Making Sense of Generative AI

CUTTING THROUGH THE HYPE FOR BUSINESS LEADERS AND CURIOUS MINDS

Dr. Dominik Hörndlein

Making Sense of Generative AI

Cutting through the hype for
business leaders and curious minds

- Reading Sample -

Dr. Dominik Hörndlein

Website: https://making-sense-of.tech

ISBN: 978-3-9827019-0-5 (paperback)
ISBN: 978-3-9827019-1-2 (e-book)

1st edition January 2025

CONTENTS

FOREWORD

Human nature is driven by our curiosity and ability to create new solutions to everyday problems. Therefore, it is understandable that new technologies, and the innovations they enable, keep fascinating us to the present day. Yet, these novelties come with uncertainty about how they can be used. Not only for our own good, but also by others against us.

The invention of the steam engine as a general purpose technology in the late 18th century provides a compelling historical parallel to our current situation with AI. When James Watt improved upon the steam engine in 1769, it sparked a revolution that would transform society. People were simultaneously fascinated and frightened by its potential applications.

On one hand, the steam engine promised unprecedented industrial power and efficiency. Visionaries saw potential for faster transportation, increased manufacturing capabilities, and economic growth. This excitement led to rapid adoption in factories, mines, and eventually railways and steamships.

However, the steam engine also instilled fear in many. Workers worried about job displacement as machines could now perform tasks that previously required human labor. Some religious leaders viewed it as an affront to divine order. There were also concerns about safety, as early steam engines were prone to explosions.

As history unfolded, both the promises and concerns proved valid to varying degrees. The steam engine indeed revolutionized industry, transportation and the global economy, ushering in the Industrial Revolution. It created new jobs and industries while rendering others obsolete. While it brought prosperity to many, it also led to exploitation of workers and environmental degradation

in some areas.

This example illustrates how transformational technologies can have wide-ranging and sometimes unforeseen consequences, both positive and negative. It underscores the importance of thoughtful implementation of new technologies.

With the latest hype around generative AI, we experience the same patterns. Only that innovation cycles have become faster than in the past, and that we face various social media channels which allow to spread information – and excitement – at an unprecedented scale and speed worldwide. This leaves many of us with the conviction that AI is going to radically change many aspects of our world. We are just not sure which ones exactly.

While some company leaders and technology experts are ensuring us that these changes are going to improve our lives significantly, there are other popular voices sharing doomsday prophecies. Who of them is right? What will be the true impacts of generative AI on our daily lives? How can we take the right decisions to implement generative AI in a value-creating way? And how do these technologies actually work?

This book won't give you the answers to all questions you might have – nobody really knows how the future will unfold. What I offer you instead are insights that let you understand better how generative AI works, what drivers impact their future evolution, how we can make use of it and where the chances and limitations of this technology lie.

I am convinced that it already makes a big difference to understand the core concepts behind generative AI. Therefore, we won't get deep into technical details, but deep enough to understand them.

This book is written for business professionals and curious minds who possess at least a basic understanding of IT. At the end of the book, you shall feel much more confident to lead discussions about this topic with domain experts and take decisions in your business.

1

FUNDAMENTALS

You cannot make sense of generative AI without looking at AI in general. AI is a domain with ongoing research for many decades now, so that concepts and breakthroughs have been building up on each other for a long time. Generative AI is just one part of it – though an important one with huge potential. Yet, it has its limitations. Many of them can be appreciated when you have a feeling how AI solutions have been created long before the advent of generative AI.

In general terms, artificial intelligence (AI) encompasses technologies that enable machines to learn, think, and perform tasks that typically require human intelligence. These technologies stretch over many domains where machine-based solutions aim to replicate and go beyond human achievements. Therefore, we begin with a brief history of AI to get a broad overview on the most prominent fields of AI in recent years. Discussing how AI is created will make clear how practitioners turn data into working solutions, closing with some more thoughts on the important role of data in this domain.

1.1 A brief history of AI

Concepts explained in this section: major milestones and concepts driving progress in AI in recent years, explanation and impact of

computer vision, deep learning, GPUs, reinforcement learning, image generation and language models.

Like other technological revolutions, advances in Artificial Intelligence often unfold in unpredictable ways. While AI has roots stretching back to the 1950s, the past years have seen significant developments in its capabilities and real-world applications.

Drivers for this acceleration are rooted in progressing capabilities of computer hardware – IT systems were able to process ever-growing amounts of data at less processing time and costs. With hardware costs decreasing, the entry barriers lowered to a level which made it affordable for private persons to run and develop AI solutions on their private computers. A growing open-source community in the AI field makes up for a second driver of progress.

A few decades back, in the 1980s, the situation looked different. Researchers in the field of AI had made bold claims about the chance that AI will soon be capable of performing any intellectual task a human can do.[1] Such claims raised expectations that could not be met, as the technologies of that area were not advanced enough. What followed was the so-called AI winter.

Let's look at some of the most significant milestones in AI development of the past decade that have shaped the AI landscape we find ourselves in today. This will create a better understanding about what these systems are capable of and introduce some concepts. We do not aim for completeness here, but rather for providing some insightful examples. Also, you don't need to be able to recall all of the names and examples given in this section – the aim is to create a basic understanding upon which we can build up on in the following sections and chapters.

Computer Vision

The field of computer vision relates to AI solutions that process visual information like images and videos. Back in 2010, major developments in this field started with the ImageNet challenge. In it, research teams, companies and whoever wanted to participate

[1] Read more here: https://redresscompliance.com/the-evolution-of-ai-tracing-its-roots-and-milestones/

were tasked with creating AI systems that recognize and categorize objects that appear in millions of images into a few dozen categories (growing up to a few hundred categories later on). Objects to be recognized in the challenge included, for example, animals like cats and dogs, objects of daily life as balloons and bath towels.

