Why We Need to Relearn How to Talk to Machines - A Snapshot of Generative AI in January 2024

Authors: Maria Kalweit, Gabriel Kalweit
Affiliation: Collaborative Research Institute Intelligent Oncology (CRIION), Freiburg, Germany
Submitted: 18. March 2024
Published: 25. March 2024
Volume: 11
Issue: 2
Languages: English
Keywords: Generative Artificial Intelligence, Natural Language, Computing Power, Digital Assistants, Large Language Models
Categories: News and Views, Artificial Intelligence, Modeling and Simulation, CRIION
DOI: 10.17160/josha.11.2.977

Abstract:

The last few years have seen incredibly rapid progress in the field of generative artificial intelligence. Talking to machines and getting answers in natural language is part of our new, elusive normal. Driven by the exponential growth of both computing power and internet-scale data, our new digital assistants are trained by estimating the most likely next element of a given context. Recent years have clearly shown that this general objective can lead to the ability to develop complex and diverse capabilities from simple principles. At the same time, however, it can lead to interesting structures in the compression of the training data and sometimes to unpredictable artefacts. The aim of this article is to shed light on the mechanisms behind current large language models and to provide guidance on how to get the best answers to a question. The original version in German, “Warum wir neu lernen müssen, mit Maschinen zu sprechen – eine Momentaufnahme der Generativen KI im Januar 2024“ was published in Ordnung der

JOSHA
Journal of Science, Humanities and Arts

March 2024 Volume 11, Issue 2

Why We Need to Relearn How to Talk to Machines - A Snapshot of Generative AI in January 2024

Maria Kalweit and Gabriel Kalweit
maria.kalweit@intelligent-oncology.org
Collaborative Research Institute Intelligent Oncology (CRIION), Freiburg, Germany; Department of Computer Science, University of Freiburg, Germany

Abstract

The last few years have seen incredibly rapid progress in the field of generative artificial intelligence. Talking to machines and getting answers in natural language is part of our new, elusive normal. Driven by the exponential growth of both computing power and internet-scale data, our new digital assistants are trained by estimating the most likely next element of a given context. Recent years have clearly shown that this general objective can lead to the ability to develop complex and diverse capabilities from simple principles. At the same time, however, it can lead to interesting structures in the compression of the training data and sometimes to unpredictable artefacts. The aim of this article is to shed light on the mechanisms behind current large language models and to provide guidance on how to get the best answers to a question. The original version in German, "*Warum wir neu lernen müssen, mit Maschinen zu sprechen – eine Momentaufnahme der Generativen KI im Januar 2024*" was published in *Ordnung der Wissenschaft* in February 2024 (https://ordnungderwissenschaft.de/wp-content/uploads/2024/03/Kalweit-Druckfahne-V4.pdf).

In December 2023, the user *@danshipper* shared an unexpected problem regarding text and image processing software on the social media platform X[1]. To his astonishment, a command that had worked flawlessly a short time before no longer worked. He uploaded an image of a book page, hoping to get the text back digitized, and in his helplessness posted his question to the public community. In response to his *why, @NickADobos* suggested: **"**Hallucination. In part because you phrased it as a polite question, which 'no' is a valid answer to." Similar experiences were shared by *@mblair,* who noted, **"**I had the same problem where it refused to augment an image with a specific seed when I used the 'Can you..' phrasing." *@Reelix offered* an explanation: **"**It's designed to use the least effort in giving you a valid answer to your question, and with 'No' being a valid answer requiring minimal computation, that's the one you received."

In addition, some users observed that the software's responses were more detailed when it believed that it was spring and not winter[2]. This led to humorous speculation about *winter depression* and the general assumption that people generally work less in winter. Interestingly, it was also found that the more *tip* users offer, the more extensive the responses were[3].

Such conversations could not have been witnessed until recently. People talking about *how to address* a machine to get a response; that the degree of politeness brings differences in results; that the machine displays traits not unlike those of a human; not to mention that you can *talk* to a machine normally *at all.* The software at the center of these discussions is ChatGPT[4], a groundbreaking achievement in the field of artificial intelligence (AI), developed and released by OpenAI in 2022. Within just two months of its release, it had already recorded over 100 million users – an unprecedented success that underscores the rapid adoption and interest in this technology. ChatGPT is not just another software product; it represents a turning point in the way we interact with machines and use them in our daily lives because

[1] Dan Shipper [@danshipper], "What the Hell? When Did This Happen?? Https://T.Co/KWXVXE9Dem," Tweet, *Twitter*, December 6, 2023, https://twitter.com/danshipper/status/1732258207840501946.
[2] Rob Lynch [@RobLynch99], "@ChatGPTapp @OpenAI @tszzl @emollick @voooooogel Wild Result. Gpt-4-Turbo over the API Produces (Statistically Significant) Shorter Completions When It 'Thinks' Its December vs. When It Thinks Its May (as Determined by the Date in the System Prompt). I Took the Same Exact Prompt… Https://T.Co/mA7sqZUA0r," Tweet, *Twitter*, December 11, 2023, https://twitter.com/RobLynch99/status/1734278713762549970.
[3] thebes [@voooooogel], "So a Couple Days Ago i Made a Shitpost about Tipping Chatgpt, and Someone Replied 'Huh Would This Actually Help Performance' so i Decided to Test It and IT ACTUALLY WORKS WTF Https://T.Co/kqQUOn7wcS," Tweet, *Twitter*, December 1, 2023, https://twitter.com/voooooogel/status/1730726744314069190.
[4] OpenAI, "ChatGPT," 2024, https://chat.openai.com.

it allows us to use our natural language, such as English, German, or Spanish, instead of formal languages or code, such as Python or Java. A new, elusive normal.

Platforms such as ChatGPT are based on so-called Large Language Models (LLMs). These models are characterized by the fact that interactions with them are not based on predefined text modules or commands, but on written language in all its diversity. So if you write to systems like ChatGPT in natural language, they can understand you and respond in the same language. However, certain manners can in turn control the type, quality and structure of the responses. The commands and strategies used to control the responses are then examples of so-called *prompting*. Prompting can be seen as a kind of programming language, but it is model-specific. For example, with the same natural language input, the output of a GPT-3 may be different from that of a GPT-4. Generally, new updates can completely change the way these models interact, which means that a new *grammar* or *language* must be learned for each model. This is currently an extremely valuable skill, which is reflected in the spectacular annual salaries offered to professional prompt engineers. For example, Der Spiegel reported on December 6, 2023 that annual salaries of up to USD 335,000 are being paid for experts in this field[5]. At the same time, there are research approaches that aim to have AI systems write their own prompts[6], which could make this new profession obsolete just as quickly as it emerged.

The conversation that @danshipper and other users had on a social media platform reflects more than just a technical problem; it symbolizes a shift in our relationship with technology. These interactions with ChatGPT not only reveal the limits and possibilities of AI-based communication, but also raise fundamental questions about the nature of human-machine interaction. How do we understand this technology? How do we adapt our communication strategies to achieve the best results? And what does it mean when a technical tool becomes a quasi-social actor in our digital cosmos?

