (Pathway Language Model from Google). These LLMs can independently generate both natural language and formal language, such as programming languages. What they have in common is that they gain a deep understanding of language and content through their complex structure and training with large amounts of text data. It is important to differentiate here that these models tend to imitate concepts and recombine trained patterns rather than really "understand" context.

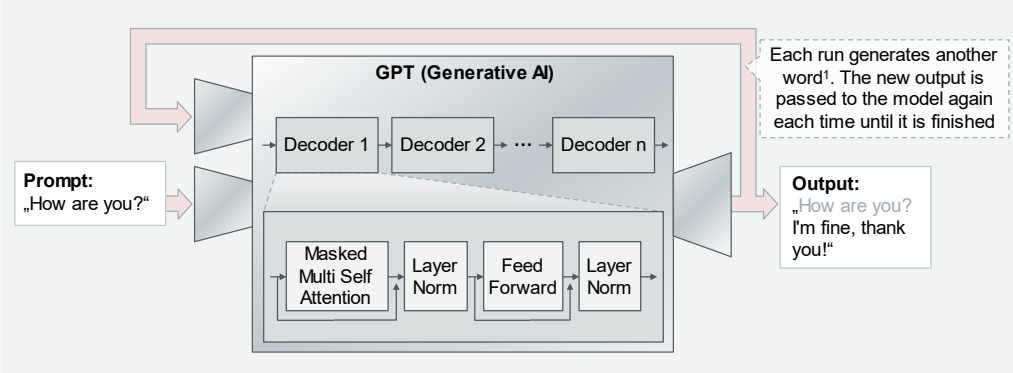

Figure 2 Functionality and structure of a Generative AI model using the simplified example of the GPT-LLMs1 with original designations of the technical components.

1.2. Increasing market momentum

The term AI was coined back in the 1950s. In recent decades, continuous developments in AI technologies, in particular the machine learning and deep learning revolution, have led to more efficient training times and more complex network models. The creation of today's large language models was made possible by the introduction of the Transformer architecture, a special form of deep learning that was made possible by the so-called self-attention mechanism. This recognizes connections in word sequences via a of the network. Popular applications such as ChatGPT from OpenAI were developed based on this technology and opened new possibilities for AI applications in various areas.

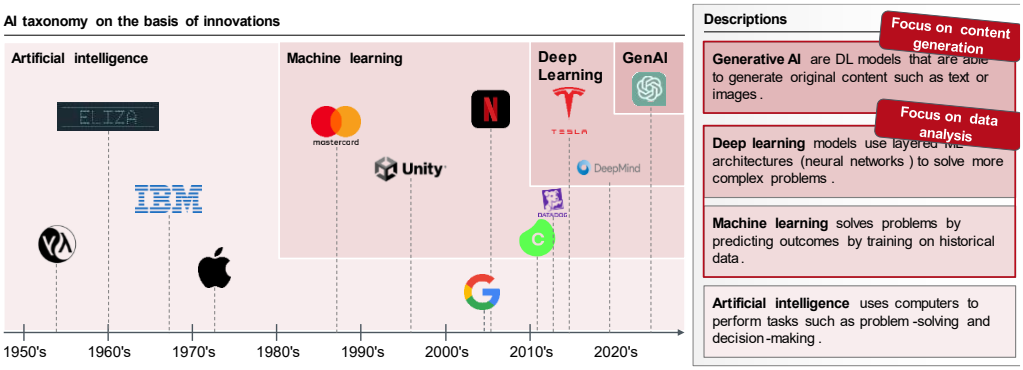

Figure 3 The AI taxonomy based on AI innovations

The hype around ChatGPT might give the impression that it is the only GenAI technology, but this is not the case: The market offers a variety of vendors and products in the field of Generative AI, which are integrated into various applications such as text generation, image synthesis and speech processing. OpenAI, with the GPT and DALL-E series, is one of the main players. Google, Microsoft and Amazon Web Services (AWS) also offer comprehensive products and services for Generative AI. The landscape of this technology is constantly evolving and includes multiple providers and products. Likewise, the integration of this technology into various applications is

[1] LLMs, like GPTs, do not actually work with words, but with so-called tokens. These are also text modules, which are usually smaller than whole words.

increasing. The GPTs platform from Open AI enables users to define GPT-driven applications via GPT as a voice interface and thus structurally solve not only individual problems, but entire problem groups. In the past, ChatGPT was used manually to extract figures, data and facts from annual reports with a series of questions, but now a program can be defined via a voice interface that automatically executes this workflow for an annual report. This user-friendly approach, which is accessible to the general public, was already the key to the success of Open AI with ChatGPT. It can be assumed that GPTs could lead to a further rise in Open AI.