This arena proved useful because it allowed teams with different approaches to realize computer vision AI while competing on comparable terms. The community could see who sets the benchmarks at the moment. Further, great new ideas get visibility immediately. The annual ImageNet challenges have since become cornerstones in advancing computer vision, proving that with enough data and computing power, AI could begin to "see" the world in ways that rivaled human perception.

In 2012, a team tried out a new software architecture based on so-called *deep neural networks* (named AlexNet). It achieved an unprecedented accuracy in this challenge, far surpassing traditional methods. This breakthrough demonstrated the power of *deep learning*, a technique that would come to dominate AI research and applications in the years to follow.

Deep learning

What is deep learning about, and what makes it such a game-changer?

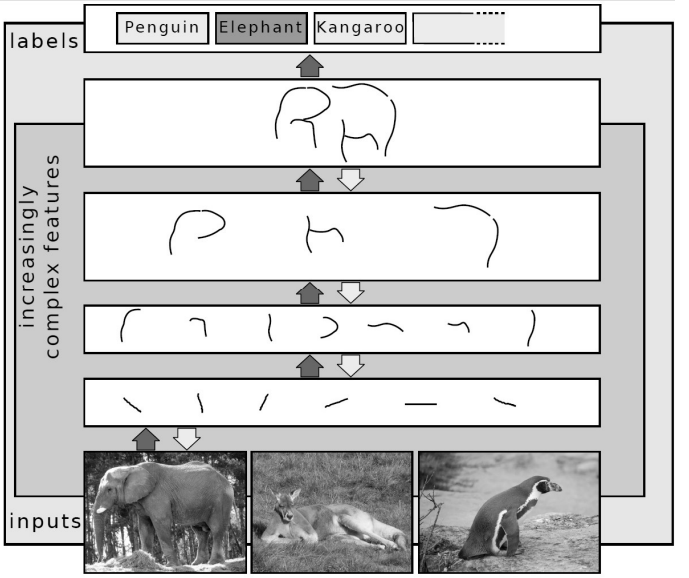
At its core, deep learning solutions were inspired by the human brain. Just like our brains are made up of interconnected neurons, deep learning systems are built with interconnected “nodes” that process information. These nodes are arranged in layers, so that information flows from the first to the last layer before the systems decides on the actual answer.

Imagine you're looking at a photo of a cat. Your brain doesn't immediately recognize it as a cat – it processes the image in stages. First, you might notice simple shapes and edges, then fur texture, then cat-like features such as whiskers or pointy ears and finally you conclude it's a cat. Deep learning works similarly.

In a deep learning model for image recognition, the first layer might detect basic elements like edges and colors. As information flows through subsequent layers, the model recognizes increasingly complex features – from simple shapes to intricate patterns. The

final layers combine all this information to make a decision, like "This image contains a cat."

As an example, the figure below illustrates how the “deep learning layers” – indicated through white boxes – learn increasingly complex feats from bottom to top.[2]



The AI isn't explicitly programmed to recognize cats or any other specific object. Also, it is not told what features work best to recognize and distinguish a cat from a dog. Instead, when fed with millions of labeled images, it learns to identify patterns on its own. This is why deep learning models can often recognize subtle patterns that even humans might miss.

In order to learn how to identify the patterns that work best, deep learning systems rely on huge amounts of data. This distinguishes them from us humans: we might not be able to process information as fast as computers, but we are great in learning new topics very

[2] The visualization is based on graphics from Sven Behnke and Wikipedia, https://en.wikipedia.org/wiki/Deep_learning#/media/File:Deep_Learning.jpg, available under CC BY-SA license.

efficiently. Once we have seen a cat, we can identify one the next time we see it. Deep learning methods need much data before they can do this.

The deep learning approach was soon applied to areas far beyond computer vision. From image and speech recognition to language translation, deep learning allowed to tackle problems that were previously thought to be too complex for machines.

As written before, one key concept resides from the fact that on a general level, these systems essentially teach themselves the best way to achieve a goal, without the need of humans explicitly telling how the single features look alike. As a drawback, humans also have a hard time to understand how exactly deep learning systems come to their conclusion. As these system taught themselves their abilities, they are often black boxes to us: they might make accurate predictions, but we can't always explain exactly why or how they arrived at a particular conclusion.

Another key aspect is their hunger for data. While traditional AI methods might struggle with making sense of vast amounts of information, deep learning thrives on it. The more data you feed these models, the better they become at recognizing patterns and making accurate predictions.

GPUs

To create such powerful AI systems based on deep learning, we need computer hardware that is able to process this vast amount of data. But the capabilities of hardware are limited. Providing an increasing amount of data to the AI increased their capabilities significantly. The trade-off, however, was that also the time it takes to create them increased accordingly.

Researchers at the University of Toronto, driven by scientific curiosity, discovered a new approach to make this training process more efficient. While CPUs – central processing units of any computer – are used to process all kind of data inside a computer, the gaming industry introduced special GPUs – graphical processing units – that were better at processing data that is used to ultimately show great visuals on the computer screen. When investigating how well these GPUs would perform when running deep learning AI instead of visualizing videos, the found out that they were much

more efficient than CPUs.[3]

The big breakthrough moment happened, again, in the ImageNet challenge of 2012. The team that created the winning AI (AlexNet) leveraged for training their AI. Only with this approach, they could succeed in creating their deep learning solution in a time that was short enough to compete in the ImageNet competition.

This was the starting point when major tech companies and research institutions started their work towards creating larger and more capable AI systems on a much broader scale than before. Powered by the capabilities of GPUs.

On a sidenote: this is also the origin of why Nvidia is one of the major players in the tech industry today. Although they were the leader in producing GPU hardware for the gaming industry before, this niche was much smaller than the AI market is today. It was luck that modern AI solutions happen to be a great battlefield for applying GPUs.

And it was the genius of Nvidia's CEO to bet the company on this trend. Seeing how the company's stock price has risen from around $ 0.35 in January 2012 over $ 2.70 in early 2017 to over $ 130 in January 2025, we can tell that the bet has paid off.

