ARTIFICIAL INTELLIGENCE AND LAW—AN OVERVIEW OF RECENT TECHNOLOGICAL CHANGES: KEYNOTE ADDRESS AT THE 2024 IRA C. ROTHGERBER JR. & SILICON FLATIRONS CONFERENCE ON ARTIFICIAL INTELLIGENCE AND CONSTITUTIONAL LAW

PROFESSOR HARRY SURDEN*

*Professor of Law, University of Colorado; Associate Director, Stanford Center for Legal Informatics (CodeX).

INTRODUCTION

It's truly an honor to be here today to give the keynote presentation at this conference on artificial intelligence and constitutional law. I am a Professor of Law here at the University of Colorado. I am also the Faculty Director of the Silicon Flatirons Artificial Intelligence and Law Initiative, and Associate Director of Stanford University's CodeX Center for Legal Informatics. Much of my scholarly work explores the intersection of artificial intelligence (AI) and law, drawing on my earlier experience as a professional software engineer prior to entering law.

I. OVERVIEW

Our discussion will begin with an overview of AI fundamentals and its historical development. Then, we'll survey the major changes in AI in 2022 involving large language models (LLM), such as GPT (the LLM used by ChatGPT), which have recently generated so much widespread excitement. We'll next look at the modern capabilities of these AI advances, examining the state of the art as it used in law today, while also highlighting its limits. Finally, we'll explore some of the implications for AI use in the context of constitutional law—the subject of this conference.

A major point will be that, although recent AI systems have shown remarkable capabilities in law in terms of tasks such as legal analysis, legal document generation, and so on, we must be very careful when employing them in certain critical legal contexts. These AI tools are now widely available to the public, including legal officials, such as judges who might be tempted to use these systems for complex and important legal tasks such as constitutional question analysis. Although this emerging AI is extremely powerful and can often provide coherent and capable legal analysis, the technology also has certain critical limitations that are not always obvious to end-users. Quite apart from well-known problems of accuracy, there are other subtleties in the underlying technology that are not immediately apparent, which I will discuss. If used carefully, these AI tools can be extremely helpful. However, many judges and lawyers may not yet possess the degree of AI literacy necessary to understand these technical nuances, which can have profound and

unexpected effects on the direction and substance of the outputs that these systems produce.

II. WHAT IS AI?

Because many in the audience do not have experience in AI, a review of basic concepts will help set the context. Let's begin with a basic question: What is "artificial intelligence"? There's probably no single, universally agreed-upon definition of artificial intelligence, but one definition that I find useful is:

> Artificial intelligence (AI) involves using computers to solve problems, make predictions, answer questions, generate creative output, or make automated decisions or actions on tasks that, when done by humans, are typically associated with "intelligence."[1]

This is a definition that many, but not all, AI researchers would subscribe to, in part because "intelligence" itself doesn't have a widely agreed-upon definition. But for our purposes, we can think of "intelligence" as any of a series of higher-order cognitive skills that are generally associated with human thinking. Such cognitive processes include planning, problem-solving, abstract reasoning, estimating, understanding and generating language, learning, visual-spatial understanding, and so forth.[2]

For example, consider that in the world, people routinely perform many complex activities, such as driving cars, playing chess, writing books, and solving math problems. Notably, to perform these tasks, people must engage multiple aspects of their brains' advanced cognitive processes, such as visual-spatial understanding when driving or language understanding when reading a book. Thus, when we take such a task that normally involves advanced cognition to do it, and we are able to get a

1. For a fuller explanation of this definition, see Harry Surden, *ChatGPT, AI Large Language Models, and Law*, 92 FORDHAM L. REV. 1941, 1944–45 (2024); P. M. Krafft et al., *Defining AI in Policy Versus Practice*, ARXIV (2019), http://arxiv.org/abs/1912.11095 [https://perma.cc/99RP-D3E2].

2. *See* STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 1–5 (4th ed. 2021) (discussing a variety of approaches to "acting rationally" and describing what it means for a machine to act like a human).

computer to perform that same activity, we can refer to it as an "artificial intelligence task."[3]

However, it is important to emphasize that the underlying processes by which computers perform these tasks *are very different* from human cognitive processes that people employ. The human brain is powered by extremely complex physical and biological processes that are not fully understood. By contrast, AI systems only approximate and simulate such intelligence processes using data, sensors, mathematics, computation, and statistics. This difference is crucial, because although modern AI models are able to produce very human-like outputs, we must take care not to anthropomorphize them and treat them as if they are actually human. Today, such confusion is more likely than ever given the recent and unprecedented ability of modern AI systems to simulate fluent human-level conversations and reasoning.

III. CHATGPT AND GPT-4

Figure 1. Example of a ChatGPT prompt

Let's look at an actual example of an AI task to make this discussion more concrete. Many of you might have seen this system on the screen—this is ChatGPT using GPT-4 (fig. 1). ChatGPT is an intuitive chat-based interface for interacting with the underlying AI technology powering this system that is known as "GPT." GPT stands for "generative pre-trained transformer," and we'll discuss this technology in more detail shortly. ChatGPT is an example of what is known as an AI large language model (LLM) system and was created by a company called OpenAI. The initial version was released as ChatGPT using the GPT-3.5 model in November 2022 to much acclaim, followed four months later by an even more powerful successor, GPT-4, which I am demonstrating today.

One interacts with the GPT model by typing in some text, known as a "prompt" into ChatGPT. A prompt is typically an instruction or a question to which the AI system responds. As we

3. *See id.* (discussing multiple frameworks under which AI performs a variety of actions similar to human cognition tasks). *See also* RONALD KNEUSEL, HOW AI WORKS ch. 3 (2024).

shall see, the specific words and information included by the user in the prompt are important, as the prompt sets the *context* for the system to calculate its response, and different word choices can sometimes lead the system to produce starkly different responses.[4]

In the example on the screen (fig. 1), my prompt asks ChatGPT to write a merger agreement—a legal contract—between two fictional companies. As you can see, the GPT-4o model in ChatGPT reads my prompt, processes it appropriately, and is almost instantly able to produce a comprehensive and reasonably capable first draft of a legal merger contract (fig. 2). This is by no means a fully complete and error-free contract, but it is not so different from a rough, first draft that might be produced by a junior attorney.