In the future, will behave similarly with LLMs as with other machine learning models : At a certain level of maturity, older models will become open-source and accessible to the public. Over time, the quality of the new models will no longer differ significantly from the open-source models. This means that it is not so much the models themselves, but the frameworks that will make the difference in the application (e.g. Google Tensorflow or Meta PyTorch).

2. Choose the right ones from the inexhaustible range of potential use cases

AI has the potential to revolutionize industries. N t is not for nothing that people talk about "disruptive technologies". In an innovation-driven market characterized by highly competitive pressure, AI-supported use cases can lead to precisely such disruptive changes. While it may be tempting to assume standardized use cases, reality shows that a case-by-case approach is required. While there are recurring use cases – the 24/7 customer support hotline, intuitive dashboards that dive deep into data and innovative image generation for marketing purposes – the key is to carefully examine possible use cases for your own company and not rely on rigid models.

2.1. Recognize use case potentials

Adapting the technology first requires a careful use case analysis.

In principle, potential use cases result from data types and objectives. The first step is to define the objectives. The application goals include:

- **Generation of data points:** Creation of new objects, e.g., through suggestions in the creative process or the production of social media posts optimized for likes. The object can be created with a single prompt (the input of a query to a GenAI model) or in an iterative process in a series of prompts, so that time expenditure can be reduced and the creative process can be expanded.
- **Interaction with the model:** Using the models as conversation partners, such as chatbots that interact with other GenAI models. LLM models can bring all patterns and thus content that was contained in the training data set into the conversation. This sets them apart from previous retrieval-based chatbots such as Alexa, which could only return predefined answers.

- **Accessing information:** Special case of interaction in which information is used for training or context, e.g., research in documents or image generation for campaigns. The model is used as an interface to the document or image. Requests can be made for content so that even large amounts of information can be consumed in a short space of time.

2.2. Determine feasibility

Once the objectives of the use case have been defined, it is necessary to determine the feasibility from various perspectives:

- **Determinism:** If you ask a GenAI model the same question twice, it can give two different answers. The models are not deterministic, which is undesirable for some use cases (e.g., the use case: blocking a credit card). The use of targeted prompting methods reduces this problem. However, the more control is desired, the more difficult it is to use.
- **Pretrained models:** It must be checked whether suitable pretrained models are available. If no model is available, the training effort increases drastically.
- **Data basis:** Although many GenAI models are highly generalizing and can therefore be used for similar applications, fine-tuning or complete training is sometimes necessary. A corresponding data basis must therefore be created.
- **Computing power:** The application of large models often requires specialized hardware and data centers - typically GPUs/TPUs. While server-based applications or suitable local hardware can easily run such models, local applications on mobile devices, for example, are quite difficult to implement. Depending on the load peaks of the intended application, it will most likely make sense to choose cloud environments.
- **Compliance:** Regulatory conditions (e.g., GDPR) can severely restrict data processing. GenAI models can be web-based and some of them continue to learn from user prompts. Once the information has been learned, it can be queried by other users through prompts and at the same time is stored in such a complex way that it cannot be specifically unlearned or deleted from the model. It is advisable to have each use case validated by a technology-trained compliance employee.

After the objectives and feasibility, the analysis focuses on the economic viability. The business case must take into account both direct and indirect costs, including training, system customization and model development and operation. If the model fulfills a purpose and can be classified as feasible and profitable, an economic use case has been identified.

3. Challenges of the operationalization of GenAI in the company

The introduction of GenAI in a company is more than just a technical challenge; it also requires far-reaching organizational adaptations and comprehensive legal consideration.

3.1. Legal

The legal challenges of GenAI are diverse and differ from jurisdiction to jurisdiction. In addition to the EU regulations currently being coordinated in the form of the AI Regulation and the AI Liability Directive, existing national regulations, particularly in the area of copyright and data protection law, influence the handling of GenAI. In the area of application of German law, there are therefore far-reaching challenges along the processing chain.