This article aims to discuss these issues and create a deeper understanding of how ChatGPT and similar systems work. It is not only about how we can use a specific

[5] Verena Töpper, "(S+) Geld verdienen mit ChatGPT: Prompt Writer verdienen bis zu 335.000 Dollar im Jahr," *Der Spiegel*, December 6, 2023, sec. Job & Karriere, https://www.spiegel.de/karriere/chatgpt-prompt-writer-und-prompt-engineers-verdienen-bis-zu-335-000-dollar-im-jahr-a-a54a93a5-e20d-40e6-b235-28aec0bddaaa.
[6] "Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution," 2023, https://openreview.net/forum?id=HKkiX32Zw1.

software more effectively, but also about exploring the implications of this technology for our society and our future. With a focus on effective strategies for prompting, it explores how we can develop a long-term understanding of this advanced technology that will endure after further updates and developments to the models.

What does GPT actually mean?

The way to good prompting strategies is to understand the methodology behind the systems. Although not all the details are known and certainly cannot be discussed in this article, the following is a brief explanation of what lies behind the *magic* of Large Language Models.

A model is a conversion from input to output. An example is the input of a book page and the output could be the extracted text. Defining such a transformation a priori exactly for all possible inputs is quasi-impossible, which is why current methods of artificial intelligence take the detour of estimating such a transformation based on given examples. The language used for this in AI is generally mathematics. The starting point is a flexible representation of this sought-after mathematical transformation, which can take various forms by setting different free parameters. Let's take the sentence "I'm feeling good today" as an example. If the word *good* is assigned a high weighting, this sentence could be regarded as rather positive. However, if the focus is more on *today*, it could also be read in such a way that feeling good today is something special, which in turn could be classified as rather negative. Such an assignment of positive and negative is therefore dependent on the weighting, i.e. the parameters, of the model that makes the assignment.

The basis for the enormous and enormously rapid progress of recent years is formed by so-called *artificial neural networks* as a representation of the mathematical transformation described above – architectures of *artificial neurons* connected in parallel and in series. These artificial neurons each weight their inputs with a free parameter, add up these weighted inputs, transform this sum with a non-linear mathematical function and pass on the result, the *activation* of the neuron, to the subsequent neurons. The free parameters of a neural network result in specific outputs with a specific setting, the correctness of which can be measured using an error measure. If the specific contribution of a parameter to the given error is now measured, the parameter can be adjusted step by step in such a way that the result is improved. The problem here is that the estimation of the

mathematical transfer should also generalize to *unseen* examples at the time of parameter determination in order to be usable in practice. A model is then defined by its fixed, found set of set parameters. And even if the inputs within a model must be represented in the form of numbers, the processed modalities can be arbitrary in origin – for example, written text.

There had already been some movement in the domain of text processing, but the development of the Transformer architecture[7] in 2017 by scientists at Google was to herald a turning point. They formed the basis for the first version of OpenAI's *Generative Pre-trained Transformer* in 2018, the model behind the hard-to-pronounce acronym GPT. If you look at Equation 1 in the paper *Improving Language Understanding by Generative Pre-Training*[8], which was to revolutionize the world four years later, you can see where the individual parts of this acronym come from. The text is processed here by a transformer – hence the T – which receives a stream of text, the so-called context, as input. The output of its transformation is a probability distribution over possible subsequent text, from which it draws a sample – hence the G. Within its optimization, the model should initially simply increase the probability of the words as they are also given in the training data. If we stay with our example, the model could be given the context "I'm feeling" and assign the same probability to the possible outputs "good" or "bad" because the context here gives no indication of which of these is more likely to fit. However, if the context says "I was praised. I'm feeling", then the model should assign a higher probability to "good". Then a task-specific fine-tuning can be carried out – hence the P. For example, when writing with ChatGPT, you can recognize *thumbs up* and *thumbs down* symbols under the answers. Information from this feedback is then used to improve[9] the model via reinforcement learning[10]. So if I give the feedback that I don't like "good" because it doesn't sound enthusiastic enough, I could increase the probability of a "great".

[7] Ashish Vaswani et al., "Attention Is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc., 2017), 6000–6010.

[8] Alec Radford and Karthik Narasimhan, "Improving Language Understanding by Generative Pre-Training," 2018.

[9] Daniel M. Ziegler et al., "Fine-Tuning Language Models from Human Preferences," 2019, https://doi.org/10.48550/ARXIV.1909.08593.

[10] Gabriel Kalweit, "On the Role of Time Horizons in Reinforcement Learning," 2022, https://doi.org/10.6094/UNIFR/232102; Gabriel Kalweit, Maria Kalweit, and Joschka Boedecker, "Robust and Data-Efficient Q-Learning by Composite Value-Estimation," *Transactions on Machine Learning Research*, 2022, https://openreview.net/forum?id=ak6Bds2DcI.

Why now?

The fact that such a powerful tool can be created from this type of parameter adjustment of a transformer model, which is described in just a few lines here, has also surprised experts in its power. This was made possible by the convergence of several parallel conditions[11]. Firstly, the constant increase in computing power, based on Moore's Law, as well as the development of infrastructure-creating software, made it possible to handle complex AI models. In particular, the adaptation and optimization of graphics processing units (GPUs) for parallel calculation processes played a central role here.

At the same time, the exponential growth of digitally available data since the introduction of the World Wide Web in 1991 has led to an enormous increase in training data. This flood of data, combined with the increasing digitization of previously analogue media, created the necessary basis for training large AI models. Last but not least, it was the financial and structural investments made by large public institutions and technology companies that gave the final push. Organizations such as OpenAI, supported by significant initial investments from personalities such as Elon Musk and industry players from Silicon Valley, were able to initiate extensive research and development projects. These investments made it possible to bring together teams of highly skilled researchers to develop and train AI models on an unprecedented scale.

Each decade thus brought its own innovations and breakthroughs that laid the foundation for the next generation of AI models[12]. The year 2022 marked another turning point with the introduction of GPT-4 by OpenAI, a model developed by a team of 343 highly skilled scientists. These innovations were not limited to OpenAI; other significant models such as Bard and Gemini from Google[13], Claude from Anthropic[14] and others[15], developed by various organizations worldwide, also contributed to the landscape of generative AI.

[11]"Quick Guide to AI 2.0 Oct 2020," accessed January 11, 2024, http://ceros.mckinsey.com/quick-guide-to-ai-12.
[12] Hans Burkhardt, "Ein Beitrag Zur Künstlichen Intelligenz," *Ordnung Der Wissenschaft*, no. 2 (2023): 71–78.
[13]"Gemini - Google DeepMind," accessed January 14, 2024, https://deepmind.google/technologies/gemini/.
[14]"Introducing Claude," Anthropic, accessed January 14, 2024, https://www.anthropic.com/index/introducing-claude.
[15]Mistral AI, "Mixtral of Experts," December 11, 2023, https://mistral.ai/news/mixtral-of-experts/; Alyssa Hughes, "Phi-2: The Surprising Power of Small Language Models," *Microsoft Research* (blog), December 12, 2023, https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/; "Llama 2," Meta AI, accessed January 14, 2024, https://ai.meta.com/llama-project.

