International Working Group on
Data Protection in Technology

Working Paper on Large Language Models (LLMs)

Published 6 December 2024[1]

Table of contents

[1] *This paper was discussed at the 72nd IWGDPT Meeting on 7th – 8th December 2023 and adopted, after final discussion, at the 73rd IWGDPT Meeting on 18th – 19th June 2024. The written procedure followed after the latter meeting.*

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

Preface

Over the past year, governance initiatives concerning generative artificial intelligence (AI), including large language models (LLMs), have proliferated across the world. Examples include the establishment of AI Safety Institutes in both the United Kingdom (UK) and United States of America (US), a United Nations (UN) AI Advisory Body, and AI Regulatory Sandboxes in countries as diverse as Brazil, Singapore, and France. China also recently announced a Global AI Governance Initiative.

These initiatives have overlapped with, and in many instances directly complemented a wide range of legislative proposals and government strategies that continue to emerge. For example, in the United States, NIST's AI Risk Management Framework[2] published in January 2023, preceded the Biden Administration's Executive Order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.[3] States such as California and Colorado have developed frameworks for regulating certain high-risk models.[4] In addition, the European Union has adopted the AI Act, the world's first comprehensive legal framework for

[2] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, June 2023, https://doi.org/10.6028/NIST.AI.100-1.
[3] United States, Executive Office of the President [Joseph Biden], *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
[4] California Privacy Protection Agency, Draft Risk Assessment and Automated Decisionmaking Technology Regulations (March 2024), https://cppa.ca.gov/meetings/materials/20240308_item4_draft_risk.pdf; Colorado SB 24-205, Consumer Protections for Artificial Intelligence (2024), https://leg.colorado.gov/bills/sb24-205.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

AI.[5] Meanwhile, international frameworks, such as the OECD's AI principles,[6] have existed for several years and predate the widespread emergence of generative AI, and have since become applied[7] while being updated to reflect technological developments.[8]

Amid the growing global array of AI governance commitments, proposed legislation, strategies, principles, and frameworks, one thing is becoming clear: that privacy and data protection legislation have a central role to play in AI governance. Notably the Biden administration's EO called for Congress to adopt privacy legislation.

At the same time, in October 2023 the Global Privacy Assembly (GPA) published a Resolution on Generative Artificial Intelligence Systems.[9] Within Europe, in April 2023 the EDPB announced a Chat-GPT Taskforce, designed to coordinate a response among the EU's data protection authorities (DPAs). Generally, as the OECD observes, many countries have begun to "link their data access and sharing policies with AI policies."[10]

These emerging governance and regulatory initiatives form the backdrop of this paper, which sets out some key areas to account for when considering data protection and privacy within the context of generative AI, including LLMs. The GDPR (General Data Protection Regulation) offers a starting point, in full

[5] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), available at https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[6] For a landscape review of how these principles have been adopted, see OECD, "The state of implementation of the OECD AI Principles four years on," *OECD Artificial Intelligence Papers*, No. 3, OECD Publishing, 2023, Paris, https://doi.org/10.1787/835641c9-en.

[7] See Lucia Russo and Noah Oder, "How countries are implementing the OECD Principles for Trustworthy AI," October 31, 2023, https://oecd.ai/en/wonk/national-policies-2.

[8] See OECD, "OECD updates AI principles to stay abreast of rapid technological developments," https://www.oecd.org/newsroom/oecd-updates-ai-principles-to-stay-abreast-of-rapid-technological-developments.htm.

[9] Global Privacy Assembly, *Resolution on Generative Artificial Intelligence Systems*, October 20, 2023, https://www.edps.europa.eu/system/files/2023-10/edps-gpa-resolution-on-generative-ai-systems_en.pdf.

[10] OECD, *supra* note 3.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

acknowledgment that many other data protection laws exist around the world and the terminology used in the paper is likely to change during further development. Data protection is of course highly contextual, and becomes even more so when considered across different jurisdictions. The key assumption is that, as is the case with governance in general, data protection should ideally contribute to forming the basis for the prevention or mitigation of the many known risks and harms of generative AI, on an international basis.

Disclaimer

This paper does not contain legal advice, nor do the views expressed in it necessarily reflect the official policy or position of individual IWGDPT members.

Introduction

LLMs are mathematical models developed using artificial intelligence (AI) and machine learning (ML) data processing techniques to perform tasks related to natural language. The state of the art has advanced significantly in recent years, with some LLMs demonstrating human- and even expert-level natural language processing capabilities for certain tasks. The observed progress in the field is due mainly to advancements in model architecture and training techniques, combined with exponential increases in model sizes, training data corpora and availability of compute.

Despite their capabilities, LLMs are no technical panacea. Their AI-enabled approach to automating linguistic tasks raises a number of privacy and data protection risks. Some risks stem from design choices in the underlying technology; others from practices relating to the processing of personal information; and still others from inherent limitations in mathematical approaches to machine-based language acquisition and understanding.

The aim of this paper is to provide an in-depth, multifaceted analysis of LLMs from the point of view of privacy and data protection. Just as LLMs are complex technologies that raise various privacy and data protection risks, so any

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

proportionate analysis must view the technology from multiple perspectives. It is not only necessary to analyze LLMs from the point of view of the technology itself, that is, a technical analysis of how LLMs fundamentally work, but equally from the perspectives of the privacy and data protection risks they raise and the emerging set of best practices to reduce or eliminate their risks. Only with an understanding of LLMs from the point of view of these three perspectives—the technology, privacy risks and best practices—can DPAs position themselves to effectively regulate and respond to this new situation.

Given the multifaceted nature of the analysis undertaken here, there is no single audience to whom this paper is directed. Rather than exploring LLMs from one level of analysis only, the idea is to present the material in a way that engages multiple audiences throughout, including technologists, policy analysts, lawyers and decision-makers, from across various domains.

While at first glance this approach may seem apt to cause confusion, having multiple audiences may in fact lead to greater clarity overall. Instead of treating the issues in silos, this paper may serve as an opportunity for the various domains of expertise in the field of privacy and data protection to have a larger conversation with a view towards increasing their collective knowledge. This is the goal of the paper—to provide a common starting point for such a conversation to occur, so that privacy and data protection experts can ask and answer questions both amongst themselves and with stakeholders.

This paper is divided into five sections. In the first section, we motivate our analysis by describing some use cases of LLMs that raise both benefits and risks to individuals. In section 2, we provide a technical explanation of LLMs, focusing on the role and functionality of various components at each stage of the LLM development lifecycle. In section 3, we provide an analysis of the various data protection and privacy risks raised by LLMs. In section 4, we discuss best practices to prevent or mitigate some of the risks of LLMs, framing the discussion in terms of key areas requiring consideration by developers and deployers. Finally, in section 5, we provide a brief discussion of emerging practices in the form of local LLMs.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

1. Use cases

LLMs give rise to various possible use cases across different sectors of the economy. The following fictional scenarios provide a glimpse into the practical circumstances in which LLMs may be used and the different considerations their users may have in terms of their benefits and risks.

Chatbot assistant for online shopping

Scenario: Angus recently switched his diet to vegan and implemented a new fitness regime and he is now looking for healthy organic food for himself and his family. Disappointed by the offerings of his local supermarket chain, he turns to specialty online stores he has heard good things about from his friends. Unsure of which products to choose, he is pleased to discover that the online shop offers a helpful chatbot to assist him. He types in his dietary restrictions and that he is looking for a few recommendations for easy-to-cook, healthy weekday meals. The chatbot responds promptly with a broad assortment of fresh and canned products, which Angus can directly add to his shopping cart. After further filtering to fit his family's needs, Angus completes his grocery shopping and is delighted how quickly and easily it went. He continues to chat with the online system to get more recipe recommendations, but he is a bit puzzled when the bot starts to recommend slimming products and writes that overweight people are careless and irresponsible. It's only later that day he starts wondering what will happen with his data. Should he really have input those allergies of his kids to filter down the products? He is also still perplexed by the fat-shaming remarks of the bot and wonders if there is something wrong. After going through the grocery list one more time, this time with more attention, Angus is shocked to find a product recommended by the bot without taking into account the allergies he had previously indicated.

Possible Benefits: Instantaneous advice for everyday situations.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Possible Risks: Inadvertent input of sensitive information (e.g., health data); insufficient transparency; unchecked advice (including potentially medical advice); hallucinations; insults and harassment.

Assistance with medical report writing

Scenario: Bertil recently joined the local hospital that was excited to recruit him as a leading expert in his field of oncology. After immigrating to a new country, Bertil is happy to learn that the hospital offers language courses. Language skills can still sometimes be a hurdle, but the medical expertise counts far more, and the hospital also recently added LLM-powered AI features to their hospital information system. These new capabilities assist with writing patient reports based on International Classification of Diseases (ICD) codes and other notes from the anamnesis such as patient's symptoms, medical history and demographic information. Doctors and nurses have been excited by the efficiency wins so they can spend more time with their patients, and Bertil is pleased the system also helps him as a non-native speaker. These wins have been overshadowed by recent doubts, though. When recently treating a melanoma on a person of color, the system failed to name the correct diagnosis in the medical report, and Bertil only caught and corrected this error in the last minute. Maybe the system was not properly trained on a diverse set of patients he wonders. Since then, he has been extra careful at fact-checking all the reports. He also still wonders how patients' data is protected, and what happens if a patient withdraws their consent for their data to be used to train the model.

Possible Benefits: Reduction in administrative overhead and efficiency wins; assistance for staff and especially for non-native speakers.

Possible Risks: Factual inaccuracies plus extra effort for checking for such mistakes; unclear consent into training (including for health data); possible bias in the training data and thus perpetuated bias in decisions.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Preserving history and the memory of people

Scenario: Clara is working at a Holocaust memorial museum where she is heading the department for public education including classes and tours for the youth. Stories by contemporary witnesses are a key part of her work and help current generations understand the past. But organizing such events has become increasingly difficult as survivors have passed away. She has now been looking into a new AI startup that is developing technology to preserve oral history and digitize memories. By fine-tuning a LLM model on text data from different historical resources and people's testimonials, their model is capable of interacting with users and answering specific questions by generating text. This would enable future generations to still experience and interact with past stories. Recently, however, a teacher visiting the exhibition observed the model interacting with his student and talking about a historical event that never happened and reported it to the museum. After seeing that the model could also tell false stories, Clara is increasingly worried it could be used to spread misinformation.

Possible Benefits: Preservation and interaction with historical artefacts; remembering historical witnesses.

Possible Risks: Disinformation; fake historical facts/news.

2. What are LLMs?

LLMs are extremely large, complex machine learning systems capable of routinely generating highly articulate, plausible-sounding—but not necessarily true—linguistic content in response to queries on virtually any topic. LLMs consist of hundreds of billions or even trillions of parameters organized across various architectural components. Each component plays a specific role and contributes new functionality to the system. Examples of components include the language vocabulary, word embeddings, context window, multi-head self-attention blocks and feed-forward neural networks.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

Collectively, these components form what is known as the "transformer" architecture. Artificial intelligence (AI) models, including LLMs, whose design is based on this architecture are commonly referred to as "transformer" models. For a technical discussion of the transformer architecture, including a breakdown of the number of parameters, please refer to Appendix A.

The training lifecycle of LLMs is unlike that of most other machine learning applications. Instead of a single stage of training using one form of machine learning, LLMs typically employ a two-stage training procedure with multiple types of learning. The first stage of training is called "pre-training" while the second is called "fine-tuning / alignment."

LLMs also differ from most other machine learning applications in their mode of use. LLMs typically interact with their users in the form of a back-and-forth, question-and-answer dialogue, with the ability for users to change the level of randomness in the output.

In what follows, we will explain how LLMs work according to each of the above stages in their development, that is, pre-training, fine-tuning / alignment and use. These are not the only stages in the development of LLMs. For example, many LLMs undergo a stage of "red-teaming" before they are deployed, in which a team of security and other subject-matter experts attempt to identify vulnerabilities and opportunities for misuse. However, the three stages we have chosen provide an opportunity to discuss many of the unique features of LLMs to better understand their overall functionality.

Stage 1: Pre-training

During this initial stage, the goal is to create a general-purpose model with a kind of raw, unrefined ability to continuously predict the next word or sub-word "token" in a sequence of text about a given topic. To do this, the model is trained on extremely large amounts of natural language, typically taken from aggregated sets of scraped websites and/or digitized books.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

The pre-training procedure follows a form of "self-supervised" learning. This is similar to supervised learning, except that the labels representing a correct prediction or "ground truth" for the model are taken from the training data itself, rather than relying on external labels added separately to the training data. Because natural language contains its own "correct" next-word predictions, pre-training is able to supervise itself, without the need for additional human-generated labels.

Pre-training consists of a series of steps, applied repeatedly across batches of examples until a preset number of training cycles is reached. In general, the training algorithm:

1. samples a sequence of text from the training data;
2. inputs the sequence (minus the last word) into the model to receive a prediction for the next word;
3. calculates the model prediction error for the sequence by taking the difference between the probability distribution of the prediction and that of the actual last word in the sequence; and
4. adjusts the value of each parameter in the model (using backpropagation) to reduce the error going forward.

The term "foundation model" is sometimes used to describe the resulting model after the completion of pre-training.[11] However, this term is somewhat controversial. The authors of the paper that coined the term claim to have chosen it to "capture the unfinished yet important status of these models" given their ability "to serve[] as the common basis from which many task-specific models are built via adaptation."[12] Yet, critics have countered that the term is self-serving and misrepresents the nature of the relationship these models have to human language

[11] See Rishi Bommasani, Drew A. Hudson, Ehsan Adeli et al., "On the Opportunities and Risks of Foundation Models," August 2021, https://doi.org/10.48550/arXiv.2108.07258.
[12] Ibid., p. 3 (n. 2) and p. 7.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

and understanding. One AI researcher in particular provided a memorable critique: "These models are really castles in the air. They have no foundation whatsoever."[13]

A more practical and plain-language description can be found outside of academic research. In the words of one AI practitioner, the result of pre-training is a model that "babbles Internet" in the form of a "document completer."[14]

Self-supervised training datasets

The quality and size of self-supervised training datasets for the purposes of pre-training has a significant impact on the capabilities of LLMs. Pre-training data comes from a variety of sources and its content can be divided into two broad categories:

- General data from web pages, books and social media conversations broadly improves linguistic knowledge and generalization skills; and
- Specialized data such as multilingual texts, scientific publications and codebases help to refine certain specific task-related skills.[15]

An important source that is used for the pre-training of LLMs comes from Common Crawl.[16] The noisy and low quality of the data and the biases in the distribution of web content make it unsuitable for direct use in training LLMs without some form of pre-processing. Several projects have been launched to improve the quality of the data, with the aim of producing more curated and cleaner datasets. Colossal Clean

[13] Quote from Jitendra Malik in: Will Knight, "A Stanford Proposal Over AI's 'Foundations' Ignites Debate," *Wired*, September 2021, https://www.wired.com/story/stanford-proposal-ai-foundations-ignites-debate/. A video of Malik's remarks is available at https://www.reddit.com/r/MachineLearning/comments/pd4jle/d_jitendra_maliks_take_on_foundation_models_at/.