Reinforcement learning

We saw that major breakthroughs were achieved by scientific curiosity – or to put it differently: try and error. Many smart people work on improving AI systems and the underlying approaches to create them. But in the end, you have to try out a lot of ideas and see if they really work.

Games provide an excellent testing ground for AI development.. Computers that play chess are available since many years already. It was in 1996 when Deep Blue was the first computer to ever win a chess game against a reigning world chess champion (Garri Kasparov).[4]

While chess is hard enough a game to master, there are games

[3] For computer vision tasks, CPUs usually take around 5 to 10 times longer than GPUs to finish.

[4] Find more details here: https://www.ibm.com/history/deep-blue

which prove a much more complex testing ground. The team of Google Deep Mind took on the ancient Chinese game Go, which is still popular to the day. We don't need to understand what Go is in detail. But basically, you have a board with 19 times 19 fields. There are two players, one who has black stones, one who has white stones. The objective is to occupy more territory on the board than the other player. You do this by placing stones on the board, one after the other, and kick off the board as many of the other player's stones as possible by surrounding them with your own stones. Though the basic idea of the game is simple, its inherent complexity was too high for traditional computing.

In 2016, the team of Deep Mind succeeded to defeat the world champion. For achieving this goal, they relied on a new way of enabling AI to learn new things, namely *reinforcement learning*.[5]

We will discuss how AI models are created in the next section in more detail. But to understand what reinforcement learning is all about, let's pick out one single aspect already at this point.

On a basic level, what does an AI system do when it identifies objects on an image (like animals in our earlier example)? It gets an image as input data, processes it, and returns what it has recognized. There is a clear input and output of data. Further, you can easily judge whether the output is correct by comparing what the AI returned to what you really see on the image.

How does this work for a game like Go? As an input, you have the configuration of how the players' stones are arranged on the board. What is the output of the AI? It shall tell you the next move you should do. Even better, the next best move to optimize your chances of winning the game against your competitor.

This is where the challenge starts. How can you judge whether a move is the best one to ultimately win the game? Before the game ends, both players will have taken dozens to hundreds of moves. Moreover, there is a strong uncertainty about how your opposing player will react to your actions, so that it is hard to predict whether your move was good or bad. Overall, the connection from the task

[5] Find more details here: https://www.alexanderthamm.com/en/data-science-glossary/alphago/

(suggest next move) to the ultimate goal (win the game) is much weaker than for the example with the image.

Therefore, how can an AI judge whether a single move is good or not? This is the challenge which is addressed by reinforcement learning.

On an abstract level, the approach it takes is reminiscent of how you would train pets. You cannot tell them directly to sit down. But you can reward them for correct behavior. If you play out the reward game well, and invest enough time, then your dog will learn what kind of behavior you expect.

For the game Go, the game plays a lot of games against itself. When it wins, it gets rewarded, when it loses it gets penalized. This starts a lengthy process where the AI improves in a trial-and-error manner.

Be aware that this approach does not directly answer the aforementioned question on how to judge whether a single move is good or not. It rather answers how its overall approach to a situation, or its strategy, unfolds.

We will revisit reinforcement learning, among other approaches to creating AI models, in the next section. This is because the concept has been fundamental to the success of generative AI – more on that later.

Creating images

These breakthroughs of reinforcement learning were covered in public media a lot, and therefore raised awareness even by many non-AI-experts. This is because it was an astonishing achievement that an AI was able to beat humans in difficult games where humans need years of training to master.

The moment when generative AI got awareness from a broader non-technical audience for the first time occurred some years later. It was when image-generating AI achieved a level where it produced images that looked almost like photos taken by a camera. For this, you only need to describe what you want to see in a picture in natural language, let the AI process your words, and after some seconds you receive an image that depicts this very content.

This also points towards an important aspect of why this achievement raised so much attention: you don’t need to be an IT

expert to formulate sentences in natural language. As long as somebody takes care of hosting the AI model together with a good user interface that is available via internet, anybody can use it. This is in contrast to earlier times when only IT experts were able to use generative AI to create images.

It was in January of 2021 when OpenAI released the first version of their DALL-E[6] product, which allowed users to create images in the aforementioned way. Around a year later, an increasing number of similar services from other vendors followed – some as paid services, like Midjourney, some as free open-source software, like Stable Diffusion.

To create images, the AI first needs to understand the meaning encapsulated in the phrases which the users write down to describe what they want. Only then can the AI start and do its "magic" to create a corresponding image. Therefore, the image generators where only possible because there were significant advancements in the way machines understand human natural language.

Language models

Processing natural language was obviously an important part of AI research for a long time. Humans interact with each other in written or spoken ways in so many different manners that having machines making sense of what we write and say opens a lot of interesting possibilities.

An essential part of handling natural language are so-called language models, which hold a model of how language is built.

This still sounds very abstract, so let's get more concrete. Before 2010, the usual approach to language models were based on statistics. When you look at an incomplete sentence, what are the most likely next words to finish it? For example, assume that you have a phrase that starts with "This morning was great because". TO continue it, probable next words that make sense could be "I", "we", "my friend" or many other nouns. On the contrary, continuing the sentence with "This morning was great because hello" does not

[6] The name DALL-E stems from a pun mixing the family name of the famous painter Salvador Dallí with Wall-E from the Disney movie.

make any sense, irrespective on how the sentence continues further.

Thus, such early language models work like this: given the words "This morning was great because" as input, the word "I" has a much higher probability to be the next word in this sentence than "hello".

What makes sense or not can be inferred from statistics. How? Looking at a lot of texts, you take out any sequences of two (or three, four, five, etc.) words that appear in the texts. Then, for any word that exists in your vocabulary, you calculate the probabilities of it being the next word. When you do this for a lot of texts, you end up with meaningful statistics to predict next words in a sentence – this is your language model.

