Fordham Law Review

Volume 92 | Issue 5 Article 9

2024

ChatGPT, Large Language Models, and Law

Harry Surden
University of Colorado Law School

Recommended Citation

Harry Surden, *ChatGPT, Large Language Models, and Law*, 92 Fordham L. Rev. 1941 ().
Available at: https://ir.lawnet.fordham.edu/flr/vol92/iss5/9

CHATGPT, AI LARGE LANGUAGE MODELS, AND LAW

*Harry Surden**

This Essay explores Artificial Intelligence (AI) Large Language Models (LLMs) like ChatGPT/GPT-4, detailing the advances and challenges in applying AI to law. It first explains how these AI technologies work at an understandable level. It then examines the significant evolution of LLMs since 2022 and their improved capabilities in understanding and generating complex documents, such as legal texts. Finally, this Essay discusses the limitations of these technologies, offering a balanced view of their potential role in legal work.

[*] Professor of Law, University of Colorado Law School. Associate Director, Stanford Center for Legal Informatics (CodeX), Stanford University. This Essay was prepared for the Symposium entitled *The New AI: The Legal and Ethical Implications of ChatGPT and Other Emerging Technologies*, hosted by the *Fordham Law Review* and cosponsored by Fordham University School of Law's Neuroscience and Law Center on November 3, 2023, at Fordham University School of Law.

INTRODUCTION

In recent years, there have been significant advancements in the field of artificial intelligence (AI). Many of these improvements have occurred within the domain of large language models (LLMs).[1] LLMs are AI systems that are designed to understand and generate human language (as opposed to AI systems specialized for other tasks, such as driving cars or detecting fraud).[2] Perhaps the best-known example of an LLM is ChatGPT from OpenAI, a chat-based AI system that can convincingly engage in dialogue, answer questions, and emulate human writing.[3]

So, what has changed? Since about 2022, there has been remarkable progress in the capabilities of LLMs, such as ChatGPT, to create and "understand"[4] complex written language texts, such as research papers, fiction stories, reasoning puzzles, and newspaper articles.[5] Crucially, these advancements also extend to understanding and generating legal documents such as contracts, statutes, motions, and court opinions.[6] In addition to these

[1] *See generally* Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei & I. Sutskever, Language Models Are Unsupervised Multitask Learners (2019) (unpublished manuscript), https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf [https://perma.cc/7TUD-38J5]; Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike & Ryan Lowe, Training Language Models to Follow Instructions with Human Feedback (Mar. 4, 2022) (unpublished manuscript), https://arxiv.org/pdf/2203.02155.pdf [https://perma.cc/MYF8-28L9].

[2] *See* Muhammad Usman Hadi , Qasem Al Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu & Seyedali Mirjalili, Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects (Dec. 7, 2023) (unpublished manuscript), https://www.semanticscholar.org/paper/Large-Language-Models%3A-A-Comprehensive-Survey-of-Hadi-tashi/24de1048791bac4972ecc16d1c3c1de23691407d [https://perma.cc/FLY8-ZD2P].

[3] *See* OpenAI, GPT-4 Technical Report (Mar. 1, 2024) (unpublished manuscript), https://arxiv.org/abs/2303.08774 [https://perma.cc/M5VX-TJLT].

[4] The word "understanding" is in quotes because it is important not to anthropomorphize these AI systems and unwittingly imply that they have cognitive abilities that resemble those of humans. Rather, as this Essay will emphasize, LLM AI systems come to their intelligent-seeming results through statistical approximations. Although it is true that they are often able to produce remarkably accurate and human-like responses, it is quite likely that, at present, they do not "understand" human language in ways that are comparable or analogous to human cognitive understanding. Thus, "understanding" in this context can be thought of as saying that the models produce statistical outputs that are responsive given the input and often approximate what a similarly situated person, who did understand the input at a cognitive level, would produce in response. Even though this ability to produce responsive and salient human-like outputs is remarkable, one must take care not to imply human-like cognition given the way current AI models work.

[5] OpenAI, *Introducing ChatGPT: Optimizing Language Models for Dialogue*, OPEN AI: BLOG (Nov. 30, 2022), https://openai.com/blog/chatgpt [https://perma.cc/8QWZ-7NKY].

[6] Daniel Schwarcz & Jonathan H. Choi, *AI Tools for Lawyers: A Practical Guide*, 108 MINN. L. REV. HEADNOTES 1, 1 (2023); Jonathan H. Choi, Amy Monahan & Daniel Schwarcz, Lawyering in the Age of Artificial Intelligence (Nov. 9, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4626276 [https://perma.cc/DW2H-8G

text analysis abilities, LLMs can now more capably synthesize and reason about facts and the physical world; problem-solve within abstract topics; and work with modalities beyond text, including images, sound, and video.[7]

Moreover, these changes have occurred rapidly. Although basic AI systems focused on human language have existed for decades, these earlier approaches had significant limitations. Often, when analyzing text, these systems would be confused by basic facts or distracted by word choices. In response, earlier AI systems would routinely produce nonsensical outputs or answers that were not responsive to the meaning of what was being asked.[8] Thus, although many researchers considered AI engagement with human language a promising field, it was still distantly elusive in practical application even as recently as 2020.

However, in a relatively quick timespan from 2022 to 2023, LLM systems crossed major thresholds in terms of quality, usefulness, reasoning ability, and generality.[9] In particular, LLMs like OpenAI's GPT-4 (released in early 2023) displayed unprecedented and unforeseen abilities to follow arbitrary written-language instructions, analyze data and text, and solve unseen problems, far surpassing the former state-of-the-art. This increase in AI language capabilities—in a comparatively short period of time—has surprised many researchers who study artificial intelligence.[10] This Essay will examine these recent, significant advancements of LLM AI technologies. It will detail how the systems work and explore their current—and potential short-term—impact on law. This Essay will argue that given the fundamental role of language and textual documents in law, these technological advancements are likely to affect the legal domain significantly.

RC]; Ronald M. Sangrund, *Who Can Write a Better Brief?: Chat AI or a Recent Law School Graduate: Part I*, Colo. Law., July–Aug. 2023, at 24. For video examples, see generally Harry Surden, *GPT-4 and Law: ChatGPT Applies Copyright Law*, YouTube (Mar. 22, 2023), https://www.youtube.com/channel/UCqsQlzP1Mj3vQmYdi42ziVw [https://perma.cc/2WDA-QSRZ].

[7] *See, e.g.*, Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang & Fei Huang, mPLUG-Owl 🦉: Modularization Empowers Large Language Models with Multimodality (Apr. 27, 2023) (unpublished manuscript), https://arxiv.org/pdf/2304.14178.pdf [https://perma.cc/PK7C-PPES]; Muhammad Maaz, Hanoona Rasheed, Salman Khan & Fahad Shahbaz Khan, Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models (June 8, 2023) (unpublished manuscript), https://arxiv.org/pdf/2306.05424.pdf [https://perma.cc/GLT2-TC8Q].

[8] *See generally* Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 Ga. St. U. L. Rev. 1305 (2019).

[9] *See* David Foster, Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play 260–64 (2d ed. 2023).

[10] The author of this Essay, who has studied AI systems for nearly twenty years, is one of the researchers surprised by how quickly AI systems performance increased between 2022 and 2023. As will be discussed, in just one year, there were improvements in terms of AI natural language understanding and problem-solving that most thought would take closer to five years or more to come to fruition.

Importantly, this Essay will also explore the limitations of these new AI systems. Although advanced LLM systems have demonstrated unprecedented levels of performance, they are not without their flaws.[11] Rather, they currently tend to perform certain tasks (e.g., summarization of documents) much better than others (e.g., answering abstract questions). These limitations make sense once one understands the underlying design of LLMs and how they work. Detailing strengths and limitations in a rapidly evolving field like AI is challenging. What may be a limitation today may be improved tomorrow, and new, unforeseen issues might emerge. Nevertheless, this Essay will endeavor to identify, based on current evidence and research, those AI limitations that seem more transient and those that are likely to persist within a five-year time frame, as well as those AI capabilities that are likely to improve in the coming years.

I. WHAT HAS CHANGED IN ARTIFICIAL INTELLIGENCE?

It is helpful to situate the recent advances in LLM technology within the larger field of AI. AI is notoriously difficult to define, and there is probably no single satisfactory definition that most researchers would agree to.[12] However, one practically useful description of AI is "[u]sing computers to solve problems, make predictions, answer questions, [generate creative output,] or make automated decision or actions, on tasks that when done by people, typically require 'intelligence.'"[13] "Intelligence," although similarly hard to define, is often used to refer to a suite of higher-order human cognitive skills, such as abstract reasoning, problem-solving, decision-making, learning, visual processing, language generation and understanding, critical thinking, and planning.[14]

Using these concepts, AI can be understood as follows: whenever we use a computer system to automate tasks that, when humans do them, require higher-order, cognitive skills associated with "intelligence"—such as driving a car, playing chess, reading books, writing emails, or creating art or music (tasks that, when performed by humans, variously engage cognitive abilities such as visual processing, planning, linguistic ability, or artistic creativity)—we can categorize such automated activities as AI tasks.[15] More broadly, AI

[11] *See generally* Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu & Pascale Fung, A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity (Nov. 28, 2023) (unpublished manuscript), https://arxiv.org/pdf/2302.04023.pdf [https://perma.cc/CP2B-A2AC].

[12] *See* Surden, *supra* note 8, at 1306–07; *see also* Samuel D. Hodge, Jr., *Revolutionizing Justice: Unleashing the Power of Artificial Intelligence*, 26 SMU SCI. & TECH. L. REV. 217, 219 (2023) ("Artificial intelligence has been defined in many ways. A Google search will yield over 1,670,000,000 references to the term." (footnote omitted)).

[13] *See* Sangrund, *supra* note 6, at 26.

[14] *See* Dana S. Nau, *Artificial Intelligence and Automation*, *in* SPRINGER HANDBOOK OF AUTOMATION 249 (Shimon Y. Nof ed., 2009); Surden, *supra* note 8, at 1306–07.