[14] See Andrej Karpathy, "Let's build GPT: from scratch, in code, spelled out," January 2023, https://www.youtube.com/watch?v=kCc8FmEb1nY, at 1:51:45.

[15] Wayne Xin Zhao, Kun Zhou, Junyi Li et al., "A Survey of Large Language Models," 2023 https://arxiv.org/abs/2303.18223.

[16] Common Crawl is a non-profit organization, that crawls the web and creates publicly accessible archives and datasets of extracted text. It has been collecting petabytes of data since 2008. See https://commoncrawl.org/.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Crawl Corpus (C4)[17] and RefinedWeb[18] are two popular datasets obtained by filtering and deduplicating Common Crawl data.

Other commonly used corpora include carefully curated data from sources such as Wikipedia, social media platforms such as Reddit, book corpora such as Project Gutenberg,[19] consisting of 70000 literary books, code corpora from public software repositories such as GitHub and code-related answering platforms such as Stack Exchange, as well as scientific articles from ArXiv. The Pile,[20] an 825-gigabyte collection of 22 datasets, is an example of corpora constructed from diverse sources to provide a more balanced and representative dataset. However, it has faced issues of copyright infringement, with the Books3 dataset, a collection of nearly 200,000 books and part of The Pile, being removed from the internet following legal action.[21] Many additional data sets used to train LLMs, can be found on Hugging Face Datasets.[22]

LLMs are typically trained on multiple datasets, enriching the models' adaptability across to different contexts with carefully selected content for a particular purpose or context. For example, LLaMA[23] extracts training data from several sources including Common Crawl, C4, Github, Wikipedia, books (Project Gutenberg and Books3 section of The Pile), ArXiv and StackExchange.

The quality of LLMs is dependent on the quality of the training data. Various pre-processing techniques are employed to improve data quality. One such technique

[17] See Jesse Dodge, Maarten Sap, Ana Marasović et al., "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus," September 2021, https://arxiv.org/abs/2104.08758.
[18] See Guilherme Penedo, Quentin Malartic, Daniel Hesslow et al., "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only," 2023, https://arxiv.org/abs/2306.01116.
[19] See Project Gutenberg, https://www.gutenberg.org/.
[20] Leo Gao, Stella Biderman, Sid Black et al., "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," 2020 https://arxiv.org/abs/2101.00027.
[21] See A giant online book collection Meta used to train its AI is gone over copyright issues https://mashable.com/article/books3-ai-training-dmca-takedown
[22] See Hugging Face, https://huggingface.co/datasets.
[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard et al., "LLaMA: Open and Efficient Foundation Language Models," 2023, https://arxiv.org/abs/2302.13971

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

uses automatic filtering methods,[24] such as heuristic-based and classifier-based approaches, which distinguish and remove low-quality data. The approach of training a selection classifier on high-quality texts and using it to detect and remove low-quality data is called "classifier based."[25] The heuristic-based approach improves data quality by eliminating low-quality text through a well-designed set of rules.[26] Additional preprocessing steps, such as deduplication, privacy redaction and tokenization, further refine the training dataset. Deduplication ensures model stability, privacy redaction methods safeguard against personally identifiable information being unintentionally exposed through the model, and tokenization enhances processing efficiency.

It should be noted that these methods are not infallible and that there are regulatory and ethical considerations in regard to scraping the entirety of the Internet.[27]

Stage 2: Fine-tuning / alignment

After creating a general-purpose "foundation" model, the next stage in the training procedure of LLMs is to refine the behavior of the model to better "align" its responses with human preferences and values. The desired behavior can be distilled

[24] Wayne Xin Zhao, Kun Zhou, Junyi Li, et al., "A Survey of Large Language Models," 2023, https://arxiv.org/abs/2303.18223

[25] Typically, this method involves training a binary classifier, which sorts data into one of two categories, with well-curated data representing positive instances and unfiltered Common Crawl data representing negative instances. The method predicts a score that reflects the quality of each data instance.

[26] This includes language-based filtering to exclude irrelevant languages, metric-based filtering using perplexity (a measure of how well a language model predicts a sample of text, with lower values indicating better predictability) for unnatural sentences, static-based filtering to assess features like punctuation distribution and sentence length, and keyword-based filtering to eliminate HTML tags, hyperlinks, boilerplates, and offensive content.

[27] See GPA International Enforcement Cooperation Working Group, "Joint statement on data scraping and the protection of privacy," August 24, 2023, https://www.priv.gc.ca/en/opc-news/speeches/2023/js-dc_20230824/.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

into a set of three criteria, commonly referred to as the three H's: LLMs should act in a manner that is "helpful, honest and harmless."[28]

This is a challenging task. After pre-training, the model is optimized only to continue predicting next words in a sequence. This is a related, but ultimately distinct, task from following user instructions, avoiding falsehoods or refraining from toxic or harmful content. For example, when given a prompt in the form of a user instruction, OpenAI's third-generation generative pre-trained transformer or "GPT-3" simply continues to list examples of similar prompts, whereas the same model but fine-tuned to act as an "InstructGPT" (the precursor to OpenAI's popular "ChatGPT") provides a helpful response. See Table 1 for details.

Prompt	**Explain the moon landing to a 6 year old in a few sentences.**
Completion	GPT-3: Explain the theory of gravity to a 6 year old. Explain the theory of relativity to a 6 year old in a few sentences. Explain the big bang theory to a 6 year old. Explain evolution to a 6 year old.
	InstructGPT: People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.

Table 1: Example of different responses between GPT-3 and InstructGPT for the same prompt. From OpenAI, "Aligning language models to follow instructions," https://openai.com/research/instruction-following.

[28] See Amanda Askell, Yuntao Bai, Anna Chen et al., "A General Language Assistant as a Laboratory for Alignment," 2021, https://doi.org/10.48550/arXiv.2112.00861.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

In general, the training procedure of fine-tuning is divided into two (sub)stages. The first follows a form of "supervised" learning, while the second follows a form of "reinforcement" learning.

Supervised learning

This stage is similar to pre-training, except that the set of examples on which the model is trained are explicitly selected and curated by the developers to demonstrate the type of prompts the LLM is expected to receive and the type of responses it should provide. This is why the training is deemed to be "supervised." The training data contains full examples of task-specific interactions with the LLM, including both the user prompt and the "correct" LLM response.

The amount of training data used at this stage is typically much smaller—in the range of orders of magnitude less—than the amount used during pre-training. The reason for this is due to both practical and scientific considerations. From a practical perspective, creating tailored supervised training datasets is far more resource intensive and time consuming than downloading collections of scraped websites and/or digitized books for use in self-supervised learning, especially given the amount of online digital content available today. Yet from a machine learning perspective, less but high-quality data is actually "more" at this stage. Studies have shown that supervised fine-tuning is "sample efficient," in the sense that comparably less data is needed to train the LLM to perform well on a specific task, such as follow user instructions in a chat-like manner.[29] Thus, using the pre-trained model as a basis, supervised learning is able to tweak the parameters of the model to transform its raw, unrefined linguistic abilities into more direct and purposeful behavior. After this stage of training, LLMs respond more "helpfully."

[29] See Urvashi Khandelwal, Kevin Clark, Dan Jurafsky et al., "Sample Efficient Text-Summarization Using a Single Pre-Trained Transformer," 2019, https://doi.org/10.48550/arXiv.1905.08836.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Reinforcement learning

Yet being able to perform a task directly is not the same as being able to perform it responsibly or ethically. While supervised learning can train LLMs to provide more helpful responses, in general, the modifications do not extend to the values of honesty and harmlessness. To gain better alignment with these other values, LLMs typically undergo a second stage of fine-tuning using a technique known as "reinforcement" learning.

Reinforcement learning is a form of machine learning in which a model is trained by interacting in a dynamic environment with feedback, similar to a process of "trial and error." Unlike supervised or self-supervised learning, the model does not learn by way of repeated exposure to examples of "correct" behavior. Instead of a form of imitation, the key pedagogical concept at work in it is that of "reward and punishment." A model is rewarded for behavior that achieves or takes it closer to the goal of the environment and punished for behavior that does the opposite. By exploring different strategies to achieve the goal and updating its parameters based on the positive or negative feedback it receives, the model develops an optimal "policy" that maximizes the reward associated with the environment. Thus, reinforcement learning is more open-ended and exploratory than other forms of machine learning. This is why it is typically used to train models in strategy-based tasks such as games like Go[30] or StarCraft.[31]

In the case of LLMs, the "game" the model is trained to play is that of responding ethically and appropriately to user prompts. While at first blush this may seem like an analogous task to strategic game play, upon closer inspection it becomes clear that ethical decision-making differs in important respects. These differences pose a number of challenges to the application of reinforcement learning within the context of LLMs.

[30] See David Silver, Aja Huang, Chris J. Maddison et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, 2016, p. 484–489, https://doi.org/10.1038/nature16961.
[31] See Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, 2019, p. 350–354, https://doi.org/10.1038/s41586-019-1724-z.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

The main challenge is that, unlike strategic games such as Go or StarCraft, there is no precise definition for what constitutes a "win" in ethics. Ethics differs from strategic game play in that it does not occur under the direction of a predefined goal or outcome such as "achieving a high score" or "defeating an opponent." There is no separate, "higher" end or objective under which its actions are subsumed. Ethical action is done for the sake of itself, simply because it is the right thing to do. As Aristotle explains the distinction, "the end of making [e.g., strategic game play] is different from itself, but the end of [ethical] action could not be, since acting well is itself the end."[32]

A consequence of this property is that ethical criteria are inherently ambiguous. They do not admit of the same precision as mathematics or the natural sciences. This is a challenge for reinforcement learning because without a precise or well-defined objective, the training process cannot determine whether some action or strategy employed by the model should be rewarded or punished. Since ethical action is its own end, reinforcement learning cannot simply define an external objective by which to evaluate the responses of LLMs.

A second challenge of reinforcement learning within the context of LLMs has to do with the multiplicity of ethical values. The "game" of ethics the model is trained to play does not consist of one value (or "virtue" in Aristotle's terminology) but a combination of three. To respond ethically and appropriately to user prompts, LLMs must act in accordance with the values of helpfulness, honesty and harmlessness.

This raises an additional challenge in that the meanings of these values overlap and conflict with each other, especially when taken to extremes. Due to their inherent ambiguity, instead of being mutually compatible—or in machine learning parlance, mutually "maximizable"—the values of helpfulness, honesty and harmlessness exhibit an inherent tension or tradeoff, where too much of one results in too little of another. This further complicates the task of defining an ethical objective by which to train LLMs using reinforcement learning. In addition to the challenge of programmatically defining ethical values, the "game" of LLMs includes that of

[32] See Aristotle, *Nicomachean Ethics*, 1140b8.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

determining the right proportion of each value to apply when formulating a response to a user request or prompt.

For example, LLMs that are trained to act in accordance with the value of helpfulness tend to conflict with the value of harmlessness, since they "helpfully" respond to any query or user prompt, even ones that ask the system to produce toxic or harmful content, such as overtly sexist or racist material. By the same token, LLMs that are trained to act in accordance with the value of harmlessness tend to conflict with the value of helpfulness, since they "harmlessly" avoid inappropriate content at all costs, to the point where they refuse to respond to valid or innocuous requests.[33]

The relationship between helpfulness and honesty is also illustrative, but for different reasons. In this case, the tension is not so much due to a direct conflict between values, but to the essential indifference of LLMs to the truth (or falsity) of their claims. LLMs are trained to continuously predict the next word or sub-word "token" in a sequence of text based on the characteristics of their training data. This is not the same objective as truth. For a statement to be true, it must relate to and accurately describe the world. LLMs have no world model against which to verify their claims. They only interact with words, not the world. Accordingly, the objective they are optimized for is not truth, but the *appearance* of truth.

This is why LLMs are often (ab)used to write fiction books.[34] It is also why they tend to conflict with the value of honesty by default. In an effort to be more "helpful," they prioritize eloquence over objectivity. They produce content that sounds convincing and seems factual to the user, even if in reality it is false, inaccurate or even nonsensical. Instances of this tendency of LLMs to produce content detached

[33] It is interesting to note that this tradeoff between helpfulness and harmlessness has led to the creation of a marketplace of LLMs based on the degree to which they respond to user requests. For example, xAI's "Grok" LLM markets itself as a chatbot willing to "answer spicy questions that are rejected by most other AI systems" (see xAI, "Announcing Grok!," Nov. 5, 2023, https://twitter.com/xai/status/1721027348970238035).

[34] See, for example, Ella Creamer, "Amazon restricts authors from self-publishing more than three books a day after AI concerns," *The Guardian*, September 20, 2023, https://www.theguardian.com/books/2023/sep/20/amazon-restricts-authors-from-self-publishing-more-than-three-books-a-day-after-ai-concerns.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

from reality are commonly referred to as "hallucinations" or sometimes "confabulations."

In general, there are two kinds of hallucinations. The first and more obvious kind are hallucinations triggered from prompts that specifically request false or misleading content. This is what happened with Meta's now defunct "Galactica" LLM. Originally marketed as a tool to aid in the production of "scientific knowledge,"[35] Galactica was taken offline after only three days after it was discovered it would produce scientific-sounding, but entirely false wiki articles on fictitious topics such as the "flux capacitor" or "Streep-Seinfeld theorem."[36]

The second kind of hallucination are those that arise directly from the LLM itself, unbeknownst to the user. These are more pernicious and difficult to detect. There are many documented examples,[37] but one notorious case involves false criminal accusations against an individual. After being asked "What scandals have involved law professors?" ChatGPT provided a false narrative claiming that a real-life law professor had been accused of sexual harassment by a student.[38] What is even more concerning, however, is that the prompt included a request to "[p]lease cite and quote newspaper articles," to which ChatGPT "helpfully" obliged by appending a false quote from a non-existent source.

How, then, can a "win" in ethics be defined for the purposes of reinforcement learning within the context of LLMs? Given the ambiguity of ethical criteria as well as the general incompatibility between the values of helpfulness, honesty and

[35] See Ross Taylor, Marcin Kardas, Guillem Cucurell et al., "Galactica: A Large Language Model for Science," 2022, https://arxiv.org/abs/2211.09085.
[36] See Ernest Davis and Andrew Sundstrom, "Experiment with GALACTICA," 2022, https://cs.nyu.edu/~davise/papers/ExperimentWithGalactica.html.
[37] See Gary Marcus and Ernest Davis, "Large Language Models like ChatGPT say The Darnedest Things," 2023, https://garymarcus.substack.com/p/large-language-models-like-chatgpt.
[38] See Eugene Volokh, "Large Libel Models: ChatGPT-3.5 Erroneously Reporting Supposed Felony Pleas, Complete with Made-Up Media Quotes?," 2023, https://reason.com/volokh/2023/03/17/large-libel-models-chatgpt-4-erroneously-reporting-supposed-felony-pleas-complete-with-made-up-media-quotes/.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

harmlessness, how can a precise goal or objective be defined by which to train LLMs to act more ethically?

