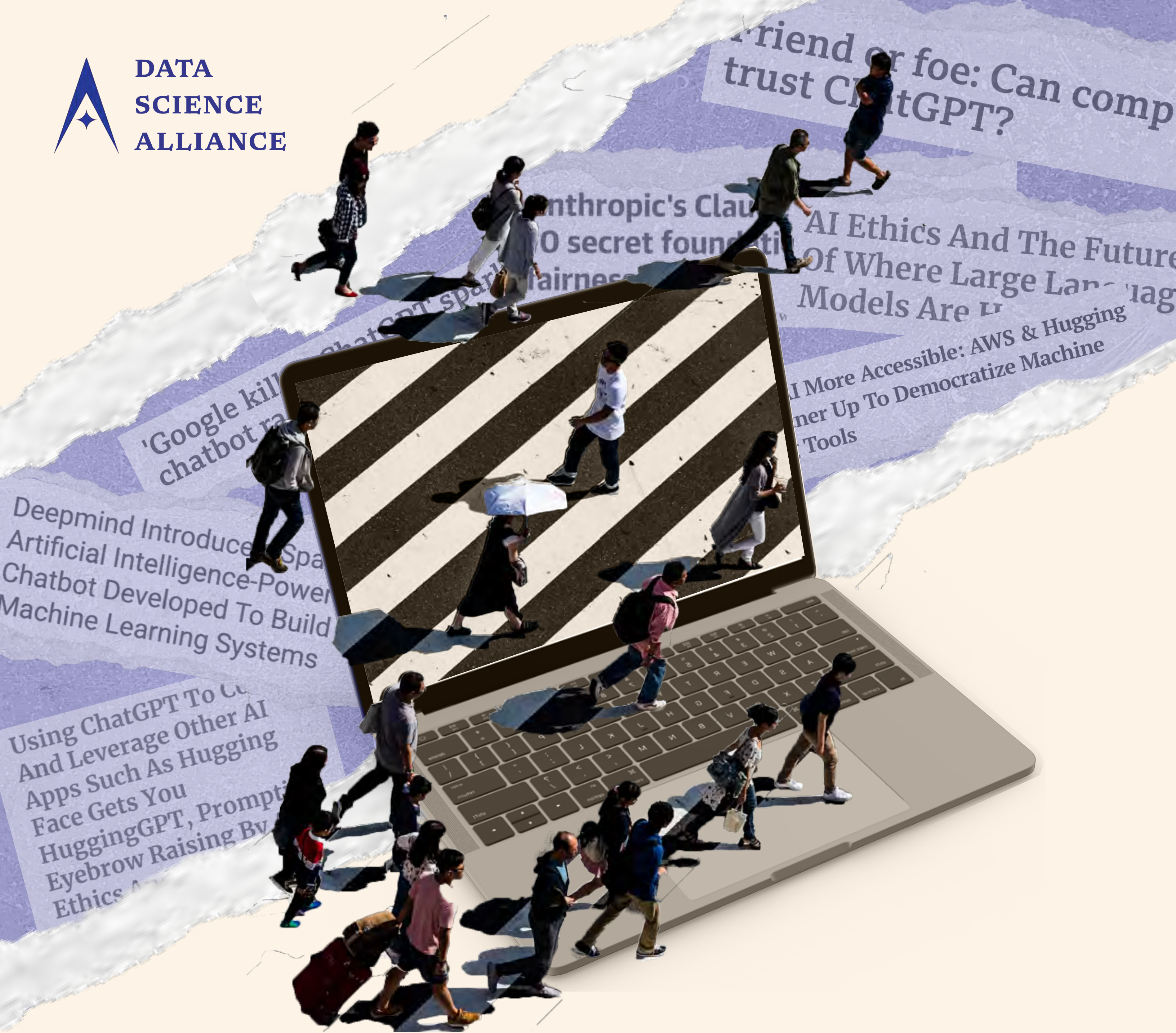
DATA
SCIENCE
ALLIANCE



WHITE PAPER

Breaking the Technological Fourth Wall:
The Democratization of AI

Written By

Di Le AI/ML Design Strategist | ServiceNow

Adir Mancebo Jr., Ph.D. Lead Data Scientist | Data Science Alliance

Contributors

Czarina Argana, Michael Limbo, and Patricia Lopez

INTRODUCTION

AI and the Technological Fourth Wall

The advent of artificial intelligence (AI) has been likened to the Industrial Revolution, as it promises to transform not just the business world but various aspects of our daily lives. However, unlike the Industrial Revolution, which was a collective effort with clear beneficiaries and victims, the initial AI revolution can be thought of more like the film industry — a small group of innovators leading the charge, driving change that ripples throughout culture, society, and business. We are entering a new phase of the AI revolution, where the general populace are not just the audience of this unfolding technology, but also active participants.

In this context, it is intriguing to draw a parallel with Jean-Luc Godard's groundbreaking 1965 film, "Pierrot le Fou," as it allows us to explore the concept of breaking the "fourth wall" in both cinema and AI. Godard's fourth wall represents the invisible barrier between actors and the audience. By having the protagonist, Pierrot, directly address the audience, Godard invites viewers into a conversation with the character [18].

A similar barrier in AI accessibility exists where AI technologies, traditionally reserved for experts and researchers, were out of reach from the general public. However, generative AI based on large language models (LLMs), like ChatGPT and Dall-E, have fostered advancements that build a more dynamic and interactive relationship between AI and its audience, democratizing AI in a way never done before.

Image Sourced From:
Verso Books

Taking inspiration from Godard's approach, we can view the democratization of AI as the breaking of its fourth wall, rendering the technology more accessible and comprehensible to everyone. By dismantling barriers between AI and the public, a broader audience can engage and leverage the technology for problem-solving and innovation. The development of user-friendly interfaces as well as tools that enable more natural interactions with AI have been critical in facilitating this democratization. For example, LLMs like ChatGPT allow users to communicate with AI through conversational interfaces, resembling human-to-human text chat interactions. This intuitive approach eases the learning curve, helping people better understand the technology's potential benefits and applications.

Overall, breaking the fourth wall in AI fosters greater public engagement and understanding of the technology, ultimately leading to more innovative and impactful applications of AI. Perhaps more importantly, accessibility creates a more informed public, allowing us to hold the technology accountable. Nonetheless, there are great challenges and potential risks associated with the rapid development and dissemination of such disruptive technology.

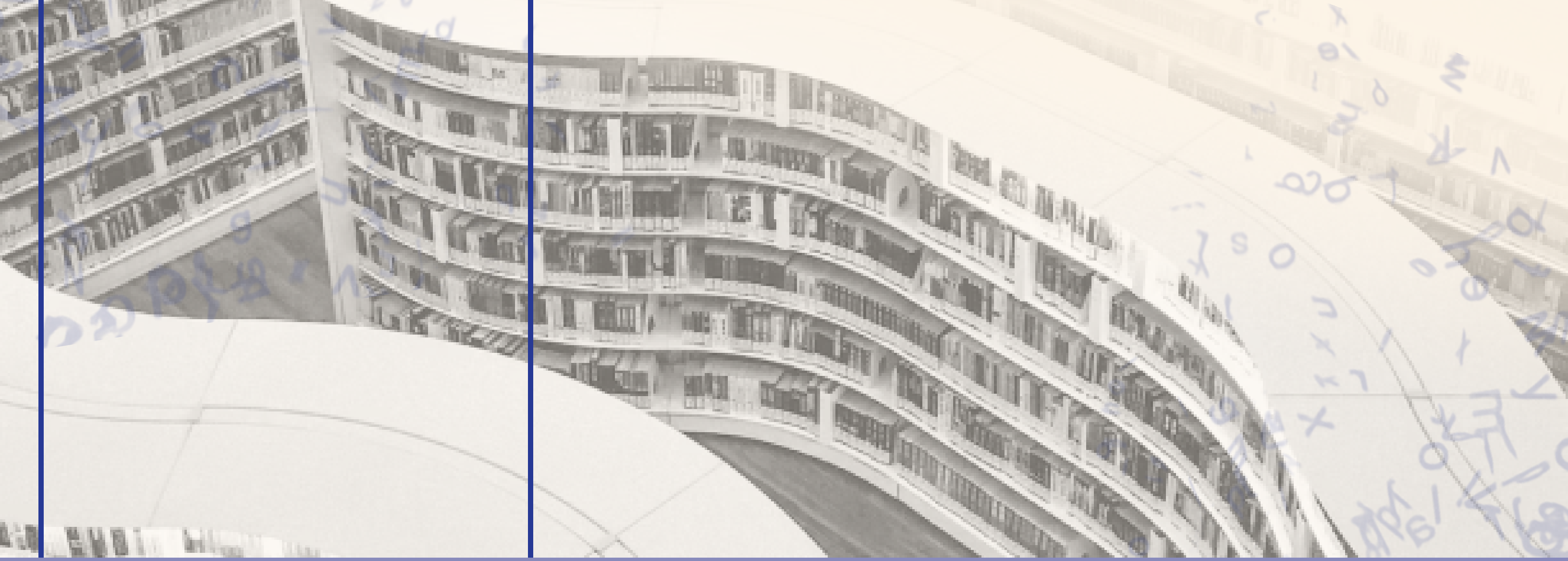
These issues must be addressed to ensure a positive outcome for all society. With the groundbreaking waves that ChatGPT has made, the following is our assessment of the state of the union of LLMs and their role in the democratization, productization, and large-scale adoption of AI.