[15] There are two caveats with this definition. First, although computers can perform certain AI tasks associated with human thinking, the methods that they use are very different

is a field that spans many related topics, including robotics, computer vision, logical reasoning, automated prediction, and the development of algorithms that learn from data, to name a few.[16]

Within the broader AI landscape, researchers most closely associate LLMs with the subfield known as "natural language processing" (NLP). NLP is the research area focused on creating AI systems that can understand and generate human language and text.[17] The term "natural language" has a specific meaning in computer science. It refers to the ordinary languages that people use to communicate with one another, such as English, French, and Spanish. The phrase "natural language" is meant to contrast with the computer science concept of a "formal language." "Formal languages" are the highly constrained and mathematically structured technical "languages," such as Python and Javascript, which are used, among other things, to program computers. Because formal languages are created for unambiguous, mechanical interpretation, they are very limited in what they can express, but computers are able to reliably process them.

By contrast, computer systems have traditionally struggled to respond sensibly to ordinary natural language sentences that many people effortlessly understand, such as "Read this lease and tell me the name of the renter." For this reason, a longstanding but challenging goal of AI has been the development of systems that can understand ordinary human communications, rather than just formal programming languages. Much of NLP research has thus focused on trying to develop techniques that would allow systems to understand and create ordinary language documents or speech, at levels that approach (or exceed) those of a similarly situated person.[18] Although today LLMs like ChatGPT can generate surprisingly human-like answers and sensibly respond to natural language instructions, this was not the case until recently.

To fully grasp the significance of these recent AI advances, it is helpful to explore the earlier limitations in NLP technology. In short, prior to 2020, NLP systems were unable to interpret written text at a level even remotely

from our own. AI systems typically solve problems using statistical associations, rules, and computational processes that look very different from human mental and cognitive processes. *See* Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han & Yang Tang, *A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development*, 10 IEEE/CAA J. AUTOMATICA SINICA 1122, 1129 (2023). To avoid anthropomorphizing, and making inapt analogies between human and computer capabilities, it is important to keep these structural differences in mind. Second, although this AI definition focuses on automating human tasks, it is important to observe that there are some tasks that AI systems can do that no human could ever realistically do. A good example is spotting a single fraudulent credit card purchase among millions of transactions—something narrow AI systems do quite capably today. Nonetheless, although imperfect, this AI definition is useful for our purposes.

[16] *See, e.g.*, Surden, *supra* note 8, at 1325–26.

[17] *See* K. R. Chowdhary, *Natural Language Processing*, *in* FUNDAMENTALS OF ARTIFICIAL INTELLIGENCE 603 (2020).

[18] *See* Surden, *supra* note 8, at 1313.

approaching that of a literate person.[19] Systems of the recent past struggled to grasp the meaning of written documents such as emails, letters, books, or instructions.[20] Rather, NLP systems of this period primarily analyzed human communication as merely word and data patterns.[21] Thus, even though these earlier NLP systems could reliably identify keywords and text patterns within documents, such as emails, they struggled to truly comprehend what emails and other natural language documents were actually about.[22]

For instance, consider how NLP systems of this period were limited with respect to legal natural language documents, such as contracts. Take a short natural language contract clause, such as, "This agreement is governed by the laws of the state of California." If one were to have given this text to an earlier NLP system, it might have been able to statistically categorize it, with reasonable reliability, as a "governing law" clause. Such systems operated by learning patterns from large document databases, in which similar contract language had been manually identified and labeled by lawyers, or other domain experts, as a "governing law" clause. NLP systems of that era would have been able to match the significant words of this clause using data patterns previously ascertained from large contract datasets in order to classify it.[23] Thus, these earlier systems excelled at, but were largely limited to, narrow tasks like categorizing natural language text based on previously detected, statistical patterns.[24]

However, the important limitation of these NLP systems was their inability to engage with the underlying meaning of these words. Notably, such NLP systems could not reliably grasp the underlying *significance* of the word patterns (such as "governed by the laws of the state of California") that they analyzed. They treated written language primarily as data, rather than words with meaning. For instance, if one had asked a system from that period, "What are the consequences of having this contract governed by California Law?," the response likely would have been nonsensical or irrelevant.[25] Such systems simply lacked the capacity to understand and coherently answer such a question and would not actually have had a sense of what a

[19] *See* Luciano Floridi & Massimo Chiriatti, *GPT-3: Its Nature, Scope, Limits, and Consequences*, 30 MINDS & MACHINES 681, 681 (2020) (observing limitations of early GPT iterations noting "GPT-3 [from 2020] does not do what it is not supposed to do, and that any interpretation of GPT-3 as the beginning of the emergence of a general form of artificial intelligence is merely uninformed science fiction").

[20] *See* STEVEN BIRD, EWAN KLEIN & EDWARD LOPER, NATURAL LANGUAGE PROCESSING WITH PYTHON 27–30, 361–67 (2009).

[21] *See id.* at 224–27.

[22] *See id.*

[23] *See id.* at 57–59.

[24] *See id.* at 230–33.

[25] *See* ChatGPT, OPENAI (Mar. 3, 2024), https://chat.openai.com/share/46aecbaf-dfcb-4c5b-bf9f-5d254aa26ca0 [https://perma.cc/Y55D-NCXJ] (text generated by ChatGPT in response to the query "What are some consequences of having a contract governed under 'California Law'?") ("California's approach to contract law includes several distinctive principles that might not be as pronounced or interpreted similarly in other jurisdictions. For example, California courts have a strong tendency towards the enforcement of contracts as written . . . .").

"governing law" provision meant to a lawyer, nor could they meaningfully have interpreted the implications of such a provision in natural language. Instead, these earlier NLP systems largely treated language as data, identifying statistical matches to specific patterns (e.g., "g-o-v-e-r-n-i-n-g"), without grasping the underlying legal or real-world *meaning* of a phrase like "governing law."[26]

Until recently, creating AI systems that could usefully understand the meaning (or "semantics") of natural language phrases, such as "This contract is governed by the laws of the state of California," seemed like an intractable technical problem. Part of the difficulty lay in the expressive flexibility of human languages. In most natural languages like English, one can convey the same idea in a nearly infinite variety of ways (e.g., "This contract must be interpreted under the legal rules and principles of California" versus "The laws of the state of California control this agreement"). One can talk about a vast range of topics, using synonyms, metaphors, nuances, and implied context, as well as by rearranging sentences in multiple, essentially equivalent ways. Despite the incredible expressive variation of human language, most *people* quite easily make sense of the vast range of possible linguistic characterizations, implications, contexts, and subtle nuances in meaning and syntax as they rapidly comprehend text or speech from others.

This earlier gap in AI language processing can be contrasted against the cognitive flexibility of the earliest human learners. Even very young children can adeptly respond to an arbitrary and varied range of linguistic expressions, ranging from "Can you draw a picture of a red circle?" to "Tell me a story about a dog." Similarly, children can effortlessly integrate common sense and facts about the world in responding to nonsensical question, such as "How many legs does an apple have?" (i.e., most children would respond appropriately with something like "Apples don't have legs!"). Thus, most people, including children, can comprehend and respond appropriately to nearly any arbitrary but meaningful communication, even if that communication is unusual in topic, requires common sense, or is outside of direct past experiences.

By contrast, AI NLP systems have not traditionally been able to handle the wide variety of syntactical variation, semantic ambiguity, and abstract complexity that is a typical part of most human written and spoken communication.[27] Indeed, these limitations are one of the major reasons that we traditionally program computers in *formal* computer programming languages—like Python, Javascript, or C—rather than in natural languages, such as English.[28] Past computer systems simply did not have the ability to understand the meaning of natural human language communication reliably enough to carry out comparable computational instructions. For instance, if we had tried to instruct computers of the recent past using ordinary English

[26] *See* BIRD ET AL., *supra* note 20, at 397–401, 445.

[27] *See id.* at 33.

[28] *See* FRANKLYN TURBAK, DAVID GIFFORD & MARK A. SHELDON, DESIGN CONCEPTS IN PROGRAMMING LANGUAGES 3–8 (2008).

instructions, such as “Make a program where a red ball bounces around the screen,” a typical system could not understand what was being asked nor produce a useful or responsive output.

Because of the gap between the natural language communication of humans and that of computers, the computer science “solution” was not to instruct computers using everyday English (or other natural languages), but rather to communicate with computers using artificially created, specialized formal “computer languages.”[29] These languages offer only a relatively small number of limited, unambiguous, and highly constrained instructions. Given these strictures, a computer can then reliably translate such a limited range of formal programming instructions into known and determinate computational actions and execute them. Thus, “formal” programming languages trade off the linguistic flexibility of natural human languages for the precision and constraint that makes computers able to reliably process and execute commands. However, these formal computer languages are much more limited in what they can express as to the complexity and nuance of real-world abstractions, ideas, phenomena, and facts compared to natural languages.

In short, until recently, NLP computer systems could not reliably answer and respond to the wide variety of natural language questions in the way that human adults and children effortlessly do. As late as 2021, state-of-the-art NLP systems were critiqued as “stochastic parrots” because they provided an illusion of understanding natural language when, in fact, most of the coherent-seeming text generated by systems from that time occurred by remixing versions of existing sentences that *people* had previously written on the internet.[30] Thus, the lack of “understanding” by NLP systems of that era became quickly apparent when they were asked unusual or common-sense questions. This inability to engage reliably and responsively with unexpected or flexible human communications made past NLP systems quite limited and useful for only very specific applications. Although these narrower NLP systems were somewhat useful in heavily language-based areas like law, their application was primarily limited to the statistical categorization of legal documents, rather than in understanding the underlying meaning or significance of the written words on a document such as a contract, motion, or lease. These limitations had largely remained in place for decades, and, for that reason, many researchers saw the ability of computers to usefully understand human language as a comparatively distant research goal.

[29] *Id.*

[30] *See, e.g.*, Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?: 🦜*, *in* FAccT ‘21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610 (2021), https://dl.acm.org/doi/pdf/10.1145/3442188.3445922 [https://perma.cc/UE4B-5W9Q].