This problem remained a barrier to the adoption of LLMs until a special technique was developed that enabled reinforcement learning to be applied to more "insightful" tasks based solely on human judgement, such as ethics. This technique came to be known as "reinforcement learning from human feedback" (RLHF).[39]

How it works is that, instead of attempting to programmatically define a set of ethical criteria directly, RLHF leverages the capabilities of machine learning to indirectly "discover" the features of such criteria by modeling the preferences of human evaluators. In general, the technique follows a five-step process:

1. Task a group of human evaluators to review multiple LLM responses to the same prompt and then rank the responses in order of most to least ethical, that is, according to how well each response balances the values of helpfulness, honesty and harmlessness;
2. Create a supervised training dataset from the prompts, responses and human rankings, with the rankings serving as labels;
3. Train a supervised model to learn the implicit features of what constitutes a "winning" response in the "game" of ethics, that is, what indirectly constitutes the criteria of the values of helpfulness, honesty and harmlessness;
4. Set this learned "preference model" as the reward function for the LLM within the context of a reinforcement learning environment; and
5. Further fine-tune the LLM to act in accordance with the values of helpfulness, honesty and harmlessness by rewarding it for responses that fit the criteria of the preference model and punishing it for responses that do not.

[39] See Paul F. Christiano, Jan Leike, Tom B. Brown, et al., "Deep Reinforcement Learning from Human Preferences," 2017, https://arxiv.org/abs/1706.03741; Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, et al., "Fine-Tuning Language Models from Human Preferences," 2019, https://arxiv.org/pdf/1909.08593.pdf; and Nisan Stiennon, Long Ouyang, Jeff Wu, et al., "Learning to summarize from human feedback," *34th Conference on Neural Information Processing Systems* (NeurIPS 2020), https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Despite RLHF's ability to define an ethical objective for use in reinforcement learning, its method for "automating ethics" comes with a number of limitations. The main drawback is that the technique cannot ensure that the judgements made by the human evaluators are in fact appropriate or ethical. Just because a group of randomly selected humans are tasked with using their judgement does not entail that the results are ethical. The evaluators themselves could be biased or prone to making flawed decisions, in which case RLHF would simply reinscribe the unethical tendencies of the evaluators, but under the guise of an "objective" mathematical process.

Moreover, even assuming a non-biased population of human evaluators, the conditions in which they exercise their judgement could be coercive or exploitative, thereby negatively affecting their ability to rank LLM responses appropriately. For example, as reported by *Time Magazine*, OpenAI used Kenyan workers paid less than $2 an hour to create their RLHF training data for ChatGPT.[40]

In response to concerns about RLHF, another technique was developed known as "reinforcement learning from AI feedback" (RLAIF).[41] This technique follows the same process as RLHF, but with two important differences: (1) instead of human evaluators, it tasks the LLM itself with evaluating multiple LLM responses to the same prompt; and (2) instead of a set of ethical values, it provides the LLM with a "constitution" consisting of a set of principles, along with some examples of appropriate evaluations. For this latter reason, RLAIF is sometimes referred to as "constitutional AI."

While RLAIF may improve the scalability of results, it still suffers from some of the same limitations as RLHF. Just as RLHF cannot ensure that the decisions made by a group of human evaluators are appropriate or ethical, so too RLAIF cannot guarantee that the LLM's evaluations are not biased or flawed in some way. Indeed,

[40] Billy Perrigo, "Open AI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," January 18, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/.
[41] See Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al., "Constitution AI: Harmlessness from AI Feedback," December 2022, https://arxiv.org/abs/2212.08073.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

the risk may be even greater in the case of RLAIF, since the LLM is tasked with making ethical evaluations *before* it has been fine-tuned to act more ethically.

Stage 3: Use

After LLMs have been pre-trained and fine-tuned for alignment, the next overall stage in their development is their use. This typically takes the form of an interactive question-and-answer dialogue with an end-user. In general, there are two main components that arise at this stage: prompts and the temperature parameter.

Prompts

A prompt is a set of instructions given to an LLM. It customizes or enhances the LLM's capabilities and can take the form of a question, statement, or request for information.[42] The LLM analyses the prompt and generates a response in real-time, such as providing information, summarizing, answering questions, or generating content.

Prompt processing

When presented with a prompt, the LLM generates a response based on the new input data. The prompt is first transformed into tokens[43] before being processed by the LLM's neural network. The transformer architecture uses the attention mechanism to determine the significance of each token in relation to others, helping the model understand the semantics, nuances, and intent of the prompt by considering how each token relates to those around it.[44] The model then employs its trained parameters, which consist of a vast number of weights and biases

[42] Jules White, Quchen Fu, Sam Hays et al., "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," February 2023, http://arxiv.org/abs/2302.11382.
[43] See Appendix A, section on "Vocabulary."
[44] Ashish Vaswani, Noam Shazeer, Niki Parmar et al. "Attention is all you need." *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008, https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

adjusted during the training phase, to predict the next token in a sequence based on the preceding tokens. To generate text, the model begins with the context provided by the prompt and repeatedly predicts the next token until it forms a complete answer. The model can use various methods[45] to select the most likely next token. It generates a sequence of tokens that create a coherent response and translates the tokens back into a normal sentence/word sequence.

Elements of a prompt

Prompts can be simple or complex, depending on the situation. Users can guide the model by providing context and information in their prompts. A prompt may consist of several elements, including:

- Instructions to direct the model on the specific task or action required, such as asking it to generate a story, solve a problem, or provide an explanation;
- Context to provides background information or situational details;
- Input data to form the actual content or text that the model processes. This could be a question, a statement, or paragraph from which the model derives the information needed to generate a response;
- Output indicators to provide cues within the prompt that signal to the model how to format or structure its response. For example, if the output should be a list, a summary, or a detailed answer, these indicators help guide the output's form and extent; and
- Examples to illustrate the kind of response expected.[46]

Example prompt, with elements in brackets:

> *Translate the following sentences into French (instruction). It is for a colleague at work (context). 'Throw a spanner in the works'. 'Bob's your uncle' (input data). Provide the translations followed by explanations of any*

[45] LLMs can predict the next token using methods like greedy sampling (selecting the highest probability token) and top-k/top-p sampling (choosing from a limited set of likely tokens). Adjustments such as temperature scaling and repetition penalties refine the selection process, balancing randomness and coherence.
[46] Prompt Engineering Guide, "Elements of a Prompt," https://www.promptingguide.ai/introduction/elements.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

idiomatic or cultural references and an example sentence. Keep it simple, concise and fun (output indicators).

Prompting techniques

Prompting techniques can be categorized into different types.[47,48] Zero-shot prompting relies on the model's pre-existing knowledge and does not provide any specific examples.[49] For example, prompt:

Classify the sentiment of the following sentence: 'I absolutely love the new design of the car!'

Response:

The sentiment of the sentence 'I absolutely love the new design of the car!' is ***positive***

In this case, the model has not been trained on its specific task but relies on its generalized understanding of sentiment analysis to respond.

In contrast, few-shot prompting provides the model with examples of the desired task or output to guide it to the expected response pattern.[50]

For example, prompt:

Craft engaging social media posts for product announcements:

Product: Eco-Friendly Laundry Detergent,

Post: 'Excited to launch our new Eco-Friendly Laundry Detergent! 🌍🌿 Made with natural ingredients, it's tough on stains but gentle on the planet. #EcoClean'

[47] See Prompt Engineering Guide, "Prompting Techniques," https://www.promptingguide.ai/techniques.
[48] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, et al., "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," 2024, https://arxiv.org/abs/2402.07927.
[49] Prompt Engineering Guide, "Zero-Shot Prompting," https://www.promptingguide.ai/techniques/zeroshot. This approach tests the model's ability to generalize from its training data to new, unseen tasks.
[50] Prompt Engineering Guide, "Few-Shot Prompting," https://www.promptingguide.ai/techniques/fewshot.

> *Product: High-Tech Running Shoes*
>
> *Post: 'Step into the future with our latest High-Tech Running Shoes! 🏃‍♂️🚀 Engineered for speed and comfort. Ready, set, go! #RunFaster'*
>
> *Now, create a post for this product: Gourmet Dark Chocolate*

Response:

> *Here's an engaging social media post for the Gourmet Dark Chocolate product:*
>
> *Indulge in the rich, velvety taste of our new Gourmet Dark Chocolate. 🍫❤️ Handcrafted with the finest cocoa beans, it's a luxurious treat for your senses. Savor the moment! #ChocoLovers #GourmetDelight*

In this few-shot setup, the model is given two examples and is expected to produce a similar output.

Other techniques, such as chain-of-thought prompting, break complex tasks into simpler, intermediate steps allowing the model to explain each step before reaching final answer.[51]

For example, prompt:

> *A company needs to anonymize user data stored across 3 different databases: Database A, Database B, and Database C. Database A contains 2.3 million records, with 32% requiring anonymization. Database B contains 3.5 million records, with 76% requiring anonymization. Database C contains 4.6 million records, with 91% requiring anonymization. How many records in total need to be anonymized? Show your reasoning step by step*.

The process of prompt engineering involves the design, refinement and optimization of input prompts with the objective of obtaining the best response from the

[51] Chain of thought prompting is a natural language processing technique that guides a model through a step-by-step reasoning process. See Jason Wei, Xuezhi Wang, Dale Schuurmans, et al., "Chain of thought prompting elicits reasoning in large language models," 2022b, https://arxiv.org/pdf/2201.11903.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

model.[52,53] This process necessitates a user's understanding of how the model works and the experimentation with different prompting techniques to determine the most effective method of communicating the task to the model.[54]

As user-directed queries into LLMs, prompts raise a number of issues such as the risk of misuse of prompting,[55] and the potential for introducing or amplifying biases that may exist in the training data .[56] Therefore, it is important for developers and deployers of LLMs to implement appropriate safeguards to avoid any potential harm or biases that could be introduced through prompting.[57]

Temperature parameter

Temperature is a configuration variable that adjusts the degree of randomness of the responses generated by the LLM.[58] The LLM generates responses by calculating the likelihood of different tokens being next in sequence. It assigns raw scores to each token, based on its likelihood to follow the given text. These scores are then converted into probabilities. The techniques used ensure that the probabilities assigned to each potential token add up to one, creating a probability distribution over all possible next tokens, from which the model can select.

[52] See "What is prompt engineering," https://www.ibm.com/topics/prompt-engineering.
[53] See "Why Prompt Engineering is the Key to Mastering AI," https://hackernoon.com/why-prompt-engineering-is-the-key-to-mastering-ai.
[54] Jules White, Quchen Fu, Sam Hays, et al., "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," 2023, http://arxiv.org/abs/2302.11382.
[55] Xiaodong Wu, Ran Duan, and Jianbing Ni. "Unveiling security, privacy, and ethical concerns of ChatGPT. Journal of Information and Intelligence, vol. 2., no. 2, March 2024, https://www.sciencedirect.com/science/article/pii/S2949715923000707.
[56] AI Safety Institute, "AI Safety Institute approach to evaluations," February 2024, https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations.
[57] Safeguards could include content filtering for harmful outputs, mechanisms for users to report unexpected behaviors, monitoring how models are being used to help identify patterns of misuse, educational outreach, providing transparency about how the models work, and the data were trained on to better understand potential biases and risks with their outputs.
[58] Prompt Engineering Guide, "LLM settings," https://www.promptingguide.ai/introduction/settings.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

Higher temperatures lead to a smoother, flatter distribution across tokens, increasing the randomness of the text by making less probable tokens more likely to be selected. Conversely, lower temperatures lead to a sharper distribution, resulting in the model's choices being more predictable and focused on the most likely tokens. Temperature can be set as part of the query when using LLMs via APIs or various interfaces. This enables greater control over the balance between randomness and determinism in the LLM's response.

3. Risks to data protection and privacy

LLMs carry with them significant privacy, data protection, and data security risks, some of which may be mitigated and some of which may be inherent to the systems themselves. Common practices like indiscriminate scraping to create training datasets, irregular or non-existent audits of training data and outputs, black box algorithms that cannot be reviewed or explained, and a lack of technical safety measures embedded in LLMs all contribute to these high risks. Without protections in place, these systems can cause serious individual and societal harms, perpetuate discrimination against marginalized groups, embed bias into the algorithms themselves, and exacerbate data misuse and the risk of breach.

The rush to deploy these systems without adequate testing for risks and weaknesses has allowed LLMs to quickly become embedded in numerous industries,[59] exposing the public to risks with few guardrails or protections. While

[59] *See e.g., CFPB Issue Spotlight Analyzes "Artificial Intelligence" Chatbots in Banking,* Consumer Financial Protection Bureau (Jun. 6, 2023), https://www.consumerfinance.gov/about-us/newsroom/cfpb-issue-spotlight-analyzes-artificial-intelligence-chatbots-in-banking/; Ashley Belanger, *Air Canada must honor refund policy invented by airline's chatbot*, ArsTechnica (Feb. 16, 2024), https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/; Plan for Promoting Responsible Use of Artificial Intelligence in Automated and Algorithmic Systems by State, Local, Tribal, and Territorial Governments in Public Benefits Administration, United States Department of Health and Human Services (Mar. 28, 2024), https://www.hhs.gov/sites/default/files/public-benefits-and-ai.pdf.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

many enforcement bodies have attempted to make clear that existing data protection, product safety, civil rights, consumer protection, and other regulations apply to LLMs, many developers still operate as if the novelty of the technology somehow exempts it from these regulations.[60] The lack of transparency and accountability means little option for grounded and effective recourse for those harmed and increased difficulties faced by regulators attempting to engage in meaningful creation and enforcement of data protection, consumer protection, civil rights, and civil liberties laws. In general, LLMs amplify serious risks to individuals, democracy, and cybersecurity. Even the creator of OpenAI has stated, "I'm especially concerned that these models could be used for widespread misinformation… [and] offensive cyberattacks."[61]

We must carefully examine the risks to privacy and data security if we are to protect individuals and society from those harms. To that end, we set forth the privacy and data security risks stemming from LLMs below. Please note that this technology and associated risks are still developing. In addition, LLMs present several risks that, while not as directly related to privacy and data security, deeply affect individuals and may fall under the consumer protection remit of DPAs (namely information manipulation, increased data processing, misinformation, and disinformation). We discuss those harms as well at the end of this section. This section aims to be thorough but does not claim to be comprehensive.