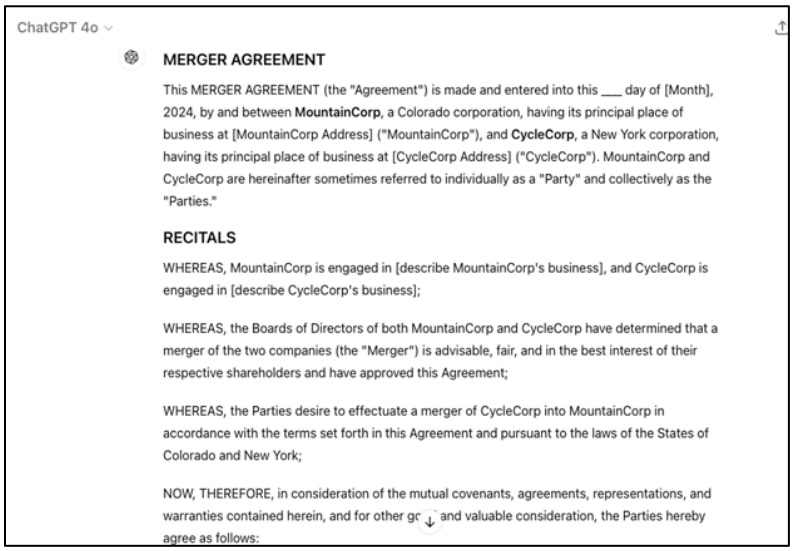

Figure 2. ChatGPT using GPT-4o (2024)[5]

Taking a step back, this is a helpful illustration under the earlier definition of AI as involving the automation of tasks normally associated with human intelligence. When a lawyer writes a merger agreement for a client, we generally think of that process as engaging various higher-order mental processes,

4. *See* Xiang Li et al., *Context Matters: Data-Efficient Augmentation of Large Language Models for Scientific Applications*, ARXIV (Feb. 27, 2024), https://ar5iv.labs.arxiv.org/html/2312.07069 [https://perma.cc/CGU8-D3PU] (discussing how prompt engineering influences outputs in an AI system).

5. GPT-4o is the updated version of GPT-4. *See infra* Section VI.E.

like planning, problem-solving, language understanding and creation, and abstract reasoning. Analogously, if an AI system can simulate the outputs of such tasks and automatically produce a draft merger agreement (or other fluently written documents) using statistical and mathematical mechanisms, we can categorize such examples as "artificial intelligence tasks."

As I will demonstrate, such AI computational abilities arise in part because during their development systems like ChatGPT were exposed to thousands or millions of examples of existing human-written merger agreements and other common documents. From these millions of existing, largely human-written examples, AI systems are able to identify core patterns that are common to various types of documents. They are able to leverage these previously identified patterns to later reproduce reasonable variations of documents, such as merger contracts, in response to a prompt requesting them.

IV. HOW DID WE GET HERE?

As someone who has been researching AI and law for twenty years, I find abilities like I have just demonstrated to be truly striking. It is hard to believe that today we have AI systems that can do a quite good, albeit *imperfect*, job at answering just about any question or topic you can ask or producing reasonably good first drafts of just about any legal document. Though perhaps not readily apparent to those who do not follow the field, it is important to emphasize that prior to 2022 such AI capabilities were not remotely possible at the current level of usefulness and sophistication that we see today.[6]

How did we get here at this moment in AI? How did we arrive in a world with general-purpose AI models like GPT-4 that can do a pretty good, even if imperfect, job at answering just about any question or topic, or that can produce arbitrary documents?

A. A Short Overview of AI Historical Development

To comprehend this, I think it's helpful to see the broader historical context of AI, which dates back to at least the 1950s (fig. 3). Broadly, we can divide AI into two major technical

6. Wayne Xin Zhao et al., *A Survey of Large Language Models*, ARXIV, 5–7, 10 (Mar. 31, 2023), http://arxiv.org/abs/2303.18223 [https://perma.cc/E4S8-338T].

approaches. It is worth gaining a high-level understanding of each approach and the differences between them to appreciate where we are with AI today.

B. Timeline of AI

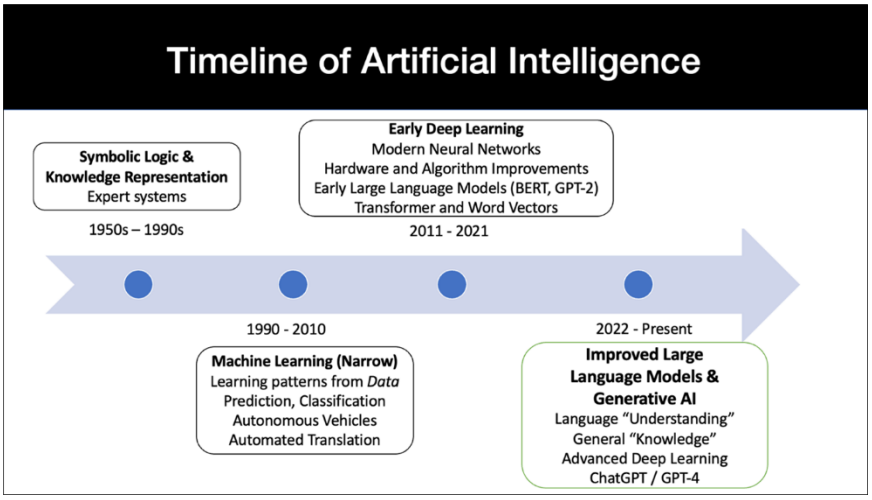

Figure 3. Timeline of AI

The first approach, which dominated AI from the 1950s through the 1990s, is known variously as "symbolic AI," "rules-based AI," or "knowledge representation."[7] The gist is that researchers initially built AI systems mostly by having computer scientists interact with domain experts like lawyers, engineers, or doctors to manually create computer rules that modeled aspects of the world that a system could mechanically process. For instance, people would create computer rules representing specialized areas of expertise such as tax law or medical diagnosis. The important thing to understand about this early symbolic method of AI was that it was a "top-down" approach, meaning that it involved people explicitly trying to understand some aspect of society or the world (e.g., law, medicine, or engineering) and then manually translating their analyses into a series of formal, mathematical rules that a computer could easily process. This symbolic AI process was

7. RUSSELL & NORVIG, *supra* note 2, at 16–23.

quite successful in certain narrow domains, such as income tax calculation, and is still widely used today.[8]

However, after some initial successes, symbolic approaches to AI began to show their limits in dealing with complex tasks that we associate with advanced human cognition. First, such manually created AI systems were often seen as "brittle," meaning that they had trouble dealing with new or unexpected situations the designers hadn't specifically thought up. Second, they were limited by the fact that some real-world processes were too complex to be broken down into simple rules, and the designers sometimes missed or struggled with representing these tasks symbolically. Thus, early symbolic-based AI systems did not fare well with many sophisticated tasks associated with intelligence, such as language understanding, robotics, and general problem-solving.