A. Large Leaps in Capability in Just One Year (2022–2023)

However, in 2022, the capabilities of AI NLP systems dramatically improved. This was precipitated by the November 2022 release of ChatGPT (GPT-3.5) from the company OpenAI.[31] ChatGPT was announced as a research preview—a chat-based interface to an underlying LLM architecture known as "GPT" (described below). ChatGPT was designed to take natural language text-based "prompts" from users, such as "What is the capital of France?" or instructions such as "Write a poem about a dog" and respond with appropriate, human-like text.[32] It was also able to create a large variety of draft documents at an unexpectedly high level of quality on just about any topic, including legal documents like court motions and contracts.[33]

What was surprising to many researchers was how strongly GPT-3.5 surpassed many of the limitations that had plagued earlier AI LLMs, such as Galactica (from Meta) and GPT-3 (OpenAI's earlier model from 2020).[34] Due to engineering innovations by OpenAI, ChatGPT displayed unanticipated AI capabilities compared to earlier LLM technology.[35]

First, ChatGPT was able to respond sensibly to nearly any arbitrary text question, input, or instruction that it was given.[36] The fundamental hurdle facing LLMs at that time was not primarily factual accuracy, but rather much more fundamental: simply getting LLMs to generate replies that were remotely pertinent to the questions or instructions posed.[37] LLMs prior to ChatGPT struggled to cope with the nearly infinite flexibility of human language.[38] When faced with inputs that were somewhat unusual or that they had not been specifically trained to handle, earlier LLM systems would, more often than not, respond with incoherent or irrelevant answers, indicating that the AI system was confused or did not understand what was being asked.[39] By contrast, ChatGPT was the first LLM to be able to respond appropriately to nearly anything that it was asked or instructed.[40] Because its answers consistently were responsive to what was asked (if not always completely

[31] *See* OpenAI, *supra* note 5.

[32] *See* Wayne Xin Zhao, Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie & Ji-Rong Wen, A Survey of Large Language Models 32–33 (Nov. 24, 2023) (unpublished manuscript), https://arxiv.org/pdf/2303.18223.pdf [https://perma.cc/7QBJ-BPYC].

[33] *See, e.g.,* Harry Surden, *GPT-4 and Law: ChatGPT Generates a Federal Motion to Dismiss Using GPT-4*, YOUTUBE (Mar. 19, 2023), https://www.youtube.com/watch?v=jjFBGrv6eiY [https://perma.cc/QZL2-8SCL].

[34] *See* Hodge, *supra* note 12, at 217–18, 240.

[35] *See id.*

[36] *See id.* at 224.

[37] *See* FOSTER, *supra* note 9, at 252–64.

[38] *See* BIRD ET AL., *supra* note 20, at 33.

[39] *See id.* at 295–98.

[40] Holly Else, *Abstracts Written by ChatGPT Fool Scientists*, 613 NATURE 423 (2023).

factually accurate), it gave the appearance to the user of being among the first AI systems to "understand" the questions or instructions given to it.[41]

For instance, a user could actually ask ChatGPT, "What is the significance of contract language like, 'This agreement shall be governed by the laws of the state of California'?," and ChatGPT would not just recognize word-patterns like previous LLMs; it could rather explain its real-world meaning.[42] This was impressive because at the time it was thought that creating an AI LLM system that consistently produced pertinent responses to whatever a user asked was unlikely to be achieved for some time.[43]

To emphasize the significance of this achievement, observe that other state-of-the-art LLMs that were released roughly contemporaneously to ChatGPT struggled to provide coherent answers to arbitrary questions. For instance, Meta's LLM, Galactica, released just a few weeks before ChatGPT, was unable to be consistently and broadly responsive (outside of a narrow area) in the way that OpenAI's ChatGPT managed to be.[44] Thus, ChatGPT's ability to produce outputs that were reliably related to the topic of inquiry, and to cope with the nearly infinite variety of natural language expression, was a major and surprising accomplishment.[45] Whereas previous LLMs quickly displayed their limits with incoherent answers in response to unexpected text, ChatGPT's generally appropriate outputs also made it one of the first LLMs to give an impression of "intelligence."[46] ChatGPT appeared to be among the first LLM systems that were able to reliably engage with the underlying *meaning* of natural language words (and not just their textual patterns)—a long elusive goal of NLP researchers.[47]

The second surprising and emergent ability that ChatGPT displayed was the ability to reason and solve problems. Some pointed out that its reasoning

[41] Jianyang Deng & Yijia Lin, *The Benefits and Challenges of ChatGPT: An Overview*, FRONTIERS COMPUTING & INTEL. SYS., May 2023, at 81.

[42] *See, e.g.*, CHATGPT, https://chat.openai.com/share/dece5c96-ef42-4c08-a844-17c4582c649a [https://perma.cc/VP7N-QZ9P] (last visited Mar. 3, 2024) (text generated by ChatGPT in response to the query "give me a very short contract governing law clause for california").

[43] *See* Hongyeon Yu, Jaehyun An, Jeongmin Yoon, Hyemin Kim & Youngjoong Ko, *Simple Methods to Overcome the Limitations of General Word Representations in Natural Language Processing Tasks*, 59 COMPUT. SPEECH & LANGUAGE 91, 93–96 (2020); Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi & Zaid Harchaoui, MAUVE: Measuring the Gap Between Neural Text and Human Text Using Divergence Frontiers (Nov. 23, 2021) (unpublished manuscript), https://arxiv.org/pdf/2102.01454.pdf [https://perma.cc/E96E-8P3B].

[44] *See, e.g.*, Will Heaven, *Why Meta's Latest Large Language Model Survived Only Three Days Online*, MIT TECH. REV., Nov. 18, 2022, https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ [https://perma.cc/R2ZU-VTED].

[45] Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, Carl Yang, Yue Cheng & Liang Zhao, Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models (Jan. 4, 2024) (unpublished manuscript), https://arxiv.org/pdf/2401.00625.pdf [https://perma.cc/A2Y6-LAXE].

[46] *Id.*

[47] *Id.*

was not always accurate and frequently contained logical flaws.[48] However, this critique, like the criticism of the occasional factual errors it produced, although true, missed a larger point. The bigger surprise was that ChatGPT displayed emergent reasoning and problem-solving abilities *at all*. This was because problem-solving and reasoning were largely not what ChatGPT had been designed to do.[49] Rather, GPT-3.5 had been built on a predecessor model, GPT-3, which had been developed primarily to produce human-quality natural language text output.[50] Such systems had been designed mainly to simulate human-like writings—like poems, stories, articles, and letters—by having previously analyzed patterns from billions of example documents from the internet.[51] Notably, however, problem-solving and reasoning were not a major part of its design. Rather, it appeared that ChatGPT's problem-solving and reasoning capabilities emerged spontaneously as an unexpected by-product of exposure to large amounts of data, including programming code, math problems, and logic puzzles. In other words, at the time most AI researchers did not expect that a probabilistic system that was specifically designed to simulate human-like written stories or articles using next word prediction would suddenly exhibit emergent (albeit imperfect) abilities to solve logic or technical problems that it had not been specifically exposed to before nor be able to reason about everyday scenarios.[52]

It is important to emphasize that GPT-3.5 was not without its flaws. Like other LLMs then and since, it would make mistakes in answers or reasoning and occasionally invent (or "hallucinate") untrue but plausible-seeming information.[53] Moreover, as many commenters have observed, one must take care not to anthropomorphize advanced LLMs like ChatGPT.[54] Words like "understanding" and "intelligence" should be taken as functional metaphors; it would be misleading to characterize these systems as "intelligent" in the same way that cognitively active humans are. At base, these LLMs are still very advanced, statistical pattern-matching machines, despite their capable output, and most experts believe that they did not "understand" their inputs and outputs in a manner comparable to a similarly situated person.[55] Researchers readily acknowledged those limitations.[56]

[48] *See* Wu et al., *supra* note 15, at 1129.

[49] *See* Van Lindberg, *Building and Using Generative Models Under US Copyright Law*, RUTGERS BUS. L. REV., Spring 2023, at 1, 9.

[50] *See generally* BIRD ET AL., *supra* note 20.

[51] *Id.*

[52] *See* Zhao et al., *supra* note 32, at 3.

[53] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua & Weiqiang Jia, Cognitive Mirage: A Review of Hallucinations in Large Language Models (Sept. 13, 2023) (unpublished manuscript), https://arxiv.org/pdf/2309.06794.pdf [https://perma.cc/E7HE-2E7U].

[54] R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy & Thomas L. Griffiths, Embers of Autoregression: Understanding Large Language Models Through the Problem They Are Trained to Solve 35 (Sept. 23, 2023) (unpublished manuscript), https://arxiv.org/pdf/2309.13638.pdf [https://perma.cc/7XUL-PLRQ].

[55] *Id.*

[56] *Id.*

However, the reason that ChatGPT attracted so much attention was that it was the first NLP system to be broadly useful. What impressed researchers was that ChatGPT displayed practical abilities for a huge range of real-world tasks. Further, it displayed unexpected emergent properties that most researchers previously thought would not be possible for many years, such as reasoning, problem-solving, and producing pertinent answers to arbitrary natural language questions or instructions.[57]

Thus, the larger story of 2022 was not about ChatGPT's accuracy limitations or whether it was "intelligent," but rather the large leap in usefulness and capabilities of LLMs that few thought possible in such a short time. One could, for example, ask ChatGPT in ordinary English about a nearly infinite variety of tasks, such as "Write a computer program to make a ball bounce around a screen"; it could produce reasonably responsive, if not always perfect, outputs. These were capabilities that previous LLMs appeared far from being able to accomplish. In just a few months, AI LLM technology took a significant leap from "not so useful" to "reasonably useful" in terms of performing tasks as well as generating, understanding, and reasoning about natural language human communication.