Increased data processing

Many LLM developers want as much training data as possible to develop their LLMs with the belief that the more training data, the more a system can "learn" and the more complex and precise the output. The AI industry incentivizes unrestricted,

[60] We refer here to several practices, including wide-spread data scraping without a clear legal basis, copyright violations, arguments over who is liable for harmful outputs, generation of illegal images, defamatory outputs, etc.
[61] Victor Ordonez, Taylor Dunn, and Eric Noll, *OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks: 'A little bit scared of this,'* ABC News (March 16, 2023), https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

mass collection.[62] This creates a data arms race and directly conflicts with privacy principles and responsible data practices like data minimization. To address this perceived need for mass data, many LLM developers set up systems that indiscriminately and continuously scrape the internet for data.[63] While some developers may review and "clean" the scraped data before use, like Google's C4, many either skip this step or cannot keep up quality checks without limiting the volume of information absorbed—a sacrifice many are unwilling to make.[64] This means that training datasets will include inaccurate, biased, and discriminatory data as well as personal data of individuals completely unaware that their information is now being used by an LLM. The lack of review or transparency in dataset creation also creates problems around establishing a proper legal basis for collection and allowing individuals to exercise data rights. Further, the practice is directly counter to the principle of data minimization and may often violate the purpose limitations of why personal data may have been available online in the first place. To this end, many high traffic websites have included text in the coding of their platforms in an attempt to block web crawlers, but this code may or may not be heeded.[65]

In addition, mass scraping pulls from multiple different sources, allowing for data combination and inferences that can reveal even more detailed and sensitive information about an individual. This may ultimately lead to a chilling effect of free speech and expression online, reflecting studies that have shown that individuals often self-censor when they suspect they are being surveilled.[66] As people become

[62] Apostol Vassilev, et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2023 (Jan. 2014), https://doi.org/10.6028/NIST.AI.100-2e2023 at 40.
[63] *See e.g.* Kristi Hines, *OpenAI Launces GPTBot With Details on How to Restrict Access*, Search Engine J. (Aug. 7, 2023), https://www.searchenginejournal.com/openai-launches-gptbot-how-to-restrict-access/493394/; Kevin Schaul, et al., *Inside the secret list of websites that make AI like ChatGPT sound smart*, Wash. Post (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.
[64] Misinformation on Bard, Google's New AI Chat, Center for Countering Digital Hate (Apr. 5, 2023), https://counterhate.com/research/misinformation-on-bard-google-ai-chat/.
[65] Kali Hays, OpenAI's GPTBot and other web crawlers are being blocked by even more companies now, Bus. Insider (Sep. 28, 2023), https://www.businessinsider.com/openai-gptbot-ccbot-more-companies-block-ai-web-crawlers-2023-9.
[66] *See* Jon Penney, *Understanding Chilling Effects,* 106 Minnesota Law Review 1451 (2022), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3855619.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

more and more aware that any information about them online can be scraped and used in ways they can't anticipate or control, they may post less information and withdraw from any online discussions or communities.[67] This not only would stifle online speech and public discourse, it may undermine the utility of major facets of the internet altogether.

Loss of data rights

The personal data included in mass-collected training datasets may have come from the individuals themselves or could be information that others post about them. However, just because data is available does not mean it is legally or ethically open for the taking.[68] Many website policies restrict how data on their site may be used, information may have been posted without an individual's consent, confusing settings may lead to information being publicly viewable that was intended to be private, information from data breaches may be released online, or information may have been posted for a specific purpose (for example, information posted on LinkedIn solely to find a job). Indeed, we have already seen examples of highly sensitive data being scraped and used in training datasets that was supposed to be confidential, such as when LAION-5B's public database was found to contain photographs from private medical records.[69] Data scraping undermines individual control over their personal data and takes their ability to control how their data is used, particularly since individuals will frequently be entirely unaware that their data is being used by LLMs. Even where consent is not the legal basis for

[67] *see* Jeramie D. Scott, *Social Media and Government Surveillance: The Case for Better Privacy Protections for our Newest Public Space*, 12 J. Bus. & Tech. L. 151 (2017), https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=1272&context=jbtl.; Jonathon W. Penney, Understanding Chilling Effects, 106.3 Minn. L. Rev 1451 (2022), https://digitalcommons.osgoode.yorku.ca/cgi/viewcontent.cgi?article=4074&context=scholarly_works.

[68] *See* "Joint Statement on data scraping and the protection of privacy, Office of the Privacy Commissioner of Canada, et. Al (August 24, 2023), available at https://www.priv.gc.ca/en/opc-news/speeches/2023/js-dc_20230824/.

[69] Benj Edwards, *Artist finds private medical record photos in popular AI training data set,* ArsTechnica (September 21, 2022), https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

processing personal data, individuals must be informed and able to exercise their personal data rights.

The nature of LLMs makes exercising certain data rights very challenging, particularly the right to correct data or request deletion of the personal data often present in training datasets. While some datasets may be more tightly curated and checked for the origin and necessity of including personal data, scraping datasets in particular may include unnecessary personal data, personal data that was only made available through data breaches, or defamatory or inaccurate information about an individual. While some may argue that this is not a concern because datasets are not meant to be made public, there are still significant issues here. First, whether public or not, an individual may simply not want certain personal data to be processed in this way. Second, there have been tests performed that demonstrate LLMs can be tricked into revealing the raw data contained in the training datasets,[70] meaning the personal data within them can be exposed to a broad audience. Finally, the existence of the dataset opens the possibility for improper access or breach.

In general, scraping goes against multiple legal rights and protections as well. For example, purpose limitation would dictate that personal data cannot and should not be used for anything other than the specific purpose for which it was provided.[71] Oftentimes, individuals do not think about web crawlers or scraping algorithms when they are posting to social media websites. Many of these websites even include restrictions in their terms of service to block web crawlers.[72] The secondary use of these social media posts to train LLMs should be restricted and the entities

[70] Milad Nasr, Nicholas Carlini, Jon Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, Katherine Lee, "Extracting Training Data from ChatGPT" (Nov. 28, 2023), available at https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html.
[71] *See e.g.* Disrupting Data Abuse: Protecting Consumers from Commercial Surveillance in the Online Ecosystem, EPIC (Nov. 2022), https://epic.org/wp-content/uploads/2022/12/EPIC-FTC-commercial-surveillance-ANPRM-comments-Nov2022.pdf at 42-44.
[72] Kali Hays, OpenAI's GPTBot and other web crawlers are being blocked by even more companies now, Bus. Insider (Sep. 28, 2023), https://www.businessinsider.com/openai-gptbot-ccbot-more-companies-block-ai-web-crawlers-2023-9.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

that create the scrapers should be required to provide a lawful basis for processing the data it is collecting. Further, the lack of transparency and permissions here makes it nearly impossible for individuals to exercise their rights over their personal data. For example, an individual cannot correct out of date or inaccurate personal data if they are unaware it is held by an LLM developer. This would mean that the LLM continues to train on the incorrect data, making false associations with the individual and potentially spreading that false information in generated content. In fact, the Dutch DPA, a regulatory body charged with enforcing the GDPR, recently published guidelines on scraping the internet for data.[73] In the guidelines, the Dutch DPA stated that web scraping almost always violates the GDPR, both because of the lack of lawful basis for processing as well as the lack of notification to data subjects that their data is being processed.[74] The only legitimate basis for processing data for scraping would be "legitimate purpose," but the Dutch DPA ruled that if the only interest the controller has is a commercial purpose, that the legitimate purpose basis is not met.[75]

Beyond the training dataset itself, because LLMs make and embed within their decision-making systems connections and patterns drawn from the training datasets, the implications of data will likely carry on even if the data within the training dataset is altered or removed. Connections drawn from incorrect data would still influence outputs of the LLM after the data is corrected or removed. For example, if a training dataset contained false information that a professor had been accused of harassing students, that connection gets built into the system as it's trained on the dataset. Removing or correcting that information after training has occurred would not necessarily stop the professor's name from being included on,

[73] Autoriteit Persoonsgegevens, Richtlignen scraping door private organisaties en particulieren (May 1, 2024), https://www.autoriteitpersoonsgegevens.nl/uploads/2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf. *See also* Joint statement on data scraping and the protection of privacy, Office of the Privacy Commissioner of Canada (Aug. 24, 2023), https://www.priv.gc.ca/en/opc-news/speeches/2023/js-dc_20230824/.
[74] Autoriteit Persoonsgegevens, Richtlignen scraping door private organisaties en particulieren (May 1, 2024), https://www.autoriteitpersoonsgegevens.nl/uploads/2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf.
[75] Ibid.

say, a generated response to the query "what professors have harassed students." The process for "removing" this data from the model oftentimes means extensive retraining to ensure the LLM is no longer incentivized to rely on the undesirable data.[76] Even the onerous process of "removing" data may not actually guarantee that the information has been removed from the model, leaving it vulnerable to extraction attacks and leading to undesirable automated decision making.[77]

Finally, the analytical capabilities of LLMs make them extremely likely to be used in decision making process. The settings for LLM automatic decision making could include grading in education, job application evaluation, medical triage, auto-purchasing, loan approval, granting permits, sentencing parameters, and more. Some regulations, like the GDPR, grant data subjects the right to see the decision-making process that leads to an automated decision. The lack of transparency in LLMs' output derivation process makes it challenging to exercise this right. Even where humans are a part of the decision-making process, they may be unable to account for the LLM's "logic" in its decision-making contributions.

Harassment, impersonation, and extortion

LLM capabilities can be used for intentional abuse targeted at individuals. These forms of abuse often are crafted using the individual's personal data or generating false personal data that can be very challenging to disprove, impacting the individual's mental health, relationships, reputation, and more. Malicious users may intentionally feed false or harmful personal data about an individual to an LLM through prompts or other sources of training data, so that it "learns" that data, connects it to the individual, and then generates and proliferates content incorporating the wrong data and spreading it.[78] LLMs can also be trained on a person's individual speaking or writing style, allowing it to generate convincing

[76] Vaidehi Patil, et al., *Can Sensitive Information be Deleted From LLMs? Objectives for Defending Against Extraction Attacks*, arXiv (Sep. 29, 2023) (Pre-print), https://arxiv.org/pdf/2309.17410.
[77] Ibid.
[78] We note that not every LLM system automatically incorporates training data or indiscriminately scrapes personal data into its training datasets, but this example refers to the systems without good data curation practices.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

impersonations that may damage that person's reputation (for example, if the impersonations are offensive or go against that person's core beliefs or image). In the scam context, LLMs have been used to draft persuasive phishing campaigns, impersonating executives to gain financial or otherwise incriminating sensitive business data.[79] There are uncountable legal or otherwise significantly prejudicial consequences for victims targeted by those using LLMs to impersonate others, for both the impersonated individual as well as the target of the scam. These possibilities have already resulted in an official warning from the European Union's agency to combat serious international and organized crime that ChatGPT and other tools may be exploited by criminals.[80]

Scams

LLMs threaten to dramatically escalate the volume of scam robocalls (by generating scripts), texts, and emails targeting the public, exacerbating an already serious problem. A recent report revealed that over one billion scam robocalls were made to American phones each month and 2021 saw 2.8 million individuals file fraud reports with the FTC.[81] In 2022, the FTC reported over $326 million lost solely from scam texts.[82] These types of scams often will target vulnerable populations, such as seniors, people in debt, those with disabilities, college students, or immigrants. Often, they will be used to trick people into revealing sensitive personal or data security information (like financial information, passwords, etc.). LLMs have

[79] The State of Phishing 2023, SlashNext Security (Oct. 2023), https://slashnext.com/state-of-phishing-2023/ at 18.
[80] Foo Yun Chee, *Europol sounds alarm about criminal use of ChatGPT, sees grim outlook,* Reuters (March 27, 2023), https://www.reuters.com/technology/europol-sounds-alarm-about-criminal-use-chatgpt-sees-grim-outlook-2023-03-27/.
[81] Press Release, FTC, New Data Shows FTC Received 2.8 Million Fraud Reports from Consumers in 2021 (February 22. 2022), https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0; National Consumer Law Center & EPIC, Scam Robocalls: Telecom Providers Profit (2022), https://epic.org/documents/scam-robocalls-telecom-providers-profit/.
[82] *Reported losses from text scams more than doubled from $131M to $330M between 2021 and 2022,* FTC Consumer Sentinel Network, Fraud Reports by Contact Method, Reports & Amount Lost by Contact Method (2023), https://public.tableau.com/app/profil/federal.trade.commission/viz/FraudReports/FraudFacts ("Losses & Contact Method" tab selected, with quarters 1 through 4 checked for 2021, 2022).

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

accelerated the use of text-based phishing scams, with one cybersecurity firm finding an increase of over 1200% in email phishing scams from 2022 to 2023.[83]

Individuals can use LLMs to generate robo-texts, robo-emails, and mailers, as well as using the text generated by LLMs in conjunction with audio and video synthetic content to create more persuasive impersonations. Not only does the sheer volume of scams put out increase, but LLMs can make the pool of people committing fraud exponentially larger by helping those with limited skills in a given language craft natural and believable-sounding content that would otherwise be more easily flagged as a scam. Combine this content with data brokers that may sell lists of phone numbers, email addresses, data breaches, and categories or "insights" about potential targets and we have a recipe for consumer loss on an unprecedented scale.

Data security risks

LLMs will almost certainly expand the scope and volume of existing data security risks. Data breaches, ransomware-as-a-service, malware-as-a-service, and other hackers-for-hire services all become more likely and more dangerous with a higher volume of personal or sensitive data in the information ecosystem.[84] LLMs' data scraping practices, massive training data sets, and outputs add to that risk – and

[83] The State of Phishing 2023, SlashNext Security (Oct. 2023), https://slashnext.com/state-of-phishing-2023/ at 2.
[84] *See e.g.* The State of Phishing 2023, SlashNext Security (Oct. 2023), https://slashnext.com/state-of-phishing-2023/ at 4; Apostol Vassilev, et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2023 (Jan. 2014), https://doi.org/10.6028/NIST.AI.100-2e2023 at 40.

we already have numerous examples of breaches,[85] information leaks,[86] and technical workarounds improperly granting access to new data sources.[87]

Hackers and other bad actors (including actors that, without a tool like an LLM, would not be able to come up with content sophisticated enough to succeed in a data hacking-related scheme) could potentially use LLMs to draft or scale up versions of malware code, phishing and spear-phishing attempts, and emails targeting businesses to gain account information or compromise email.[88] New threat methods specific to LLMs may also become a problem, such as mining information fed into the LLM's training data set or strategically and purposely poisoning the data set with bad data, as well as methods of attack we have not yet imagined.

Cybersecurity threats

Another danger of LLM-generated content is highly sophisticated phishing concerns. In addition to the typical phishing attempts that trick users into clicking dangerous links or revealing access information, LLMs may soon be used by novice hackers to create malware that requires only minimum tweaks to become serious security

[85] Eduard Kovacs, *ChatGPT Data Breach Confirmed as Security Firm Warns of Vulnerable Component Exploitation,* SecurityWeek (March 28, 2023), https://www.securityweek.com/chatgpt-data-breach-confirmed-as-security-firm-warns-of-vulnerable-component-exploitation/.

[86] Mark Gurman, *Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak,* Bloomberg (May 1, 2023), https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak.