- Data protection and copyright play a central role in **input**, in the form of the input of personal data and the use of copyrighted works in the prompt. In the corporate context, there is also the potential risk of losing the protection of trade secrets due to a lack of protective measures if these are used by employees in the prompt.
- During **processing,** the general question of admissibility according to sector-specific requirements arises first. In addition, any co-determination rights of the works council must be observed when using GenAI. Questions of product safety and liability may also arise. Even today, misguided AI can trigger liability claims under general civil law.
- In terms of **output**, the biggest challenge again lies in the area of copyright, as in the absence of personal intellectual creation by an AI, no copyright arises. This inevitably means that no licenses can be granted to AI-generated works. AI works are generally in the public domain. There are also other pitfalls to avoid, such as antitrust law in cases of coordinated automated pricing or labor law, which requires personal performance.

In addition to the existing national regulations, the EU AI Liability Directive, which is intended to resolve the attribution of unlawful acts in the case of fault-based liability claims in relation to "black box" AI, is eagerly awaited. Also of interest is the EU AI Act, which takes a risk-based approach, enables preventive risk assessments and prescribes different protection and security measures depending on the potential risk.

It remains to be seen how standards already in the legislative process will develop and how standard law will adapt to the new challenges. The addressees of these standards, especially in a business context, are therefore advised to take care of a comprehensive compliance system in addition to the technical dimension at an early stage in order to be able to use the full potential of GenAI in a legally secure manner.

3.2. Technical

Challenges in the use of GenAI can also be observed in technical areas:

- A key challenge in the technical area is **data availability**. The efficient operation of AI depends on the quality and quantity of data. In addition to technical knowledge, data protection and ethics are crucial. Data challenges include data bias and security. Creating robust data pipelines and governance frameworks is therefore essential.
- When designing the AI setup in the company, **choosing the right model** is key. This determines the system decisions and depends on the type of problem, data quality, computing capacity and application objective. The model architecture and optimization influence performance. Constant adjustment and balancing of accuracy, fairness and interpretability is necessary.
- The third challenge is **integration into existing IT structures**. Many organizations have legacy systems that are not AI-oriented. Integrating AI into such systems requires planning and expertise. Smooth data flow, connectivity and compatibility are essential.

To use AI successfully without jeopardizing operational stability, these complex challenges must be mastered. The first prerequisite for this is the technical basis that already exists in the company. Technical efforts may need to be made before more far-reaching AI objectives can be pursued. A corresponding analysis is recommended so as not to experience a sobering realization in the middle of the process. How great leaps can be made on the existing technological basis depends in turn on the existing skills and experience in the company. Here too, it is advisable to carry out an upstream analysis in order to avoid having to identify gaps in competence afterwards.

3.3. Organizational

While technology is advancing rapidly, organizations are often unable to adapt their structures, strategies and workforce fast enough to take advantage of the full potential of new innovations. "Martec's Law" visualizes the growing gap between the rapid development of technology compared to slower organizational development, which results in organizations struggling to embrace and integrate technological opportunities.

The readiness of the organization is a basis for the use of AI technologies. However, there are obstacles to organizational development in the context of AI:

- The main problem is often **incomplete data governance**, which can trigger legal, ethical and data protection issues. Such incidents can damage a company's reputation.
- There is often a lack of understanding of AI. **Insufficient knowledge** and blind trust in AI lead to misinterpretations and missed opportunities.
- **Internal organizational resistance** to AI inhibits its implementation and reduces added value.

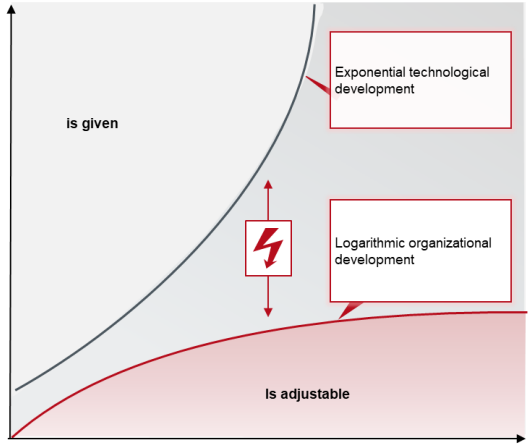

Figure 4 Visualization of Martec's Law

To bridge the gap between technological progress and organizational development, a data-oriented culture is crucial. It should be remembered that to automate an activity through the use of AI methods, knowledge of the expertise and thus the specialist himself is required, who is ultimately to be supported or replaced by the AI component. A lack of understanding or a threatened job reduces the willingness to provide this support. The definition of suitable project structures together with simultaneous education, communication and the creation of appropriate incentive structures can make the difference for the success of the project.