Just seven months after the release of the original GPT-3.5, in March 2023, OpenAI released a much more advanced LLM: GPT-4.[58] Although technical details were never formally released by OpenAI, GPT-4 was rumored to be significantly larger than its predecessor GPT-3.5.[59] Notably, GPT-4 improved in every AI ability across the board compared to the original GPT-3.5. GPT-4 exhibited huge gains over GPT-3.5 in terms of factual accuracy and reasoning ability, as well as in the ability to understand and generate complex texts, create computer programs, and analyze and process abstract ideas.[60] As of the writing of this Essay, GPT-4 is still the most advanced LLM system to date, although several competitors are in the works and OpenAI is currently creating the next generation version within its GPT family of LLM systems, GPT-5.[61] Thus, much of the discussion below will refer to GPT-4, as it is currently the state of the art in LLM AI technology, rather than the original GPT-3.5, which is the less capable version available to the public for free.

[57] *Id.*

[58] *See* Achiam et al., *supra* note 3, at 7.

[59] It is important to distinguish GPT-4 from GPT-3.5. OpenAI offers a free version of ChatGPT—that is powered (as of the end of 2023) by GPT-3.5. By contrast, GPT-4, the most powerful and capable AI version, is only available to ChatGPT Plus subscribers. OpenAI, *Introducing ChatGPT Plus*, OPENAI: BLOG (Feb. 1, 2023), https://openai.com/blog/chatgpt-plus [https://perma.cc/5NPS-9M6W]. This is important to clarify because people often conflate the less powerful abilities of the free GPT-3.5, which makes frequent mistakes, with the much more powerful GPT-4, which they may not have used. As of December 2023, all users can, without paying, access a version of the more powerful GPT-4 AI model through Microsoft's Bing Chat and Copilot website. *See* Hodge, *supra* note 12, at 230.

[60] *See* Achiam et al., *supra* note 3, at 7–10.

[61] As of December 2023, Google has announced an AI system called "Gemini Ultra" that they claim is competitive with GPT-4. Sundar Pichai & Demis Hassabis, *Introducing Gemini: Our Largest and Most Capable AI Model*, GOOGLE: KEYWORD (Dec. 6, 2023), https://blog.google/technology/ai/google-gemini-ai/ [https://perma.cc/TZY6-M7QP].

Today, LLMs like GPT-4 have shown impressive capabilities in law that were thought to be nearly impossible only a few years ago.[62] For example, GPT-4 can (albeit sometimes imperfectly) engage in legal reasoning about law and facts, analyze or generate contracts, summarize legal cases, draft patents, write motions, and answer questions about legal opinions or documents.[63] Although the results are occasionally unsatisfactory, and sometimes contain errors, just the fact that these systems can perform reasonably at these—and many other—legal tasks at all is astonishing, given the recent technical limitations that had made such flexible and responsive AI natural language capabilities seem distantly out of reach. Moreover, there is reason to believe that many of the issues of accuracy with current LLM systems are likely to be reduced in upcoming technological iterations.[64]

To understand the potential uses of LLMs like GPT-4 in law, it is helpful to understand how they work. The next part will endeavor to explain this at a detailed but understandable level.

B. How Large Language Models Work

Many readers are likely familiar with the interface of ChatGPT. Users interact with ChatGPT by entering a "prompt," which is the term for a user-provided text input, such as an instruction (e.g., "Write a poem about dogs" or "Summarize this document" or "Write a merger agreement") or a question ("What is the capital of France?" or "What are the factors of copyright fair use?"). The prompt provides ChatGPT with the context needed to start producing the next words of its response. ChatGPT (which is the "user interface") will send the user's prompt to an underlying LLM system, such as GPT-4 (the more advanced, paid version), or GPT-3.5 (the free, less advanced version), to start computing a relevant series of sentences that are pertinent to the user's instruction or question.[65]

At a high level, LLMs like ChatGPT are essentially advanced AI word-prediction[66] systems. Somewhat counterintuitively, when ChatGPT appears to be answering a question or producing a document (such as a poem or a legal motion), it is actually using its internal AI system to predict the

[62] *See* FOSTER, *supra* note 9, at 264.

[63] *See, e.g.*, Ashley Binetti Armstrong, Who's Afraid of ChatGPT?: An Examination of ChatGPT's Implications for Legal Writing (Jan. 23, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4336929 [https://perma.cc/JC2F-ZLWE]; *The Legal AI You've Been Waiting For*, CASETEXT, https://casetext.com/cocounsel/ [https://perma.cc/6SUH-NWYH] (last visited Mar. 3, 2024).

[64] *See* Schwarcz & Choi, *supra* note 6, at 3 n.14.

[65] Because many of the most advanced LLMs require intensive computation, they usually reside at distant commercial data centers with powerful computers, and they send and receive information over the internet. *See* Bai et al., *supra* note 45, at 29.

[66] As a technical matter, most LLMs actually predict "tokens"—word subparts rather than complete words. *See* FOSTER, *supra* note 9, at 146–49. Tokens are small units of text that a model uses internally to process longer words more easily; they are usually subparts of larger words. *Id.* at 134–37. For example, the word "computer" might be broken down internally by the LLM into the tokens "com," "put," and "ter." *Id.* However, to avoid complication, I will discuss LLMs as predicting words, even though they predict tokens.

most appropriate word to generate next, one word at a time.[67] Even when generating lengthy documents or comprehensive answers, it still operates by generating the text incrementally, predicting one word at a time, based on what has been asked or instructed.

How does the model choose what word to produce next as it is generating a response? It begins by examining the prompt, such as “Write a poem about dogs.” This prompt is put inside the model's "context window.” We can think of the context window as like the model’s short-term memory—a collection of context words that indicate to the model what the conversation is about and that the AI system then uses to make an educated guess about what word should come next, given the patterns of grammar and writing and also the underlying meaning of the user’s prompt. As the model answers the user’s question one word at a time, the context window grows incrementally to include the words that the model itself has produced so far in response (e.g., if AI were asked to “Write a poem about dogs,” and the model has partially answered “Dogs like,” the last word in its growing context window would be “like”). Thus, before ChatGPT chooses what word to produce next, it first examines the entire context window—what the user has asked it to do in the original prompt, plus the words it has just produced in response—to figure out the most appropriate next word.

The model has a large vocabulary of more than 50,000 of the most common “words” to choose from when identifying the next word (e.g., “attorney,” “beach,” “play,” or “zebra”).[68] When choosing the next word to add to the text it has already produced (“Dogs like to”), GPT-4 uses its advanced AI algorithm to probabilistically identify the one word, from its 50,000 word vocabulary list, that is likely the most appropriate given the words it has already produced and what the user originally asked. It may seem strange that the model looks at the previous words that it itself has already produced in its response so far (e.g., “Dogs like to”) to help it predict the next appropriate word to generate (e.g., “play”), but that is the essence of how LLMs such as ChatGPT function.[69]

For instance, imagine that the system has received a user question as a prompt, such as “What is the capital of France?” The prompt provides the initial context to guide the LLM. The system will run this entire context window through its AI algorithm and, after a sophisticated computation, assign a probability to each one of the over 50,000 possible next words in its vocabulary list (such as “apple,” “cat,” or “Paris”) in light of the appropriateness of each possible word given the context. Imagine that, after this AI analysis of the context words, the next word it gives the highest relevance probability score to is “Paris.” “Paris” is a good choice for the next word to generate because it would factually and succinctly answer the prompt.

[67] *See* Wu et al., *supra* note 15, at 1124.

[68] *See* Radford et al., *supra* note 1, at 4.

[69] *See* Jingshan Huang & Ming Tan, *The Role of ChatGPT in Scientific Communication: Writing Better Scientific Review Articles*, 13 AM. J. CANCER RSCH. 1148, 1149–50 (2023).

As will be discussed, the model has previously learned, after examining millions of existing written webpages and articles, that the word "Paris" is statistically very likely to closely follow earlier context words like "capital" and "France" in written, English sentences. However, because the predicted outputs are probabilistic in nature, the same prompt might result in a slightly different path to a comparable answer, if asked at another time. For instance, if given the same prompt ("What is the capital of France?") another time, the system might pick a different, most probable first word of the answer such as "The" (rather than "Paris"), on the way to ultimately producing a comparable, but wordier, response such as "The capital of France is Paris."[70] This is because, during training, the model was exposed to multiple valid ways to answer a question. These ranged from single-word responses to other responses that partially reformulated the original question beginning with a "The," and the model internalized these and other common question-answering patterns. Thus, the model can answer the same question in slightly different ways because there is some intentional randomness injected each time as to the ultimate format and words that the model chooses. This randomness helps the model produce varied and more creative text across sessions.

It is important to emphasize the iterative nature of GPT systems. ChatGPT generates its responses incrementally, one word at a time, by adding a new word to its answer based on what it itself has produced so far. So, the earlier response sequence might look like: "*The*," "The *capital*," "The capital *of*," "The capital of *France*," "The capital of France *is*," "The capital of France is *Paris*." The iterative nature means that it continually feeds the original prompt, plus the words it has produced so far ("The capital of France *is*")—the context—back into the system to keep generating the next most likely word based on the incrementally building context ("Paris"). This is why ChatGPT's answers appear slowly, one word at a time, as the system keeps inputting its own produced words back to itself, to determine the word that it is going to produce next.[71] The choice of next word is based on conditional probability: the earlier words of the user's prompt influence the probability of certain later words being produced in the output.

To illustrate conditional probability, consider the initial prompt "Write a poem about a dog." There are important context words in this prompt, such as "dog," that guide the model to be more likely to subsequently choose dog-related words from its vocabulary—such as "barks" or "plays," rather than unrelated words like "window" or "computer"—as it incrementally writes the poem ("*A*", "A *dog*", "A dog *barks*"). If the model had already produced the partial three-word poem "A dog barks," choosing a fourth word commonly related to "barks," like "loudly," becomes much more probable out of the over 50,000-word vocabulary than, say, an unrelated word such as "telephone." This is the essence of conditional probability—given the

[70] SINAN OZDEMIR, QUICK START GUIDE TO LARGE LANGUAGE MODELS: STRATEGIES AND BEST PRACTICES FOR USING CHATGPT AND OTHER LLMS (2023).