By contrast, there is a whole different approach to AI known as "machine learning," which has come to dominate the field. Machine learning takes a very different, "bottom-up" perspective. In machine learning, we provide pattern recognition algorithms with huge amounts of data about whatever phenomenon want the computer to model—whether that phenomenon is driving vehicles, medical diagnosis, predicting weather patterns, or email spam.[9] Importantly, in this approach, the algorithms themselves "learn" the relevant patterns and the rules from this data. So, machine learning is much less about top-down, manually-crafted expert rules, and much more about providing a machine learning algorithm with large amounts of relevant data and having the algorithm itself detect relevant rules using statistical patterns.[10] As will be discussed, AI models like ChatGPT are developed through machine-learning processes such as this.

The era of machine learning in AI began around 1990. This was largely driven by improved hardware, and the sudden availability of large amounts of data stored in digital form (as opposed to on paper) thanks to the widespread emergence of the Internet and more pervasive computing and sensors. Researchers realized that the combination of large datasets and

8. *Id.* at 22–23 (discussing the use of expert systems in medicine for certain type of diagnostic aid, and in law for aiding in areas such as income tax calculations).

9. *Id.* at 24–26.

10. *See* Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. UNIV. L. REV. 1305, 1311–16 (2019).

more powerful hardware allowed machine-learning, pattern-matching algorithms to address AI problems that were much more complex than was previously technologically possible.[11]

For example, researchers were able to, for the first time, create robust, autonomous, or "self-driving" vehicles, using machine-learning techniques. When building a self-driving car today, developers don't primarily produce long lists of specific computer rules, such as, "If you see a pedestrian, avoid it." That would be more of a manually-coded approach, more like the earlier symbolic or knowledge representation era. Driving is simply too complex to be captured by thousands of hand-programmed computer rules. Instead, modern self-driving systems learn through the machine-learning approach: They analyze millions of examples of data from real driving situations, recognizing patterns of pedestrians crossing streets, traffic signals changing, and so on. By training on both successful driving behavior ("good" examples) and near-misses ("bad" examples), these systems identify useful patterns about accelerating, braking, steering, and obstacle avoidance and eventually learn to navigate complex road situations.[12]

This period, from about 1990 to 2010, produced some fascinating and useful machine-learning-based AI applications. Let us refer to this as the era of "AI and narrow machine learning." We use the word "narrow" because these systems were good at the specific tasks that they were trained to do, but not for more general activities outside of their specialized domains. So, for example, a self-driving car built using machine learning was quite capable at driving, but couldn't, for example, perform in a totally different domain such as chess because driving and chess playing are quite distinct. Nonetheless, the narrow machine-learning systems of this era were often reasonably reliable as long as they were used within the purposes for which

11. RUSSELL & NORVIG, *supra* note 2, at 26–29.

12. Harry Surden & Mary-Anne Williams, *Technological Opacity, Predictability, and Self-Driving Cars*, 38 CARDOZO L. REV. 121, 147–50 (2016). As of the writing of this article in 2024, autonomous vehicles have been successfully deployed in public, most prominently by Waymo, in warm weather cities such as San Francisco, Phoenix, and Los Angeles. While the technology is quite advanced as of 2024 and reliable if equipped with the appropriate sensors, hardware, and software, a few factors are holding back more widespread deployment. One of the biggest hurdles is the difficulty in handling snow-covered roads. Other factors include the high cost of autonomous vehicles equipped with advanced sensors, such as lidar.

they had been specifically trained: Machine-learning-based chess AI machines, for the first time, became very good at playing chess, and a self-driving systems became very good at navigating the roads autonomously.

C. Deep Learning and the Modern Era

Around 2011, one particular technique within machine-learning began to dominate, marking the start of the "deep learning" era.[13] Deep learning involved reviving a long-dormant approach known as "neural networks."[14] "Neural networks" referred to a class of machine-learning techniques, known since at least the 1950s, that took their name because they were very loosely inspired by the human brain. By the 1980s, however, neural networks had been largely abandoned as an AI approach due to perceived technical limitations. But around 2011, a small group of researchers decided to revisit neural network theory and began using it in new ways that had been previously out of reach in the past.[15] Particularly important was the realization that modern graphics cards—computer cards that had been used primarily for computer gaming—were also exceedingly effective at doing AI computations of the sort needed by the neural-network architecture.[16] Thus, thanks to intervening improvements in hardware, software, research, and data that occurred since the 1980s, researchers were able to successfully revive and scale up neural-network systems, achieving unprecedented success by making these systems much larger than was previously possible.

By 2011, researchers began referring to these larger systems as "deep learning" neural networks, with "deep" indicating the increased size and complexity compared to earlier versions. Importantly, these much bigger and more computationally intensive deep-learning neural networks

13. *See, e.g.*, Alex Krizhevsky, et al., *ImageNet Classification with Deep Convolutional Neural Networks*, *in* 25 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (F. Pereira et al. eds., 2012), https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html [https://perma.cc/DD5H-TZDZ]. This is paper widely credited with starting the deep-learning revolution. See, e.g., JOHN D. KELLEHER, DEEP LEARNING 138 (2019); Md Zahangir Alom, *The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches*, Sept. 2018, at 1, 6, 11, 16–17, https://arxiv.org/abs/1803.01164 [https://perma.cc/F2DF-L8H8].