Early applications of these simple language models were the auto-complete functionalities that you found when typing on your smartphone. For this purpose, the auto-completion does not need to have a deeper understanding of what you are writing - predicting the next word correctly with a reasonable probability is good enough.

This approach, however, is too simplistic for many applications. As you surely know, language is more than just a concatenation of words. Words and sentences have a meaning, and this meaning can differ widely when the context changes in which they are used. As an example, the sentences "You are right" and "You are correct" have the same meaning, so that the words "right" and "correct" can be used interchangeably here. In other contexts, however, the word "right" can be used in the sense of "the opposite to left" – inferring a completely different meaning.

Therefore, to make sense of language, you cannot stop at just using words but have to bring the meaning behind words on a more abstract level. This is what research followed up on, with different approaches to embed the raw words of a text into representations of what they mean. A lot of such research progressed this field, with one of most remarkable achievements leading to the invention of what we know today as *attention mechanism*.

Look at the sentences you are just reading. They are building up on each other. Some sentences would not make a lot of sense to you if you read them alone – you might need to be aware of context from a few pages earlier. Also, not every word in a sentence is equally important to understand it. Filling words like "also" might make a

text more readable, but they are not relevant for understanding it in many cases.

Then, in order to make language models better in understanding the meaning behind what we say and write, it would make sense to focus their efforts on those words that are important, wouldn't it? This is where the attention mechanism steps in.

Basically, it is a means to tell the AI model a) which words are most relevant, and b) which words are somewhat connected to each other with regards to the overall meaning.

Let's make the last part more specific. Consider that a language model is processing the following text:

> The animal didn't cross the street. This is because it was too tired.

In this sentence, the word "it" refers back to "animal". Both words actually stand for the very same object in this small piece of text. Therefore, the attention mechanism helps the AI recognize this connection. This may sound trivial to us, but making this connection clear boosts performance of language models.

As a second example, "it" and "tired" are correlated meaning-wise, while the words in "This is because" carry far less relevance to the meaning of the sentence. Accordingly, when processing the word "it", the AI should somehow consider the words which are relevant as well and filter out those that are not.

This shall clarify the importance behind the attention mechanism: to enable the language model to process words as connected entities, instead of treating each word in isolation.

Practical applications for this approach were translation solutions which translate text from one language into another. They improved significantly in quality with the rise of the attention mechanism. Which makes sense when you think about it. While earlier language models were treating text as a statistical sequence of words, the newer language models are better in extracting the structure which holds the meaning behind a sentence.

When you translate this paragraph from English into your mother tongue (or any other language) – would you translate it word by word? Or wouldn't you first try to catch its meaning before starting

to write down the translated sentence? There are many further applications of this concept beyond translation which we will discuss in more breadth in later chapters.

The invention of the attention mechanism goes back to 2014 and the years that followed. A big break-through moment took place in the Summer of 2017 when researchers from Google published the article "Attention is all you need".[7] Besides the introduction of the attention mechanism, the authors also introduced the so-called *transformer* architecture for constructing the AI part behind a language model. We'll go into more details on how transformers work in the next chapter. For now, it suffices to know that transformers are the components in an AI model that apply the previously explained attention mechanism. This allowed transformer-based AI to capture context and meaning better than AI relying on previous approaches.

Transformers keep on being an important approach to artificial intelligence, with its main concepts still being used. You see it in popular applications like ChatGPT, where GPT is short for generative pre-trained transformer. GPT is *generative*, because it generates new text. It is *pre-trained*, because the underlying AI is trained on data. And it utilizes *transformer*-based approaches in its AI.

In a nutshell

We have discussed multiple concepts that marked the cornerstones of AI progress in recent years. All of these are still relevant today, as generative AI builds up on them. Therefore, we will revisit all of these concepts throughout the book with regards to their relevance for generative AI.

1.2 How AI models are created

Concepts explained in this section: high-level process of how AI solutions are created.

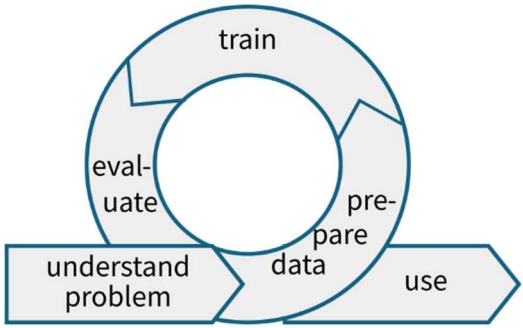
When we examine how media covered the topic of generative AI in

[7] Find the article here: https://arxiv.org/abs/1706.03762

the months after the release of ChatGPT in late 2022, one was easily tempted to conclude that there is something fundamentally new going on. From a product perspective, there truly really were new possibilities opening up through the availability of generative AI. From the view point of an AI practitioner, however, it was just the steady evolution of research and improvement going on for many years and decades.

On a basic level, generative AI models are still created by the same mechanisms that allowed to create non-generative AI long before the year 2022. Therefore, it's good to have a rough understanding of these mechanisms and processes. This will clarify how AI learns, and hopefully demystify the supposed "magic" happening inside AI a bit. Explaining the process in a way that you are able to conduct them yourself would go far beyond what a single book can deliver - and it's not our intention here. After reading this section, however, you will be better equipped to identify potential limitations of AI models and can comprehend how they learn.

We will go through a usual process of creating an AI model, from first idea to productive usage. We will not be exhaustive and complete in our discussion, we will skip steps that are mostly of technical nature. The goal here is to provide an understanding of crucial steps in shaping AI solutions. Find below a visualization of these process steps.



Besides a description of the process parts, we will further use the example of predicting sales of different products in a web shop for the sake of making the concepts more tangible.

Understand the problem

Like all digital solutions, an AI must solve a problem of end-users or business teams in a company. Therefore, any AI model creation starts with understanding this problem and what it exactly means for the users.

When a common understanding of the overall situation has been reached between business teams and developers, potential solutions are ideated and broken down into single steps.