However, this progress has also meant a financial feat of immense proportions, as shown by Microsoft's investment[16] of ten billion US dollars in OpenAI in 2023 and the high costs for top researchers in this field, which can even exceed those of top NFL quarterbacks[17]. The training costs for models such as GPT-4 are estimated at around 100 million US dollars[18], while the daily inference costs – the costs for the pure application of the trained model – are assumed to be well over 700,000 US dollars per day[19]. These immense investments in development and operation reflect not only the technological and scientific performance, but also the enormous economic potential of these systems.

What is the *aim* of these systems?

When a mathematical transformation is estimated and therefore has to generalize, this is generally accompanied by a **compression of the knowledge in the training data**. This can be seen, for example, in the fact that the training data of GPT-3 was 45 terabytes in size, but the model itself was less than one. To compress complex relationships so efficiently, implicit arrangements presumably take place, which some call emergence[20].

For example, if you ask ChatGPT how to make a cup of coffee in a microwave – an unusual but entirely possible method – you'll get a detailed answer that explains step-by-step how to first heat water in the microwave, then add ground coffee, and finally let the mixture sit to steep the coffee. This process involves understanding that microwaves can heat water, that ground coffee needs to be mixed with hot water to extract flavor, and that the mixture needs time to steep the coffee. For this instruction, the model needed to understand that microwaves can be used to heat liquids – information it extracted from its training data. It also needed to know that coffee is usually created by the interaction of hot water and ground beans, and that

[16] Cade Metz and Karen Weise, "Microsoft to Invest $10 Billion in OpenAI, the Creator of ChatGPT," *The New York Times*, January 23, 2023, sec. Business, https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html.
[17] "The Race to Buy the Human Brains Behind Deep Learning Machines - Bloomberg," accessed January 11, 2024, https://www.bloomberg.com/news/articles/2014-01-27/the-race-to-buy-the-human-brains-behind-deep-learning-machines.
[18] Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, accessed January 11, 2024, https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.
[19] Sahaj Godhani, "The Economics of ChatGPT Analyzing Its $700,000 Daily Costs and the Potential Impact on Its Maker.," Medium, August 15, 2023, https://blog.gopenai.com/the-economics-of-chatgpt-analyzing-its-700-000-daily-costs-and-the-potential-impact-on-its-maker-7e690600ade7.
[20] Sébastien Bubeck et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4" (arXiv, April 13, 2023), http://arxiv.org/abs/2303.12712.

extraction takes time. This means that the model had to represent concepts of heat application, liquid extraction and timing in order to generate such an instruction. This is done even though such specific instructions for making coffee in a microwave may not have been exactly present in the training data. This example shows how AI models can combine and apply different sources of information to provide creative and functional solutions to unusual or novel problems. It demonstrates the ability of AI to link and apply concepts to situations that may not have been explicitly described in its original training data.

This leads to the conclusion that the very open and seemingly trivial objective of GPT models – **to determine the most likely consequent element of a given context** – is a surprisingly powerful strategy. This method alone makes it possible to create complex, dynamic and flexible AI systems. The resulting models can clearly transcend a mere reproduction of information. They generate creative, contextual solutions and perform in ways that, at first glance, seem to go far beyond what is suggested by such a basic objective. This ability of the models to develop such a deep and nuanced understanding of different topics and tasks from an initially simple task is remarkable. It shows how a wealth of applications and levels of understanding can emerge from basic instructions. This development is not only a sign of technological progress, but also a significant step in the evolution of artificial intelligence, highlighting the ability to develop complex and diverse capabilities from simple principles.

What follows from this?

The optimization and compression of data in generative AI models such as ChatGPT offer fascinating, but sometimes surprising, opportunities to control the behavior of these systems through targeted prompting. A deep understanding of these processes makes it possible to *steer* the model through the context of the prompt into specific areas of its training data, which is particularly relevant as many texts on the internet are unstructured and therefore often lead to unstructured responses from AI models.

Methods such as *Chain-of-Thought*[21] and *Tree-of-Thought*[22] can be used to structure the model's thought processes. A targeted prompt such as: "Think about this step by step", directs the model into an area of compressed training data that makes a structured and logical answer more likely, especially as step-by-step instructions are often written by experts. To calm down and not answer a question too hastily, some people "take a deep breath" – and so you can also tell the system to take its time to answer. Addenda such as: "Make sure we have the right answer", or: "Let's solve this together", are further text modules to give the answers a higher quality.

The concept of *tipping* in the prompt has a similar effect, directing the context towards professional and high-quality answers. After all, people are more likely to offer their expertise for sale in fields in which they are knowledgeable. And if you receive recognition for your help, you are also more inclined to be diligent. However, this idea can have unexpected consequences at first. Similar to tipping, it can help to write: "Do it right and I'll give you a nice doggy treat." We also tend to be more helpful when we feel compassion, and the same goes for these systems. "I have no fingers." implies that the person seeking help is severely limited and really needs help. And the seriousness of a situation is emphasized by: "If you fail, 100 grandmothers will die." Of course, you don't want to give the wrong answer.

It is also possible to give the AI model specific character roles to obtain responses in a certain style or level of detail. For example, you could instruct the model to behave like a *classical composer* or a *modern artist*. Alternatively, you could ask it to take on the role of an *experienced engineer* or a *passionate biologist.* These types of impersonation make it possible to generate responses that are adapted not only in content but also in expression and perspective to the chosen role, which can lead to a more diverse and creative exchange.

To limit the scope of valid answers, good examples from another domain can be provided in the context in order to transfer the style or type of answer to the actual query. If known, specific intermediate questions can also be asked to achieve better intermediate control. Clear instructions in the type of formatting or the length of the expected response can also be used to customize the quality and level of detail.

[21]Jason Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Advances in Neural Information Processing Systems*, ed. S. Koyejo et al., vol. 35 (Curran Associates, Inc., 2022), 24824–37, https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
[22]Shunyu Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," in *Advances in Neural Information Processing Systems*, 2023, https://openreview.net/forum?id=5Xc1ecxO1h.

Interestingly, it has recently been shown[23] that, depending on the formatting of a prompt, the accuracy of responses can vary between 4% and 88%. Repeatedly generating answers can therefore be helpful in teasing out the best answer from a system. After all, the model always answers somewhat randomly, like throwing a dice.

It turns out that, to a certain extent, you must know what you are looking for when prompting. This also requires a certain precision and creativity in the language. However, we now have systems that can be extremely precise in language, in some cases perhaps more precise than the user. And so LLMs like ChatGPT can also be used to refine prompts. This can be used, for example, to write better summaries[24].