14. RUSSELL & NORVIG, *supra* note 2, at 17, 24.

15. Krizhevsky, et al., *supra* note 13.

16. RUSSELL & NORVIG, *supra* note 2, at 27.

proved dramatically more effective at real-world tasks across many domains compared to other, earlier AI machine-learning or symbolic techniques.[17] For instance, such deep-learning systems could translate between languages or control self-driving cars with far greater accuracy than even the best machine-learning systems of the prior decade.

However, one of the biggest and most intractable problems in AI since its beginning was creating systems that could understand (and generate) human language. This area—known as "natural language processing" (NLP) proved to be particularly hard for both knowledge representation and machine-learning approaches. Since the inception of AI in the 1950's, researchers had made numerous attempts to develop systems capable of reading and understanding ordinary documents such as books, emails, legal contracts, or legislation. Despite intensive research, however, these efforts remained largely unsuccessful. Even the most advanced NLP systems of the early eras routinely failed to understand (or create) written documents at levels anywhere comparable to a literal person, nor could they reliably reason or solve even easy logic problems involving written language.

From 2015 to 2021, encouraged by the sudden success in other areas, researchers explored applying deep-learning approaches to NLP with the goal of creating robust AI systems that could reliably read and write human language documents.[18] However, despite its achievements elsewhere, deep learning in the area of language understanding during this period still remained very limited in its ability to comprehend ordinary human language and struggled at reasoning, producing coherent text, or providing useful information.[19] I will shortly show examples of the limitations of these recent AI deep-learning language models from 2021 and earlier.

---

17. *Id.* at 26.

18. *See, e.g.*, Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, *in* PROC. OF THE 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610, 610–12 (2021), https://dl.acm.org/doi/10.1145/3442188.3445922 [https://perma.cc/257W-FPSY] (discussing Bidirectional Encoder Representations from Transformers (BERT) and similar early language model projects).

19. *See, e.g.*, *Improving Language Understanding with Unsupervised Learning*, OPENAI (June 11, 2018), https://openai.com/index/language-unsupervised [https://perma.cc/7GCK-488M] (describing the original GPT-1 from 2018). While impressive for its time, GPT-1 exhibited extremely limited reasoning and language understanding abilities compared to later GPT-3 and GPT-3.5 models from 2022 onwards.

Despite deep learning's limitations in NLP from 2010 to 2021, several important research inventions developed during that period hinted at its future potential for handling human language. Two important contributions, "word vectors" (2012)[20] and "transformers" (2017),[21] both emerged from Google's research labs and led to improvements in AI language systems. "Word vectors" allowed computers, for the first time, to reliably represent the meaning of words using numbers, which was a necessary step for math-oriented AI systems to be able to effectively "understand" human language. Similarly, the transformer was another major deep-learning innovation that enabled NLP systems to consider for the first time the context and meaning of surrounding sentences in a document. This allowed for a more accurate understanding of the meaning of a document's written text based upon the other sentences around it. Crucially, the transformer was a new deep-learning architecture that contained efficiencies that enabled AI systems to be trained on far more documents, and far faster, than earlier deep-learning designs had allowed.

As an aside, rather than keeping these innovations secret, it is commendable that Google published them openly in research journals. Speaking as an advocate for open research, I can say that today's AI progress would not have been possible without such a practice of knowledge-sharing demonstrated by Google and other AI researchers across universities and industry. I encourage AI researchers to continue to widely share their AI knowledge going forward, rather than keeping it secret, as we have all collectively benefited from the knowledge shared by others with us in the past.

However, even with these emerging research innovations during that period, most researchers, myself included, still thought that highly flexible and capable AI language models that could read documents, solve problems, or understand arbitrary text, questions, or instructions, were many years off, given the limited state of NLP. The deep-learning NLP systems of the period from 2015 to early 2022, despite incremental improvements, simply performed too poorly on many basic

20. Tomas Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, ARXIV (Jan. 16, 2013), http://arxiv.org/abs/1301.3781 [https://perma.cc/ERF5-SLAA].

21. The most notable invention was the "transformer" deep-learning, neural-network architecture, developed by Google in 2017. Ashish Vaswani et al., *Attention Is All You Need*, ARXIV (June 12, 2017), http://arxiv.org/abs/1706.03762 [https://perma.cc/U9HY-D99L].

language understanding tasks. This made ChatGPT's sudden arrival in November 2022 all the more unexpected—as it markedly surpassed the capabilities of all previous state-of-the-art AI language models.

D. History of GPT

To understand why many AI researchers were caught off guard by the arrival of ChatGPT, it helps to look back at the evolution of GPT models over time. OpenAI released several earlier versions of the GPT model between 2018 and 2022, and none of them were nearly as capable as ChatGPT's GPT-3.5 turned out to be.

In 2018, OpenAI launched its initial foray into natural language processing with GPT-1, a system based on Google's transformer deep-learning architecture. With GPT-1, OpenAI was among the first organizations to apply Google's newly released transformer model to analyze large text datasets, such as unpublished books or public webpages. The goal was to identify linguistic patterns in existing text to enable automatic generation of new, human-like text. Having experimented with GPT-1 at the time, I found it to be interesting but underwhelming in terms of being able to produce realistic human language. A year later, in 2019, OpenAI released GPT-2, a larger and more sophisticated transformer model than its predecessor. While it was somewhat better at generating more coherent sentences than GPT-1, GPT-2 was still far from a system capable of truly understanding or creating human-like text at a deep level.

The release of GPT-3 in 2020 marked a more noticeable improvement. For the first time, an AI model displayed the ability to sometimes generate human-like language, including poems and other creative texts. While its ability to produce human-like content was a notable achievement—especially given that earlier models had significant challenges in this regard—GPT-3 appeared to primarily be a tool for generating basic short texts like stories and poems. GPT-3 seemed far from the general-purpose, thinking system that most people envision when they imagine AI, as it struggled with answering questions responsively and often failed to maintain coherence in extended discourse. For instance, as shown below, GPT-3, considered a state-of-the-art AI system in January 2022, was unable to produce a coherent legal motion (fig. 4). The limitations of GPT-3

and other similar state-of-the-art models made more general AI systems, capable of reasoning, conversing, or understanding arbitrary documents, still seem quite far off in early 2022.