Let's go for an example to make this clear. Assume that we are working for a company with a big web shop, selling various types of goods via the internet. We have a big physical storage where goods are stored for a short time before they get sent to the customers. However, our colleagues have a hard time to predict how much of each good shall be kept in stock – having too little delays the delivery times for customers, having too much costs money because you have to manage more goods.

What would you do? For understanding the exact situation, you could ask how things are exactly handled today. Heading for how accurate today's predictions are, how these predictions are made, if there are goods for which predictions work much better than for others, what data is exactly used for calculating the predictions, if there are other potentially relevant data sources which are not considered today, what kinds of customer segments we have in our web shop, how well we understand the buying preferences of customers in each segment, and so much more … You see how quickly a seemingly easy challenge gets broken down into a plethora of aspects that raise complexity.

Further on, don't underestimate the "language barriers" between the business and developers involved in here. While the challenge might seem obvious to the business team, the developers are likely to miss context from day-to-day work. Further, they might have never spoken to an end user in person, so that they don't know about all the small inefficiencies with which users have to spend their time on. Also, developers use different wordings in their work compared to business teams, so that a sentence that is well understood by business people might lead to a different interpretation for developers. All of this raises the chance that IT

solutions are not aligned with real customers' challenges – not only when AI is involved.

The point is: for creating an AI that does a good job, you need be formulate this job as concrete as possible. If somebody cannot explain the problem which he or she is solving, don't expect him or her to provide valuable solutions.

Prepare the data

AI needs data to work, but not just any data – the content must be relevant. It shall be of good quality (meaning that the data may not have wrong information). Completeness is important – if the data mostly covers behavior of customers below the age of 35, any AI solution you build for the exemplary web shop will be predicting rather bad information for your older customers.

As an outsider, it is hard to assess whether the used data for creating an AI has been chosen wisely. If you do find out what kind of data has been used, however, reflecting on what kind of information is available inside the data gives you good hints towards what the AI that is created with this data can or cannot predict.

Getting back to our example to make things clear. For building an AI that predicts orders for your web shop, assume that you have access to data that tells you what items your users clicked on when browsing your website. Which already is helpful. But the exact information you can get out of it is merely: which items sounded interesting to which customer? More relevant would be data that tells you what customers actually did order. This does not mean that the first kind of data is not relevant – the contrary is the case. Still, you should look for more data that helps you get closer to the actual answer you want to solve.

On a general note, you can often read that it is crucial to have as much data is possible. Looking at major American tech companies, we see that each of them has millions, if not even billions, of users who regularly use their applications. This gives them access to user data on a large scale that no small competitor can keep up with. Obviously, the more data you have, the more behavioral patterns of you customers can be included in the data, which ultimately powers your AI.

Yet, size is not everything. Quality and relevance are just as

important. If you get access to huge amounts of data, but you don't know exactly how much wrong information it carries or where it is coming from, then you have a good chance of creating an AI that predicts information that does not suit well to the specific problem you want to solve. Having a smaller amount of data set, but with higher quality in turn, is often favorable. For the context of generative AI, we discuss this point in more depth in a later chapter.

The key takeaways here are: data is crucial for creating any AI, so that it has to be selected and prepared with much care. This is why the data preparation can consume up to 80% of the time in many projects. The amount of data is important, but the quality, relevance and completeness to your problem are as well.

Train and evaluate the AI

The steps to select the right kind of AI model and actually train it are of very technical nature, so that we won't discuss them here. On a high level, AI developers create systems that extracts the patterns and correlations from your data in a way that you can apply them on new data. As an example, if you have collected and prepared good data that carries information on how many and which customers in specific customer segments were to buy a product X in the past, you can predict how likely it is that a new customer will buy the product X in the future.

How good is this AI in reality? When trying to solve complex business problems, you had to simplify and break down the problem into smaller tasks you can solve. There are always chances that you have overseen a task, or that you have oversimplified the business problem. Therefore, it is crucial to test how well an AI performs in real life.

How would you do this? You don't use all data you prepared for training the AI. Instead, you hold back a small portion - e.g. 20% of it – and call it test data. While the other 80% are training data on which the AI gets trained, you can use the test data to measure how close the real values are to the values that were predicted by the AI.

In most cases, they are not good from the very beginning. You have to test the AI models accuracy regularly, and find the cases in which it behaves particularly bad. Then, you must deduce why things got wrong in these specific cases, and try to improve your

approach to address these points of improvement. Over time, you will reach a point where the AI gets good enough so that you can really use it in a way that creates value for you.

In a nutshell

You can say that it's the data which drives the quality of an AI, and a specific (business) problem which defines the context in which the AI acts. Further, the better you understand the context, the better you will be able to make use of the data and create good AI models. Moreover, knowing the data on which an AI has been trained will allow you to get a better feeling for what an AI can do and what it cannot.

1.3 The role of data

Concepts explained in this section: how data and data processing influence AI quality.

In the following, we will review important aspects about the role of data.

Spotting patterns

What, exactly, does an AI do with data so much better than we humans can? As a short answer: it is very good in spotting patterns and correlations inside data. And it is able to do this at scale.

This capability to spot patterns should not be misinterpreted as if the AI were really intelligent and "knows" what it is looking at. Let's get specific to understand this better. Recall when we discussed the topic of computer vision in the first section? An example I gave there included an AI that was shown many images with animals on it. From these, it learned which images show cats, dogs, or whatever other animal. We discussed that powerful AI models actually learned to identify increasingly complex structures in the images – starting with simple geometric shapes like edges, circles and putting these together in increasingly complex structures.

Then, whenever specific structures are identified on an image (like pointy ears, appearing together with a small nose, a long tail, etc.), the AI can deduce that there's a cat in the image. It recognizes patterns in the image that usually are present when a cat appears.