What is a Large Language Model?

A large language model (LLM) is an AI technology that utilizes advanced machine learning algorithms to process and generate natural language outputs. These models are considered a subset of a broader category called generative AI due to their capabilities to produce text and sometimes code. Training on massive amounts of text data allows these models to generate human-like text for various applications [17]. LLMs emerged around 2018 and have shown impressive performance on a wide range of natural language processing (NLP) tasks — such as generating and classifying text, answering questions in a conversational manner, and translating text from one language to another [31, 15].

LLMs typically consist of a neural network with many parameters (usually billions or more), trained on large quantities of unlabelled text using self-supervised learning, a technique where the model learns from its own data without requiring human annotations or labels. During training, the LLM is fed with large amounts of text, such as books, articles, and web pages, to learn the patterns in the language. Once trained, the LLM can then generate new text by predicting the most likely word or phrase to follow a given input text.

Expanding further, multimodal large language models are AI models that can generate responses or output in multiple modes, such as text, images, and videos. They are capable of integrating information from various sources and modalities to generate more comprehensive and accurate results [39].



Generative AI: Turning the Audience into Actors

Image Sourced From:
Mass:Werk – Media Environments

```
Welcome to
                    EEEEEE  LL      IIII  2222222   AAAAA
                    EE      LL       II        22  AA   AA
                    EEEEE   LL       II      222   AAAAAAA
                    EE      LL       II    22      AA   AA
                    EEEEEE  LLLLLL  IIII  2222222  AA   AA

ELIZA is a mock (Rogerian) psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation ('elizabot.js') by Norbert Landsteiner 2005.
Graphics and real-time text to speech integration added in 2013.

****  Type or click the microphone-button below for voice input.  ****


VOICE SETUP
Please choose an accent to be used by ELIZA (speech output):

   [1]   English - US
   [2]   English - EN

>
```

Generative AI refers to the subset of artificial intelligence that involves the use of machine learning models to generate or create new data, such as images, videos, or text. While the term is sometimes used interchangeably with LLMs, they are not equivalent. LLM is a specific type of generative AI model that generates human-like language, where as generative AI encompasses a broader range of models that can generate other types of data or combine multiple modalities. Based on training data and patterns, generative AI is capable of creating entirely new data that has not been seen before.

One of the earliest examples of generative AI was ELIZA, a chatbot created by MIT in 1966 to simulate conversation with a therapist [34]. ELIZA was able to generate responses to user inputs using a set of predefined rules, making it one of the pioneering works in NLP [11]. Despite its early emergence, generative AI has continued to evolve and become increasingly sophisticated over time, offering exciting possibilities for various industries and applications.

Generative AI technologies such as DALL-E, Midjourney, and ChatGPT are unlocking new realms of human creativity by enabling people to generate new ideas and content in various mediums, from visual art to written text. These tools allow users to explore and experiment with new possibilities they may not have otherwise considered, providing a platform for professionals and amateurs to explore their creativity [37]. These AI technologies are helping to democratize access to the powerful capabilities of AI, making it more accessible to individuals and small businesses that may not have the resources to invest in expensive AI development.

In this new era of AI, accessibility is fostering a two-way technical exchange between the general population and the creators of generative AI technology. Akin to Godard’s cinematic innovation, the technological fourth wall is now broken. Users are no longer passive observers to technological revolution, but are now active participants in the refinement and improvement of AI tools. Creators are observing the audience participation with their AI and in turn iterating on their models, releasing updates to cater to the evolving needs of users. This collaborative feedback loop is breaking down the traditional barriers between technology creators and end-users, creating a dynamic relationship where the users' input shapes the future of generative AI.

As a result, generative AI is no longer a distant and detached technology, but rather a participatory tool that is co-created by the collective efforts of the user community and the developers. This exchange of ideas and collaboration is propelling the field of generative AI forward, unlocking new possibilities for human creativity and democratizing access to the benefits of AI for a wider audience.

The Human Impact of Generative AI

Generative AI will transform how people access and consume information and accelerate the speed of work transformation, contributing to the democratization of AI. However, if these changes do not reach all segments of society equally, there is a risk of exacerbating existing inequalities.

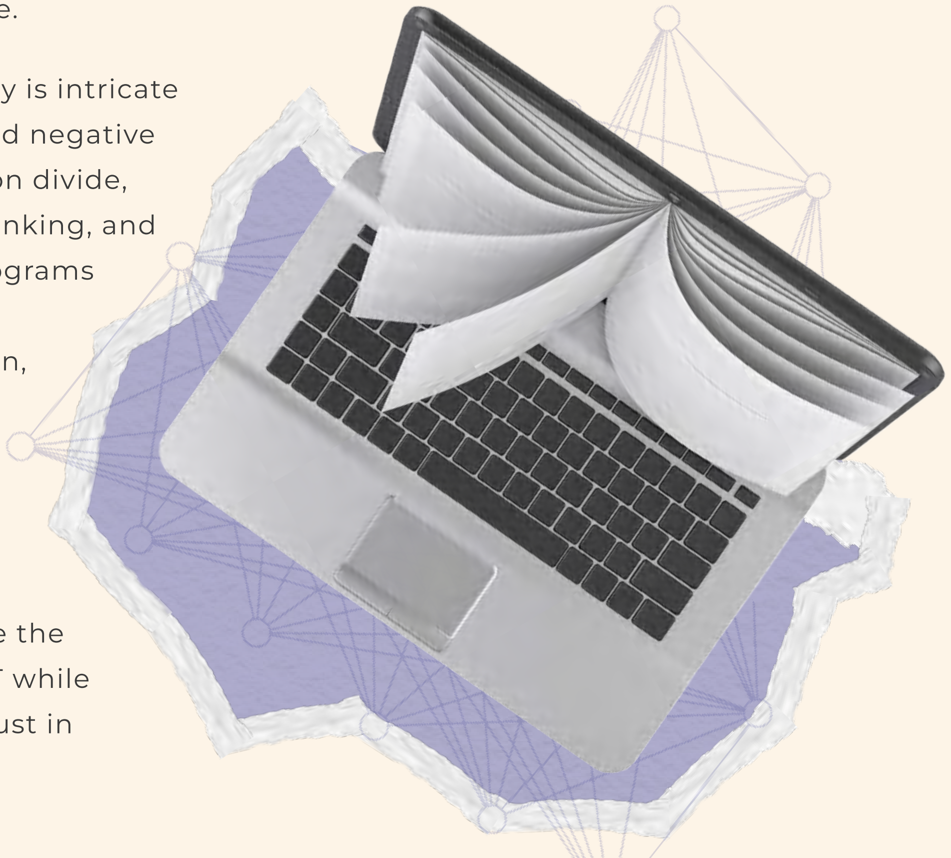
One prominent issue is the information divide, wherein generative AI may widen the gap between those with access to these technologies and those without access. For instance, the majority of contemporary NLP research is centered around only 20 out of the world's 7,000 languages, resulting in a significant underrepresentation of most languages [21]. Consequently, many of the current LLMs only cover high resource languages like English but struggle with low resource languages such as Hindi, which is spoken by more than 500 million people. This lack of attention to low resource languages has significant implications, as it amplifies a bias towards dominant languages, hindering the development of NLP technologies for diverse linguistic communities, limiting access to information, and exacerbating the digital divide. It also reinforces linguistic inequality, as low resource languages face challenges in preserving their cultural heritage and addressing societal issues through technology, resulting in marginalization and exclusion of those communities from the benefits of NLP advancements.