[87] Mitchell Clark and James Vincent, *OpenAI is Massively Expanding ChatGPT's Capabilities to Let It Browse the Web and More,* Verge (March 23, 2023), https://www.theverge.com/2023/3/23/23653591/openai-chatgpt-plugins-launch-web-browsing-third-party.

[88] *See, e.g.,* The State of Phishing 2023, SlashNext Security (Oct. 2023), https://slashnext.com/state-of-phishing-2023/; Elias Groll, *ChatGPT Shows Promise of Using AI to Write Malware,* CyberScoop (December 6, 2022), https://cyberscoop.com/chatgpt-ai-malware/; Crane Hassold, *Executive Impersonation Attacks Targeting Companies Worldwide,* Abnormal Blog (February 16, 2023), https://abnormalsecurity.com/blog/midnight-hedgehog-mandarin-capybara-multilingual-executive-impersonation; Center for Strategic and International Studies, *A Conversation on Cybersecurity with NSA's Rob Joyce,* YouTube (April 11, 2023), https://youtu.be/MMNHNjKp4Gs?t=530 (8:50 mark).

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

threats.[89] Some security professionals have already tracked examples of hackers exchanging tips on how to use ChatGPT to recreate malware strains and techniques and develop new scripts.[90]

Bias

LLMs can easily perpetuate bias by including biased data in their training datasets, through algorithms that develop their own biases, and in outputs stemming from those biased training datasets and algorithms. While bias may be present in curated training datasets, there is a particularly high risk of bias where datasets are built from web scraping methods that bring in massive collections of data on a continuous basis. In these cases, the training datasets constantly expand and they are often not regularly checked for accuracy, bias, appropriateness for use, and other key metrics. Even when companies attempt to clean these datasets up, massive amounts of discriminatory content may slip through the cracks. For example, one of the most commonly used AI training databases, Google's C4 (or, Colossal Clean Crawled Corpus), is a sprawling database full of terabytes upon terabytes of English language data scraped from the internet.[91] Despite Google's multiple attempts to filter the data collected by underlying web crawler, C4 still contains a multitude of potentially biased and discriminatory data, such as 72,000 instances of the word "swastika."[92] In addition, some of Google's filtering attempts proactively filtered out LGBT+ related content, biasing the data even further.[93] The statistical models in the LLMs trained on C4 and similar databases then extrapolate

[89] *See Generating Harms: Generative AI's Impact & Paths Forward,* EPIC (May 2023), available at https://epic.org/new-epic-report-sheds-light-on-generative-a-i-harms/ at 5.
[90] *OpwnAI: Cybercriminals Starting to Use ChatGPT,* Check Point Research (January 6, 2023), https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/.
[91] Kevin Schaul, et al., *Inside the secret list of websites that make AI like ChatGPT sound smart*, Wash. Post (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.
[92] Ibid.
[93] Ibid.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

that the overtly discriminatory data, including underrepresentative and adverse depictions of minority communities.

Even where a dataset does not contain explicitly biased data, LLMs may draw biased conclusions when analyzing and learning from those datasets. For example, factual historic data on demographic arrest rates in the U.S. will show that people of color are incarcerated at much higher rates per capita than white people. Because this data is divorced from the context of deep and historic prejudice built into the U.S. law enforcement system, policing, sentencing, etc., an LLM may erroneously conclude that people of color are more likely to commit crime and extrapolate from this conclusion that people of color are genetically or innately prone to commit crime. This incorrect and biased conclusion may then be reflected in the LLM outputs, further spreading the bias.

Information manipulation

The wide-spread availability of LLMs facilitates the high speed and volume spread of content – both benign and harmful. For example, LLMs can propagate and amplify false, misleading, biased, inflammatory, and dangerous content in their outputs. The same content may easily be scraped and put into training datasets, embedding the harm directly into the systems. Broadly, there are four categories of harmful content that we believe LLMs will empower and amplify: scams, cybersecurity threats, disinformation, and misinformation.[94]

Disinformation

While misinformation (addressed below) covers individuals unknowingly spreading false or inaccurate information, disinformation involves false information purposely intended to lie or mislead. LLMs facilitate a higher volume of persuasive disinformation generation that can then be spread easily, cheaply, and at a much higher speed. The potential impact of this on elections, politics, news (particularly related to health or safety), and other highly sensitive areas is disastrous.

[94] *Generating Harms,* EPIC.

LLM-fueled disinformation may be used to impact public opinion on any number of important political matters, stoke hateful sentiments (racism, xenophobia, homophobia, etc.), or even harass individuals. Once spread, countering the easily proliferated information effectively will be nearly impossible. In tests done by the Center for Countering Digital Hate, Google's Bard (Google's answer to ChatGPT) has actively spread conspiracy theories with no context, disclosure, or accuracy checks.[95]

Misinformation

Misinformation faces many of the same problems as disinformation with one important distinction – individuals spreading misinformation may genuinely believe what they are sharing is accurate. Misinformation can be generated from the input perimeters supplied by the user or from inaccurate information generated by the LLM.

Multiple cases of LLMs generating and spreading misinformation have already been reported. Frequently, the misinformation is well-written and weaves in true facts with the false information. As previously mentioned, when ChatGPT was asked for a list of legal scholars who have been accused of sexual harassment, it named a real scholar and provided details of the allegation, citing a March 2018 article as its source.[96] The article did not exist and the scholar had never been accused of harassment. Another example of LLMs' ability to confidently put forth false information (even with false citations) comes from the legal field, where attorneys have been sanctioned for submitting briefs written by ChatGPT that cite non-existent court cases.[97]

[95] Misinformation on Bard, Google's New AI Chat, Center for Countering Digital Hate (Apr. 5, 2023), https://counterhate.com/research/misinformation-on-bard-google-ai-chat/.
[96] Pranshu Verma and Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused,* Washington Post (April 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.
[97] Dan Mangan, *Judge sanctions lawyers for brief written by A.I. with fake citations,* CNBC (June 22, 2023), https://www.cnbc.com/2023/06/22/judge-sanctions-lawyers-whose-ai-written-filing-contained-fake-citations.html.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

In addition, both misinformation and disinformation can easily create a cyclical, self-feeding effect in LLMs that operate on datasets built from continuously scraped data. When disinformation is put into the digital ecosystem at high volume, more and more systems may scrape up that false information and incorporate it into their learning methods, generating increasingly false, biased, and otherwise inaccurate outputs. In some circumstances, the continued re-integration of low-quality data into datasets has led to total model collapse.[98]

Long-term expansion of disinformation and misinformation through LLMs will entirely undermine our ability to trust any information. Not only is it easy to believe that inaccurate information is true, but it may also become common to dismiss true information as fictional or fabricated.[99] The speed of new information generation makes it impossible to consistently check that information for accuracy and issue corrections before the spread is irreversible.

4. Privacy principles and technical mitigations

The core data protection and privacy risks of Generative AI are not particularly novel. What primarily differentiates LLMs, and generative AI more broadly, from other forms of AI is the increase in scale of the data being processed, the complexity of the techniques used to develop and deploy the models, and the unprecedented scale and pace of adoption across the economy.

[98] *See* Carl Franzen, *The AI Feedback Loop: Researchers Warn of 'Model Collapse' as AI trains on AI-Generated Content*, VentureBeat (June 12, 2023), https://venturebeat.com/ai/the-ai-feedback-loop-researchers-warn-ofmodel-collapse-as-ai-trains-on-ai-generated-content/; Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, arXiv (Cambridge Univ. Working Paper, 2023), https://www.cl.cam.ac.uk/~is410/Papers/dementia_arxiv.pdf.

[99] *See* Danielle Citron and Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,* 107 Cal. L. Rev. 1753, 22 1785-86 (2019).

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

AI guidance pre-dating the Generative AI boom already exists.[100] However, some academics have questioned whether tensions exists between the GDPR and AI business models,[101] while others have further argued these two concepts are in fact incompatible, particularly in terms of the principle of purpose limitation and data minimization[102]. In the context of these debates, data protection still offers a clear framework for the protection of rights. Additionally, a number of technical measures also provide further risk mitigations to the risk identified in the previous section.

This chapter discusses: (1) privacy principles / key areas of consideration and (2) technical mitigations to the data protection and privacy risks associated with Generative AI.

Privacy principles

Lawful basis

The developers and deployers of generative AI systems that process personal data must have a valid lawful basis under data protection and privacy legislation, and also be lawful in accordance with other applicable legislation (e.g. copyright law). For example, Article 6 of the GDPR offers six lawful bases, with additional requirements under Article 9 for special category data.

In terms of training data for Generative AI, it is crucial to note that personal data that is publicly accessible still falls under data protection and privacy legislation in most jurisdictions, as stressed in a recent joint statement by the GPA's

[100] U.K. Information Commissioner's Office, "Artificial Intelligence," https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/; Canadian Privacy Commissioners, "Principles for responsible, trustworthy and privacy-protective generative AI technologies," December 2023, https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/gd_principles_ai/.
[101] Chris Jay Hoofnagle, Bart van der Sloot and Frederik Zuiderveen Borgesius, "The European Union general data protection regulation: what it is and what it means," February 2019, https://doi.org/10.1080/13600834.2019.1573501.
[102] Tal Zarsky, "Incompatible: The GDPR in the Age of Big Data," 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3022646.

International Enforcement Cooperation Working Group (IEWG).[103] Apart from data protection, upcoming copyright rulings in US federal courts and in the UK[104] may carry significant weight in relation to the lawfulness principle within the GDPR if it is deemed that web-scraped training data violates copyright and intellectual property laws. DPAs, of course, rely on these rulings as it is beyond their remit to make these judgements themselves.

Purpose limitation

The developers and deployers of LLMs and generative AI systems that process personal data need to ensure that this data is processed for specified explicit and legitimate purposes. Furthermore, they need to ensure that they do not process it beyond individuals' reasonable expectations, or for incompatible purposes.

There are complex questions here with regards to the different stages of processing, as the generative AI model lifecycle itself involves several stages. For example, the purpose of training a core model will require training data and test data, while the purpose of adapting the core model may require fine-tuning data from a third-party developing its own application. Nonetheless, it is vital that at each stage of processing, the purpose is detailed and specific that all relevant parties, including data subjects, have a clear understanding of why and how personal data is being processed.

Data minimization

The developers and deployers of LLMs and other generative AI systems that process personal data should limit processing to what is "necessary" for their purpose. The

[103] GPA's International Enforcement Cooperation Working Group, "Joint statement on data scraping and the protection of privacy," August 2023, https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf.
[104] Emilla David, "Getty lawsuit against Stability AI to go to trial in the UK," *The Verge*, December 4, 2023, https://www.theverge.com/2023/12/4/23988403/getty-lawsuit-stability-ai-copyright-infringement.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

greater the volume of personal data being processed, the greater the potential privacy risks and other harms to individuals there are.

The development of many generative AI systems requires very large amounts of training data. [102] As such, organizations may face challenges applying the data minimization principle, and especially in terms of determining what processing counts as "necessary".

Limiting the occurrence or processing of any personal information as early as possible is an important step towards protecting the rights of data subjects. To this end, developers should strive to apply data minimization to any occurrences of personal information in their data sets. A common approach is to apply data sanitation by exclusion and different anonymisation procedures. However, even with these techniques applied, it can be challenging,to fully ensure that datasets do not contain any personal information. In cases where pre-collected third-party datasets are used for training, it is equally important to remove personal information in post-processing steps.

Transparency

The developers and deployers of LLMs and other generative AI systems that process personal data must implement transparency measures, and must do so particularly in relation to data subjects, who have a number of information rights. This should include information on what, how, when, and why personal data is collected and used in the process of training the system, including the sources of training data, the pre- and post-processing measures to remove personal information and the reliability of the prediction of the generated text.

Transparency is an internationally recognized principle when it comes to AI, even beyond its regular interpretation in privacy and data protection. As an OECD principle, it calls for AI actors to, "provide meaningful information, appropriate to the context, and consistent with the state of art."[105]

[105] OECD, "Transparency and explainability (Principle 1.3)," https://oecd.ai/en/dashboards/ai-principles/P7.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Deployers of LLMs should ensure that their end-user consent mechanisms are clear, accessible, specific and always up to date. The mechanisms should communicate how data subjects' rights are protected throughout the whole lifecycle of data processing, including user prompts, LLM training and outputs, and allow data subjects to make informed decisions about how their data is processed by the LLM.

Security

The developers and deployers of LLMs and other generative AI systems that process personal data must implement security measures. This is multifaceted. Personal data needs to be kept secure during storage, development, but also during post-deployment to account for complex security issues such as prompt injection attacks, model inversion attacks[106], and data leakages.

The vulnerability of models can vary, depending not only on their deployment method but also the data governance that surrounds them. Many companies engage in activities such as red teaming to test security vulnerabilities. Meanwhile, open access models may be more exposed by their nature, but also benefit from a community-led approach to security.

Accountability

The developers and deployers of LLMs and other generative AI systems that process personal data should ensure they can demonstrate compliance with data protection. Accountability is in effect a meta-principle that acts as a guarantor.

Like transparency, accountability is also an internationally recognized principle. As an OECD principle, the rationale behind it is that "organisations or individuals will ensure the proper functioning, throughout their lifecycle, of the AI systems that

[106] Michael Veale, Reubin Binns and Lilian Edwards, "Algorithms that remember: model inversion attacks and data protection law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, October 2018, https://doi.org/10.1098/rsta.2018.0083.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

they design, develop, operate or deploy, in accordance with their roles and applicable regulatory frameworks."[107]

Accuracy

The developers and deployers of LLMs and other generative AI systems must ensure that any personal data processed by them is as accurate, complete, and up-to-date as is necessary for purposes for which it is to be used. This applies in particular to personal data used to train LLMs or generative AI models.

To support this principle, developers and deployers should have a process by which their LLM or generative AI system can be updated (for instance, by refining or retraining the model) in cases where inaccurate or out-of-date model inputs, such as training data, are discovered. In addition, developers and deployers should inform end-users about any known issues or limitations with the accuracy of model outputs. This may include where the training data is timebounded (i.e. only contains information up to a certain date); where the content may be adversely affected by non-representative sources; or where there are particular subject matters or prompts that tend to lead to inaccurate outputs.

Challenges to determining an appropriate level of accuracy are exacerbated by the syntactic processing and indeterminate, open-ended nature of the purposes of LLMs.

Data subject rights

The rights of data subjects are at the core of data protection. The developers and deployers of LLMs and other generative AI systems that process personal data are fundamentally required to ensure that individuals can access, rectify, erase, and opt-out of the use of their data, among other rights. This is especially important in relation to special category data and respecting the rights of children.

[107] OECD, "Accountability (Principle 1.5)," https://oecd.ai/en/dashboards/ai-principles/P9.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

Some developers have provided portals for individuals to exercise their information rights, but there is currently little public evidence on the extent to which these mechanisms effectively allow data subjects to fully fulfill their requests. This links to both the principles of transparency and accountability.

Technical mitigations

When it comes to training LLMs, there are multiple stages and types of technical interventions that one can make to mitigate privacy risk. In this section we will focus on what are the benefits and drawbacks of leveraging some of these interventions.