[71] *See* Zhao et al., *supra* note 32, at 25.

previous context words, “A dog barks,” the probability of choosing “loudly” (or another word related to “barks”) as the next word increases. By contrast, had the model instead chosen a different, but relevant, third word—“plays” instead of “barks” (“A dog plays”)—the system would have instead statistically been more likely to choose a different fourth word that commonly modifies the word “play,” such as “fetch” or “outside” (e.g., “A dog plays fetch.”).

These patterns—that words like “dogs,” “play,” and “fetch” or words like “Paris” and “France” frequently appear together—were learned by the model previously in training as it examined millions of human-written documents about dogs, France, and innumerable other topics.[72] Related words frequently appear in context together in written text, and the model eventually learns these relations during training as it sees them together many times. Later, as the trained model answers questions, both the initial prompt and the words that the model itself has already produced end up influencing the probability of the subsequent words in the response.

The fact that this relatively simple, one-word-at-a-time prediction system ends up producing coherent and responsive human-like text is astonishing. This was partly why researchers were surprised about ChatGPT’s capabilities when it was released. It was not obvious that taking an AI system that simply predicts one word at a time based on the probability of earlier context words—and scaling that system up significantly in size—would allow that larger system to suddenly be able to produce complex documents, engage in reasoning, solve problems, and answer complex questions, in ways that earlier, smaller next-word prediction LLM systems were unable to do.

Importantly, the “path” that the LLM starts going down at the beginning with the words it initially predicts can heavily influence the content of its ultimate answer. Because these systems produce a response one word at a time, they can only look “backward” at the text that they have already produced (along with the user’s original prompt).[73] Such systems currently cannot see “forward”—they do not know what sentences they will ultimately produce in the later, not yet created portions of their ultimate response (although this limitation is likely to be reduced in upcoming LLM systems). In other words, as they create a document word by word, they cannot look paragraphs ahead to see the future sentences that will appear in the final document or answer; all they can see is the last word and previous words that they produced. This limitation in GPT systems being able to look backward but not forward is very important to understanding how to use LLM capably—a point discussed below.

[72] *See* Xiang Li, Haoran Tang, Siyu Chen, Ziwei Wang, Anurag Maravi & Marcin Abram, Context Matters: Data-Efficient Augmentation of Large Language Models for Scientific Applications (Dec. 22, 2023) (unpublished manuscript), https://arxiv.org/pdf/2312.07069v2.pdf [https://perma.cc/7HA3-Z2AT].

[73] *See id.*

C. The Meaning of "GPT" in GPT-4 and ChatGPT

ChatGPT can produce rather sophisticated answers to questions, create complex documents, and generate human-like text. But how does it do such intelligent-seeming work, when all it does is predict the next most likely word, given what it has been asked and the text that it has already produced itself? To understand how this remarkable prediction works, it is helpful to explore the underlying GPT architecture. One helpful place to start is by examining a foundational question: what do the letters "GPT" in GPT and ChatGPT stand for? The answer to this question will help us understand how GPT achieves its sophisticated, intelligent-seeming output, merely by next word prediction.

II. GPT: GENERATIVE PRETRAINED TRANSFORMER

The letters "GPT" in ChatGPT stand for "Generative Pretrained Transformer."[74] Let's break down each of these words individually.

A. "Generative"

The phrase "generative artificial intelligence" is an umbrella term that refers to the use of AI systems in areas that are normally associated with human creativity, such as human language understanding and generation or the creation of music, art, and video.[75] The term "generative" is used to distinguish systems like ChatGPT that create human-like text from other AI systems that are focused on issues outside of human creative contexts—for example, AI systems that are designed to predict future events, classify documents or images, control self-driving cars, spot credit card fraud, or identify consumer preferences.[76] Thus, the term "generative" in GPT emphasizes that the purpose of the system is focused on an activity normally associated with human creativity: the generation (and understanding) of ordinary, written language.[77]

B. "Pretrained"

The term "Pretrained" in GPT is one of the most important concepts to grasp in order to understand how such an AI system actually produces its outputs. "Pretraining" refers to the computer science process of exposing an NLP system like GPT-4 to enormous amounts of written documents, including a substantial portion of the internet, as well as many books and

[74] *See* Wu et al., *supra* note 15, at 1123.

[75] Jennifer Haase & Paul H.P. Hanel, *Artificial Muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity*, J. CREATIVITY, Dec. 2023, at 1, 1–2.

[76] *Id.*

[77] Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, Donguk Kim, Sung-Ho Bae, Lik-Hang Lee, Yang Yang, Heng Tao Shen, In So Kweon & Choong Seon Hong, A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? (Mar. 21, 2023) (unpublished manuscript), https://arxiv.org/abs/2303.11717 [https://perma.cc/7BCT-UWNV].

other references, in order to teach it the general process of generating human-like language. To help understand this, it is useful to briefly examine the AI concepts of "machine learning" and "self-supervised" learning.[78]

"Machine learning" is a subfield of AI that is focused on developing algorithms that can learn useful statistical patterns from data.[79] Once a machine-learning AI algorithm has identified patterns from a dataset, these learned patterns can then be used for tasks like prediction or automation going forward. A familiar example of machine learning, outside of the GPT context, involves "spam" email detection. These algorithms examine large amounts of emails that have been categorized (or "labeled") as "spam" or "wanted" by users.[80] They learn from these examples statistically relevant telltale indicators of spam, such as phrases like "send cash." After examining thousands of emails that have been labeled as "spam" and "wanted," a machine learning algorithm might learn that a phrase like "free money" in an email is significantly more likely to signal that the email is spam rather than wanted. From detected patterns like this, such machine-learning AI systems very reliably learn how to filter out future spam emails. The key idea in machine learning is that the algorithm itself learns the relevant pattern by examining large amounts of past example data and identifying statistically relevant indicators, as opposed to having rules (e.g., if "free-money," then "spam") manually and laboriously entered by programmers.

The process of "pretraining" in GPT is a form of machine-learning, and it is similar in concept, although different in form, from the email process just described. Pretraining involves exposing GPT's machine learning algorithm to language patterns that already exist in webpages, books, and other written texts.[81] Implicit in written documents like Wikipedia articles, contracts, blog posts, and emails are the rules of grammar, punction, semantics, and syntax that humans can naturally understand. By exposing GPT systems to large amounts of text data during pretraining, the system begins to learn both patterns of human language—including grammar, punctuation, and words that are commonly related to one another—and facts about the world.[82]

This pretraining approach, known as "self-supervised" training, was one of the big breakthroughs that has led to our current era of AI performance. As described, machine-learning systems learn from examples, such as analyzing patterns in examples of spam emails and wanted emails. Machine-learning systems in the past had to have their examples manually annotated by people, which made such "labeled" data time-consuming.[83]

For example, machine-learning spam detection systems learn from the manual flagging of millions of email systems when users hit the "spam" button upon receiving a spam, thereby labeling the spam email as a

[78] *Id.*

[79] IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING (2016).

[80] CHIP HUYEN, DESIGNING MACHINE LEARNING SYSTEMS: AN ITERATIVE PROCESS FOR PRODUCTION-READY APPLICATIONS 104–05 (2022).

[81] *See* Zhang, *supra* note 77, at 12.

[82] *See id.*

[83] *See id.*

human-verified example of spam that the system can examine. This process of an algorithm learning from examples which have been (mostly) manually annotated is known as supervised learning because the humans are supervising the algorithm by giving it examples that a person has verified as "spam" by taking a definitive action such as hitting a "spam" button. From these human-supervised and verified examples, machine-learning algorithms can then learn associated patterns between words and labels.[84] In the past, a major bottleneck was that human-generated, manually curated text data from which NLP systems would be able to learn language patterns in a supervised manner would have been typically limited and quite expensive to generate.[85]

In the early 2010s, clever researchers realized something fundamental about existing text documents, such as Wikipedia pages, contracts, articles, and emails: each written document *implicitly* has its own approximately correct "label" inside each written sentence.[86] Every time a human writes a grammatically correct sentence, such as, "The capital of France is Paris," they are implicitly providing an example of useful, labeled language that can be analyzed for patterns.

In other words, these researchers realized that there was already a vast amount of *implicitly* verified language data—existing web pages written by people—whose significance earlier scientists had partially overlooked.[87] This removed the need to create new and costly hand-labeled machine-learning data. By taking advantage of the trillions of "correct" examples that people had created as a by-product of writing web pages and other documents, GPT systems could be trained simply by repeatedly analyzing one document after another—learning, for instance, that words like "dog" and "barks" often follow each other. Thus, self-supervised pre-training involves exposing a GPT system to large amounts of text documents so that the system can learn statistically what words are likely to follow other words, given the billions of "verified" examples provided in existing, human-written webpages, books, and other documents.

How is the system "'trained"' to better predict words given the prompt and what it has produced already? Internally, GPT systems are just a series of billions or trillions of numbers, known as parameters. These parameters are what guides the system's predictions to select one word ("Paris") versus others ("tree" or "zebra") among its over 50,000-word vocabulary.[88] Before being pretrained, the parameters are random, so a GPT system at this early stage would be extremely bad at producing coherent text or answering questions. If given the prompt "What is the capital of Paris?" at this point, overwhelmingly it would predict a nonresponsive next word, such as "tree." Rather, a system such as GPT must be taught how to respond in a sensible

[84] *See* CHIP HUYEN, *supra* note 80, at 96.
[85] *See generally* BIRD ET AL., *supra* note 20; *see also* Zhang, *supra* note 77.
[86] BIRD ET AL., *supra* note 20.
[87] *Id.*
[88] *See* FOSTER, *supra* note 9, at 311–12.

and syntactically correct manner, through the process of repeatedly examining large amounts of existing text written by people.[89]

The pretraining process essentially involves GPT repeatedly being given documents, such as Wikipedia articles, with individual words intentionally "hidden" from the algorithm. GPT then tries to predict what the missing word is based on the previous context words in the sentence. When the prediction is incorrect, it adjusts its internal parameter numbers to increase the accuracy in future predictions. For example, if the Wikipedia page about France contains the phrase, "The capital of France is," the GPT algorithm will make its best guess (without looking at the hidden word) as to what the next word in the sentence is likely to be, given the previous context words in the sentence and the current numerical values of its billions of parameters.