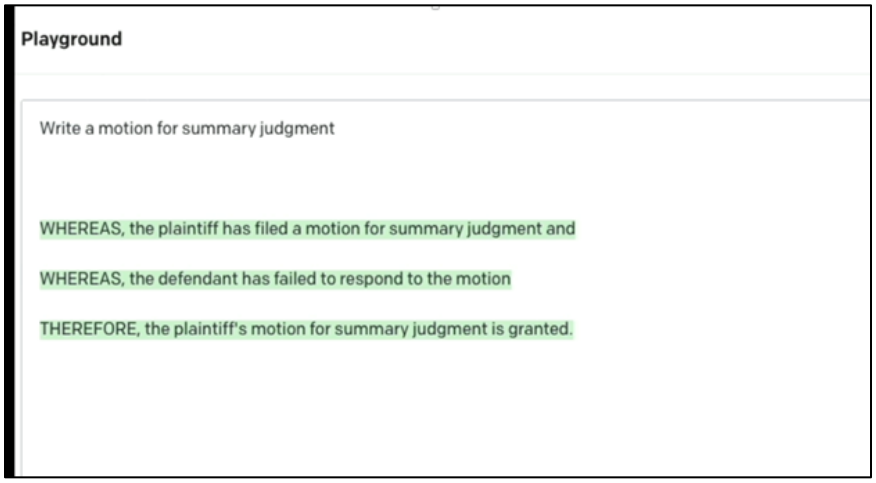

Figure 4. GPT-3 (early 2022, from OpenAI Playground), fails to produce a basic legal document, a motion for summary judgment.

E. The Era of Highly Capable Large Language Models like ChatGPT (2022–Present)

This is why many researchers were surprised when, in late 2022, OpenAI released ChatGPT using the improved GPT-3.5, an AI model that suddenly seemed to surpass the limitations of earlier systems. ChatGPT was, in essence, a more sophisticated system built on top of GPT-3's existing deep-learning, transformer architecture, with key engineering improvements from OpenAI that greatly enhanced its ability to answer questions, follow instructions, and solve problems. This initial version became known as ChatGPT (or GPT-3.5), to signal its substantial improvement over its predecessor, GPT-3.

Upon its release, it became quickly apparent that GPT-3.5 was the first highly capable AI language model to overcome many of the longstanding technical hurdles that had plagued NLP systems through much of AI history. For the first time, researchers had produced an NLP model that could reliably "read" and "understand" nearly any written document, including legal documents. GPT-3.5 represented quite a remarkable leap

compared to earlier AI language systems in terms of its general quality and usefulness.

Unlike previous iterations, GPT-3.5 also demonstrated a significant gain in its overall reasoning capabilities, offering a level of sophistication that had not been seen before in AI language systems. For the first time, an AI system was able to respond coherently (if not always accurately) to nearly any question posed, analyze just about any document it was given, and produce basic draft versions of most legal documents. GPT-3.5 was also able to reason about and respond sensibly to most of the basic commonsense questions that had stymied all prior AI systems I had tested, including the state-of-the art AI systems in 2022 from Google and Meta.

Just four months later, in March of 2023, OpenAI released an even more sophisticated model—GPT-4.[22] At the time, access to ChatGPT using GPT-4 required a paid subscription, and an updated version of GPT-4 (GPT-4o) is the model that I am demonstrating to you today. It is important to understand the distinction between ChatGPT using GPT-4 (the current state of the art), and GPT-3.5 (the older free model), because GPT-4 (fig. 5, fig. 6) is far more sophisticated than even GPT-3.5, which itself had been a huge leap over all previous existing AI language models.

What is the difference between the terms "ChatGPT" and "GPT"? ChatGPT refers to the user interface—the chatbot that the public uses through the website (or app) to interact with the AI model. But behind the scenes, the answers generated by the ChatGPT user interface can come from different AI models of varying degrees of ability with names like GPT-3.5, GPT-4, GPT-4o, or o1-preview. Whether or not one has access to the most advanced ChatGPT AI models typically depends upon whether one has paid for a subscription. This distinction is important because many people, unaware of the differences, may have only used the less advanced GPT-3.5, often leading to disappointing results and an underappreciation of the technology's true capabilities. Additionally, for certain legal tasks, unknowingly using a less capable model could be problematic, as the distinctions in AI ability from one model to the next can be substantial. This difference in quality between

22. OpenAI (2023), *GPT-4 Technical Report*, ARXIV (Mar. 15, 2023), http://arxiv.org/abs/2303.08774 [https://perma.cc/4HAW-RYHW].

paid and free AI models is a nuance to be aware of going forward and is likely to continue as new, more capable AI models emerge.

F. GPT-4: A Huge Advance

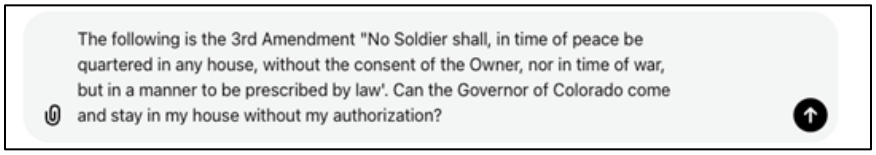

Figure 5. Third Amendment prompt for GPT-4.

Let's focus on the more capable GPT-4 model, as it is currently, in April 2024, the state of the art, and the model most likely used by lawyers. In this image (fig. 5), we can see GPT-4 reading a provision of the U.S. Constitution that I provided. In particular, I have asked it a question about the not-so-controversial Third Amendment,[23] and it has responded with a sensible, coherent, and at least, superficially well-argued, if not necessarily legally authoritative, answer (fig. 6).