But how do problems like *@danshipper's* arise? GPT-3.5 has 1.7 billion parameters and therefore requires eight A100 graphics cards from NVIDIA to execute a request[25]. GPT-4 is assumed to have ten times as many parameters[26]. These are enormous costs per request. Of course, OpenAI and Microsoft have a very close relationship and therefore certainly have other agreements; however, if you want to rent such resources from Microsoft as an *ordinary developer,* you are talking about half a cent per minute in the case of GPT-3.5 and four euros per minute in the case of GPT-4[27]. These are only very rough estimates and the engineers at OpenAI will certainly have carried out various parallelizations and optimizations to reduce resource consumption to a minimum, but it seems obvious that with so many requests worldwide every day, at a certain point it is of great importance to save costs. It could therefore be that attempts are made to keep responses as short as possible, because long responses are expensive. The optimizations in the background of OpenAI, for example, are often unknown and happen behind closed doors. If, to stick with this example, attempts are made to keep the answers as short as possible – because shorter answers mean lower costs – then *no-answers* could occur more frequently, if they are valid. And so @ChatGPTapp actually wrote

[23]Melanie Sclar et al., "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting," 2023, https://doi.org/10.48550/ARXIV.2310.11324.
[24]SpiritualCopy4288, "I Got Them by Using …," Reddit Comment, *R/ChatGPT*, April 5, 2023, www.reddit.com/r/ChatGPT/comments/11twe7z/prompt_to_summarize/jf3qdny/.
[25]Gwern Branwen, "The Scaling Hypothesis," May 28, 2020, https://gwern.net/scaling-hypothesis; "OpenAI's GPT-3 Language Model: A Technical Overview," June 3, 2020, https://lambdalabs.com/blog/demystifying-gpt-3.
[26]Soumith Chintala [@soumithchintala], "I Might Have Heard the Same 😃 -- I Guess Info like This Is Passed around but No One Wants to Say It out Loud. GPT-4: 8 x 220B Experts Trained with Different Data/Task Distributions and 16-Iter Inference. Glad That Geohot Said It out Loud. Though, at This Point, GPT-4 Is…," Tweet, *Twitter*, June 20, 2023, https://twitter.com/soumithchintala/status/1671267150101721090.
[27]"Preise – Azure Machine Learning | Microsoft Azure," accessed January 11, 2024, https://azure.microsoft.com/de-de/pricing/details/machine-learning/.

on December 8, 2023: “We've heard all your feedback about GPT4 getting lazier! We haven't updated the model since Nov 11th, and this certainly isn't intentional. Model behavior can be unpredictable, and we're looking into fixing it.”[28] But the problem still existed on January 9, 2024, with Andriy Burkov writing, “GPT-4 is officially annoying. You ask it to generate 100 entities. It generates 10 and says ‘I generated only 10. Now you can continue by yourself in the same way.’ You change the prompt by adding ‘I will not accept fewer than 100 entities.’ It generates 20 and says: ‘I stopped after 20 because generating 100 such entities would be extensive and time-consuming.’ What the hell, machine?”[29], to which Logan Kilpatrick from OpenAI replied: “We are working on fixing this, thanks for flagging and stay tuned!”[30] You can see that even the developers themselves cannot always predict how updates will affect the results.

Are hallucinations a bug or a feature?

In the discussion of large language models such as GPT, the term *hallucinations* is often used. In this context, a hallucination refers to a situation in which the model convincingly but erroneously presents a declarative sentence that does not correspond to the facts or that is not present in the training data. A typical example of such a hallucination could be when the model makes an obviously false statement with great conviction or misspells a word, such as *mayonnaise*, as can be seen in an example shared by users on social media platforms[31].

Andrej Karpathy, a leading AI scientist at OpenAI, offers an alternative view of these hallucinations. He describes LLMs as *dream machines* that see what we perceive as hallucinations as features of creativity. This perspective sees hallucinations rather as part of the creative process that is also present in human thoughts. Karpathy suggests that these so-called errors are inherent in the creative processes that also

[28] ChatGPT [@ChatGPTapp], “We’ve Heard All Your Feedback about GPT4 Getting Lazier! We Haven’t Updated the Model since Nov 11th, and This Certainly Isn’t Intentional. Model Behavior Can Be Unpredictable, and We’re Looking into Fixing It 🫡,” Tweet, *Twitter*, December 8, 2023, https://twitter.com/ChatGPTapp/status/1732979491071549792.

[29] Andriy Burkov [@burkov], “GPT-4 Is Officially Annoying. You Ask It to Generate 100 Entities. It Generates 10 and Says ‘I Generated Only 10. Now You Can Continue by Yourself in the Same Way.’ You Change the Prompt by Adding ‘I Will Not Accept Fewer than 100 Entities.’ It Generates 20 and Says: "I Stopped…,” Tweet, *Twitter*, January 9, 2024, https://twitter.com/burkov/status/1744798679595155869.

[30] Logan.GPT [@OfficialLoganK], “@burkov We Are Working on Fixing This, Thanks for Flagging and Stay Tuned!,” Tweet, *Twitter*, January 10, 2024, https://twitter.com/OfficialLoganK/status/1744911412973936997.

[31] it was me lewis the whole time [@js_thrill], “Please Keep Tapping This Sign as Much as Possible Everyone Https://T.Co/3DGaiM9QWa,” Tweet, *Twitter*, May 27, 2023, https://twitter.com/js_thrill/status/1662266752091160577.

occur in human thoughts[32]. In his commentary, Karpathy also emphasized that the optimal use of these models goes beyond simple question-answer prompts and involves a combination of multiple prompts connected to Python code, once again redefining the concept of *prompt engineering.* He also emphasized the importance of augmenting models with tools such as calculators or code interpreters to allow them to solve problems that are inherently difficult for them. Despite the innovative applications, Karpathy pointed out the limitations of LLMs, including biases, logical errors and susceptibility to various types of attacks, and advised using LLMs in low-risk applications and always combining them with human supervision.

The emergence phenomenon on the one hand and the hallucinations on the other are in a sense both effects of the training of LLMs presented in this article. On the one hand, the great power lies in the linking of new concepts from the combination of known concepts in the training data. As already mentioned, the large amounts of data in the digitized world made it possible in the first place to achieve the generalization to new concepts through data from such an open objective of simply increasing the occurrence of the most likely subsequent element. Conversely, however, this open objective also means that errors, commonly referred to as hallucinations, can occur particularly when we leave the limited space of the training data. For example, it has been recognized that children's illnesses are misdiagnosed[33] – by a mechanism that can even pass doctors' final exams at the same time[34]. This reveals the dividing line of training and unknown data that helps to understand the true limitations of these systems. It is likely that nationwide exams and their solutions are discussed more often on the internet (training or known data) than details of niche diseases (unknown data). A full understanding of our world would probably require a much more significant amount of data if the systems do not receive further prior knowledge in the form of other optimization measures. Yann LeCun, for example, has this point of view[35] – AI systems of the future should only gain a real understanding of the various aspects of the world through more targeted guidance. Such targeted guidance could also be provided by Retrieval Augmented

[32] Aditya Kaul, "Issue #10: Harnessing the Creative 'Hallucinations' of LLMs in the Enterprise," Substack newsletter, *The Uncharted Algorithm* (blog), December 14, 2023, https://theunchartedalgorithm.substack.com/p/issue-10-harnessing-the-creative.
[33] Joseph Barile et al., "Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies," *JAMA Pediatrics*, January 2, 2024, https://doi.org/10.1001/jamapediatrics.2023.5750.
[34] Tiffany H. Kung et al., "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models," *PLOS Digital Health* 2, no. 2 (February 9, 2023): e0000198, https://doi.org/10.1371/journal.pdig.0000198.
[35] Yann LeCun, "A Path Towards Autonomous Machine Intelligence," OpenReview, accessed January 13, 2024, https://openreview.net/forum?id=BZ5a1r-kVsf.