Imagine that, after running the phrase "The capital of France is" through its internal algorithm, the system incorrectly guessed that the next word is likely "tree." Because the system can now access the full text of the Wikipedia document, it actually has the "correct" answer for the hidden word it attempted to guess. For training purposes, the "correct" word is simply the next word, now unhidden, that the human author originally wrote in the document. For example, if the document reads, "The capital of France is Paris," and the word "Paris" was hidden, the system would now know that "Paris" should have been the correct word to predict. However, in this example, observe that the system's current parameters guided it to a wrong next-word prediction, "tree." Its internal numbers must therefore be changed to make the system more likely to produce a more relevant word next time. This is how the system "learns" over time to be better at producing more responsive and human-like language.

The fact that GPT can, during training, look up the actual, correct next word in the document after having previously guessed is the essence of self-supervising learning. Every written webpage is implicitly labeled with thousands of correct answers: whatever the next word the author of the page actually chose. Thus, if the author wrote "The capital of France is *Paris*," then this labels the word ("Paris") as approximately "correct" given the previous context words "The capital of France is," at least compared to most of the other 50,000 possible but irrelevant "incorrect" words in the vocabulary such as "tree," "attorney," or "zebra." The system can automatically teach itself how to improve—without human intervention—by comparing its incorrect guess ("tree") to the correct word provided within the document ("Paris"). From there it can incrementally "learn" a useful pattern—that "Paris" is somewhat more likely to follow previous context words such as "The capital of France is."[90]

The model learns by adjusting its parameters over and over again for each word in a document every time it makes a wrong guess. Using mathematical

[89] *See, e.g.,* Katikapalli Subramanyam Kalyan, *A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4*, NAT. LANGUAGE PROCESSING J., Mar. 2024, at 1.

[90] *See* Zhang, *supra* note 77, at 10–15.

techniques,[91] the AI system can look internally among its billions of parameters and determine which of its internal numbers had incorrectly guided it to guess the word "tree" given an input like "The capital of France is." Once it identifies those parameters that led it to predict the wrong word in light of that context, it can incrementally demote the weights of those specific parameters, diminishing their influence in subsequent predictions around similar context words like "capital" and "France." Additionally, since it now knows that the right answer should have been "Paris" given the prompt "The capital of France is," the model can also identify those internal parameters that would have been more likely to have guided it to the correct answer given that context, "Paris," and effectively promote those specific parameters, increasing their influence on subsequent predictions involving similar context words like "France" and "capital."

Thus, during training, as GPT analyzes the sentences of billions of books, web pages, and other documents, it continually aims to guess the next word in existing sentences, word by word, until it reaches the end of that document. Each new word guess is an opportunity to teach the system how to produce better and more accurate outputs by tuning the numbers that would have led to the correct predicted word, by numerically increasing their impact, and by numerically reducing the impact of any parameters that led to the wrong word prediction, given that context. After having gone through such a single learning update, and given its newly updated numerical parameters, the model would be incrementally more likely to predict the word "Paris" after seeing the phrase "The capital of France is" than it had been previously.[92]

In sum, "pretraining" refers to the process of teaching an AI model how to understand and generate human-like text by exposing the model to a large portion of the internet (e.g., sources like Wikipedia and Reddit), newspapers, research papers, and books such as textbooks or works of fiction. After being pretrained on billions of webpages, books, contracts, legal opinions, and other text documents, the AI system's billions of parameters have become appropriately adjusted to reliably predict the next words, given nearly any selection of prompting words.[93] This is because, in such a large selection of billions of documents, just about any topic has been previously written about multiple times, and an LLM system will be exposed to and learn the topical, linguistic, and formatting patterns within.

Stepping back for a moment, one can see the reason that modern AI NLP systems, such as ChatGPT, are referred to as "large language models." This is primarily because they have been trained with so many documents and contain such a huge number of internal mathematical parameters that are capable of learning immense numbers of patterns and guiding the prediction of the next relevant word given the previous words in the context. This is important to note, because many of the improvements that have occurred in

[91] The main mathematical techniques are known as "gradient descent" and "back-propagation."

[92] *See* Zhang, *supra* note 77, at 10–15.

[93] *See* Wu et al., *supra* note 15, at 1131.

AI systems like GPT have simply been due to continual increases in the size and scale of these systems. As described, there have been several earlier variants of GPT systems, GPT-1 (2018), GPT-2 (2019), GPT-3 (2020), GPT-3.5 (2022), and GPT-4 (2023).[94] With each new release, the system grew much bigger in terms of the numbers of parameters; further, each time the scale has increased, it has led to increased capabilities and, in some cases, new unexpected emergent properties like reasoning and problem-solving.[95]

Of course, increase in size is not the only reason that AI LLM systems have been getting better. There have also been significant engineering, algorithmic, training, and hardware advances that have contributed to the increase in capabilities.[96] But it is worth emphasizing that simply increasing the size of the system is one factor that has consistently led to improvements in quality of outputs. Thus, as future LLM systems increase in scale beyond GPT-4, they are likely to improve in their abilities as well.

It is important to note that training large scale, state-of-the-art LLMs like GPT-4 is today typically very expensive and slow. Because pretraining LLMs involves analyzing billions or trillions of words of text documents for patterns, one word at a time, the process often takes months, requires huge amounts of computers and graphical processing units, and results in expenses of hundreds of millions of dollars. For this reason, pretraining highly capable large language models has thus far been limited to a few, well-resourced corporations, such as Google, OpenAI, and Meta.

Nonetheless, it is fascinating that pretraining a system like GPT-4 on a large body of webpages and books can ultimately teach it to produce coherent, human-like text. What is more remarkable is that such a system, designed only to produce human-like outputs, word by word, also developed the ability to engage in problem-solving and reasoning, almost as a by-product of learning to reliably predict the next most likely word.

C. "Transformer"

The final word in GPT, and among the most important innovations, is the term "transformer." The term "transformer" refers to a deep-learning AI architecture developed by Google in a 2017 breakthrough paper entitled, "Attention Is All You Need."[97] This new approach uses what is known as "self-attention," a mechanism that has been fundamental to the advances in NLP, such as those demonstrated by GPT-4 and other state-of-the-art AI models. GPT-4 and other GPT models are built on the transformer architecture.[98]

[94] *See id.*

[95] *Id.*

[96] *Id.*

[97] See Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin, Attention Is All You Need (June 19, 2017) (unpublished manuscript), https://arxiv.org/pdf/1706.03762v2.pdf [https://perma.cc/PNP6-PZHK].

[98] *See* Achaim et al., *supra* note 3.

To understand the transformer, it is helpful to first understand "deep learning" and "neural networks." GPT-4 and other modern LLM systems fundamentally use an algorithmic approach to machine learning known as a "neural network" (although it is worth noting that there are many other machine-learning approaches besides neural networks). Specifically, neural networks are AI computational models that take loose inspiration from the human brain (although they operate quite differently). They process data using a system of interconnected "nodes" or "neurons."[99] The billions of "parameters" mentioned earlier that guide ChatGPT to predict one word over another are essentially the strength of connections (or "weights") between one numerical node and another.

In essence, neural networks are just extremely flexible pattern detectors: they can approximate just about any underlying pattern, if given enough data to examine. Neural-network approaches to AI are notable for their ability to process data in a layered manner, one layer after the other, with each subsequent layer able to add different, higher levels of abstractions on top of the previous layer.[100] Because of this layering ability, neural-network approaches stand out from other machine-learning techniques for their capability to represent high-level conceptual abstractions that help in modeling the complexity of human language and thought.[101]

Within neural networks, there is an approach known as "deep learning," which essentially involves taking ordinary neural networks and scaling them up to huge sizes, resulting in neutral networks with billions of "neurons" or parameters with many "deep" layers. Such large deep-learning systems have proven to be extremely effective at AI tasks but have become practicable only in the last decade, thanks to algorithmic advances and advances in hardware engineering.[102] Google's transformer is one particular deep-learning neural-network design that stands out as an exceptionally successful architecture.

Prior to 2017 and the invention of the transformer design, there were two significant bottlenecks holding back LLMs. First was the ability for existing NLP systems to analyze contextual data in a prompt, outside of a narrow range.[103] Crucially, in order for an NLP to properly respond to a user's question or instruction in a prompt, it has to understand the *context* of what is being asked by the user.[104] For example, if an LLM system is asked "Can you tell me about apples?," it has to be able to determine the context—if the user is talking about the fruit or the computer/electronics devices (e.g., Apple iPhone).

[99] Rene Y. Choi, Aaron S. Coyner, Jayashree Kalpathy-Cramer, Michael F. Chiang & J. Peter Campbell, *Introduction to Machine Learning, Neural Networks, and Deep Learning*, TRANSLATIONAL VISION SCI. & TECH., Jan. 2020, at 1, 14.

[100] *Id.* at 7.

[101] *See* PEDRO CUENCA, APOLINARIO PASSOS, OMAR SANSEVIERO & JONATHAN WHITAKER, HANDS-ON GENERATIVE AI WITH TRANSFORMERS AND DIFFUSION MODELS (2024).

[102] *Id.*

[103] *See* Wu et al., *supra* note 15, at 1125–26.