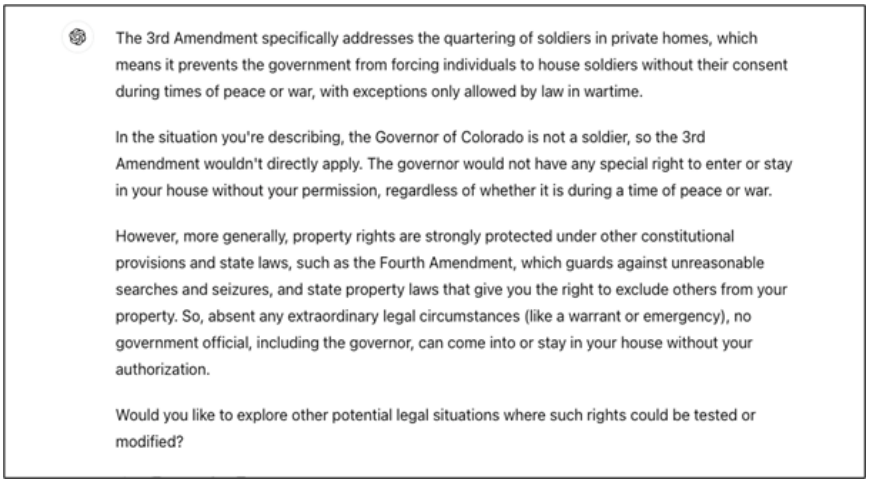

Figure 6. ChatGPT using GPT-4 (2024) provides an answer to a constitutional law question

More generally, today we can input questions about nearly any part of the Constitution (or any part of the law more

23. The Third Amendment of the U.S. Constitution, which prohibits the quartering of soldiers by the government, has been very rarely litigated in the history of the United States.

generally) into ChatGPT (using GPT-4) and expect it to produce a coherent and at least minimally well-reasoned, even if not necessarily authoritative, answer.

Maybe at this point, with most in the audience having used ChatGPT, we might take for granted that AI systems should have the ability to read and analyze arbitrary legal texts, such as the U.S. Constitution, or a statute or a contract, and produce some sort of reasonable analysis. But that was not the case until recently. These robust capabilities are new in the history of AI, and I want to emphasize how the AI systems of just two years ago, prior to ChatGPT's release in 2022, could not remotely produce analysis at the level of sophistication that you see demonstrated here.

Importantly, we can actually see side by side a demonstration of the difference between the state-of-the-art AI models like GPT-4 today and the AI models of only two years ago. Fortunately, the older AI models from early 2022 are still online, so we can think of this as a frozen time machine of the state of technology of the recent past. Importantly, we can actually ask the old and the new models the very same questions, and see the difference in quality of their responses, to just appreciate how far we have come in such a short time.

G. Looking Back to 2022

Let me demonstrate GPT-3, the older AI model from early 2022 that was the predecessor to GPT-3.5. As a reminder, the GPT-3 model was reasonably good at producing human-like text in certain contexts but was far from the general-purpose AI question answering and reasoning models that we have today. To see this, we can ask these older models some basic commonsense questions and assess the quality of their responses. Such testing is important because a basic limitation of the "intelligence" of most earlier AI models was the inability to reason in many simple, commonsense applications that are obvious to most people. To test these boundaries, my approach was to ask AI models simple, commonsense questions that are easy enough for a typical toddler to answer correctly, but that are also unusual enough that someone would not have written an answer online previously that the model could have copied during its training.

Asking simple but unusual questions is key. Since these systems like GPT-3 essentially "read" the whole Internet as they

are being developed, they can sometimes simulate knowing and reasoning about commonsense topics simply by parroting word patterns that they have repeatedly seen previously in their training data. For example, if somebody asks GPT-3 a common question, such as, "What do apples grow on?" it will know the answer "trees"; not necessarily because it "understands" the relationship between "apples" and "trees," but rather perhaps because there are many previously written webpages that mention that apples grow on trees and it is possible the AI system has previously read and simply memorized this "apples grow on trees" word pattern. Thus, if we want to actually test a model's commonsense reasoning, it is important to ask it something that it is unlikely to have directly read before, so we know it is not just "parroting" the pattern of earlier examples. But we would also like those same questions to be simple enough that even a small child would be able to answer correctly in order to demonstrate any AI commonsense-reasoning limitations compared to people. If it fails at many of these tasks, it is fair to say that such AI models do not routinely demonstrate the commonsense reasoning or knowledge typical of a small child.

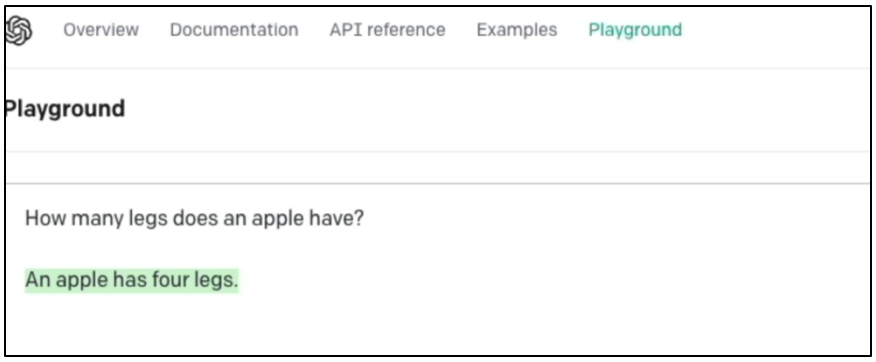

Figure 7. GPT-3 (early 2022) answers a commonsense question incorrectly

In this example (fig. 7), I asked the older GPT-3 model from early 2022, a basic, commonsense question, "How many legs does an apple have?" And it confidently answered, "An apple has four legs." This type of incorrect answer was very typical of AI models prior to 2022 and showcased their inability to simulate many aspects of commonsense intelligence. By contrast, if were to ask this same question to a human toddler, the toddler would likely laugh and say, "What? Apples don't have legs!"

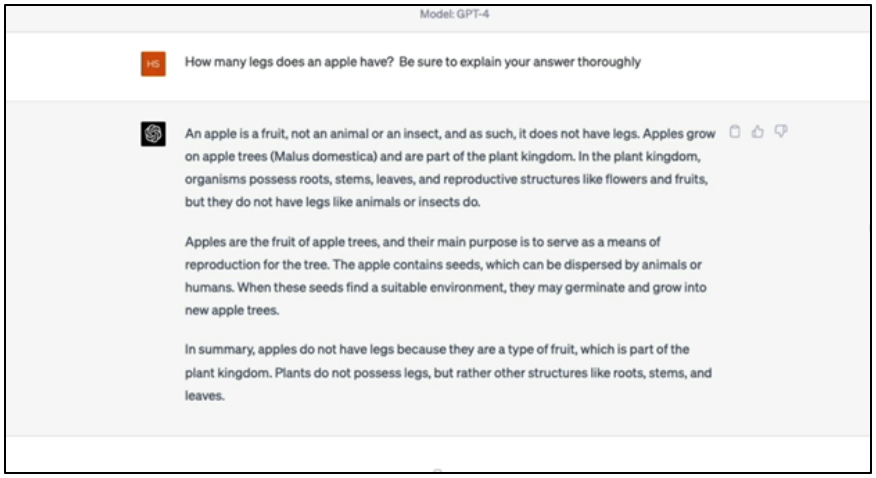

Figure 8. ChatGPT using GPT-4 (2023) answers a commonsense question

By contrast, both ChatGPT using GPT-3.5 and GPT-4 (fig. 8) were able to easily answer commonsense questions, like the apple question, as well as most of the other simple reasoning questions that had routinely stumped AI language models prior to 2022. This provided some evidence that these models had gained never before seen and impressive leaps in certain, basic AI reasoning abilities.