For language models, this behavior holds true as well. As humans, we learn to speak when we are babies. We learn to form sentences that show correct grammar and carry a meaning. In contrast, an AI actually learns to predict the next words that are most likely to continue a given text. It doesn't learn grammatical rules in school, nobody explains it whether a sentence makes sense in the context of the discussion. It is given a lot of texts, and from these it learns how the appearances of all words are correlated with each other.

It's fascinating to see how this very basic capability – identify correlation of words inside text – translates into how we use large language models today – to review text, giving us advice, and so much more. We dedicate the full next chapter to discussing this better. For now, the key aspect is: AI is great in spotting patterns and correlations inside data.

Efficiency and generalization

Besides its capabilities in identifying patterns, AI solutions are also able to do this at scale. This is because they run on computers, which are much faster in processing data (in form of bits and bytes) than we humans are.

Yet, they can only do what they have been created for initially. And nothing else. Which should not take the fascination away from what AI can do – for generative AI, we see it every day what things it enables us to achieve that haven't been possible only a few years back. Still, to understand the limitations of AI, keep in mind: it can just do exactly what it has been built for. Even if it can do this at blazing fast speed and on a vast amount of data, it cannot extend its scope by itself.

But wait. Based on our example of the web shop, we have discussed that an AI is trained on user data in the past and applied to the data of new customers to create useful insights of their future behavior. It extended its scope to cover the behavior of new, previously unknown customers. Didn't it?

So far that's true. And still you should only apply this AI in the domain it has been created on. When the AI has mainly learned from the behavior of customers below the age of 35, it might fail for the context of older customers. If you only have data from European

customers, it may or may not be successful for American customers. You must know the context and scope in which the AI can be used.

In real life, context can also shift. Reflect on your own behavior – do you buy the same things in the same way as you did ten years ago (or maybe even only some months ago)? Likely not. Personal behavior and preferences change. Not only for single persons, but for whole societies and customer groups. An AI that performed well one year ago might behave worse today because customers changed their behavior. This is known under the name *model drift*. This means that the quality of AI models can degrade over time because the context that it was trained on drifts and changes.[8]

In a nutshell

AI models are machines that excel in spotting complex patterns they have learned from data, and they are very efficient in doing this. Yet, they are bound to patterns they have learned from training data.

But enough of theory about patterns in data. Let's get more concrete and talk about what this all means for generative AI.

[8] We discuss model drift in more depth in section 6.4.

2

GENERATING TEXT

Now that we have a better understanding of what AI is and where it comes from – what turns an AI into generative AI? The subtle answer is that it *generates* new things. While "classical" AI is mostly used to predict or classify data, generative AI is perceived as more. For example, when you prompt a large language model to write a bedtime story for your kids about a certain topic, it will generate a novel text for you that hasn't been there in this exact way before.

But there's more to it than the creation of novel texts and images. In this and the next chapter, we dive deeper into what potentials generative AI has to offer. This one focuses on text, and the next one will discuss images.

We keep the spirit from the last chapter and explain how things work without getting lost in technical details. This will serve the goal of understanding the potentials and limitations that generative AI has to offer on a broader level.

The chapter starts by explaining how large language models work, how they are created and what features make them most relevant to us. By getting through the most prominent options to make use of them, these concepts will become more tangible and it will set the foundation to discuss some of the most promising use cases that are realized these days later on. At the end, a reflection about the relevance of size in LLMs will close the chapter and lead

over to the generation of images in the next chapter, which will broaden the scope from text to this further medium.

2.1 How large language models work

Concepts explained in this section: the most relevant aspects and components that power LLMs, a simplified view on why LLMs work so well.

When we talk about generative AI today, we often refer to large language models (LLMs). But what exactly makes a language model "large", and why does it matter?

Large language models earn their name from two key factors: the enormous amount of data they're trained on and their massive size in terms of parameters. To put this into perspective, consider GPT-3.5 from OpenAI, which is one of the most well-known LLMs.

GPT-3.5 was trained on a diverse dataset of text from the internet, including websites, books, articles, and social media posts. The exact volume is not publicly disclosed, but estimates are at around 225 billion words. This vast amount of text covers a wide range of topics, styles, and formats. This allows the model to learn from and generate coherent text on almost any subject.

The size of an LLM is typically measured by the number of parameters it has – think of these as the "knobs" the model can adjust as it learns. To give a concrete example, GPT-3 is built up of 175 billion parameters. To put that in context, its predecessor, GPT-2, had only 1.5 billion parameters, which was considered large at the time of its release in 2019. The next version of the LLM, called GPT-4, was released in 2023, and is estimated to consist of around 1.8 trillion parameters. This shows how fast the language models grew in size over the timespan of only 4 years.

Not only the size of parameters of these models grew fast. The amount of text on which they were trained on grew as well – from around 10 billion words for GPT-2, over 225 billion words for GPT-3 to an estimated 10 trillion words used for GPT-4.

These numbers sound humongous, and they are. To put them into more relatable terms – the English version of Wikipedia contains articles with around 3.9 billion words in total (as of December 2023). An average adult would need around 30 years to

read every single Wikipedia page when doing nothing else for 24 hours a day, 7 days a week.[9]

Model version	Year of release	Parameters in the model	Training data, compared to English Wikipedia
GPT-2	2019	1.5 billion	2.5 * Wikipedia
GPT-3	2021	175 billion	58 * Wikipedia
GPT-4	2023	1 800 billion	2 500 * Wikipedia

There's a direct correlation between the size of the model and the amount of text (or more general: the amount of data) needed to train it effectively. The larger a model is, the more data it requires to reach its full potential because it contains more parameters that need to be optimized. This relationship has led to a sort of arms race in the AI world, with companies and researchers constantly pushing the boundaries of model size and training data volume.

However, this scaling behavior has almost reached its limits. This is because most parts of the data available in the internet have already been used to train LLMs, so that there is not much additional text data available that could be added to the training data.[10] This makes it hard and increasingly expensive to get larger sets of data on which to train new models. But when you aim for training larger models, you rely on increasing the amount of training data.[11] Training larger models on larger data sets then directly translates into higher costs to create the AI. All of this makes the business case behind such an effort hard to justify.