[104] *Id.*

Crucially, clues from what a user actually means are contained in the input *context*—the surrounding explicit and implicit text that a user provides. So, for example, if a user asks, "Can you tell me about apples?," but in the preceding sentence had said, "I am looking for a new phone," an LLM must be able to tell, by context, that the user was asking about electronics. Conversely, if the user had previously said instead, "I am writing a paper about popular fruits," the LLM must be able tell, from the input *context*, that the user was referring to the fruit. Problematically, most LLM approaches prior to 2017 had difficulty analyzing context words necessary to make the full meaning understandable, and this was particularly true as the relevant context words got further and further away from one another in a user's prompt.[105] For example, if a user wrote, "I am writing a paper about popular fruits," and then included several other sentences of other text in between in the input before saying, "Can you tell me about apples," the text distance between "fruits" at the beginning of the input and "apples" at the end—known as a long-range dependency—proved a technical challenge for NLP systems at the time.[106]

The second major challenge for architectures at the time had to do with computational inefficiency. Recall that pretraining required an LLM system to be trained on billions of pages of written text.[107] However, a problem plaguing the architectures before the transformer was that such training could only be conducted very slowly, given the techniques at the time. Thus, the earlier techniques at the time had difficulty processing at a reasonable rate the billions of pages of written text necessary to create high-quality NLP systems.[108]

Google's transformer architecture solved both of these issues. First, the "transformer" architecture allowed LLMs like GPT to look at the entire context of the user input, even words that were far away, and determine which contextual words were most helpful in figuring out the more accurate next word.[109] For instance, if a user wrote, "I am interested in learning what the capital of France is. I have never been to France before. Can you tell me?," the transformer architecture made it so that the AI system could learn to focus on the important words in the prompt, such as "capital" and "France" and "tell me," while ignoring the other information that was less relevant to the question. This ability to take into account the crucial context words in a user's prompt, even among words that were located distantly from one another in the input, proved crucial to huge improvements in AI capabilities in being able to respond sensibly.[110] LLMs such as GPT-4, built using the transformer architecture, could learn the ability to focus on relevant context words in a prompt to help it produce more relevant answers, and this capability is known as "self-attention" or simply "attention." This

[105] *Id.*

[106] *See generally* FOSTER, *supra* note 9.

[107] *Id.*

[108] *See* OZDEMIR, *supra* note 70.

[109] *Id.*

[110] *Id.*

self-attention, which provided an efficient and effective mechanism for an LLM to consider all the words provided by the user in order to figure out what the user was actually asking from the context, represented a major breakthrough.

Second, the transformer architecture also helped solve the computational bottleneck, which had prevented training on large amounts of data. Due to clever design, the attention mechanism allowed the system to analyze large chunks of data in *parallel*, rather than one word at a time, leading to parallel processing of data and huge efficiencies.[111] It was these huge architectural efficiencies, in combination with advances in hardware, that partially allowed for language models to become so large and, in turn, display unprecedented language generation abilities compared to earlier NLP systems.[112]

D. Word Vectors

The transformer also incorporated another key, but earlier, innovation in NLP: "word vectors" or "word embeddings." Word vectors are a way of representing the meaning of words using lists of numbers.[113] This was a huge breakthrough in NLP. Previously, NLP systems had largely dealt with written text as mere data patterns, but such systems could not understand the *meaning* of the words that they analyzed.[114] For the first time, word vectors allowed the relative meaning of words to be reliably encoded in lists of numbers, with words of similar meaning numerically grouped together. Related words like "dog" and "bark," "Paris" and "France," and "Apple" and "electronics" could be numerically linked to one another. Moreover, these semantic relationships between words could be learned automatically during training, across nearly any context, rather than being manually curated by people. The fact that the meaning of words could be represented numerically meant that mathematical systems, such as deep-learning neural networks, could reliably process them. For the first time, systems could produce outputs that took into account the underlying meaning of the word (e.g., "Apple" means electronics), rather than just seeing a word as a meaningless pattern of letters as NLP systems had previously (e.g., "A-p-p-l-e").

The transformer model operated by taking a user prompt sentence such as, "I am looking for a new phone. Can you tell me about apples," and converting each of those words to word vectors. Using the "attention mechanism" of the transformer, LLMs such as GPT mathematically nudge the numerical values of the different word vectors so that they encode the context of the words around them.[115] So, if the input sentence enters GPT-4,

[111] *See* FOSTER, *supra* note 9.

[112] *Id.*

[113] Tomas Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space (Sept. 7, 2013) (unpublished manuscript), https://arxiv.org/pdf/1301.3781.pdf [https://perma.cc/H79C-JJKC].

[114] *See generally* OZDEMIR, *supra* note 70.

[115] *See* Zhao et al., *supra* note 32, at 4.

it will be processed by its transformer-based neural net, and the sentence will be mathematically nudged in the direction of the word "phone" so that the AI model knows to be more likely to output related words like "electronics," "cellular," or "Android," rather than fruit-related words like "banana," "peel," or "eat." This was one of the huge breakthroughs that allowed LLMs to understand the meaning of words ("Governing law is California"), and not just to see words as data patterns (e.g., "g-o-v-e-r-n-i-n-g").

E. Instruction Tuning and Reinforcement Learning

In building the GPT family of LLMs, including GPT-4, OpenAI capitalized on Google's transformer and word-vector inventions. However, OpenAI made many significant engineering advances of their own to produce a system as capable as GPT-3.5 and then GPT-4.

Two advances are worth describing. The first is known as "Instruction Fine Tuning."[116] Earlier GPT models, like GPT-3 (released in 2020), were surprisingly good at producing human-like text, such as poems or news articles, when given a prompt.[117] This text-generation ability was impressive for the time. But few thought of this as useful, general AI, as earlier systems such as GPT-3 routinely responded with nonsensical answers to simple questions or instructions and had difficulty including basic facts or reasoning.[118] For this reason, given the state of the art in 2020–2022, many thought that AI systems that could respond to any arbitrary input, and could reason, were far off.[119]

But the engineers at OpenAI had an insight: what if models like GPT-3, which had the core ability to generate human-like language, could be coaxed into following instructions and producing more responsive, useful answers? To do this, they pursued an approach known as "Instruction Fine Tuning."[120] As mentioned, LLM systems today undergo a huge pretraining process, which involves billions of documents, often takes weeks or months, and is quite expensive.[121] However, once pretrained, an AI model might have a reasonably good grasp on human language generation but not be able to follow instructions or answer questions.[122] What if you could simply nudge this general model in the right direction so that it could harness its own

[116] *See* Wu et al., *supra* note 15, at 1126.

[117] *See* Hodge, *supra* note 12, at 225.

[118] *See* Wu et al., *supra* note 15, at 1122–26.

[119] *See* Bai et al., *supra* note 45.

[120] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai & Quoc V. Le, Finetuned Language Models Are Zero-Shot Learners (Feb. 8, 2022) (unpublished manuscript), https://arxiv.org/pdf/2109.01652.pdf [https://perma.cc/9C8K-N8BD]; Maarten Bosma & Jason Wei, *Introducing FLAN: More Generalizable Language Models with Instruction Fine-Tuning*, GOOGLE RESEARCH: BLOG (Oct. 6, 2021), https://blog.research.google/2021/10/introducing-flan-more-generalizable.html [https://perma.cc/EH36-HJWT].

[121] *See* Andrew M. Dai & Quoc V. Le, Semi-supervised Sequence Learning 7 (Nov. 4, 2015) (unpublished manuscript), https://arxiv.org/pdf/1511.01432.pdf [https://perma.cc/4BV-SV8Y].

[122] *See id.*

general language abilities to be more useful? That is the process known as "fine-tuning," which involves a much more limited but focused training of an LLM that has *already* been pretrained, using a smaller yet higher quality and focused dataset. This additional fine-tuning, on top of an existing pretrained model, is designed to elicit certain abilities that exist in the AI system but are latent and difficult to develop.[123]

To make GPT-3 a more capable AI system, OpenAI decided to try to fine-tune its GPT-3 model in two ways. First, it created a dataset involving thousands of questions and instructions, paired with thousands of responsive and reliable answers to those questions and instructions.[124] To do this, OpenAI hired hundreds of contractors who laboriously created meticulous answers to questions.[125] Then, OpenAI further trained GPT-3 on this much smaller set of gold-standard question and answer pairs, to try to teach it to learn further how to follow instructions and respond sensibly to questions.[126] Incredibly, this approach turned out to work, as ChatGPT (which was built on GPT-3), after receiving these additional instruction-following examples, learned how to produce responsive (rather than nonsensical answers) to nearly any user question, instruction, or prompt.[127]

The other major innovation from OpenAI in creating ChatGPT came from an approach that they helped develop known as Reinforcement Learning from Human Feedback (RLHF).[128] The gist of RLHF is to incorporate direct human input into the learning process, allowing the model to improve based on human evaluations of its performance.[129] The process works by having the AI model produce multiple possible answers to a given prompt and having human raters pick the best and worst among the different versions.[130] After doing this repeatedly, OpenAI created a separate AI system "reward model" that was capable, based on examining the data from human ratings, to begin to automatically rate different output versions as good or bad by itself.[131] This other AI reward model was then used to further fine-tune GPT-3 to produce outputs that humans were likely to prefer. These two innovations by OpenAI helped make the original ChatGPT (released in November 2022) much more effective and capable of producing results that were much more intelligent-seeming than any previous NLP system. And it was this release that ushered in the modern state of advancements in AI NLP, reasoning, and understanding that exists as of early 2024.

[123] *See id.* at 3.
[124] *See* Ouyang et al., *supra* note 1.
[125] *Id.*
[126] *Id.*
[127] *Id.*
[128] *Id.*
[129] *Id.*
[130] *Id.*
[131] *Id.*

III. LLMs IN LAW AND THEIR LIMITATIONS

Today, as mentioned, the state-of-the-art LLM systems like GPT-4 are capable of tasks such as: (1) producing reasonable, draft-quality legal documents, such as contracts or motions; (2) performing basic legal analysis; and (3) answering questions about legal documents, such as statutes, legal opinions, or contracts.[132]

However, despite that promise, it is important to emphasize that there are currently some significant issues that limit LLMs' use within law unless done with extreme care. For one, many uses of modern AI systems like ChatGPT require inputting documents or background information into the prompt or uploading relevant documents. For instance, if one wanted to create a draft merger contract using GPT-4, one would need to input background information about the deal so that the LLM could have the appropriate context to create a useful document. However, attorneys must be very careful to avoid inputting private, privileged, or sensitive information, unless an AI organization expressly provides enterprise-grade security and privacy guarantees. Numerous AI systems available to consumers, such as the free version of GPT-3.5, indicate that information contributed by users could be integrated into the system for ongoing AI training. Thus, without express guarantees from an AI provider, private client data or privileged information uploaded directly into the prompt of an LLM like ChatGPT might be incorporated into the future knowledge base of the AI system, potentially violating client confidentiality or privilege or exposing sensitive client information. Additionally, many such consumer-grade systems expressly indicate that human reviewers might review user-uploaded content. By contrast, many enterprise-grade systems that incorporate LLMs like GPT-4 in the backend indirectly, such as Lexis+ AI and Westlaw CoCounsel, do provide guarantees of security and privacy, assuring that humans will not see uploaded input and that AI systems will not be trained on any user-uploaded data.