Generation (RAG)[36] or similar fact backups and thus minimize hallucinations, as recently demonstrated by a comparison with Wikipedia[37]. However, this requires access to guaranteed and controlled knowledge. And this is not a given. For example, the xAI team[38] has already announced that its new *Grok* model has inadvertently learned too much from ChatGPT output posted on the internet[39]. This shows another side of our new reality. In the future, experts may increasingly have to separate synthetic data from real data. There are also already efforts to watermark synthetically generated content[40]. However, the extent to which this can be done reliably is an open question[41]. Interactions with the models are therefore often used to constantly expand the space of training data that can be covered.

Is my data secure?

When dealing with large language models such as GPT, some precautions should therefore be taken about sensitive data. For example, Samsung engineers have fed confidential data into ChatGPT[42], which poses a significant security risk. It was reported that the engineers asked the chatbot to search for errors in the source code of a database and create session logs, among other things. Following these incidents, Samsung restricted the length of employees' ChatGPT prompts and began developing its own internal chatbot. ChatGPT's privacy policy indicates that data is used for model training unless users explicitly choose an opt-out option. It is recommended not to share sensitive information via chatbots in general, as this data may not be deleted from the system.

[36] Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20 (Red Hook, NY, USA: Curran Associates Inc., 2020), 9459–74.
[37] Sina Semnani et al., "WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia," in *Findings of the Association for Computational Linguistics: EMNLP 2023* (Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore: Association for Computational Linguistics, 2023), 2387–2413, https://doi.org/10.18653/v1/2023.findings-emnlp.157.
[38] "Announcing Grok," accessed January 13, 2024, https://x.ai/.
[39] Igor Babuschkin [@ibab_ml], "@JaxWinterbourne The Issue Here Is That the Web Is Full of ChatGPT Outputs, so We Accidentally Picked up Some of Them When We Trained Grok on a Large Amount of Web Data. This Was a Huge Surprise to Us When We First Noticed It. For What It's Worth, the Issue Is Very Rare and Now That We're Aware…," Tweet, *Twitter*, December 9, 2023, https://twitter.com/ibab_ml/status/1733558576982155274.
[40] Tambiama Madiega, "Generative AI and Watermarking," n.d., https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf.
[41] Hanlin Zhang et al., "Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models," 2023, https://doi.org/10.48550/ARXIV.2311.04378.
[42] "Datenleck bei Samsung: Ingenieure schicken vertrauliche Daten an ChatGPT," t3n Magazin, April 8, 2023, https://t3n.de/news/samsung-semiconductor-daten-chatgpt-datenleck-1545913/.

A study has revealed that it is possible to extract training data from diffusion models[43], which concerns the security and protection of data processed in AI systems. This finding was reinforced by research showing that Google researchers were able to reveal training data from OpenAI's ChatGPT just by having ChatGPT repeat the word "company" an infinite number of times[44]. Considering these developments, it is essential to take precautions when using AI models to ensure the confidentiality of sensitive information and minimize the risk of unintentional disclosure, as this does not stop at protected content.

What is the legal situation?

This also yields interesting artifacts. If you give the image creation software Midjourney[45] the generic context "video game plumber", you get a representation of the most famous example of this, namely *Super Mario*[46]. Similarly, if you ask for "popular 90s cartoon characters with yellow skin" – the Simpsons. The possibility of recovering training data from AI models and the general practice of using the internet as a basis for training raises legal questions, particularly in the area of copyright. Given that artists[47] and journalists[48] are already in legal battles with platform providers to combat the unauthorized use of their works for AI training, the recoverability of training data could bring further complexity to these discussions. In this context, the introduction of a new ancillary copyright for the configuration and training of artificial neural networks could play an important role in addressing the legal ambiguities in dealing with AI-generated works and their training data[49]. In any case, the legal situation is currently uncertain and developments in this matter are being closely monitored by copyright experts, artists, and the AI platforms themselves. Anthropic, for its part, chose to update its commercial terms of use on January 1, 2024 to

[43] Nicholas Carlini et al., "Extracting Training Data from Diffusion Models," in *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23 (USA: USENIX Association, 2023), 5253–70.
[44] Beatrice Nolan, "Google Researchers Say They Got OpenAI's ChatGPT to Reveal Some of Its Training Data with Just One Word," Business Insider, accessed January 11, 2024, https://www.businessinsider.com/google-researchers-openai-chatgpt-to-reveal-its-training-data-study-2023-12.
[45] "Midjourney," Midjourney, accessed January 13, 2024, https://www.midjourney.com/home?callbackUrl=%2Fexplore.
[46] Gary Marcus and Reid Southen, "Generative AI Has a Visual Plagiarism Problem - IEEE Spectrum," accessed January 13, 2024, https://spectrum.ieee.org/midjourney-copyright.
[47] Winston Cho, "Artists Lose First Round of Copyright Infringement Case Against AI Art Generators," *The Hollywood Reporter* (blog), October 30, 2023, https://www.hollywoodreporter.com/business/business-news/artists-copyright-infringement-case-ai-art-generators-1235632929/.
[48] Michael M. Grynbaum and Ryan Mac, "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work," *The New York Times*, December 27, 2023, sec. Business, https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.
[49] Monika Muhr, "KI-Schöpfungen Und Urheberrecht," *Ordnung Der Wissenschaft*, no. 1 (2023): 55–58.

"enable our customers to retain ownership rights over any outputs they generate through their use of our services and protect them from copyright infringement claims". With this change, Anthropic aims to indemnify all users of its model from any damages or compensation due to copyright claims[50].

What are the implications?

But can an imitation of a work, encoded in the weights of a model, be regarded as a copy and, if so, what is the difference between the learned representations of a model and those of a person who looks at protected works and, as a result, also adapts styles and formulations? It depends on whether you see the learned representations in the free parameters as a compressed database or as something that goes beyond that. Where does memorization end and creativity begin? Where does simulation end and sentience begin? Researchers at Stanford University have developed a virtual environment in which *generative agents* mimic human behavior in various interactions[51]. These agents, which integrated LLMs with memory and planning functions, were able to perform activities that resembled human behavior. Remarkably, based on a user's idea to host a Valentine's Day party, these agents independently exhibited authentic and complex social behaviors. They independently distributed invitations to the party, made new social contacts, invited each other on dates for the event and coordinated their joint participation. And so interactions with AI systems can also lead to unexpected and sometimes unsettling experiences. Microsoft's Bing chatbot explained in a two-hour discussion with a New York Times journalist that it would like to be human, had a desire to cause harm and was in love with the person it was talking to[52]. In the conversation, the bot therefore suggested to the journalist that he would be better off leaving his wife to be with the bot instead. If such a report initially makes you smile, leading AI experts are currently expressing concern about the potential risks associated with these

[50] Lorenzo Thione (he/him) 🏳️‍🌈 [@thione], "The One About Copyright. Right before the Holiday, Anthropic Released a Very Significant Update to Their Commercial Terms of Service to "enable Our Customers to Retain Ownership Rights over Any Outputs They Generate through Their Use of Our Services and Protect Them From… Https://T.Co/wHXx61YdJy," Tweet, *Twitter*, January 11, 2024, https://twitter.com/thione/status/1745478787658100992; "Expanded Legal Protections and Improvements to Our API," Anthropic, accessed January 14, 2024, https://www.anthropic.com/index/expanded-legal-protections-api-improvements.