This does not mean that we have reached the end of the evolution of large language models. There are other ways to move forward, which we will explore in a later chapter.

---

[9] On average, an adult reads around 250 words per minute. Doing some math leads you to the approximated 30 years for reading 3.9 billion words.

[10] We will discuss the scaling behavior in more depth in section 7.1.

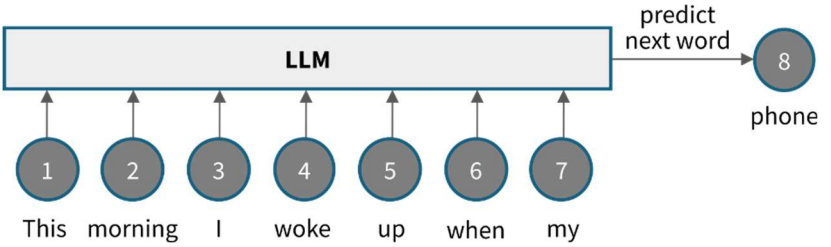
[11] This statement will get more clearer throughout this chapter.

The technical perspective

How does an LLM generate new text? Let's get through the technicalities without getting too deep.

As an exemplary case, consider the following short phrase and use an LLM to complete it: "This morning I woke up when my". As an LLM, we take the GPT-2 model. The later models like GPT-3, GPT-4 or models from any other creator are more complex, but the basic concept stays the same. Therefore, using the GPT-2 model is just fine for an illustrative purpose.

We feed the sentence "This morning I woke up when my" into the LLM, and it tells us which word will likely be the next one in the sentence – in our example, it is the word "phone".



How does GPT-2 do it? The logical steps are

- embed: take the single words and embed them into a more abstract representation,
- attention: apply attention mechanism to catch how the words are related to each other,
- neural network: apply mathematical calculations to translate abstract representations of words into new ones,
- predict: calculate the most likely next word in these abstract representations and translate it back into one can read.

In the following, we discuss what concretely happens in all these single steps.

Remaining pages of second chapter skipped.

Read more on how LLMs work, how they are trained, the aspects that make them truly useful to us, how you can optimize them to your needs and what critical aspects you need to pay attention to in chapter 2 of the book.

3

GENERATING IMAGES

While the generation of text is the most prominent and widely used application of generative AI, its capabilities are not limited to text. Especially the generation of images has the ability to have big impact on economy and society.

The working mechanisms behind image generators are quickly explained on a very high level: you describe what you want to see, and the AI creates an image that matches this description. Furthermore, you can adjust it to your needs by changing details in this description.

In this chapter, I will explain how images are generated by AI in more depth. Herein, we can build up on the insights from the previous chapters. Afterwards, it will be easier to also understand how other types of media, like videos, are generated.

We start with discussing the important mechanisms and show how these work together to create images through AI. Next, we get deeper into the process for training such AI models. After addressing limitations that these image generators show today, we discuss how we can optimize our use of AI to mitigate them and create pictures that align well with our needs. Based on these insights, we are able to move on and understand how video generation works. Finally, we close the chapter with an overview on popular services for image and video generation that help you with getting started quickly.

3.1 How image generating models work

Concepts explained in this section: the most relevant aspects and components that power image generating AI, diffusion models, encoder-decoder models, denoising images to create content.

In this section, we will go through the relevant mechanisms of how *diffusion models* work, as these have developed into the standard AI architecture for generating images.

Diffusion

Diffusion models take their name from the physical process called *diffusion*. What does this term mean?

As an example: take a glass of water, and let a drop of blue ink fall into it. In the beginning, you will see that the blue color is exactly at the place where you let it drop. Over time, the ink spreads out and mixes throughout the water. Thus, you can say that the color drop diffuses from the original area into the whole glass until you cannot tell any more where the drop originally was.

The water becomes undrinkable as the ink has dispersed throughout it. In order to get the water clean again, you would like to have something like "reverse diffusion": have the ink move back into its original place so that you can remove it from the water. Physically, this is not possible, of course. But it's a good analogy to what we will be doing with images.

Noisy images

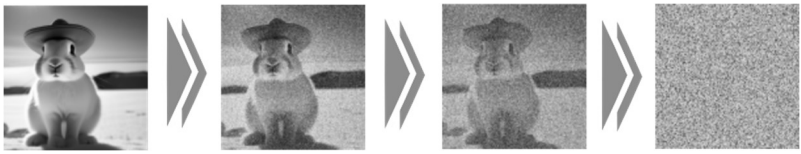
A second concept you need to understand is *noise* in the context of images.

When you take an image with your camera at bright daylight, you will get an image with clear color and the objects will be visible sharply. Now, take the same image at a later time around sunset when there is less light arriving from the sun. Likely, you will get an image which is sprinkled with a lot of tiny little dots where the color is not perfect. This is because with less light, the camera has less information on the exact color of the object. In consequence, these dots which make the image imperfect are noise.

As a next step, both concepts are applied on images together.

Assume that you have an image in good quality. Then, you can come up with an algorithm that randomly adds a little noise to it. This will make it a bit worse, but you will still be able to tell what is on the image. If you have the algorithm add more noise, the content of the image will become increasingly harder to recognize. Until you reach a point where you have lost all information on the original image, and you are left with noise only.

As an example, you can look at the image below. Starting from the original image of a rabbit on the left side, we add increasingly strong noise as we move to the right. Consequently, it becomes more diffuse until the rabbit cannot be recognized any longer.



Denoising images

But why do we talk about making images noisier? This is because diffusion models do the exact opposite: they take a noisy image and denoise it until you have a proper image on which you can recognize objects.

Just like we removed the ink from the glass of water with "reverse diffusion" in our analogy, we want to remove the noise from the image to recover the objects and make them visible again.