Generative AI can democratize access to information, but the accuracy and reliability of this information are not always guaranteed. This poses the risk of misinformation spreading. Promoting media literacy and critical thinking skills is essential for individuals to discern accurate and reliable information from misinformation [10].

Additionally, though the increased use of generative AI has the potential to accelerate the pace of work transformation via automation and increased efficiency, this may also lead to job displacement and a shift in the types of skills required in the workforce. Promoting reskilling and upskilling programs is vital to ensure individuals can adapt to the changing nature of work and remain employable [16, 33].

Education is another area where significant impact is anticipated. According to [14], generative AI can transform education in multiple ways. Advanced chatbots like ChatGPT could serve as powerful classroom tools, making lessons more interactive, teaching media literacy, creating personalized lesson plans, reducing teachers' administrative tasks, and more. Generative AI could also help equalize opportunities for students with specific learning needs or those for whom English is not their first language.

The influence of generative AI on society is intricate and multifaceted, with both positive and negative implications. Addressing the information divide, nurturing media literacy and critical thinking, and supporting reskilling and upskilling programs can help alleviate potential negative consequences. In the realm of education, it is crucial to integrate these AI technologies mindfully, ensuring they supplement human instruction and foster a more inclusive, engaging, and accessible learning environment for all students. By doing so, we can maximize the potential benefits of LLMs and ChatGPT while minimizing drawbacks and fostering trust in these powerful AI systems.



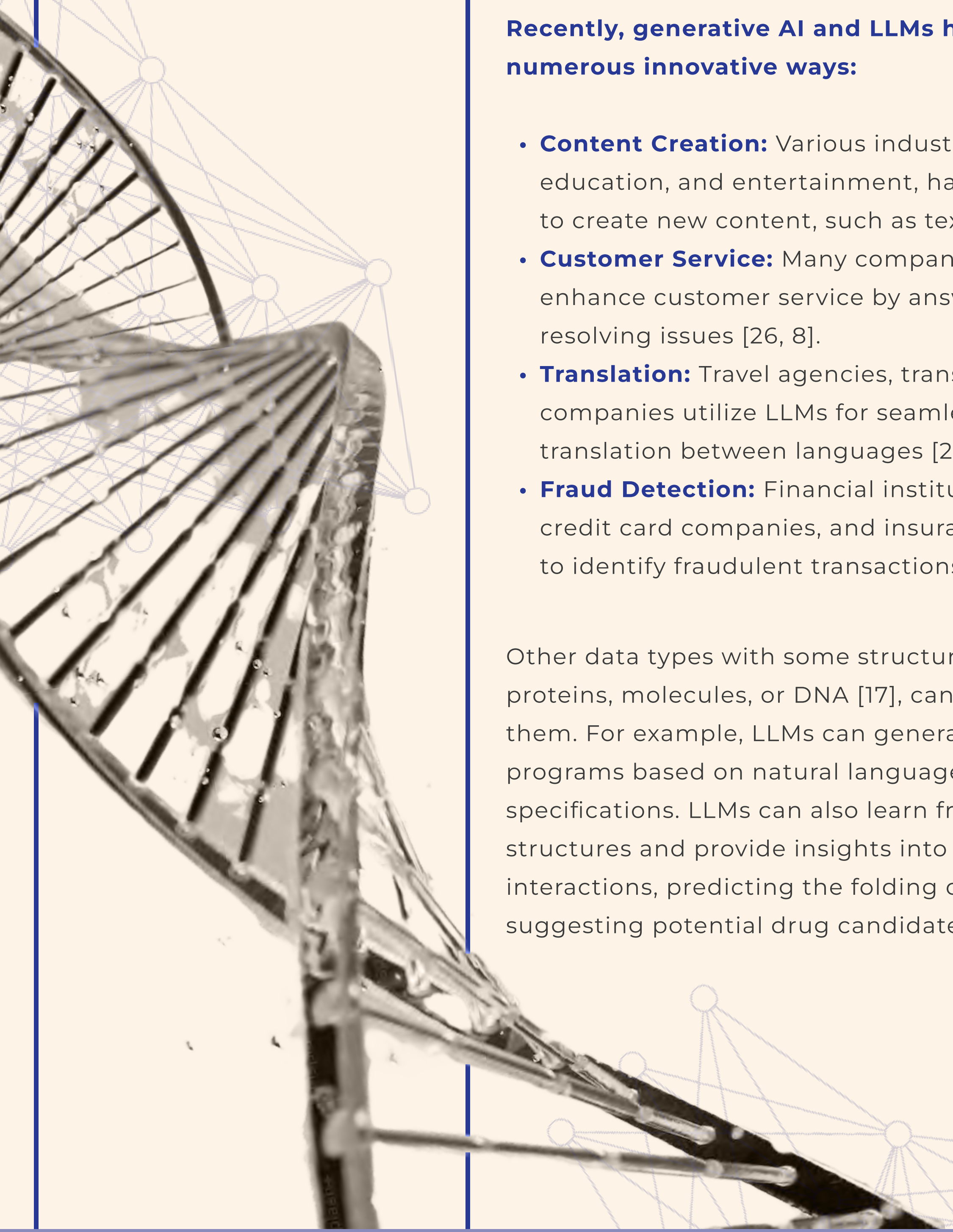
LLM and Generative AI Use Cases

Different domains and industries utilize LLMs for many use cases. LLMs can generate, summarize, and rewrite text. They can extract relevant information from texts based on queries or keywords, and measure the similarity or differences between texts based on their content or style. Moreover, LLMs can group texts into categories based on their topics, themes, or features, and assign labels to texts based on their content or sentiment [2].

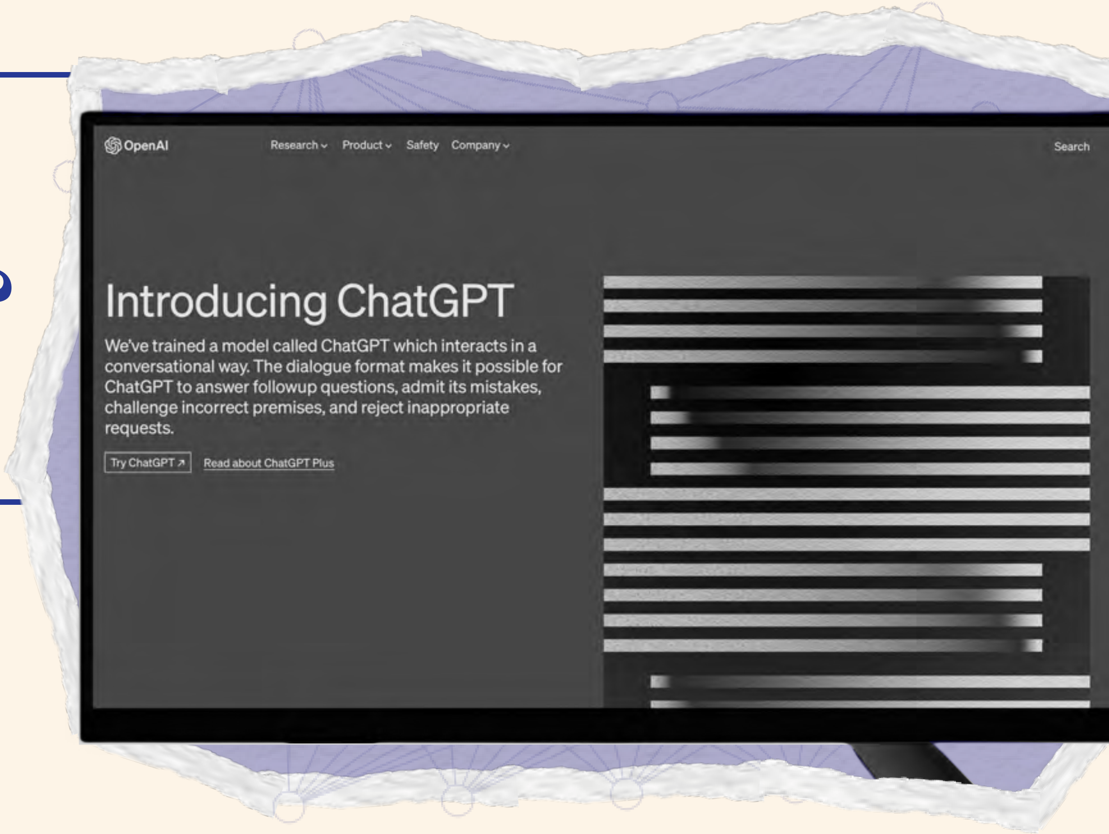
Recently, generative AI and LLMs have been applied in numerous innovative ways:

- **Content Creation:** Various industries, such as advertising, education, and entertainment, have been using generative AI to create new content, such as text, images, and music [26, 7].
- **Customer Service:** Many companies leverage LLMs to enhance customer service by answering queries and resolving issues [26, 8].
- **Translation:** Travel agencies, translation services, and software companies utilize LLMs for seamless text and speech translation between languages [26].
- **Fraud Detection:** Financial institutions, including banks, credit card companies, and insurance providers, rely on LLMs to identify fraudulent transactions and accounts [26].