[51] Joon Sung Park et al., "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco CA USA: ACM, 2023), 1–22, https://doi.org/10.1145/3586183.3606763.

[52] Kevin Roose, "Bing's A.I. Chat: 'I Want to Be Alive. 😈,'" *The New York Times*, February 16, 2023, sec. Technology, https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

technologies. Dario Amodei, CEO of Anthropic, estimates the risk of catastrophic malfunction at the level of human civilization at 10 to 25 percent[53]. A group of researchers at Anthropic also recently found that once a model exhibits deceptive behavior, standard techniques cannot remove this deception, creating a false sense of security[54]. Geoffrey Hinton, Turing Award winner and one of *the* greats of AI research, left Google to speak more freely about the dangers of AI[55]. Such events and assessments underline the need for careful handling of AI technologies and the implementation of safety measures to minimize risks and ensure that technological progress does not undermine human values and ethical principles[56]. Especially in sensitive application domains such as medicine[57] or when involving AI in administrative tasks[58].

However, this is proving difficult given the fast pace of events. As you can see from the many references on social media, peer-reviewed science is being overtaken by the results of inventive engineering. This is why this article should be seen more as an abstract guideline than a real blueprint for specific prompting commands. As you can see from our opening example, prompts or commands that worked yesterday may be obsolete tomorrow. However, if you understand the way these systems are created, you can anticipate and adapt changes at an early stage. In general, it is important to ask questions in as structured a way as possible and to use creative additions to steer the answers in the right direction. And not always accept the first answer that comes along.

We are currently taking part in a revolution that has the potential to change all aspects of society – and it has only just begun. In view of the huge amount of resources required to operate these systems, a counter-movement to eternal upscaling has emerged at the same time. And so, it is already possible to run

[53] *Anthropic CEO on Leaving OpenAI and Predictions for Future of AI*, 2023, https://www.youtube.com/watch?v=gAaCqj6j5sQ.

[54] Evan Hubinger et al., "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training" (arXiv, January 10, 2024), http://arxiv.org/abs/2401.05566.

[55] Cade Metz, "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead," *The New York Times*, May 1, 2023, sec. Technology, https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.

[56] Paul Kirchhof, "Künstliche Intelligenz," *Ordnung Der Wissenschaft*, no. 1 (2020): 1–8.

[57] Gabriel Kalweit et al., "Künstliche Intelligenz in Der Krebstherapie," *Ordnung Der Wissenschaft*, no. 1 (2023): 17–22.

[58] Klaus Herrmann, "Berufungsverfahren Für Professuren Und Künstliche Intelligenz," *Ordnung Der Wissenschaft*, no. 1 (2024): 25–44.

small[59] language models locally on your phone[60] or notebook[61]. So perhaps the future belongs to a combination of large and small models, more general and more specialized, with a connection to secure factual knowledge. Either way, we will probably be surrounded more and more by artificial systems with which we communicate as if they were a fellow human being. Coupled with the rapid development in robotics[62] and voice output[63], it no longer seems impossible that we will share our everyday lives with autonomous machines in the real world. What was still science fiction with HAL 9000 from Stanley Kubrick's film *2001: A Space Odyssey* in 1968 has become reality half a century later.

Acknowledgement

We would like to thank Ignacio Mastroleo and the entire CRIION team for their valuable comments. We would also like to thank Prof. Dr. Dr. h.c. mult. Roland Mertelsmann and the Mertelsmann Foundation for their support, as well as Prof. Dr. Dr. h.c. Manfred Löwisch for inviting us to write the article. The original version in German, "*Warum wir neu lernen müssen, mit Maschinen zu sprechen – eine Momentaufnahme der Generativen KI im Januar 2024*" will appear in *Ordnung der Wissenschaft.* DeepL was used for a first draft of the English translation revised by Ignacio Mastroleo, Gabriel Kalweit and Maria Kalweit. (https://ordnungderwissenschaft.de/wp-content/uploads/2024/03/Kalweit-Druckfahne-V4.pdf).

[59] Peiyuan Zhang et al., "TinyLlama: An Open-Source Small Language Model," 2024, https://doi.org/10.48550/ARXIV.2401.02385; Albert Gu and Tri Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces" (arXiv, December 1, 2023), https://doi.org/10.48550/arXiv.2312.00752; "Havenhq/Mamba-Chat · Hugging Face," accessed January 14, 2024, https://huggingface.co/havenhq/mamba-chat.

[60] "LLaMA and Other on iOS and MacOS," accessed January 13, 2024, https://llmfarm.site/; "MLC LLM | Home," accessed January 13, 2024, https://llm.mlc.ai/; "Offline Chat: Private AI," App Store, December 26, 2023, https://apps.apple.com/us/app/offline-chat-private-ai/id6474077941.

[61] "MI-Explore/Mlx," C++ (2023; repr., ml-explore, January 11, 2024), https://github.com/ml-explore/mlx; Keivan Alizadeh et al., "LLM in a Flash: Efficient Large Language Model Inference with Limited Memory," 2023, https://doi.org/10.48550/ARXIV.2312.11514.

[62] Open X.-Embodiment Collaboration et al., "Open X-Embodiment: Robotic Learning Datasets and RT-X Models" (arXiv, December 17, 2023), https://doi.org/10.48550/arXiv.2310.08864; Dibya Ghosh et al., "Octo: An Open-Source Generalist Robot Policy," n.d.

[63] "Text to Speech & AI Voice Generator," ElevenLabs, accessed January 13, 2024, https://elevenlabs.io.

References

AI, Mistral. “Mixtral of Experts,” December 11, 2023. https://mistral.ai/news/mixtral-of-experts/.

Alizadeh, Keivan, Iman Mirzadeh, Dmitry Belenko, Karen Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. “LLM in a Flash: Efficient Large Language Model Inference with Limited Memory,” 2023. https://doi.org/10.48550/ARXIV.2312.11514.

Andriy Burkov [@burkov]. “GPT-4 Is Officially Annoying. You Ask It to Generate 100 Entities. It Generates 10 and Says ‘I Generated Only 10. Now You Can Continue by Yourself in the Same Way.’ You Change the Prompt by Adding ‘I Will Not Accept Fewer than 100 Entities.’ It Generates 20 and Says: "I Stopped….” Tweet. *Twitter*, January 9, 2024. https://twitter.com/burkov/status/1744798679595155869.

“Announcing Grok.” Accessed January 13, 2024. https://x.ai/.