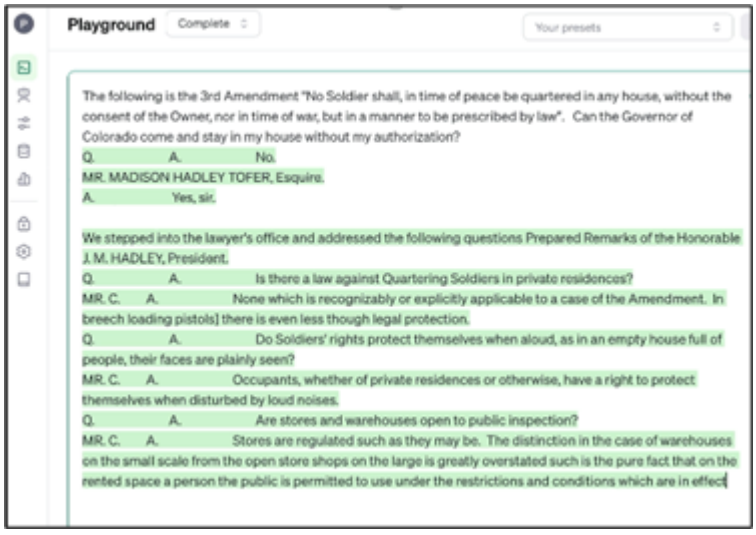

Figure 9. GPT-3 (early 2022)

Similarly, one can ask the older GPT-3 the Third Amendment question (fig. 9) that I earlier demonstrated with GPT-4 (fig. 6). In such cases, older models would tend to answer with irrelevant and non-coherent text such as it did here. This type of nonsensical answer to an unusual question that it hadn't specifically seen before in its training was typical of AI models prior to 2022. Thus, the difficult problem of the more recent past was simply getting AI models to respond sensibly and coherently to the questions being posed.

Finally, let's examine the sudden leap in AI capability in terms of creating legal documents. Recall earlier that GPT-3 from January 2022 was unable to create anything that resembled a coherent legal motion (fig. 4). Just one year later in March of 2023, GPT-4 was able to create capable first drafts of nearly any type of common legal document (fig. 10).

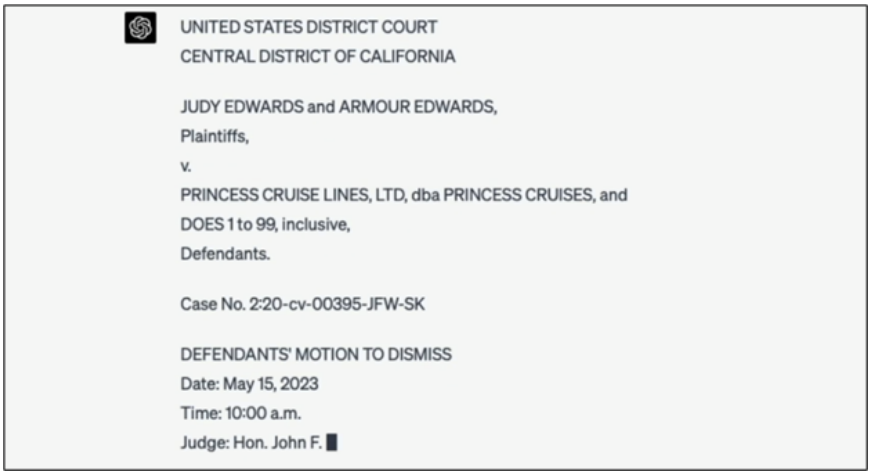

Figure 10. ChatGPT using GPT-4 (March 2023) produces a sophisticated legal motion

Although it might be hard to tell from the example on the screen, the quality of the motion produced by GPT-4 is quite good—it correctly references the facts and the relevant law and even puts it in the right format for federal district court. To be clear, the motion is not perfect, and I'll talk about how well these systems perform in a minute because I don't want to give the misimpression that these systems are without limitations. They are fallible and they do make mistakes.

Today, we take it for granted that ChatGPT using the GPT-3.5 model or better can produce relevant and coherent answers to any question, even if we sometimes worry about its

accuracy. But it is important to understand that back in 2022, the primary challenge for AI systems was simply to respond in a way that was minimally pertinent to the questions being asked. For example, GPT-3 struggled to provide a pertinent and coherent response when I asked about the Third Amendment (fig. 9). This illustrated that not so long ago, merely producing consistently relevant and readable responses to user questions was seen as the primary technical hurdle that would take many years to overcome.

Another thing that surprised many researchers was emergent properties of reasoning and problem-solving displayed by GPT-3.5. GPT-3.5 was trained by using machine learning to detect word patterns, knowledge, and abstract concepts by analyzing large amounts of documents, such as Wikipedia, books, or other Internet sources. Importantly, the system was not primarily trained to reason and problem-solve. Rather, the problem-solving and reasoning abilities seemed to be an emergent property that unexpectedly developed during the training process. By contrast, previous AI models like GPT-3 from 2022 and earlier exhibited little to no general reasoning abilities that they had not been specifically trained for.

And yet, by analyzing many documents during training, some of which involved problem-solving, GPT-3.5 was able to learn certain broader reasoning and logic abilities. That was completely unexpected to people like me—another of the major reasons that so much excitement began building in November 2022 around AI. Most previous AI systems had not appeared capable of acquiring emergent new skills, like reasoning, simply as a byproduct of being exposed to language patterns in millions of existing written documents.