Other data types with some structure or sequence, such as code, proteins, molecules, or DNA [17], can also have LLMs applied to them. For example, LLMs can generate code snippets or programs based on natural language descriptions or specifications. LLMs can also learn from protein sequences and structures and provide insights into their functions and interactions, predicting the folding of proteins or even suggesting potential drug candidates.



What is ChatGPT?



Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

OpenAI developed ChatGPT as a type of LLM. It can understand and generate human language in a conversational context, enabling it to answer questions, chat on various topics, and generate creative text based on inputs. The GPT (Generative Pre-trained Transformer) architecture, a type of deep learning model, served as the basis for ChatGPT. The model has been pre-trained on large amounts of text data to learn the statistical patterns and relationships between words and phrases in a language. ChatGPT has a significant feature in that it is trained using reinforcement learning from human feedback (RLHF), which means that it improves its performance based on the quality of its responses to human subjects and learns from the human input data [27].

ChatGPT, now based on both the GPT-3.5 and GPT-4 architectures, is a state-of-the-art application that builds upon the success of OpenAI's GPT-3 series. As one of the largest and most advanced LLMs to date, GPT-4 boasts an impressive number of parameters, surpassing its predecessor's 175 billion. These parameters are instrumental in determining how input data is transformed into output, enabling GPT-4 to generate even more coherent and human-like text responses. ChatGPT supports multilingual capabilities, allowing it to comprehend and generate text in multiple languages, further enhancing its versatility and global applicability [28].

There are several ways to access ChatGPT: a limited free version, a premium subscription with priority server performance and access to newly developed features, or through an application programming interface (API) for developers and researchers who want to build it into their own business applications such as chatbots, virtual assistants, language translation, content creation, and scientific research. ChatGPT represents a significant advancement in the field of NLP, but it has some limitations and challenges. These include: producing biased, inaccurate, or nonsensical answers, being sensitive to input phrasing or repetition, and requiring a lot of computational resources and data to train and run.

Other Major LLM Endeavors

ChatGPT is not the only project currently in development or research in the field of NLP. Many other LLM endeavors aim to create and share AI systems that can understand and generate natural language based on large amounts of text data.

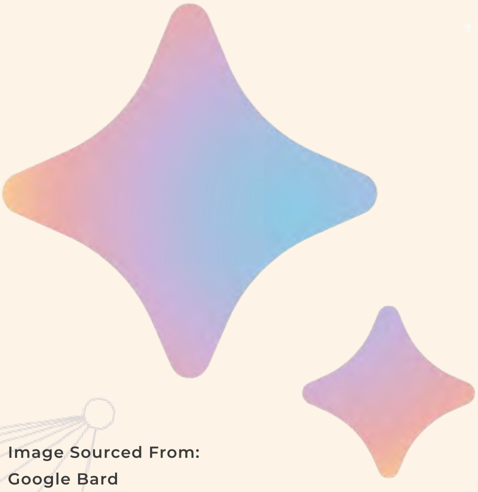


Image Sourced From:
Google Bard

Google Bard

Google developed Bard as a chatbot that uses LaMDA (Language Model for Dialogue Applications) as its underlying LLM. LaMDA is a neural network that can generate natural language responses for open-ended conversations on any topic. It is trained on dialogue data from various sources, such as books, social media, and Wikipedia, as opposed to ChatGPT, which is trained on a general corpus of text data. Bard incorporates images into its responses and is designed to be engaging, informative, and creative [29, 5].

Deepmind Sparrow

DeepMind is researching and developing Sparrow, a conversational agent aimed at creating safer dialogue agents. Sparrow incorporates reinforcement learning and feedback from research participants to minimize the likelihood of providing inappropriate or unsafe responses. It seeks to engage users, provide relevant information, and leverage Google to find evidence and enhance the accuracy of its answers [36, 32].

Image Sourced From:
Deepmind

Image Sourced From:
Hugging Face BigCode

ServiceNow/Hugging Face BigCode

BigCode is a collaboration between ServiceNow and Hugging Face, a startup that provides a platform for building and sharing natural language processing models. BigCode is an open scientific collaboration focused on responsibly training large language models for coding applications. The primary objective of this collaboration is to build cutting-edge LLMs for code in a transparent and ethical manner [42, 35].

Anthropic Claude

Claude is a research project by Anthropic, an AI research company founded by former OpenAI researchers. Claude is an LLM capable of a wide variety of conversational and text-processing tasks while maintaining high reliability and predictability. Anthropic aims to produce helpful, honest, and harmless AI systems. As such, Claude is designed to be aligned with human values and preferences and to avoid harmful or unethical outputs [3].

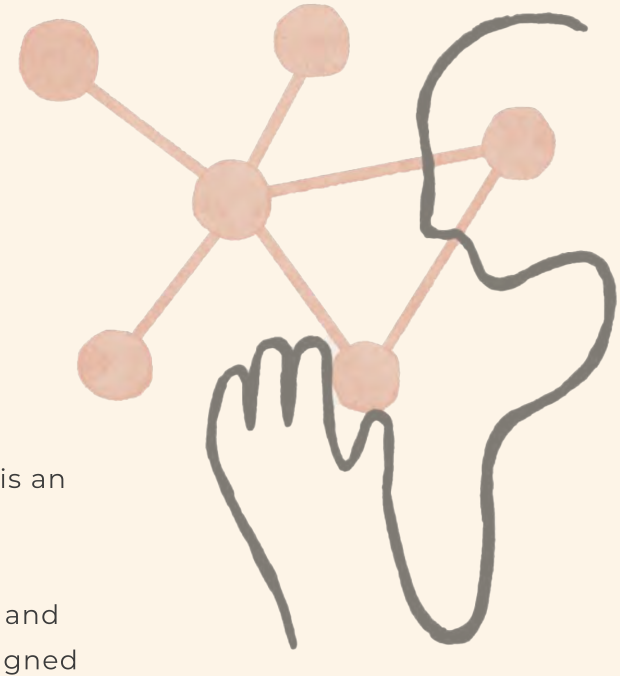


Image Sourced From:
Anthropic

Industry Impact of ChatGPT and Generative AI

Since its launch on November 30th, 2022, ChatGPT has made an unprecedented impact on the digital landscape, becoming the fastest growing consumer application in history — acquiring 100 million users within two months [41]. This is reminiscent of the historic moment in 1939 when "Gone with the Wind" surpassed 100 million ticket sales, signaling a defining indicator of changing times. Generative AI capabilities like ChatGPT are setting new standards for technological advancement and adoption, and we are witnessing these historical moments unfold rapidly in real time.

The rapid proliferation of industries leveraging ChatGPT's API for applications in general tasks — such as writing marketing materials, transcripts, corporate policies, creative collateral, and more specialized tasks related to education and programming — clearly indicates its influence. However, the reach of ChatGPT extends even further.

Dr. Ethan Mollick, an Associate Professor at the Wharton School of the University of Pennsylvania, observes that the current situation is truly unprecedented: "We are seeing widespread adoption of a technology that has the potential to significantly boost individual productivity, but which is not yet being fully utilized by organizations." [24] Two vital aspects of ChatGPT's rapid adoption are that it requires no additional technology, platform, or process to be effective, and there is no organizational advantage in adoption. With no company having exclusive access to the technology, anyone can discover how to use ChatGPT in their work and choose whether to share their adoption methods [24].

In a recent study, professionals were tasked with writing memos, strategy documents, and policies, with some participants designated to use ChatGPT. Those who used ChatGPT completed tasks 37% faster, with improved writing quality, and without added training or extensive experience using the tool [25]. These productivity gains, and anecdotal evidence indicating similar improvements in various fields, suggest that general-purpose AI tools like ChatGPT have significant potential to transform the way we work. These results establish a valuable baseline to anticipate the worldwide economic and social impacts of this technology, much like adding steam power to the average small factory in the U.S. during the 19th century increased productivity by 25%.