Second, attorneys might be aware of limitations as to the quality of the AI outputs. Although much better than systems of the recent past, current AI systems are not as reliable as one might want them to be, especially if not used appropriately. For one, such systems occasionally produce inaccurate but plausible-seeming facts or references, such as nonexistent legal case names or fake academic articles.[133] There have been numerous incidents of attorneys submitting ChatGPT-generated court documents with reasonable-sounding but nonexistent case law citations. Second, sometimes the models have flaws in their analysis or reasoning, leading to incorrect conclusions despite the premises, or make basic common-sense errors.[134] For this reason, current models should be treated with care and treated as *draft* tools, rather than final product generators. Users must carefully double-check the outputs to ensure that they are correct. A helpful analogy

[132] *See* Schwarcz & Choi, *supra* note 6.

[133] *See* Ouyang et al., *supra* note 1.

[134] *Id.*

might be to think of work produced by GPT-4 as akin to a preliminary draft from a diligent yet inexperienced third-year law student intern—it provides a useful foundation that can cut down on initial drafting time but requires a rigorous vetting process for accuracy, logic, factuality, and coherence before it can ultimately be presented to a client or a court.

There are, however, reasons to believe that some of these reliability problems will be reduced, if not completely eliminated, in the near future. One promising approach to improving reliability is known as prompt augmentation. Prompt augmentation involves not just asking the AI model a basic question in a prompt, but rather including in the prompt information that is likely helpful to the model in determining a correct answer.[135]

Importantly, if one augments the prompt with useful, relevant information, systems are much more likely to produce accurate results to factual or analytical questions as compared to simply asking the model a question. Augmenting a prompt is the idea of adding extra information before or after the prompt that GPT-4 can look at when it is deciding to predict its response. Contrast this with a traditional, non-augmented prompt, which might be "What are the factors of fair use in copyright?" This is the typical way most users interact with ChatGPT, by just asking a question or giving an instruction without additional relevant information. If you are asking in this manner, you are having ChatGPT respond by "looking" internally into its own compressed "database" of information. By having the model look internally, it is more likely to hallucinate.[136]

Imagine instead, however, that you included in the prompt information that the model could use to give you a correct answer. For instance, consider that in your prompt (before or after), you had copied and pasted segments from an authoritative treatise about copyright, along with your original question about copyright. In doing so, you have effectively augmented the prompt with additional useful information. Such augmentation makes it vastly more likely that the model will respond accurately than if you had simply asked the same prompt.[137]

The reason that this improvement occurs can be understood if we look back to our earlier discussion of the transformer "self-attention" mechanism. Recall that this mechanism allows AI to incorporate relevant context. By giving the model a document that you know, somewhere within, likely contains authoritative information that will be helpful to answering your question, you are giving the model the *context* it needs to produce a more accurate answer. In short, an AI model like GPT-4 will be able to look backward at the informative additional context that you provided in the prompt, will consider this in light of your question, and then will be much more likely to predict the sequence of relevant words that is accurate and responsive to your question.

[135] *See* Zhao et al., *supra* note 32.

[136] *Id.*

[137] *Id.*

For this reason, one area in which current LLMs like GPT-4 excel is in document summarization and analysis. If one were to input the text of a legal opinion, court document, or law review article directly into the prompt and ask GPT-4 about the inputted text, it is generally very reliable and useful. The major limitation to be aware of is the "context window." In short, the context window refers to the size of the text that an LLM like GPT-4 can reliably examine in whole. For instance, initially LLMs like GPT-4 could only examine documents inputted in a prompt that were about 2,500 words, or about ten pages of text. More recently, text content windows have expanded to approximately 8,000 words, or about thirty pages of text. However, if one tries to input a document for summarization that is longer than the context window, the LLM may either produce unreliable results because it is not able to examine the entirety of the document with its self-attention or might refuse to analyze it at all. However, if one inputs a document into the prompt that is shorter than the maximum context window, advanced LLMs like GPT-4 are quite reliable at summarizing and answering questions about the document.

The reason that LLMs like GPT can only analyze text of certain, limited length has to do with computation—each additional word that a system like GPT-4 adds to its context window requires additional computer processing equal to the square of that one word. Thus, to have ChatGPT increase its context window by just 1,000 additional words, from 8,000 to 9,000 words (allowing it to look at documents about ten pages longer than before), it would require approximately a million times (1,000 x 1,000) more computation. However, due to technical advances, many of the current limitations in context window length are likely to improve in the near future. Google has released experimental models like Gemini Pro 1.5 that have reliable context windows of one million words or more, due to technological improvements that have reduce the computation necessary to do so.[138]

In addition to users manually adding relevant documents to the prompt, such prompt augmentation is starting to happen automatically. Many current systems employ techniques such as "retrieval-augmented generation" (RAG) that automatically add relevant context to the prompt context window for the LLM so that it can produce more reliable results. Such systems typically examine a user query; reach out to reliable databases such as Google, Westlaw, or LexisNexis; gather documents likely to be relevant to the prompt; and then automatically "augment the prompt" on behalf of the user.[139] In other words, these RAG systems do not rely on a user with a copyright question to manually retrieve a page with copyright facts and then manually cut and paste that into the prompt before asking. Rather, all of this processing and inclusion of information, which dramatically increase the accuracy and relevance of the response, is beginning to happen automatically, behind the scenes, and hidden from the user.[140] However,

[138] *See* Pichai & Hassabis, *supra* note 61.

[139] *See* Zhao et al., *supra* note 32, at 61.

[140] *See* Zhang et al., *supra* note 77.

because the relevant information is being included in the LLM's context window, it can use "self-attention" to determine the relevant information it has been provided, thereby making its answers much more reliable, than had the AI model simply been asked a legal question without providing it with relevant legal, contextual information.

More broadly, it has turned out that prompting itself can influence the quality of an AI output. For example, asking a system to first "list the facts relevant to solving a problem, and then solve the problem step by step," rather than asking it directly to solve a problem, turns out to improve accuracy dramatically. Again, in part, this has to do with context and the transformer self-attention mechanism. When one prompts an AI model to first list the relevant facts and considerations, one is essentially giving the model *more context* that it can consider. Recall that LLMs can only look backward at the text that they have already produced; they cannot look forward to predict the words several sentences in the future that they have not yet produced. By inducing the model to first output information that is likely to help itself, such as a list of facts that are likely relevant to the correct answer or a list of logical steps that are likely to lead to the right answer, the model can then look backward at the relevant information or process that it itself has produced. It can then leverage its self-attention mechanism to incorporate the information that it has produced to help itself, and it will subsequently produce a much more reliable answer, because it has been induced to augment itself with relevant factual or process information before answering. As described earlier, this technique works because text already produced by the model influences the probability of later words a model produces.[141] So prompting the model to first produce information likely to be useful to solve the problem before having the model actually produce a final result helps the model statistically nudge itself down the path to more likely get a correct answer.

In a similar manner to RAG, optimal prompting is also starting to be hidden from the user so that systems, behind the scenes, automatically choose optimal prompts without requiring that the user know the details of prompt mechanisms. As hardware becomes faster, more and more of these optimizations that today require manual input are likely to be performed automatically without the user's knowledge, and accuracy of outputs are likely to increase.[142]

Thus, although LLMs like GPT-4 have shown remarkable progress in their ability to perform certain discrete legal tasks compared to prior technologies, caution is warranted. This technology is still quite new, and attorneys must be aware of the underlying technological and architectural limitations if using them. However, it is likely that, in upcoming years, some of these limitations will be reduced by hardware and software advances. It is generally better to use legal-specialized systems built on top of advanced AI systems like GPT-4 for legal work, rather than using LLMs directly. Specialized legal AI systems, such as Lexis+ AI or Westlaw CoCounsel, are designed to insulate

[141] *See* Zhao et al., *supra* note 32, at 1.

[142] *See id.* at 82.

attorneys from having to be aware of the underlying technological details of systems like GPT-4 and to provide security, privacy, and training guarantees.

However, it is important to know that even these specialized legal systems still today struggle with similar accuracy issues mentioned above that are endemic to even the most advanced AI LLM models like GPT-4, so extreme care and rigorous verification of output is still necessary. Nonetheless, in recent years, AI has shown remarkable and unexpected progress in finally providing useful tools around the understanding, analysis, and creation of natural language legal documents, such as contracts, legal motions, patents, statutes, and legal opinions.

CONCLUSION

The rapid advancements in AI, exemplified by LLMs like GPT-4, signal significant changes in law and the legal profession. As these systems grow more sophisticated, they are likely to reshape how legal work is done and require a thoughtful approach to ensure that they are responsibly and ethically used. Although these AI NLP tools have gained unprecedented abilities in understanding, analyzing, and creating draft legal documents compared to just a few years ago, it is important for legal practitioners to use these tools carefully and responsibly. These systems can sometimes inadvertently perpetuate biases from their training data. Moreover, the inner workings of these systems are sometimes opaque, making it difficult to understand why they produced the results or text that they did. This lack of transparency and interpretability can raise concerns about accountability and trust in legal decision-making.

To ensure the appropriate use of LLMs in law, it is thus important for users to develop a basic understanding of their inner workings, as well as their capabilities and limitations. For this reason, this Essay has aimed to provide an approachable explanation as to how they work. However, while taking these concerns seriously, if used appropriately, the remarkable AI advances of the past few years also raise the hope of enhancing the practice of law and increasing access to justice.