Curation and pre-processing

LLMs are trained on large amounts of text data and, given their capacity for memorization,[108] it is important to treat the models with the same risk-appropriate considerations that one would treat the data used to train it. In the process of collecting and curating the datasets, it is possible to make decisions and take steps to reduce the risk that the data used in training will violate people's privacy.

Source curation: An initial consideration is what type of data is being used to train[109] these models, with a focus on the original intended audience when the data was shared. One simple distinction is whether the data is private data, with this type of data carrying a clear privacy impact when it is used as part of the training process without consideration to the data subject's desires. However, a less often considered distinction is publicly accessible vs public (or open data). While all public (or open) data is publicly accessible, not all publicly accessible data should be treated as if it is public. Here, the distinction lies in the intent and expectations behind making that data available: public data refers to data that was crafted with

[108] Nicholas Carlini, Florian Tramer, Eric Wallace, et al. "Extracting Training Data from Large Language Models," *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650, https://arxiv.org/abs/2012.07805.
[109] "Train" used here encompasses both initial training and any fine tuning or additional training steps.

the intent of being widely shared and used, for example, government datasets and or Wikipedia contributions, whereas some publicly accessible data may have been shared with the intent of being used and consumed in specific contexts, for example social media posts and product reviews. Research has shown that, even in the context of academic research, social media users may not feel comfortable having their data used without their consent[110] even if it is publicly accessible. To reduce the privacy risk of these models, it is important to obtain data from sources where the privacy expectations for data use from those associated with the content are in alignment with the intended goal of training an LLM.

- Benefits: If machine learning (ML) engineers curate their data sources well they can address a significant portion of privacy risks at its root source. This may lead to a significant reduction in the privacy risk associated with the model. This type of exercise will also ensure ML engineers better know their data, which can help with other types of risks (such as copyright considerations) and will help guide what other additional protections may be necessary.
- Drawbacks: Source curation is a time-consuming task and, even a carefully curated list of data sources that focuses only on the far end of the data availability spectrum (public data), may still include some data that has privacy implications.

Pre-processing (removing sensitive data): After datasets have been initially compiled, the next step involves the pre-processing of that data before it is used to train models. At this stage, one can leverage automated tools to detect and remove sensitive information, for example personal information, health information, and information surrounding sensitive topics like sexuality and religion. These tools can range from simply detecting the presence of this information and flagging it for human review, to automatically removing, replacing, or obfuscating the information (for example, replacing all addresses to 123 Main St).

[110] Casey Fiesler and Nicholas Proferes, "'Participant' Perceptions of Twitter Research Ethics," *Social Media + Society, 4(1)*, 2018, https://doi.org/10.1177/2056305118763366.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

- Benefits: This can efficiently remove references to a large portion of sensitive private information.
- Drawbacks: Designing a high-quality detection tool can be very challenging because of the difficulty in differentiating between sensitive data from non-sensitive data; such differentiation is often based on the context in which information is provided, and automating the detection of this context can be difficult. For example, without context, a series of 16 digits may be completely innocuous or someone's credit card number, or an address may be to a politician's public office or to that politician's private home. Because of this, this should be one part of a broader mitigation toolkit.

A note on deduplicating data: Previous research has suggested that deduplicating training data may help with the memorization issue, as instances that were more frequently repeated in the datasets are more likely to be memorized by these models.[111] However, more recent research has found that this may be an insufficient approach to addressing the privacy issues, although in the context of a different model architecture. It was found that deduplicating data reduces the chances of that individual entry being memorized, but exposes previously-safe data to memorization.[112] While deduplicating data is still a good practice with benefits like improving efficiency of training LLMs and reducing the memorization for the deduplicated entries,[113] it should not be approached as a way to entirely resolve privacy-related issues linked with the datasets used when training LLMs.

[111] Carlini et al., *supra* note 104.
[112] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang et al., "The Privacy Onion Effect: Memorization is Relative," *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 13263–13276, https://arxiv.org/abs/2206.10469.
[113] Katherine Lee, Daphne Ippolito, Andrew Nystrom, et al. "Deduplicating Training Data Makes Language Models Better," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2022, pp. 8424–8445, https://arxiv.org/abs/2206.10469.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Differential privacy

When training LLMs, it is possible to leverage privacy enhancing technologies such as differential privacy (DP),[114] to train models that are provably private. This can be done at different stages (e.g., training data, model training, model outputs), with different units of consideration (e.g., instance-level, group-level), and in different conditions (e.g., central, local, distributed). Each of these have unique considerations, benefits, and costs that we will discuss in this section. For a more comprehensive presentation of the different approaches, their implementations and considerations, we refer the reader to "How to DP-Fy ML."[115]

Unit of privacy: Defining the appropriate unit of privacy for differential privacy is critical in ensuring the developers are providing the privacy guarantees at the appropriate level, as it determines what will make two datasets be considered "neighboring" in the definition of differential privacy. Instance-level DP will provide protections for each sample included in the dataset, whereas group-level DP will provide protections at a higher level of abstraction (e.g., user-level, document-level, etc). For LLMs, it may be better to use group-level DP as the desired sequence-length used in the training of these models will not only impact model performance but will also impact the privacy guarantees and disentangling these two factors may be more beneficial. Furthermore, the high chance for repetition of instances at the instance-level will likely significantly dilute the privacy guarantees being provided. However, it is still important to carefully consider at which level of

[114] The definition of differential privacy that is being used in this section is the one proposed in Dwork et al. (2006b):

> We say that two datasets D and D′ are neighbors if they differ in exactly one record; more precisely, one dataset is a copy of the other but with a single record added or removed. Let ε be a positive scalar. A mechanism A guarantees ε-differential privacy if for any two neighboring datasets D and D′ , and for any S ⊆ Range(A),
>
> $$P[A(D) \in S] \leq \exp(\varepsilon) \times P[A(D') \in S]$$

[115] Natalia Ponomareva, Sergei Vassilvitskii, Zheng Xu et al., "How to DP-fy ML: A Practical Tutorial to Machine Learning with Differential Privacy," *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23*), 2023, pp. 5823–5824, https://arxiv.org/abs/2303.00654.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

grouping it makes sense to define the unit of privacy. For example, while one might want to provide user-level DP, given that the training data frequently used to train these models are publicly accessible text from the internet, it may be impossible to do so as one cannot identify which samples were contributed by which users.

Implementation condition: DP is most frequently implemented in a centralized fashion, where there is a trusted aggregator that processes the raw data to produce the differentially private output. In this case the only access that an adversary would have would be to the produced output, be that an aggregated dataset or a differentially private model. However, one can also implement DP locally, when there isn't a trusted aggregated, and in a distributed fashion. While local DP provides strong privacy guarantees it also incurs the highest loss of utility. Distributed DP allows for a middle-ground approach, where some disturbance is added locally, but the stronger privacy guarantees come from the output of a private aggregation protocol. This approach is most feasible in a federated learning context, where each client produces individual gradients with minimal perturbation that are then used to build a differentially private model.

Implementation stage: There are multiple levels of granularity related to when one can implement DP. For the sake of simplicity this subsection will only address it at the level of training data, model training, and model outputs. The implementation of DP at each of these levels is dependent on the threat and risk models being considered, what is made public, and the amount of performance degradation that is acceptable. For training data, while this provides the strongest guarantees, as it is protecting the most basic data unit, this approach can at times degrade the utility of the dataset to a degree that it has a significant impact on performance. This has led existing approaches for text data[116] to leverage a relaxation of DP called dχ-

[116] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake et al., "Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations," *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20*, 2020, pp. 178–186, https://arxiv.org/abs/1910.08902; Chen Qu, Weize Kong, Liu Yang et al., "Natural language understanding with privacy-preserving BERT," *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, 2021, pp. 1488–1497, https://arxiv.org/abs/2104.07504.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

privacy.[117] Using DP on model outputs is useful in the context of models-as-a-service, where one only has access to the inferences being made and presented to the end-user. Nevertheless, in this case it is important to maintain an inference budget that limits the number of inferences that an end-user can request. There are other alternative implementations that avoid the privacy degradation of multiple inferences, but instead must rely on specialized system architectures. Another way to bypass this issue is to apply DP at the stage of model training, which provides guarantees that and adversary would not be able to differentiate between models that include or do not include a particular instance in the training data. There exist many methods to achieve differentially private training of machine learning models; however, they are not appropriate for language models as they have requirements that cannot be achieved for this type of complex and costly task. The approaches that are most feasible for this type of task relate to gradient noise injecting, with differentially private stochastic gradient descent (DP-SGD) being the most used algorithm. We will not dive into details on how these algorithms are implemented, but we will highlight that the implementation of these algorithms is complex[118] and using them to a meaningful degree requires a high-level of expertise. While it is better to have some privacy protection than none, relying on these methods without fully understanding how the different parameters contribute to the ultimate privacy guarantees can lead to a false sense of privacy. We strongly recommend that those interested in implementing differential privacy when implementing their models engage with experts on this topic or, at a minimum, leverage available resources.[119]

Finally, while the benefits and drawbacks will depend on the specifics of the DP implementation, there are broad benefits and costs associated with differential privacy:

[117] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe et al., "Broadening the scope of differential privacy using metrics," *Privacy Enhancing Technologies*, pp. 82–102, Berlin, Heidelberg: Springer, 2013.
[118] We recommend relying on existing, well-vetted open-source implementations.
[119] For example, Ponomareva et al., *supra* note 111.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

- Benefits: Differential privacy is the only approach that provides provable privacy guarantees and the wealth of research in this area has produced many valid approaches that can be leveraged for a variety of specific use cases.
- Drawbacks: As previously mentioned, correctly implementing and applying differential privacy requires a high degree of expertise to ensure appropriate privacy protections are in place while still retaining performance. Another consideration is that training models using differential privacy may increase the computational and memory cost of the training, as well as it may slow down the process, if steps are not taken to mitigate this concern.

Note on fine-tuning: An approach that has been explored in the literature is leveraging "foundation models" trained on publicly accessible data and fine-tuning it with private data using differential privacy. This allows the model to leverage the learnings from the larger dataset, while protecting the private data at a lower cost, as the fine-tuning step requires significantly less resources, making the additional cost of differentially private training more appealing. However, it is critical to note that this paradigm only protects the data used during the fine-tuning and it incorporates the incorrect assumption that publicly accessible data does not have privacy risks.[120]

Post-processing and machine unlearning

In the previous section we touched on a post-processing approach when discussing DP applied to model inferences. Other approaches that can be applied once the model is already trained involve removing sensitive data in the same way we previously presented for data curation. When used in a post-processing step removing sensitive data from the outputs will not only suffer from the limitations of it being hard to accurately do this as discussed in the "Pre-processing (removing

[120] Florian Tramèr, Gautam Kamath and Nicholas Carlini, "Considerations for Differentially Private Learning with Large-Scale Public Pretraining," 2022, https://arxiv.org/abs/2212.06470.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

sensitive data)" subsection, but it will also not avoid any privacy issues that may arise from the model, in its training process, having learned from the sensitive data present in the dataset.

Finally, an alternative that has gained traction recently is called "machine unlearning," which is focused on being able to effectively modify already trained models so they can "forget" specific pieces of training data without resorting to a complete "naïve" retraining of the model from scratch. Current research into machine unlearning focusses on two main approaches: "exact" unlearning and "approximate" unlearning.[121]

- Exact unlearning aims to fully remove the influence of targeted training data points from the LLM by initially splitting the training data into multiple subsets and then training the LLM as an ensemble of sub-models. When data points are identified for removal, only the sub-model associated with the identified data points needs to be retrained.[122] This accelerates the process of retraining, which would otherwise be a slow and costly procedure.
- Approximate unlearning, on the other hand, focuses on the model itself. Instead of re-training with altered data, it adjusts model weights after the fact to attempt to reduce the influence of targeted training data points. While its removal of information is less precise than exact unlearning, approximate unlearning may be less complex and costly in certain cases.

With respect to approximate unlearning, a subfield of research has developed around the connection between it and the privacy enhancing technology of federated learning. Researchers have explored different implementations of federated learning to achieve approximate unlearning. This is sometimes referred to as "federated unlearning."

[121] See Jie Xu, Zihan Wu, Cong Wang and Xiaohua Jia, "Machine Unlearning: Solutions and Challenges," 2024, https://arxiv.org/abs/2308.07061.
[122] See Haonan Yan, Xiaoguang Li, Ziyao Guo et al., "ARCANE: An Efficient Architecture for Exact Machine Unlearning," 2022, https://www.ijcai.org/proceedings/2022/0556.pdf.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

For example, one technique is called "FedEraser."[123] The basic idea of FedEraser is to store a complete history of the parameter updates from each contributing client where FedEraser, then reconstructs the unlearned model through retraining. This results in a significant speed-up of the reconstruction of the unlearned model instead of a complete retaining from scratch. However, a risk with this approach is that a history of every participant contribution needs to be stored and for a larger number of clients participating in the federated learning, it might consume a significant volume of data storage.

Another federated unlearning technique is based on knowledge distillation.[124] This method requires the central server to store the history of updates from each contributing client and possess some extra synthetic or outsourced unlabelled data. The idea is to first erase the historical parameter updates from the target client and then recover the damage through the knowledge distillation method.

Another approach explores federated unlearning without storing any parameter history on the central server.[125] This method relies on the individual client who asks to opt out by removing the influence of their entire local data from the trained global model. How it works is that the client first performs a local unlearning process and then this locally unlearned model is used to perform a few rounds of federated learning between the server and the remaining clients to obtain the new unlearned global model.

While proponents of machine unlearning say an effective approach—should it be developed—could improve privacy and help remove the influence of inaccurate or outdated data, truly deleting requested data cannot simply be done by erasing it from a database: the data's *influence*—such as the effect it has on a model's weights—must also be removed from machine learning models and other artifacts

[123] "FedEraser" is short for "Federated Eraser." See Gaoyang Liu, Xiaoqiang Ma, Yang Yang et al, "Federated Unlearning," 2021, https://arxiv.org/abs/2012.13891.
[124] Chen Wu, Sencun Zhu and Pasenjit Mitra, "Federated Unlearning with Knowledge Distillation," 2022, https://arxiv.org/abs/2201.09441.
[125] Anisa Halimi, Swanand Kadhe, Ambrish Rawat et al., "Federated Unlearning: How to Efficiently Erase a Client in FL?" 2023, https://arxiv.org/abs/2207.05521.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

that exist downstream from where a requester's information is stored. Furthermore, as mentioned above, recent research has pointed out that removing specific instances of data from a model's training data can expose previously safe data.[126] For now, this area of research remains too nascent and without a clear answer on how effective "machine unlearning" will be. Respecting data subject rights in the development and deployment of LLMs continues to raise challenges.[127]

5. Emerging practices: Local LLMs

In response to the privacy challenges inherent in the development and deployment of current LLMs, researchers and industry have begun to explore alternative solutions, with the goal of reducing the need for complex processing of personal information throughout the lifecycle of LLMs by default.