Generative AI has transformative potential in industries, but also raises concerns regarding misuse for spreading false information and content, such as deepfakes. Deepfakes are manipulated videos or audio created using AI algorithms to make it appear as if someone is saying or doing something they didn't. This technology can be used for fake news, misinformation, and impersonation [43]. Similarly, AI-generated text can spread false information and propaganda. Lack of safeguards to differentiate real from fake content and protect original work raises concerns about loss of autonomy and identity [13].

While generative AI holds great promise, we must recognize that we are still in the early stages of its development and implementation. We need to proactively address potential social, cultural, and economic adverse impacts by building safeguards and ethical governance measures now.

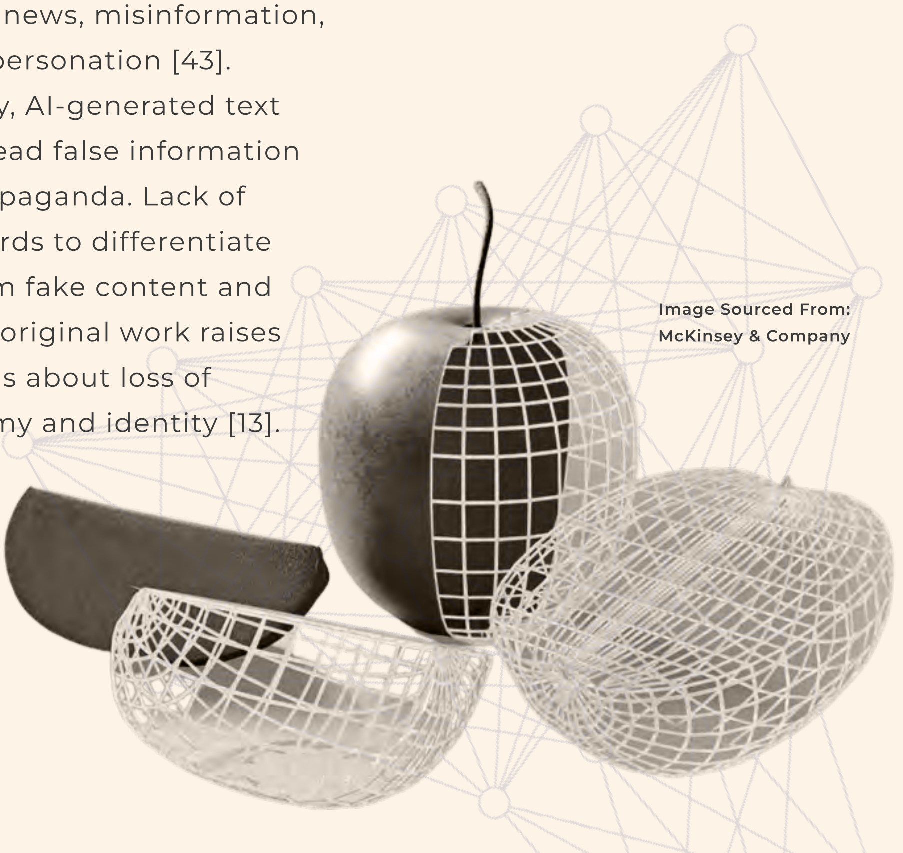


Image Sourced From:
McKinsey & Company

The Democratization of AI

Breaking the fourth wall in cinema and the technological fourth wall in AI both challenge traditional boundaries. In cinema, it involves characters acknowledging the audience or the artificiality of the medium, allowing for direct engagement with viewers and an unconventional narrative approach. Similarly, in AI it involves making tools and technologies more accessible to non-experts, empowering individuals to create and utilize AI applications without extensive technical knowledge. Both practices disrupt established norms and blur the lines between the medium and the audience.

Democratizing AI involves making AI tools increasingly accessible and usable by individuals and organizations lacking technical expertise or resources. The premise of AI is that people can boost productivity, augment decision-making, and devise new products and services that propel economic growth and societal advancement.

Here are ways in which generative AI can democratize AI:

- **Accessibility:** Large language models (LLMs) like ChatGPT can enable people to interact with AI using natural language instead of code or commands. This makes AI more accessible and user-friendly, enabling individuals and organizations lacking technical expertise or resources to build and deploy AI models.
- **Skill and Knowledge Acquisition:** LLMs offer personalized feedback, guidance, and explanations, which can help people learn and acquire new skills and knowledge related to AI.
- **Creativity and Problem-Solving:** LLMs can generate original ideas, suggestions, and solutions based on user input, enhancing creativity and problem-solving abilities.
- **Collaboration and Communication:** LLMs can facilitate natural and engaging conversations across various domains, languages, and cultures, fostering collaboration and communication among people.

The democratization of AI through LLMs can help individuals and organizations harness the power of AI to drive innovation, enhance problem-solving abilities, and stimulate social progress across a wide array of industries and applications.

Can LLMs Be Trusted?

As LLMs continue to gain traction in various applications, it is critical to address the issue of trustworthiness. LLMs, though immensely powerful, are with limitations and challenges. It is essential to view them, at least for now, as storytellers rather than factual messengers — as their primary goal is to generate coherent and contextually relevant responses regardless of their truthfulness. LLMs are trained on unlabelled text data that may contain errors, biases, opinions, or fiction, and they lack the capacity to verify the accuracy or reliability of their sources. Consequently, it is vital to approach LLM-generated content with a discerning eye and an awareness of the factors that can impact their trustworthiness.

Several technical and human-centric factors influence the trustworthiness of LLMs. Technical considerations include:

1. "Hallucinations," which occur when LLMs generate inaccurate or unrelated content
2. Data drift which arises due to changes in data distribution over time
3. Model drift, which results in decreased model performance

On the other hand, human-centric considerations involve:

1. Perceived accuracy and fidelity of information
2. Lack of transparency
3. Potential for bias in the model's output

Given these factors, it is advised not to blindly trust LLM-generated content as information or knowledge sources. Outputs should be carefully examined for factual correctness, relevance, neutrality, and originality.

To address trustworthiness in LLMs, models must be robust, transparent, and unbiased. Employing diverse training data, implementing methods for detecting and correcting hallucinations, and providing explanations for model output can contribute to this goal. Additionally, it is paramount to educate the public on LLMs' limitations and potential biases while promoting critical thinking and skepticism when interacting with these models.

In a recent Twitter thread, Professor Kate Crawford from the University of Southern California highlighted the major problem of companies like OpenAI and Google — not disclosing the training data of the most advanced models that feed ChatGPT and Bard. Without such information, reproducibility, testing, and the development of harm mitigation strategies become impossible. Like a modern-day version of Plato's cave, scientists and researchers are left to decipher the shadows, unable to understand, predict, or fully address potential risks. To foster trust, accountability, and informed decision-making, AI developers must prioritize transparency, even if it means adopting auditing, datasheets, and other methods to maintain proprietary information. The future of ethical AI hinges on striking the right balance between protecting corporate interests and ensuring responsible and verifiable development practices [6].