4. Checklist on the way to Generative AI

The relevance of GenAI for companies is becoming increasingly clear. If companies do not want to fall behind their competitors and want to exploit their full potential, the question arises as to the next strategic steps. Instead of prematurely implementing isolated solutions whose costs often exceed their benefits, management should pursue a coherent, holistic approach. This can be broken down into five essential steps:

- **Strategic alignment:** Formulate a clear position on GenAI at management and company level. In concrete terms, this means defining views and strategies in relation to AI and GenAI. AI and GenAI should be part of the business or IT strategy. It may make sense for the company to develop its own AI strategy or expand existing strategies.
- **Create a data basis:** AI requires structured training data from the company. Is the data landscape ready for the age of AI? Data should be accessible, normalized and integrable, and the legal framework for its use should also be clarified. The data landscape should therefore be evaluated in terms of its suitability for GenAI and modernized accordingly.

- **Create a basis for development:** GenAI solutions can be developed On-Prem[2] or purchased ready-made. In-house developed on-prem solutions are interesting because they offer flexibility and can save costs. At the same time, special developer skills and technical infrastructures must be created as a basis. Modern platforms such as EPAM's AI DIAL can help to bridge the gap between on-prem and market solutions and accelerate implementation.
- **Create a culture:** Employees must adapt GenAI and undergo further training on new AI-supported processes and how to use the technology. In order for it to become part of the corporate culture, incentives for adaptation must be created and the fear of being replaced by GenAI must be overcome. Employees can then develop their skills and knowledge and meet GenAI service providers on an equal footing.
- **Select use cases and develop proofs of concept (PoC):** There are numerous potential use cases for your company. It is important to identify and evaluate many possible use cases. A sound understanding of the technical possibilities and implementation of GenAI is essential for the development and evaluation of use cases. Once a shortlist or specific use case has been selected, PoCs must be created for final validation.

In practice, it has been shown that opportunity-driven Gen-AI projects without the described basis often do not bring the desired added value or even fail in practical implementation. Conversely, even a short strategy project at the appropriate decision-making level can help to holistically anchor the technological potential of this technology in the corporate strategy and thus contribute to sustainable competitiveness. Managers are therefore required to provide the necessary expertise for internal discussion.

[2] On-premises (abbreviated to "on-prem") describes when software (and hardware) is installed, operated and managed directly on a company's devices, rather than in the cloud and by third-party providers

Dr. Reinhold-Julius Heitmann is an Expert Director at CORE. He has realized various AI and analytics projects in the banking, insurance and medical sectors. Julius not only advises our clients, but also develops hands-on software for CORE that makes our day-to-day consulting work more efficient. Based on his wide-ranging experience, Julius can point out strategic perspectives as well as accompany concrete implementations.

Mail: julius.heitmann@core.se

Denis Hohan is a Transformation Fellow at CORE. With his focus on data science (especially AI), IT organization, project management and process mining, he supports our clients in business-critical technology transformations. He already gained practical experience in consulting, data science and software development during his studies in business informatics.

Mail: denis.hohan@core.se

Tatsiana Bychkouskaya is a Senior Transformation Manager at CORE. Her areas of expertise include project management, digital transformation, payments and IT infrastructure projects. With her background in technology and innovation management, she supports our clients in the development of strategies, the implementation of complex IT transformations and the design of payment systems and products.

Mail: tatsiana.bychkouskaya@core.se

Jonas Heger is an Expert Manager at CORE. As a lawyer trained in Germany, he brings a legal perspective to strategic transformation processes as well as AI and analytics projects right from the conception phase. He works in interdisciplinary teams to develop practicable solutions in day-to-day consulting, away from purely legal opinions. Jonas primarily supports clients in the banking, automotive and manufacturing sectors.

Mail: jonas.heger@core.se

Many thanks to our co-authors:

Dominik Siebert (CORE), Mauritz von Lenthe (CORE) and Muskaan Multani (CORE)

Addresses

COREtransform GmbH
Kurfürstendamm 194
10707 Berlin | Germany
https://core.se/

Phone: +49 30 263 440 20

office@core.se

COREtransform GmbH

Limmatquai 1

8001 Zurich | Helvetia

https://core.se/

Phone: +41 44 261 0143

office@core.se