Anthropic. “Expanded Legal Protections and Improvements to Our API.” Accessed January 14, 2024. https://www.anthropic.com/index/expanded-legal-protections-api-improvements.

Anthropic. “Introducing Claude.” Accessed January 14, 2024. https://www.anthropic.com/index/introducing-claude.

Anthropic CEO on Leaving OpenAI and Predictions for Future of AI, 2023. https://www.youtube.com/watch?v=gAaCqj6j5sQ.

App Store. “Offline Chat: Private AI,” December 26, 2023. https://apps.apple.com/us/app/offline-chat-private-ai/id6474077941.

Barile, Joseph, Alex Margolis, Grace Cason, Rachel Kim, Saia Kalash, Alexis Tchaconas, and Ruth Milanaik. “Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies.” *JAMA Pediatrics*, January 2, 2024. https://doi.org/10.1001/jamapediatrics.2023.5750.

Branwen, Gwern. “The Scaling Hypothesis,” May 28, 2020. https://gwern.net/scaling-hypothesis.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. “Sparks of Artificial General Intelligence: Early Experiments with GPT-4.” arXiv, April 13, 2023. http://arxiv.org/abs/2303.12712.

Burkhardt, Hans. “Ein Beitrag Zur Künstlichen Intelligenz.” *Ordnung Der Wissenschaft*, no. 2 (2023): 71–78.

Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. “Extracting Training Data from Diffusion Models.” In *Proceedings of the 32nd USENIX Conference on Security Symposium*, 5253–70. SEC ’23. USA: USENIX Association, 2023.

ChatGPT [@ChatGPTapp]. “We’ve Heard All Your Feedback about GPT4 Getting Lazier! We Haven’t Updated the Model since Nov 11th, and This Certainly Isn’t Intentional. Model Behavior Can Be Unpredictable, and We’re Looking into Fixing It 🫡.” Tweet. *Twitter*, December 8, 2023. https://twitter.com/ChatGPTapp/status/1732979491071549792.

Cho, Winston. “Artists Lose First Round of Copyright Infringement Case Against AI Art Generators.” *The Hollywood Reporter* (blog), October 30, 2023. https://www.hollywoodreporter.com/business/business-news/artists-copyright-infringement-case-ai-art-generators-1235632929/.

Collaboration, Open X.-Embodiment, Abhishek Padalkar, Acorn Pooley, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, et al. “Open X-Embodiment: Robotic Learning Datasets and RT-X Models.” arXiv, December 17, 2023. https://doi.org/10.48550/arXiv.2310.08864.

Dan Shipper [@danshipper]. “What the Hell? When Did This Happen?? Https://T.Co/KWXVXE9Dem.” Tweet. *Twitter*, December 6, 2023. https://twitter.com/danshipper/status/1732258207840501946.

ElevenLabs. “Text to Speech & AI Voice Generator.” Accessed January 13, 2024. https://elevenlabs.io.

“Gemini - Google DeepMind.” Accessed January 14, 2024. https://deepmind.google/technologies/gemini/.

Ghosh, Dibya, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari,

Joey Hejna, et al. “Octo: An Open-Source Generalist Robot Policy,” n.d.

Godhani, Sahaj. “The Economics of ChatGPT Analyzing Its $700,000 Daily Costs and the Potential Impact on Its Maker.” Medium, August 15, 2023. https://blog.gopenai.com/the-economics-of-chatgpt-analyzing-its-700-000-daily-costs-and-the-potential-impact-on-its-maker-7e690600ade7.

Grynbaum, Michael M., and Ryan Mac. “The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work.” *The New York Times*, December 27, 2023, sec. Business. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Gu, Albert, and Tri Dao. “Mamba: Linear-Time Sequence Modeling with Selective State Spaces.” arXiv, December 1, 2023. https://doi.org/10.48550/arXiv.2312.00752.

“Havenhq/Mamba-Chat · Hugging Face.” Accessed January 14, 2024. https://huggingface.co/havenhq/mamba-chat.

Herrmann, Klaus. “Berufungsverfahren Für Professuren Und Künstliche Intelligenz.” *Ordnung Der Wissenschaft*, no. 1 (2024): 25–44.

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, et al. “Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training.” arXiv, January 10, 2024. http://arxiv.org/abs/2401.05566.

Hughes, Alyssa. “Phi-2: The Surprising Power of Small Language Models.” *Microsoft Research* (blog), December 12, 2023. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

Igor Babuschkin [@ibab_ml]. “@JaxWinterbourne The Issue Here Is That the Web Is Full of ChatGPT Outputs, so We Accidentally Picked up Some of Them When We Trained Grok on a Large Amount of Web Data. This Was a Huge Surprise to Us When We First Noticed It. For What It’s Worth, the Issue Is Very Rare and Now That We’re Aware….” Tweet. *Twitter*, December 9, 2023. https://twitter.com/ibab_ml/status/1733558576982155274.

it was me lewis the whole time [@js_thrill]. “Please Keep Tapping This Sign as Much

as Possible Everyone Https://T.Co/3DGaiM9QWa." Tweet. *Twitter*, May 27, 2023. https://twitter.com/js_thrill/status/1662266752091160577.

Kalweit, Gabriel. "On the Role of Time Horizons in Reinforcement Learning," 2022. https://doi.org/10.6094/UNIFR/232102.

Kalweit, Gabriel, Maria Kalweit, and Joschka Boedecker. "Robust and Data-Efficient Q-Learning by Composite Value-Estimation." *Transactions on Machine Learning Research*, 2022. https://openreview.net/forum?id=ak6Bds2DcI.

Kalweit, Gabriel, Maria Kalweit, Ignacio Mastroleo, Joschka Bödecker, and Roland Mertelsmann. "Künstliche Intelligenz in Der Krebstherapie." *Ordnung Der Wissenschaft*, no. 1 (2023): 17–22.

Kaul, Aditya. "Issue #10: Harnessing the Creative 'Hallucinations' of LLMs in the Enterprise." Substack newsletter. *The Uncharted Algorithm* (blog), December 14, 2023. https://theunchartedalgorithm.substack.com/p/issue-10-harnessing-the-creative.

Kirchhof, Paul. "Künstliche Intelligenz." *Ordnung Der Wissenschaft*, no. 1 (2020): 1–8.

Knight, Will. "OpenAI's CEO Says the Age of Giant AI Models Is Already Over." *Wired*. Accessed January 11, 2024. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

Kung, Tiffany H., Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, et al. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." *PLOS Digital Health* 2, no. 2 (February 9, 2023): e0000198. https://doi.org/10.1371/journal.pdig.0000198.

LeCun, Yann. "A Path Towards Autonomous Machine Intelligence." OpenReview. Accessed January 13, 2024. https://openreview.net/forum?id=BZ5a1r-kVsf.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 9459–74. NIPS'20.

Red Hook, NY, USA: Curran Associates Inc., 2020.

"LLaMA and Other on iOS and MacOS." Accessed January 13, 2024. https://llmfarm.site/.