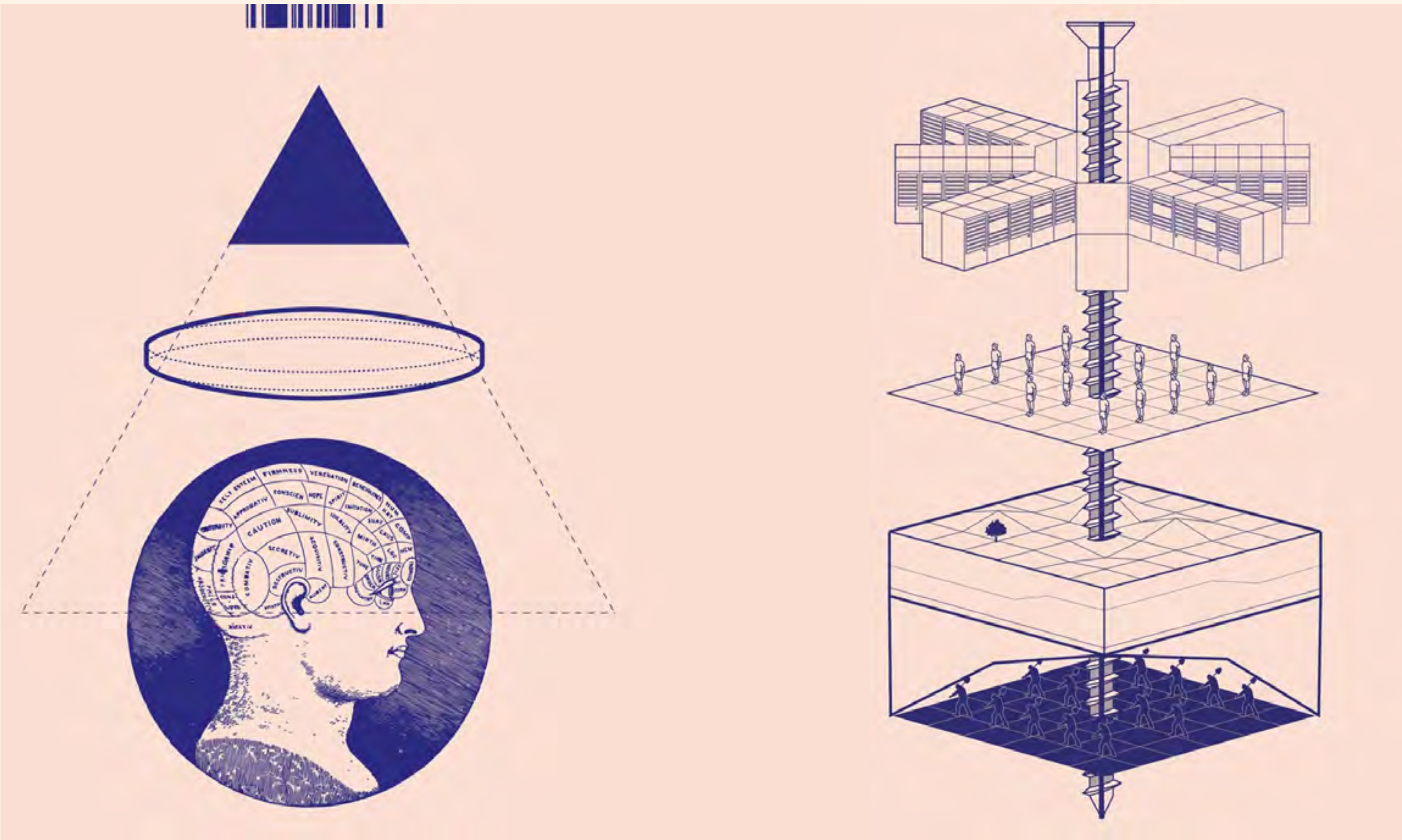
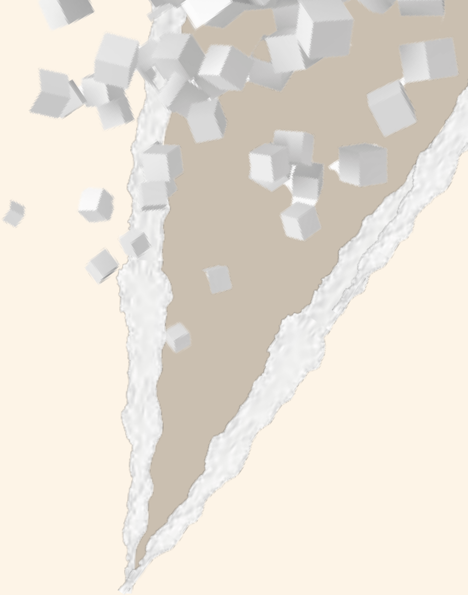
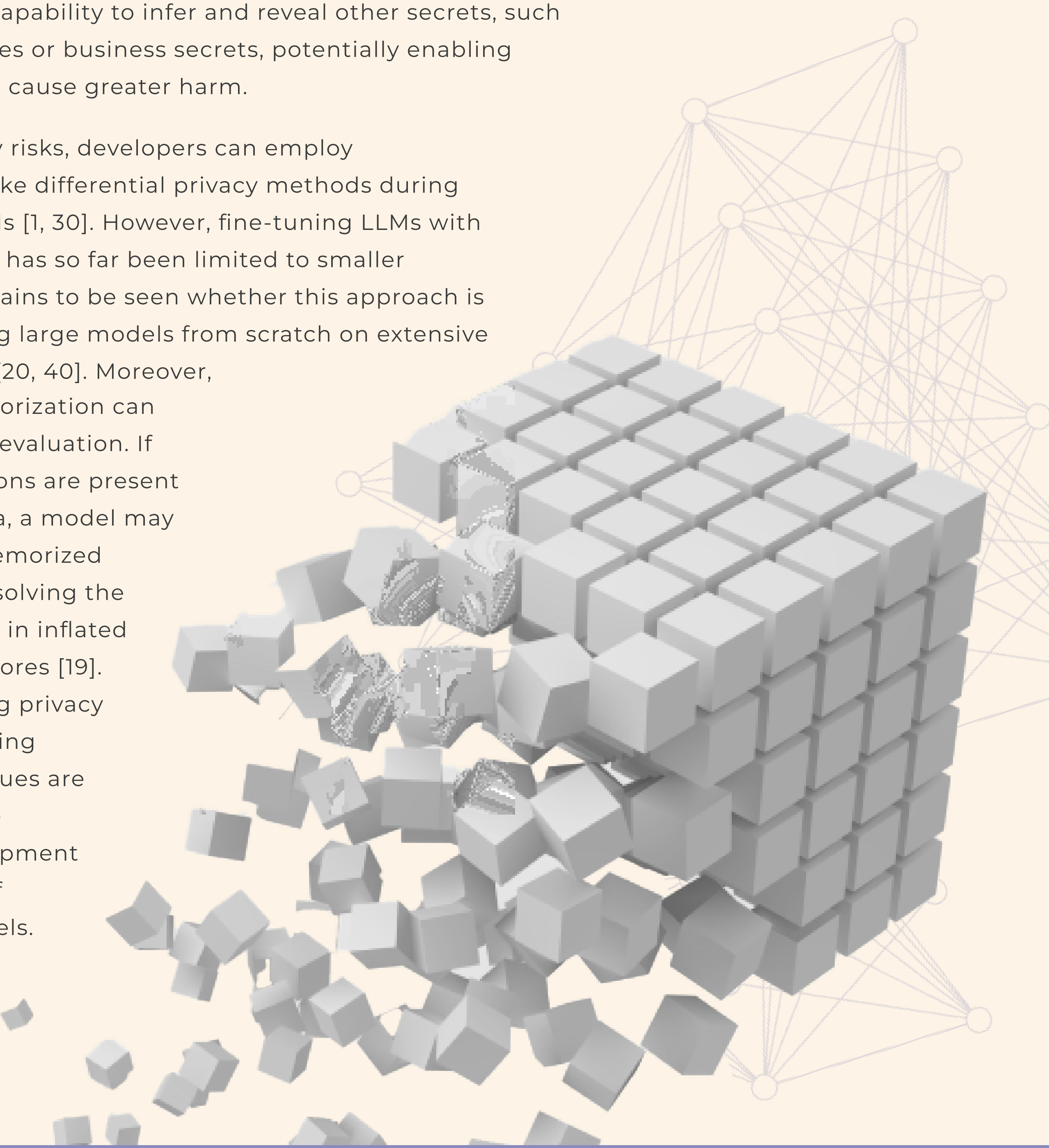


Image Sourced From: MIT Technology Review, Kate Crawford, "Atlas of AI"



Another pressing issue around generative AI is privacy [38]. LLMs can inadvertently "remember" and leak sensitive information if such data is present in their training corpus, leading to privacy violations [4]. These leaks can occur even without malicious intent, and the disclosure of private information can cause psychological and material harm, akin to the effects of doxing [9]. For instance, GPT-2 has been observed to provide personally identifiable information, such as phone numbers and email addresses [22], while the GPT-3-based tool Co-pilot leaked functional API keys [23]. As LLMs continue to evolve, they may even develop the capability to infer and reveal other secrets, such as military strategies or business secrets, potentially enabling malicious actors to cause greater harm.

To mitigate privacy risks, developers can employ algorithmic tools like differential privacy methods during the training of LLMs [1, 30]. However, fine-tuning LLMs with differential privacy has so far been limited to smaller models, and it remains to be seen whether this approach is suitable for training large models from scratch on extensive web text datasets [20, 40]. Moreover, training data memorization can pose problems for evaluation. If benchmark questions are present in the training data, a model may simply repeat a memorized answer instead of solving the question, resulting in inflated or distorted test scores [19]. As such, addressing privacy concerns and refining evaluation techniques are crucial steps in the responsible development and deployment of generative AI models.



Protecting Human Values in the Era of LLMs and Generative AI

Several safeguards, processes, and governance methods can be built into generative AI to protect human ingenuity, identity, and autonomy. Some of these include:

Public Engagement

It is important to engage the public in developing and using generative AI systems. This will help to ensure that these systems are developed in a way that is responsive to the public's needs and concerns. It will also help to build trust in generative AI systems.