Remaining pages of third chapter skipped.

Read more on how image generating AI works, how it is created, how these underlying concepts drive video generating AI, how you can optimize these AI models to your needs and what critical limitations you need to pay attention to in chapter 3 of the book.

The remaining chapters focus on how companies leverage generative AI in real-life applications already today, what challenges you will face in realizing your own solutions, how you can get started to drive your own implementations, and how the future of AI might unfold and will likely transform our business and private life.

GLOSSARY

In this glossary, you find the technical terms that were used throughout this book, together with a short description of what they mean.

Agentic Solutions: AI systems that can independently perform sequences of actions to accomplish tasks, often by combining multiple capabilities like understanding text, making decisions, and interacting with other software.

Attention mechanism: A technique that helps AI models focus on the most relevant parts of input data, similar to how humans pay attention to specific words in a sentence to understand its meaning.

API (Application Programming Interface) : A set of rules that allows different software applications to communicate with each other, enabling them to exchange data and functionality.

Artificial General Intelligence (AGI) : A hypothetical type of AI that would match or exceed human capabilities across a wide range of cognitive tasks, rather than excelling at just specific tasks.

Bias: In AI, a tendency of a model to produce certain outcomes more frequently than others, which can be either intentional (helping the model make useful predictions) or problematic (leading to unfair or discriminatory results).

Chain-of-thought: A technique where AI models break down

complex problems into smaller steps, similar to how humans show their work when solving math problems.

Computer Vision: The field of AI that enables computers to understand and process visual information from images or videos.

Deep Learning: A type of machine learning where artificial neural networks learn from large amounts of data through multiple processing layers, inspired by how the human brain works.

Diffusion Models: A type of AI model that creates images by gradually refining random noise into clear pictures, similar to how an artist might start with a rough sketch and progressively add more detail.

Embedding: The process of converting words, images, or other data into numbers that AI models can process, while preserving meaningful relationships between different pieces of data.

Few-shot Learning: A technique where an AI model learns from just a few examples of a task, similar to how humans can understand new concepts from just a few examples.

Fine-tuning: The process of taking a pre-trained AI model and further training it on specific data to make it better at particular tasks.

Foundational Model: A large AI model trained on vast amounts of general data that can be adapted for many different specific tasks, serving as a foundation for building more specialized AI applications.

GPU (Graphics Processing Unit): A specialized computer chip originally designed for rendering graphics but now widely used for AI calculations due to its ability to process many calculations simultaneously.

Hallucination: When an AI model generates false or misleading

information while presenting it as fact, often occurring when the model tries to answer questions beyond its training data.

Instruction-tuning: The process of training an AI model to follow specific commands and respond appropriately to different types of requests, helping it better understand and execute user instructions.

Large Language Model (LLM) : An AI model trained on vast amounts of text data that can understand and generate human-like text, such as GPT or Claude.

Latent Space: A mathematical representation where AI models process data in a compressed form, making computations more efficient while preserving important features of the original data.

Model Drift: When an AI model's performance degrades over time because the real-world data it encounters differs from its training data, requiring updates or retraining to maintain accuracy.

Multi-modal AI: AI systems that can work with multiple types of input data (like text, images, and sound) simultaneously.

Neural Network: A computing system inspired by biological brains, consisting of interconnected nodes that process information in layers to recognize patterns and make decisions.

Parameter: A variable within an AI model that gets adjusted during training to help the model make better predictions; larger models have more parameters.

Pre-training: The initial training phase of an AI model on large amounts of general data before it's fine-tuned for specific tasks, similar to how humans learn general knowledge before specializing in a field.

Prompt Engineering: The practice of carefully crafting inputs to AI models to get better and more accurate outputs, similar to knowing the right way to ask a question to get a helpful answer.

RAG (Retrieval-Augmented Generation): A technique that combines an AI's ability to generate text with the ability to look up and use specific information from a knowledge base.

Reinforcement Learning: A type of machine learning where an AI learns by trying actions and receiving rewards or penalties, similar to how animals learn through trial and error.

Reinforcement Learning from Human Feedback (RLHF): A technique where AI models are improved by having humans rate their outputs, using these ratings to learn which responses are most helpful and appropriate.

Scaling Law: The observation that AI model performance typically improves as the model size, amount of training data, and computing power increase.

Temperature: A setting in generative AI models that controls how random or creative their outputs are, with higher values leading to more varied and creative responses.

Token: The basic unit of text that AI models process, which can be a word, part of a word, or a character; models typically charge based on the number of tokens processed.

Transformer: A type of AI architecture that powers many modern language models, using attention mechanisms to process input data effectively.

Vector Database: A specialized database that stores data as mathematical vectors, making it efficient to find related information based on meaning rather than just matching keywords.

ABOUT THE AUTHOR

Dr. Dominik Hörndlein has spent over a decade bridging the gap between business needs and technological innovation. As an AI strategist and product owner, he helps businesses harness the power of artificial intelligence to solve real-world challenges. His journey spans roles as an entrepreneur developing AI-powered mobile solutions, a project manager leading data science teams, and a technology consultant guiding major digital transformations. With a Ph.D. in Physics and extensive experience implementing AI solutions, Dominik combines deep technical knowledge with a gift for making complex concepts accessible to diverse audiences. In 'Making Sense of Generative AI', he draws on this unique perspective to help business leaders and curious minds understand and leverage the transformative potential of AI technologies.

LinkedIn: https://www.linkedin.com/in/dr-dominik-hoerndlein/
Website: https://making-sense-of.tech

WANT TO LEARN MORE AND GET STARTED WITH IMPLEMENTING GENERATIVE AI?

The book is available in all major book stores as e-book or paperback version, such as:

Amazon – Kobo – Apple Books
Google Play – Waterstones – Foyles

Limited time offer:

get the e-book at 70% off – only until May 31st 2025
at Amazon, Apple Boks and Google Play