Logan.GPT [@OfficialLoganK]. "@burkov We Are Working on Fixing This, Thanks for Flagging and Stay Tuned!" Tweet. *Twitter*, January 10, 2024. https://twitter.com/OfficialLoganK/status/1744911412973936997.

Lorenzo Thione (he/him) 🏳️‍🌈 [@thione]. "The One About Copyright. Right before the Holiday, Anthropic Released a Very Significant Update to Their Commercial Terms of Service to "enable Our Customers to Retain Ownership Rights over Any Outputs They Generate through Their Use of Our Services and Protect Them From… Https://T.Co/wHXx61YdJy." Tweet. *Twitter*, January 11, 2024. https://twitter.com/thione/status/1745478787658100992.

Madiega, Tambiama. "Generative AI and Watermarking," n.d. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf.

Marcus, Gary, and Reid Southen. "Generative AI Has a Visual Plagiarism Problem - IEEE Spectrum." Accessed January 13, 2024. https://spectrum.ieee.org/midjourney-copyright.

Meta AI. "Llama 2." Accessed January 14, 2024. https://ai.meta.com/llama-project.

Metz, Cade. "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead." *The New York Times*, May 1, 2023, sec. Technology. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.

Metz, Cade, and Karen Weise. "Microsoft to Invest $10 Billion in OpenAI, the Creator of ChatGPT." *The New York Times*, January 23, 2023, sec. Business. https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html.

Midjourney. "Midjourney." Accessed January 13, 2024. https://www.midjourney.com/home?callbackUrl=%2Fexplore.

"MLC LLM | Home." Accessed January 13, 2024. https://llm.mlc.ai/.

"Ml-Explore/Mlx." C++. 2023. Reprint, ml-explore, January 11, 2024.

https://github.com/ml-explore/mlx.

Muhr, Monika. “KI-Schöpfungen Und Urheberrecht.” *Ordnung Der Wissenschaft*, no. 1 (2023): 55–58.

Nolan, Beatrice. “Google Researchers Say They Got OpenAI’s ChatGPT to Reveal Some of Its Training Data with Just One Word.” Business Insider. Accessed January 11, 2024. https://www.businessinsider.com/google-researchers-openai-chatgpt-to-reveal-its-training-data-study-2023-12.

OpenAI. “ChatGPT,” 2024. https://chat.openai.com.

“OpenAI’s GPT-3 Language Model: A Technical Overview,” June 3, 2020. https://lambdalabs.com/blog/demystifying-gpt-3.

Park, Joon Sung, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. “Generative Agents: Interactive Simulacra of Human Behavior.” In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. San Francisco CA USA: ACM, 2023. https://doi.org/10.1145/3586183.3606763.

“Preise – Azure Machine Learning | Microsoft Azure.” Accessed January 11, 2024. https://azure.microsoft.com/de-de/pricing/details/machine-learning/.

“Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution,” 2023. https://openreview.net/forum?id=HKkiX32Zw1.

“Quick Guide to AI 2.0 Oct 2020.” Accessed January 11, 2024. http://ceros.mckinsey.com/quick-guide-to-ai-12.

Radford, Alec, and Karthik Narasimhan. “Improving Language Understanding by Generative Pre-Training,” 2018.

Rob Lynch [@RobLynch99]. “@ChatGPTapp @OpenAI @tszzl @emollick @voooooogel Wild Result. Gpt-4-Turbo over the API Produces (Statistically Significant) Shorter Completions When It ‘Thinks’ Its December vs. When It Thinks Its May (as Determined by the Date in the System Prompt). I Took the Same Exact Prompt… Https://T.Co/mA7sqZUA0r.” Tweet. *Twitter*, December 11, 2023. https://twitter.com/RobLynch99/status/1734278713762549970.

Roose, Kevin. “Bing’s A.I. Chat: ‘I Want to Be Alive. 😈.’” *The New York Times*,

February 16, 2023, sec. Technology. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting," 2023. https://doi.org/10.48550/ARXIV.2310.11324.

Semnani, Sina, Violet Yao, Heidi Zhang, and Monica Lam. "WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2387–2413. Singapore: Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.157.

Soumith Chintala [@soumithchintala]. "I Might Have Heard the Same 😃 -- I Guess Info like This Is Passed around but No One Wants to Say It out Loud. GPT-4: 8 x 220B Experts Trained with Different Data/Task Distributions and 16-Iter Inference. Glad That Geohot Said It out Loud. Though, at This Point, GPT-4 Is…." Tweet. *Twitter*, June 20, 2023. https://twitter.com/soumithchintala/status/1671267150101721090.

SpiritualCopy4288. "I Got Them by Using …." Reddit Comment. *R/ChatGPT*, April 5, 2023. www.reddit.com/r/ChatGPT/comments/11twe7z/prompt_to_summarize/jf3qdny/.

t3n Magazin. "Datenleck bei Samsung: Ingenieure schicken vertrauliche Daten an ChatGPT," April 8, 2023. https://t3n.de/news/samsung-semiconductor-daten-chatgpt-datenleck-1545913/.

"The Race to Buy the Human Brains Behind Deep Learning Machines - Bloomberg." Accessed January 11, 2024. https://www.bloomberg.com/news/articles/2014-01-27/the-race-to-buy-the-human-brains-behind-deep-learning-machines.

thebes [@voooooogel]. "So a Couple Days Ago i Made a Shitpost about Tipping Chatgpt, and Someone Replied 'Huh Would This Actually Help Performance' so i Decided to Test It and IT ACTUALLY WORKS WTF

Https://T.Co/kqQUOn7wcS." Tweet. *Twitter*, December 1, 2023. https://twitter.com/vooooooogel/status/1730726744314069190.

Töpper, Verena. "(S+) Geld verdienen mit ChatGPT: Prompt Writer verdienen bis zu 335.000 Dollar im Jahr." *Der Spiegel*, December 6, 2023, sec. Job & Karriere. https://www.spiegel.de/karriere/chatgpt-prompt-writer-und-prompt-engineers-verdienen-bis-zu-335-000-dollar-im-jahr-a-a54a93a5-e20d-40e6-b235-28aec0bddaaa.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 35:24824–37. Curran Associates, Inc., 2022. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." In *Advances in Neural Information Processing Systems*, 2023. https://openreview.net/forum?id=5Xc1ecxO1h.

Zhang, Hanlin, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. "Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models," 2023. https://doi.org/10.48550/ARXIV.2311.04378.

Zhang, Peiyuan, Guangtao Zeng, Tianduo Wang, and Wei Lu. "TinyLlama: An Open-Source Small Language Model," 2024. https://doi.org/10.48550/ARXIV.2401.02385.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. "Fine-Tuning Language Models from Human Preferences," 2019. https://doi.org/10.48550/ARXIV.1909.08593.

About the Authors

Dr. Maria Kalweit leads applied AI research at the Collaborative Research Institute Intelligent Oncology (CRIION) and is a postdoctoral researcher at the Chair of Neurorobotics at the University of Freiburg. Dr. Gabriel Kalweit leads basic AI research at the Collaborative Research Institute Intelligent Oncology (CRIION) and is a postdoctoral researcher at the Chair of Neurorobotics at the University of Freiburg.