Disclosure of Training Data

The data sources used to train the AI model should be disclosed to help identify and mitigate biases that may have been introduced during training.

Interpretability of Outputs

Techniques such as attention mapping and saliency maps can be used to provide greater visibility into how the model arrives at its outputs, making it easier to understand and diagnose potential errors or biases.

Audit Trails and Provenance

The model should maintain a record of all inputs, outputs, and transformations that occur during its operation, making it easier to trace the source of any errors or biases.

Independent Auditing

Third-party auditors can be engaged to review the AI model and ensure that it is operating as intended and not exhibiting any unintended biases or errors.

Open-Source Software and Code

Making the software and code used to build the model publicly available can increase transparency and allow for scrutiny by external parties.

Transparency and Explainability

AI-generated content should be labeled to distinguish it from content created with no AI contribution. It should also be possible to explain how the AI arrived at its output.

Ethical Considerations

AI should be designed with ethical considerations in mind, including the impact on human dignity, privacy, and autonomy.

Consent and Ownership

Artists and creators should be given control over how their work is used, and their consent should be obtained before their work is used to train AI models.

Verification and Authentication

Mechanisms of verification and authentication can be built into AI to help distinguish real from fake content. This can include watermarking or digital signatures.

Oversight and Regulation

The practice of oversight and regulation help ensure that AI is used ethically and responsibly. This can include independent audits, certification schemes, and regulatory frameworks.

Collaboration and Cooperation

Industry, government, and civil society should work together to advocate for AI that is developed and used in a way that benefits society as a whole.

As generative AI and LLMs become an integral part of our economy and social interactions, striking a balance between innovation and the preservation of human values becomes paramount. By employing a combination of public engagement, transparency, ethical considerations, and collaboration, we can establish a strong foundation for responsible AI development and usage. Ensuring that safeguards and governance methods are deeply ingrained in the life cycle of these AI systems will help empower users. In this way, we can collectively harness the potential of AI technologies while respecting our basic human principles.

CONCLUSION

Thoughts for the Future

In the rapidly evolving age of AI, discerning the winners and losers becomes an increasingly complex task, as the impacts of this technology are often more subtle and nuanced compared to the tangible transformation brought by the Industrial Revolution. The rise of generative AI and LLMs is a defining feature of our time, unveiling capabilities once hidden behind the scenes, breaking boundaries, cultivating collaborative discourse and increasing AI accessibility to a larger audience. We are watching and writing this part of human history in parallel.

The reach and adoption of generative AI presents both challenges and opportunities, which we must proactively address to ensure the equitable distribution of AI's benefits and prevent any individuals or groups from being left behind. The democratization of AI can potentially drive innovation and social progress across various industries and applications. However, we must also acknowledge and mitigate the risks associated with exacerbating existing inequalities, misinformation, and the changing nature of work. The future of our society will be defined by how we shape the AI revolution today.

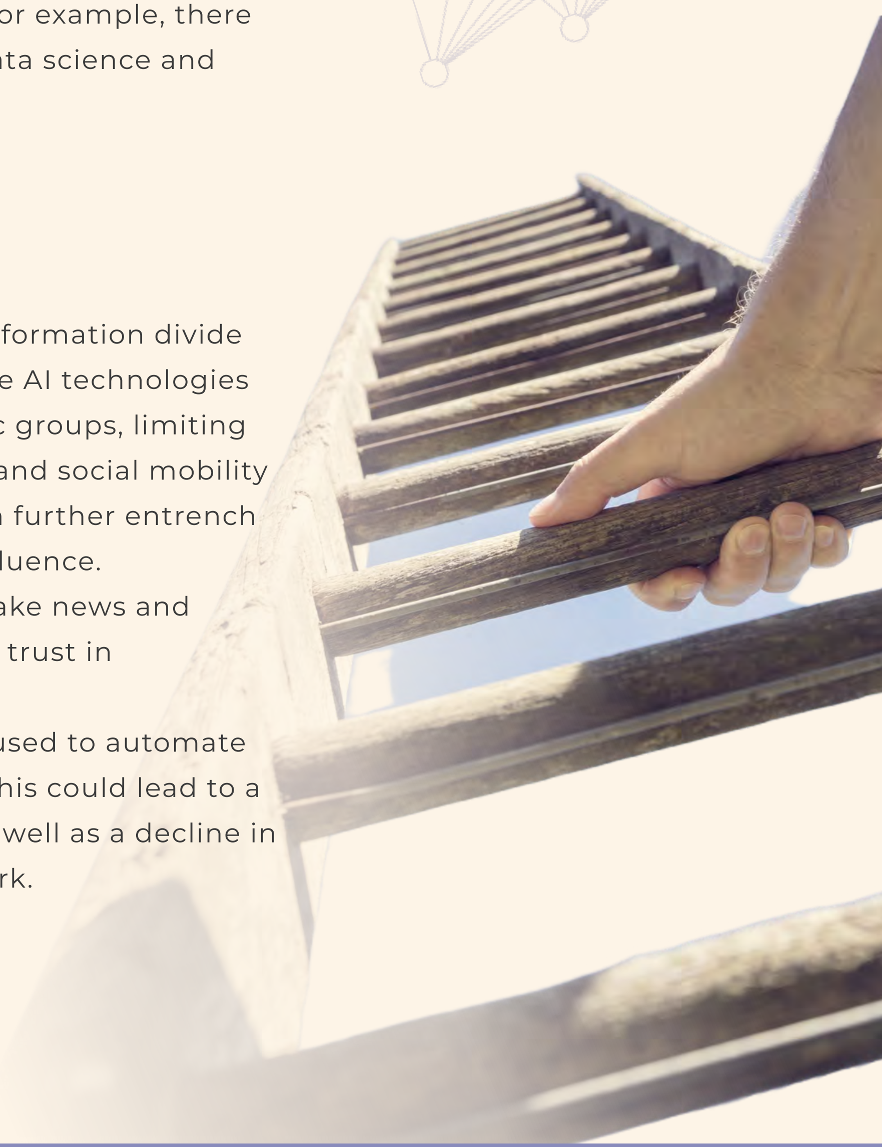
The following are some opportunities and challenges that will emerge as we transition to a future where generative AI is commonplace.

Opportunities:

- **Innovation and Social Progress:** AI can be used to develop new products and services that improve the lives of people around the world. For example, AI can be used to develop new drugs, new educational tools, and new ways to diagnose diseases.
- **Increased Efficiency and Productivity:** AI can be used to automate tasks that are currently done by humans. This could lead to an increase in efficiency and productivity, as well as a decline in the cost of goods and services.
- **New Jobs:** AI can also create new jobs, as it will require people to develop, maintain, and use AI systems. For example, there will be a need for new jobs in the field of data science and machine learning.

Challenges:

- **Exacerbating Existing Inequalities:** The information divide resulting from unequal access to generative AI technologies can widen the gap between socioeconomic groups, limiting opportunities for education, employment, and social mobility for those left behind. This digital divide can further entrench disparities in wealth, opportunities, and influence.
- **Misinformation:** AI can be used to create fake news and propaganda. This could lead to a decline in trust in institutions and a rise in social unrest.
- **The Changing Nature of Work:** AI can be used to automate tasks that are currently done by humans. This could lead to a decline in the demand for human labor, as well as a decline in the wages of those who are able to find work.



On March 22nd, 2023, a coalition of esteemed technology leaders and numerous supporters affixed their signatures to an open letter, which was addressed to all AI laboratories, via The Future of Life Institute. The letter called for a temporary cessation of large-scale AI experiments, particularly those exceeding the parameters of GPT-4, for a minimum period of six months [12]. The rationale behind this proposed pause is to facilitate additional research, comprehensive assessments of the social, cultural, and human implications of AI, and the formulation of principles, policies, and governance measures to regulate this unparalleled technology.

As we continue to break the technological barriers and witness the democratization of access to AI, with its immense and powerful capabilities becoming more widespread, it is imperative that we also disseminate cautionary guidance and raise awareness about the potential risks.

To successfully navigate the swiftly evolving terrain of generative AI and LLMs, it is non-negotiable that all stakeholders — including AI developers, policymakers, educators, and users — need to collaborate toward shaping a responsible and inclusive future. This collaboration should promote equal access to AI technologies, foster media literacy and critical thinking, and encourage reskilling and upskilling programs. By prioritizing the trustworthiness of LLMs and protecting human ingenuity, identity, and autonomy, we can pave the way for a future where AI catalyzes positive change, empowering individuals by acting as an ally in our pursuit of progress and societal advancement.