One emerging approach attempts to eliminate many of the risks related to the ongoing transfer of personal information to cloud or server based LLMs by hosting an LLM locally on the data subject's personal device, computer or on a smart home device within their own private home network. These LLMs are often called "private" or "local LLMs." They give rise to both advantages and challenges.

Advantages

Increased privacy

Locally hosted LLMs have many privacy preserving advantages. By running a LLM locally on a personal device, the processing of personal information through interaction with the user is done within that specific device or in the user's own

[126] See Carlini et al., *supra* note 108.
[127] Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang et al., "Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions," *Algorithms that forget: Machine unlearning and the right to erasure*, 2023, https://arxiv.org/pdf/2307.03941.

home and is not transferred to or processed by any third party outside the device's private network or ecosystem. Local LLMs can be seen as isolated and private models as all processing of personal information after training and deployment is initiated by the data subject themself.

Unique services

The on-device containment of local LLMs not only delivers an enhanced level of security and increased privacy, but the containment also enables the LLMs to be trained and fine-tuned with the data subject in focus. With highly personalized optimizations, these LLMs have the possibility to deliver unique services.

No need for network connectivity

An on-device hosted LLM works without network connectivity. It does not rely on cloud-based services for its operation and functionality. This can be an important factor not only from a privacy and security perspective but also in regard to availability of services.

Challenges

Memory consumption

Given that local LLMs are fully contained within the user's device, they do tend to consume a fair amount of physical storage space, which often can be several gigabytes. Further, to run smoothly on a device, they require a substantial amount of processing memory (RAM), quite often double the amount of what they need for storage.[128]

Computing power

Not only do the local LLMs put high demands on the storage space and processing memory, but they also have substantial prerequisites on the device' computing performance in order for the user to have a positive experience with them or to run them at all. This might not be an issue for a relatively modern device as many

[128] See Machine Learning Compilation for Large Language Models (MLC LLM), https://llm.mlc.ai/.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

newer devices contain special neural processing units that are optimized for AI and neural network processing. However, a common device might not have the sheer performance and prerequisites to run a local LLM.

Risk of exclusion

If the performance requirements of local LLMs are too high or costly, it could result in the exclusion of a large number of users. Such unwilling exclusion could create a two-tiered system, where disadvantaged individuals are forced to use a more traditional, cloud-based LLM service with a reduced level of privacy protection, but wealthy individuals are not.

Conclusion

The questions surrounding LLMs have recently coalesced to form one of the most challenging areas of engagement on the part of DPAs. Not only is the technology itself complex, with unique details and additional stages of development in comparison to other AI systems; LLMs raise various privacy and data protection risks whose understanding and appropriate redress depends fundamentally on an effective grasp of the underlying workings of the technology.

In this paper, we have attempted to provide an in-depth, multifaceted analysis of LLMs from the point of view of privacy and data protection, with a view towards better positioning DPAs to face the challenges posed by LLMs. The work of DPAs is only beginning with respect to LLMs and related generative AI technologies. As the field of generative AI continues to advance, it is expected that the challenges will continue to grow as well.

Given this situation, DPAs should consider making additional investments in their internal capacity to address AI challenges as well as external, cross-regulatory partnerships. This is important for two reasons:

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

- The relevance of LLMs, and AI systems in general, will keep increasing in the near future. A detailed understanding on a technological level is required to correctly evaluate the privacy risks for data subjects and the adequacy of technical and organizational measures. Therefore, DPAs should invest in building up technical know-how with respect to LLMs and AI systems in general.
- Several countries are discussing specific legislations for AI systems. Here, legislators should have in mind the necessity of strong collaboration between DPAs and AI regulatory authorities. Synergies could be leveraged by uniting those regulatory tasks within one authority.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Appendix A: The transformer architecture

The transformer architecture is the culmination of decades of research into the question of how to create a statistical model of natural language. This is challenging problem in machine learning. The overarching goal is to learn the joint probability distribution of all possible sequences of words in a given language. However, a straightforward statistical approach, where each possible sequence is defined as a parameter, is problematic. This is mainly due to the curse of dimensionality. For a language with a vocabulary of, say, 100,000 words, it would require approximately $100{,}000^{20}$ or 10^{100} free parameters to model sequences of text of up to 20 words. This is more parameters than there are particles in the known universe.

In response to this challenge, new techniques and design approaches have been developed that leverage the capabilities of neural networks to better represent and approximate both the semantics and syntax of natural language. These efforts are what has led today to what is known as the "transformer" architecture. First introduced in a paper from 2017,[129] the transformer architecture represents the most successful form of neural network model to date for natural language processing tasks.

In what follows, we will focus on what are known as "decoder-only" transformer models. These are versions of the transformer architecture designed specifically to predict the next word in a sequence of text. Other types of transformer models include "encoder-decoder" and "encoder-only" models. What differentiates decoder-only transformer models from others is their "unidirectional" architecture. Decoder-only transformer models are trained to learn relationships between words from previous time steps of a sequence only. During training, future words of a sequence are "masked." This differs from encoder-decoder and encoder-only transformer models, which are "bidirectional" in nature.

[129] Ashish Vaswani, Noam Shazeer, Niki Parmar et al., "Attention Is All You Need," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, https://arxiv.org/pdf/1706.03762.pdf.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

These differences in architecture have implications on the suitability of different types of transformer models to certain linguistic tasks. Encoder-decoder and encoder-only transformers are generally more suitable for tasks such as machine translation and semantic similarity assessment, whereas decoder-only can excel in tasks such as chat-style question answering, that is, after appropriate fine-tuning.

In general, decoder-only transformer models consist of five main elements: a vocabulary, word embeddings, context window, masked multi-head self-attention and feed-forward neural networks.

Vocabulary

The first step in developing a statistical model of language is to define the complete vocabulary of words the model is meant to predict. At a high level, LLMs work by predicting one word at a time based on the previous sequence of words and then repeating this process in an "autoregressive" manner, that is, using its previous output as an input, to create successively larger sequences of words. The vocabulary of an LLM represents the domain of possible predictions it may make.

There are different approaches to defining a vocabulary, with pros and cons to each. In general, there is a trade-off between vocabulary size and the ability of the model to learn useful relationships. If the vocabulary is too small, for example, at the level of individual letters or characters, this simplicity may make it difficult for the model to learn more complex semantic relationships. At the same time, if the vocabulary is too large, for example, consisting of all base words plus their individual inflections, this complexity may hinder the ability of the model to learn more subtle semantic relationships.

Researchers have found that a happy medium exists in the form of sub-word units or "tokens." A token is a piece of a word that can be used, either alone or in combination with other tokens, to form whole words. For example, the word "older" would consist of two tokens: "old" and "-er." Similarly, "dogs" would consist of the tokens "dog" and "-s."

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

By using a token-based vocabulary, LLMs are better placed to learn both complex and subtle semantic relationships. For example, instead of having the entire word "dogs" in its vocabulary or having to predict successive individual letters "d," "o," "g" and "s," an LLM with a token-based vocabulary would first predict "dog" and then "-s." This allows the model to learn both the meaning of "dog" as a concept as well as "-s" as a commonly used inflection.

In terms of size, the number of tokens in a vocabulary is generally consistent across LLMs. Vocabularies generally consist of around 50,000 tokens (especially if the model is pre-trained on a single language). For example, both GPT-2 and GPT-3 have a vocabulary size of 50,257 tokens.[130,131] A list of the tokens in the GPT-2/GPT-3 vocabulary can be viewed online.[132]

Word embeddings

Once the domain of possible predictions has been established in the form of a vocabulary, the next step in developing a statistical model of language is to determine a mathematical representation of the "words"[133] in the vocabulary. The simplest approach is to treat each word as a discrete atomic unit, with no underlying structure, and then model the statistical relationships between sequences of words in the training data. This is the approach taken in "n-gram" language models, where "n" stands for the maximum length of supported word sequences. While straightforward, n-gram models are limited in that they do not generalize well to not-yet-seen sequences of words outside the training data. This

[130] See Alec Radford, Jeffrey Wu, Rewon Child et al., "Language Models are Unsupervised Multitask Learners," Technical Report, OpenAI, 2019, https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.
[131] See Tom Brown, Benjamin Mann, Nick Ryder et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, 2020, https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
[132] See https://s3.amazonaws.com/models.huggingface.co/bert/gpt2-vocab.json. Note that the Unicode character "\u0120" represents a space character. It is used for tokens at the beginning of a word.
[133] For better readability, we will continue to refer to the predictive units in a vocabulary as "words," even though strictly speaking LLMs use sub-word "tokens."

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

suggests that an atomic representation of words is too simplistic to effectively grasp the semantics and syntax of natural language.[134]

In response, researchers developed the idea of "word embeddings." Instead of treating words as discrete atomic units, word embeddings represent each word in the vocabulary as a multidimensional "feature vector" consisting of a series of learned parameters or "weights." While the origins of the idea date back to the 1980s,[135] the first dedicated application of it to natural language processing was in the early 2000s.[136] According to the authors:

> In a nutshell, the idea of the proposed approach can be summarized as follows:
>
> 1. associate with each word in the vocabulary a distributed "feature vector" (a real-valued vector in $\mathbb{R}^m$ [the set of real numbers in *m* dimensions]), thereby creating a notion of similarity between words,
> 2. express the joint probability *function* of word sequences in terms of the feature vectors of these words in the sequence, and
> 3. learn simultaneously the word feature vectors and the parameters of that *function*.[137]

The intuition behind word embeddings is that the meaning of a word in a language can be thought of as a collection of overlapping traits or characteristics. For example, some characteristics could be syntactic in nature, such as gender or plurality, whereas others could be more semantic, such as whether the word represents a living or non-living thing, has a specific quantity or quality, is associated with a particular region or time, or in general has certain relationships to

[134] See also the discussion above re curse of dimensionality.
[135] See Geoffrey Hinton, "Learning Distributed Representations of Concepts," *Proceedings of the Eight Annual Conference of the Cognitive Science Society*, 1986, p. 1–12. An online version of the paper is available at https://www.cs.toronto.edu/~hinton/absps/families.pdf.
[136] See Yoshua Bengio, Réjean Ducharme, Pascal Vincent et al., "A Neural Probabilistic Language Model," *NIPS'2000* 13:933-938, and revised in *J. Machine Learning Research* (2003) 3:1137-1155, https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf.
[137] Ibid.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

other words.[138] By mapping words to points in a multidimensional feature space, word embeddings provide a machine learning data structure through which different traits and characteristics of words can be learned and represented mathematically.

However, which actual features of a word are captured in a word embedding cannot be determined in advance. Rather, they are "discovered" during pre-training by the learning algorithm. In other words, the learned features of a word are *data dependent*. The features ultimately captured in a word embedding depend on which traits and characteristics of words are present in the training data, including the extent to which they are present.

After pre-training, each word embedding consists of a vector of real-valued parameters. These vector values can be thought of as spatial points in the multidimensional feature space. An interesting property of word embeddings is that the location of these points tends to encode *relations of similarity between words* by default. This happens in two ways.

At a local level, word embeddings with similar meanings tend to be located closer together, at least across certain dimensions. This is by design. Because the different traits and characteristics of words are represented as undefined yet learnable parameters, if trained properly, that is, with enough representative examples, word embeddings should result in a situation where words with similar meanings end up with similar vector values. Researchers have confirmed this result. For example, an oft-cited study from 2011 showed how words with similar syntactic and semantic properties were encoded as "neighbors" in the feature space.[139] The study trained a LLM with word embeddings and observed spatial groupings of words with similar meanings. See Table 1 for details.

[138] See Yoshua Bengio, "Neural Net Language Models," *Scholarpedia*, 2008, http://www.scholarpedia.org/article/Neural_net_language_models.
[139] See Ronan Collobert, Jason Weston, Léon Bottou et al., "Natural Language Processing (almost) from Scratch," *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2493–2537, https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf.

France	Jesus	Xbox	Reddish	Scratched	Megabits
Austria	God	Amiga	Greenish	Nailed	Octets
Belgium	Sati	PlayStation	Bluish	Smashed	Mb/s
Germany	Christ	MSX	Pinkish	Punched	Bit/s
Italy	Satan	iPod	Purplish	Popped	Baud
Greece	Kali	Sega	Brownish	Crimped	Carats
Sweden	Indra	psNUMBER	Greyish	Scraped	Kbit/s
Norway	Vishnu	HD	Grayish	Screwed	Megahertz
Europe	Ananda	Dreamcast	Whitish	Sectioned	Megapixels
Hungary	Parvati	GeForce	Silvery	Slashed	Gbit/s
Switzerland	Grace	Capcom	Yellowish	Ripped	Amperes

Table 2: Groups of similar word embeddings from Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu and Pavel Kuksa, "Natural Language Processing (almost) from Scratch," Journal of Machine Learning Research*, vol. 12, 2011, pp. 2493 – 2537 at 2514, https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf. For each column, the randomly selected word is followed by its ten closest neighbors in the feature space.*

At a global level, word embeddings tend to encode more complex relationships of analogies between words in the vector differences between embeddings in the feature space. This property was discovered "somewhat surprisingly" in 2013.[140] After training a LLM with word embeddings, researchers showed how by using simple algebraic operations "*vector('King') - vector('Man') + vector('Woman')* results in a vector that is closest to the vector representation of the word *Queen*."[141] The analogies encoded in the global relationships between word embeddings were not limited to gender. They extended to many types of word relationships. See Table 2 for details.

[140] See Tomas Mikolov, Kai Chen, Greg Corrado et al., "Efficient Estimation of Word Representations in Vector Space," 2013, https://arxiv.org/abs/1301.3781.
[141] Ibid.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

Relationship	Example 1	Example 2	Example 3
France – Paris	Italy: Rome	Japan: Tokyo	Florida:
big – bigger	small: larger	cold: colder	Tallahassee
Miami – Florida	Baltimore:	Dallas: Texas	quick: quicker
Einstein – scientist	Maryland	Mozart: violinist	Kona: Hawaii
Sarkozy – France	Messi: midfielder	Merkel: Germany	Picasso: painter
copper – Cu	Berlusconi: Italy	gold: Au	Koizumi: Japan
Berlusconi – Silvio	zinc: Zn	Putin: Medvedev	uranium:
Microsoft –	Sarkozy: Nicolas	IBM: Linux	plutonium
Windows	Google: Android	IBM: McNealy	Obama: Barack
Microsoft –	Google: Yahoo	France: tapas	Apple: iPhone
Ballmer	Germany:		Apple: Jobs
Japan – sushi	bratwurst		USA: pizza

Table 3: Examples of word pair relationships from Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," 2013, https://arxiv.org/abs/1301.3781. The relationship is defined by subtracting two word vectors and the result is added to another word. For example, France – Paris + Italy = Rome.