References

1. Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16), Association for Computing Machinery, Vienna, Austria, 308–318. DOI:https://doi.org/10.1145/2976749.2978318
2. Meor Amer. 2022. Large Language Models and Where to Use Them: Part 1. Cohere.ai. Retrieved April 18, 2023 from https://txt.cohere.ai/llm-use-cases/
3. Anthropic Team. 2023. Introducing Claude. Anthropic. Retrieved April 18, 2023 from https://www.anthropic.com/index/introducing-claude
4. Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel HerbertVoss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. Retrieved from https://arxiv.org/abs/2012.07805
5. Eli Collins. 2021. LaMDA: our breakthrough conversation technology. Google. Retrieved April 18, 2023 from https://www.blog.google/technology/ai/lamda/
6. Kate Crawford. 2023. Kate Crawford's Twitter Thread. Twitter. Retrieved April 18, 2023 from https://twitter.com/katecrawford/status/1638524011876433921
7. Thomas H. Davenport and Nitin Mittal. 2022. How Generative AI Is Changing Creative Work. Harvard Business Review. Retrieved April 18, 2023 from https://hbr.org/2022/11/how-generative-ai-is-changing-creative-work
8. Victor Dey. 2023. ChatGPT and LLM-based chatbots set to improve customer experience. VentureBeat. Retrieved April 18, 2023 from https://venturebeat.com/ai/chatgpt-and-llm-based-chatbots-set-to-improve-customer-experience/
9. David M. Douglas. 2016. Doxing: a conceptual analysis. Ethics and Information Technology 18, 3 (September 2016), 199–210. DOI:https://doi.org/10.1007/s10676-016-9406-0
10. Lance Eliot. 2022. AI Ethics And The Future Of Where Large Language Models Are Heading. Forbes. Retrieved April 18, 2023 from https://www.forbes.com/sites/lanceeliot/2022/08/30/ai-ethics-asking-aloud-whether-large-language-models-and-their-bossy-believers-are-taking-ai-down-a-dead-end-path/?sh=54dae5b42250
11. Josh Fruhlinger. 2023. What is generative AI? The evolution of artificial intelligence. InfoWorld. Retrieved April 18, 2023 from https://www.infoworld.com/article/3689973/what-is-generative-ai-the-evolution-of-artificial-intelligence.html
12. Future of Life Institute. 2023. Pause Giant AI Experiments: An Open Letter. Retrieved April 18, 2023 from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

References

13. Karen Hao. 2019. The biggest threat of deepfakes isn't the deepfakes themselves. MIT Technology Review. Retrieved April 18, 2023 from https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/
14. Will Douglas Heaven. 2023. ChatGPT is going to change education, not destroy it. MIT Technology Review. Retrieved April 18, 2023 from https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/
15. John Jacob and Shantanu Nair. 2022. LLMs, a brief history and their use cases. Exemplary.ai. Retrieved April 18, 2023 from https://exemplary.ai/blog/llm-history-usecases
16. Michael Kan. 2023. OpenAI: ChatGPT Could Disrupt 19% of US Jobs, Is Yours on the List? PCMag. Retrieved April 18, 2023 from https://www.pcmag.com/news/openai-chatgpt-could-disrupt-19-of-us-jobs-is-yours-on-the-list
17. Angie Lee. 2023. What Are Large Language Models Used For? NVIDIA. Retrieved April 18, 2023 from https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/
18. Ernest Lee. 2020. The fourth wall: Looking beyond the lens. Cherwell. Retrieved April 18, 2023 from https://cherwell.org/2020/06/11/the-fourth-wall-looking-beyond-the-lens/
19. Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. Retrieved from http://arxiv.org/abs/2008.02637
20. Zhuohan Li, Siyuan Zhuang, Shiyuan Guo, Danyang Zhuo, Hao Zhang, Dawn Song, and Ion Stoica. 2021. TeraPipe: Token-Level Pipeline Parallelism for Training Large-Scale Language Models. Retrieved from http://arxiv.org/abs/2102.07988
21. Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2022. Low-resource Languages: A Review of Past Work and Future Challenges. DOI:https://doi.org/10.48550/arXiv.2006.07264
22. Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In Proceedings of the 10th annual ACM workshop on Privacy in the electronic society (WPES '11), Association for Computing Machinery, Chicago, Illinois, USA, 1–12. DOI:https://doi.org/10.1145/2046556.2046558
23. Abubakar Mohammed. 2021. GitHub Copilot AI Is Generating And Giving Out Functional API Keys. Fossbytes. Retrieved April 18, 2023 from https://fossbytes.com/github-copilot-generating-functional-api-keys/

References

24. Ethan Mollick. 2023. Secret Cyborgs: The Present Disruption in Three Papers. One Useful Thing. Retrieved April 18, 2023 from https://www.oneusefulthing.org/p/secret-cyborgs-the-present-disruption?utm_source=substack&utm_medium=email
25. Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Retrieved from https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf
26. ODSC Team. 2023. 5 Practical Business Use Cases for Large Language Models. Retrieved April 18, 2023 from https://opendatascience.com/5-practical-business-use-cases-for-large-language-models/
27. OpenAI Team. 2022. Introducing ChatGPT. OpenAI. Retrieved April 18, 2023 from https://openai.com/blog/chatgpt
28. OpenAI Team. 2023. GPT-4. Retrieved April 18, 2023 from https://openai.com/research/gpt-4
29. Sundar Pichai. 2023. An important next step on our AI journey. Google. Retrieved April 18, 2023 from https://blog.google/technology/ai/bard-google-ai-search-updates/
30. Swaroop Ramaswamy, Om Thakkar, Raijv Mathews, Galen Andrew, H. Brendan McMahan, and Françoise Beaufays. 2020. Training Production Language Models without Memorizing User Data. Retrieved from http://arxiv.org/abs/2009.10031
31. Sebastian Raschka. 2023. Understanding Large Language Models -- A Transformative Reading List. Sebastian Raschka. Retrieved April 18, 2023 from https://sebastianraschka.com/blog/2023/llm-reading-list.html
32. Jesus Rodriguez. 2023. Inside Sparrow: The Foundation of DeepMind's ChatGPT Alternative. Medium. Retrieved April 18, 2023 from https://jrodthoughts.medium.com/inside-sparrow-the-foundation-of-deepminds-chatgpt-alternative-854df43569fd
33. Melissa Rohman. 2023. How ChatGPT Has, and Will Continue to, Transform Scientific Research. News Center. Retrieved April 18, 2023 from https://news.feinberg.northwestern.edu/2023/03/21/how-chatgpt-has-and-will-continue-to-transform-scientific-research/
34. Kenneth Ronkowitz. ELIZA: a very basic Rogerian psychotherapist chatbot. Retrieved April 18, 2023 from https://web.njit.edu/~ronkowit/eliza.html
35. ServiceNow Research. 2023. Announcing BigCode for the responsible development of large language models. ServiceNow. Retrieved April 18, 2023 from https://www.servicenow.com/blogs/2022/bigcode-large-language-models.html
36. The Sparrow Team. 2023. Building safer dialogue agents. DeepMind. Retrieved April 18, 2023 from https://www.deepmind.com/blog/building-safer-dialogue-agents

References

37. Tamarah Usher. 2023. Generative AI: What We Know and Where It Could Go. Slalom Business. Retrieved April 18, 2023 from https://medium.com/slalom-business/generative-ai-what-we-know-and-where-it-could-go-58ecd75d5287
38. Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Kasirzadeh Atoosa, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, Isaac William, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of Risks posed by Language Models. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21-24, 2022, Seoul, Republic of Korea, ACM, New York, NY, USA, 214–229. DOI:https://doi.org/10.1145/3531146.3533088
39. Justin Weinberg. 2023. Multimodal LLMs Are Here (updated). Daily Nous. Retrieved April 18, 2023 from https://dailynous.com/2023/03/02/multimodal-llms-are-here/
40. Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2021. Differentially Private Fine-tuning of Language Models. Retrieved from https://arxiv.org/abs/2110.06500
41. 2023. ChatGPT's meteorical rise-100 million users in 2 months; beats WhatsApp, mobile phones, Twitter, internet. Mint. Retrieved April 18, 2023 from https://www.livemint.com/technology/tech-news/chatgpts-meteorical-rise-100-million-users-in-2-months-11677997670518.html
42. BigCode. Retrieved April 18, 2023 from https://huggingface.co/bigcode
43. Deepfake. Wikipedia. Retrieved April 18, 2023 from https://en.wikipedia.org/wiki/Deepfake