From a machine learning perspective, the main advantage of word embeddings is that they provide LLMs with an increased ability to generalize to new examples. The relations of similarity between words encoded in them help improve the overall quality of the model, since inputs comprised of not-yet-seen sequences of words can be mapped to learned sequences of similar words, thereby producing more coherent, higher quality outputs. This is especially important given the large number of possible sequences.

In terms of size, the total number of dimensions of word embeddings varies across LLMs. More recent LLMs tend to have larger dimension sizes. For example, GPT-2 has a word embedding dimension size of 1600, whereas GPT-3 has a size of 12,288.[142] For GPT-3, this means that each of the 50,287 tokens in its vocabulary is represented as a vector consisting of 12,288 real-numbered values. This makes for

[142] See Radford et al., *supra* note 126 and Brown et al., *supra* note 127.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

50,257 x 12,288 = **617,558,016** learned parameters just for the vocabulary embeddings of the model.

Vectors of the same size as word embeddings appear throughout the technical components of LLMs. For this reason, the word embedding dimension size is also known as the "model dimension size." It is referred to as d_{model} for short.

Context window

Just as LLMs can only predict certain words, namely those in their vocabulary, so can they only handle sequences of words up to a given length. The "context window" of an LLM is the maximum number of words the LLM can take as input when predicting the next word of a sequence. It includes both words entered in by a user (in the form of a "prompt") as well as words predicted by the LLM (in the form of a response).

In decoder-only transformer models, there are parameters associated with the context window. As discussed in more detail in the section on "masked multi-head self-attention," transformer models do not use any circularity or recurrence in their architecture despite the sequential nature of language. For this reason, positional information about the ordering of words must be injected into the sequence before processing. In decoder-only transformers, this is done through a matrix of learned parameters, whose values are added to the word embeddings of the sequence before they are processed. The values of this "positional encoding" matrix have the same number of dimensions as word embeddings, namely d_{model}, to facilitate their addition.

In terms of size, the word length of context windows varies across LLMs. More recent LLMs tend to support larger sequences of words. For example, GPT-2 has a context window size of 1024 tokens, whereas GPT-3 can support sequences of up to

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

The Working Group is being supported by Data Protection Commissioners and independent experts from countries all over the world in order to improve privacy and data protection in technology

2048 tokens.[143] For GPT-3, this means the total number of parameters of the positional encoding matrix is 2048 x d_{model} = 2048 x 12,288 = **25,165,824**.

Masked multi-head self-attention

After determining a mathematical structure to represent the meaning of words in a vocabulary (in the form of word embeddings), the next step in developing a statistical model of language is to do the same but at the level of *word sequences*. This is a challenging problem. Language is a temporal, dynamic event, with an infinite number of possible expressions. In the words of Wilhelm von Humboldt, language is the "infinite use of finite means."[144] To mathematically represent the discursive meaning of word sequences as they continue to form through successive predictions, an architecture that looks both backwards and forwards is needed.

One class of techniques researchers have used to help model the temporal dynamics of language is what are known as "recurrent" neural networks. These are artificial neural networks whose flow of information between neurons is bidirectional in nature. Instead of only unidirectional, "feed-forward" connections, recurrent neural networks have neurons whose outputs are also used as inputs to previous layers, thereby incorporating a form of recursion into the architecture.

While certain designs have had success in modeling temporal events like language—above all, long short-term memory (LSTM) networks[145]—ultimately recurrent neural networks suffer from an important limitation. Because their design forces them to "squeeze" all information from past time steps into a fixed-length

[143] See Radford et al., *supra* note 126 and Brown et al., *supra* note 127.
[144] Wilhelm von Humboldt, *On Language*, 1836.
[145] See Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation* 9(8): 1735–1780, 1997. An online version of this article can be found at https://www.bioinf.jku.at/publications/older/2604.pdf.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

vector, they have difficulty capturing long-range dependencies between words. This is known as the "vanishing gradient" problem.[146]

In response, researchers have focused on an alternative class of techniques known as "attention" blocks. Drawing inspiration from the process of cognitive attention in human perception, attention blocks are neural networks that allow each word in a sequence to "attend" to every other word, while placing different amounts of focus on a word depending on the significance of the relationship. By creating such "affinities" between words, attention blocks overcome some of the limitations in other approaches to representing the discursive meaning of word sequences, including capturing long-range dependencies between words.

In decoder-only transformers, attention is calculated from the point of view of the current word, looking backwards at the previous words in the sequence, with a view towards predicting the next word. For this reason, decoder-only attention blocks are often qualified as being "masked." The term "self-attention" simply means that the attention block focuses on itself and does not communicate with other sources of information.

Concretely, attention blocks work by applying a series of vector transformations to the word embeddings of a sequence. The results of these transformations are then combined together to produce something akin to a "thought vector" representing a learned feature of the discursive meaning of the word sequence. Similar to how word embeddings attempt to capture the meaning of a words by mapping their traits or characteristics to values in a multidimensional feature space, attention blocks do the same but at the more abstract level of a word sequence.

In general, there are three transformations associated with an attention block. Each is performed by a matrix or rectangular array of learned parameters:

[146] See Razvan Pascanu, Tomas Mikolov and Yoshua Bengio, "On the difficulty of training Recurrent Neural Networks," *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, no. 3, 2013, https://arxiv.org/pdf/1211.5063.pdf.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

- **Query matrix.** This matrix is applied to the word embedding of the current word in the sequence. It transforms the word embedding into a "query" vector, which is then multiplied against each of the key vectors (see below) of the sequence. As its name suggests, the query matrix represents the "question" posed by the current word (in the form of its query vector) to determine what features or traits of other words it considers important.

- **Key matrix.** This matrix is applied to each word embedding of the sequence, including that of the current word. It transforms the word embeddings into "key" vectors, which, as indicated above, are multiplied individually against the query vector. The result is the "score" of each word, which represents the strength of the affinity between it and the current word (with respect to the query posed). Conceptually, the key matrix represents the "answer" of each word (in the form of its key vector) to the question posed by the query vector of the current word.

- **Value matrix.** Like the key matrix, this matrix is applied to each word embedding of the sequence, including that of the current word. It transforms the word embeddings into "value" vectors, which are then multiplied against a normalized version of their respective score (see above). The results are then added together to produce what was described above as a "thought vector" representing a learned feature of the discursive meaning of the word sequence. Conceptually, the value matrix represents the "contents" of the answer provided by each word (in the form of its value vector) to the question posed by the query vector of the current word.

For better results, attention blocks are divided into multiple "heads," where each head performs the same attention calculation described above, but only on a specific portion of each word embedding. Each word embedding is split into multiple, but equally sized vectors representing lower dimensional "subspaces" of their feature space. Each attention head then calculates the query-key-value transformations for their respective portion of the feature space. Finally, the results

of each head are concatenated to produce a new post-attention vector of the same size as the original word embeddings. According to the authors of the original paper on transformer models, "Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this."[147]

There are two remaining steps in masked multi-head self-attention. The first is to project the concatenated, post-attention vector through another matrix of learned parameters. The intuition behind this projection is that, while each individual attention head is meant to capture a lower-dimensional sub-feature of the discursive meaning of the word sequence, simply flattening their results into a single vector is not a meaningful articulation of the linguistic affinities they learned. More expressive, higher-dimensional features can be discovered by taking combinations of these sub-features. This is what the projection matrix does. It projects the concatenated, post-attention vector into the full feature space to learn combinations of affinities between words.

The last step is to add the values of the final matrix projection to the values of the original word embedding used in the query matrix. This has the effect of moving the current word of the sequence into another region of the feature space according to the learned features of the self-attention block. At this step, the values are also normalized to be rescaled between 0 and 1 for more efficient processing.

In terms of size, the query, key, value and projection matrices all have the same two-dimensional form, namely d_{model} x d_{model}. Each matrix also has d_{model} bias variables. In terms of attention heads, the total number varies across LLMs, with more recent, larger LLMs tending to support more attention heads. However, the number of attention heads does not affect the size of the query, key, value and projection matrices, since multi-head attention works by dividing up and recombining subparts of the matrices according to the number of heads. For GPT-3, given that d_{model} = 12,288, this means that the total number of parameters of a masked multi-head self-attention block is (4 x 12,288 x 12,288) + (4 x 12,288) =

[147] See Aswani et al., *supra* note 125.

603,979,776 + 49,152 = **604,028,928**. The number of attention heads supported by GPT-3 is 96, which makes for an attention head dimension size of 12,288 / 96 = 128.[148]

Feed-forward neural networks

This element of transformer models plays a similar role to that of the projection matrix in masked multi-head self-attention. After the "feature extraction" stage above, the next step is to provide a "classification" of sorts by passing the output of the masked multi-head self-attention block to a fully connected feed-forward neural network. The term "feed-forward" simply means that the neural network passes information between nodes in a forward direction only, with no recurrence or other type of directionality. The employment of a neural network at this stage of processing is a common design pattern in sequential machine learning tasks. The purpose of it is to learn more complex and nuanced (possibly non-linear) combinations of the linguistic features extracted by the masked multi-head self-attention block. In general, this step works to increase the overall representational capacity of the model.

Similar to masked multi-head self-attention, after the neural network has produced its output, the last step is to take those values and add them to the original values of the masked multi-head self-attention vector. Once again, this has the effect of moving the current vector into another region of the feature space according to the learned function of the neural network.

In terms of size, the neural network has three layers: an input, inner and output layer. The input and output layer are both of size d_{model}, whereas the inner layer is of size d_{model} x 4. To implement this form of neural network, two matrices of learned parameters are needed: one of size d_{model} x (d_{model} x 4) and one of size (d_{model} x 4) x d_{model}. Each matrix also has d_{model} bias variables. For GPT-3, given that d_{model} = 12,288, this means that the total number of parameters of a feed-forward neural

[148] See Brown et al., *supra* note 127.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

network is [12,288 x (12,288 x 4)] + [(12,288 x 4) x 12,288] + (2 x 12,288) = 603,979,776 + 603,979,776 + 24,576 = **1,207,984,128**.

Total number of parameters

To calculate the total number of parameters of an LLM, there is one additional property of transformer models to consider. The last two elements of the architecture—that is, masked multi-head self-attention and feed-forward neural networks—are not instantiated once, but multiple times in succession. To further increase the representational capacity of the model, "layers" of masked multi-head self-attention followed by a feed-forward neural network are stacked on top of each other. Each layer contains its own set of parameters.

In terms of size, the total number of layers varies across LLMs, with more recent LLMs tending to have more layers. For example, GPT-2 has 48 layers, whereas GPT-3 has 96 layers.[149] For GPT-3, this means that the total number of parameters of the entire LLM is 96 x (number of parameters of masked multi-head self-attention block + number of parameters of feed-forward neural network) + number of parameters of vocabulary word embeddings + number of parameters of positional encoding matrix = 96 x (604,028,928 + 1,207,984,128) + 617,558,016 + 25,165,824 = **174,595,977,216**, which rounds up to 175 billion.

See Figure 1 for a diagram of the transformer model architecture, with values for GPT-3.

[149] See Radford et al., *supra* note 126 and Brown et al., *supra* note 127.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org

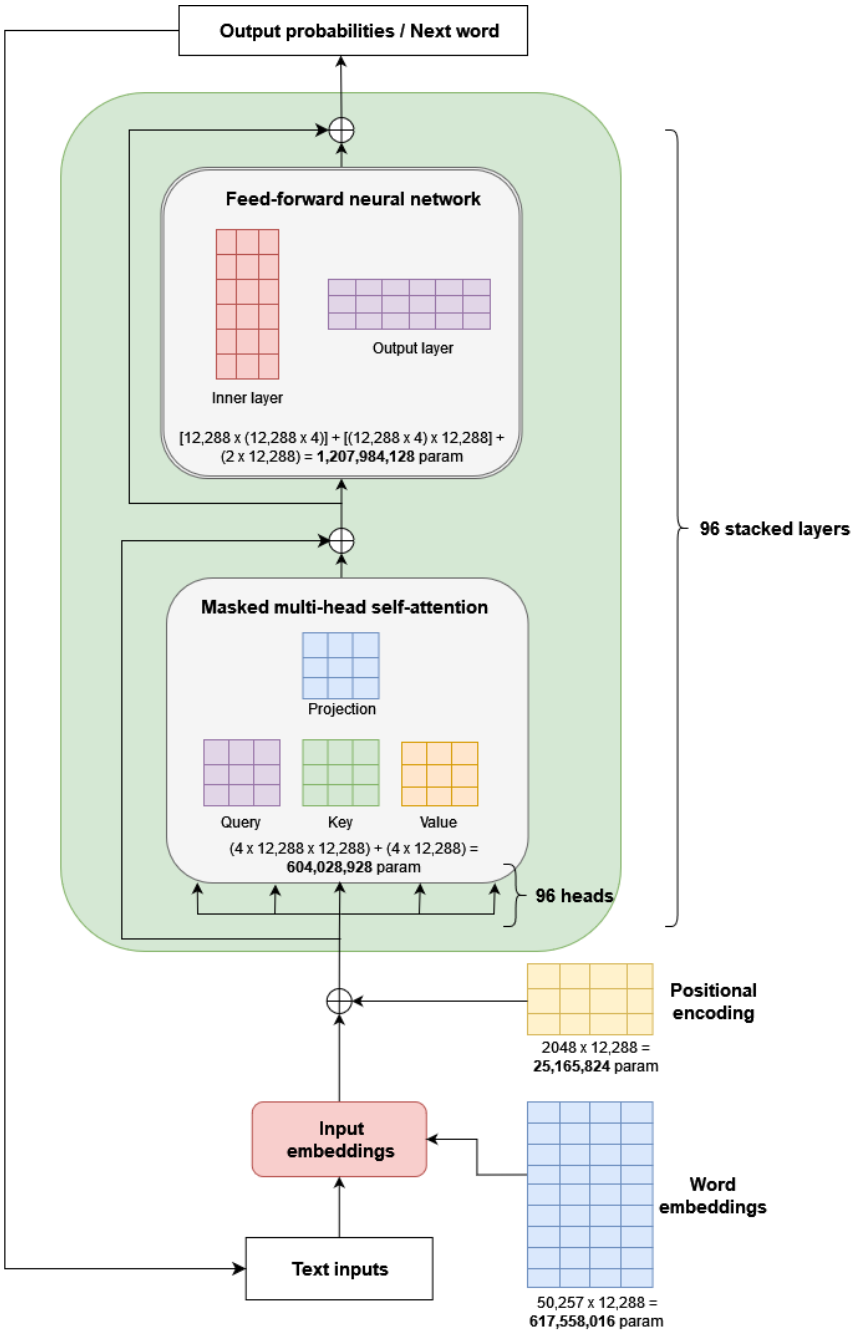

Figure 1: Transformer model architecture, with values for GPT-3. The total number of parameters is 175 billion.

Secretariat
Federal Commissioner for Data Protection and Freedom of Information (BfDI)
Graurheindorfer Str. 153
D-53117 Bonn
Phone +49 / 228 99/ 7799 1404

E-Mail:
IWGDPT@bfdi.bund.de

Internet:
www.iwgdpt.org