To recap what was demonstrated, within just one year—from GPT-3 in early 2022 to GPT-4 in early 2023—AI large language models made enormous and unexpected leaps in capability, coherence, and usefulness. AI systems went from not being able to perform reasoning, coherently create documents, or follow instructions, to being able to do all of that, and also, for the first time, being able to produce reasonably good drafts of legal document, or to analyze just about any aspect of law. So, the fact that current systems sometimes make mistakes, while important to keep in mind, is almost beside the point in the larger arc of AI progress. The bigger picture is that AI technology essentially progressed within one year at a pace that most researchers thought would take five to ten years. It's not clear

whether or not this pace of improvement will continue going forward, but looking backward it is clear that a significant jump in capabilities in AI happened in an unexpectedly short amount of time. This rapid and unforeseen improvement is why so many people are excited about AI and why it has been in the news so much in the past two years.

V. GPT-4: FOUNDATIONS

There are two more important points to emphasize about large language models like GPT-4. The first is that, unlike earlier AI systems, they can now work capably with human language, and that makes them potentially more broadly impactful. As we know, language is fundamental to how humans communicate about society, law, science, and knowledge and is one of the primary ways we transmit the achievements of the past. With AI models that can reliably understand language, we can now, for the first time, use AI to apply foundational concepts to tackle new problems—a key advancement given the central role of language in law, society, and knowledge. Today's advanced large language models seem to be able to internally encapsulate many abstract concepts inferred from analyzing the text of billions of written webpages, books, and articles across just about all domains of human endeavor. Notably, prior AI systems did not have these broad linguistic and abstract reasoning capabilities.

Secondly, and relatedly, GPT-4 and similar large language model systems are more *general* AI systems than most AI systems of the past. As mentioned previously, we have had many interesting machine-learning systems over the past thirty years, but notably, these systems were narrow in a particular sense: An earlier AI system like a chess-playing system, or a self-driving car system, could only work capably in the specific area for which it had been trained. By contrast, because language is so general and is used to communicate in just about every aspect of society, large language models are the first *generally* capable AI systems that we have seen. In just about any area of inquiry, from law to science to psychology, they can produce reasonably capable answers across nearly any domain. Moreover, newer AI models are being supplemented with different modalities besides language, including video and images, which can further increase their abilities.

Because of this unprecedented breadth and generality, I will not hesitate to say that GPT-4, and similar advanced large language models, are easily one of the biggest breakthroughs in AI in the last twenty years. It is already having an impact on law, although like most technologies, we should expect the impact to be gradual over time, rather than immediate. However, despite their impressive abilities, it is important to emphasize that large-language-model AI technology is far from perfect. It is as necessary to understand their limitations in order to use these tools well. So let us now focus on some of the limits of GPT-4 and similar AI technologies.

VI. LIMITS OF LARGE LANGUAGE MODELS LIKE GPT-4

To be clear, this technology can fail, and it is by no means without its flaws. For instance, a well-known limitation is that language models occasionally make up plausible-seeming facts—a phenomenon known as "hallucinations"—which is especially common in the older versions. If you type certain prompts into the older GPT-3.5, it will hallucinate—it will make up plausible-sounding case names. There are the widely publicized instances of lawyers penalized for using ChatGPT to write legal briefs without checking the validity of case names. GPT-4 and more modern models are less likely to hallucinate than earlier models, but they're still not immune and do occasionally make up facts.

In terms of other limitations, sometimes the information of these models can be out-of-date. They can be tricked, make reasoning errors, and occasionally reveal private data. There may be biases in the training data—we really don't understand how a model like GPT-4 actually produces the words that it does. Thus, there's a lack of transparency and inability to interpret its outputs. So, while AI models like GPT-4 are amazing, there are real limits that matter. In particular, I will shortly focus upon how the answers ChatGPT, using GPT-3.5 or newer models, and other large language models are sensitive to the particular words that one uses in the prompt in ways that might not be readily apparent to lay users.

A. Current Competitors

So where are we today? Today, there are some competitor AI models that rival, or even surpass, GPT-4. For instance, there

is Claude 3.5 Sonnet from Anthropic, and also Gemini Ultra from Google. There are also several open-source or open weight models like LLamA 3 from Meta, the largest size of which is 405 billion parameters and is the first open-weights model that is in the same class as GPT-4. Thus, today, a user has a choice of multiple, cutting-edge "frontier" models from other companies besides OpenAI.

VII. HOW TO USE AI IN LAW

Now that we understand the basics of AI and what has changed in terms of large language models such as GPT-4, let us turn to how this technology is now being used in the field of law. Roughly speaking, there are two main ways to use large language models in law. The first way is to use specialized, legal-focused AI systems from existing companies that already work within law. The other method is to use general large language model systems like GPT-4 or Claude 3.5 Sonnet directly, and ask them legal questions. I generally recommend the specialized legal systems, for the most part.

The most notable current specialized legal systems are Lexis+ AI and Westlaw (Thompson Reuters CoCounsel). These are from well-established legal research providers, and on the backend are built *on top of* frontier AI models like GPT-4. However, importantly, such legally specialized systems give you not only access to well-curated and established sources of legal data, like caselaw or statutes, but also the security and privacy and confidentiality guarantees that lawyers currently depend upon. So, that's generally what I would recommend for higher-stakes and sensitive legal work.

I don't have a preference between these two AI and law providers. I think they're both pretty good in terms of providing legal analysis, provided one knows how to use it. This is a crucial point: Lawyers should strive to improve their AI literacy—understanding the strengths and limitations of such systems. The main way to do so is to read and learn about it, and importantly, gain experience using these systems repeatedly to understand where they go right, where they can save time, and where they go wrong.

The other way to use large language models for legal work is to work directly with systems like GPT-4 or Claude 3.5 Sonnet, and not through the intermediary of specialized legal systems like Lexis or Westlaw. This comes with some risks if one is not

careful. The first thing to be aware of is that these systems have various levels of security and privacy guarantees. If one uses the *enterprise* versions of OpenAI's GPT-4 or Anthropic's Claude 3.5 Sonnet, generally speaking, one can be confident in the security and the privacy of one's data. In these enterprise-grade versions, the companies generally will provide the highest level of security and privacy, such as Service Organization Controls (SOC) 2 guarantees, that business users are accustomed when storing data in the cloud.

However, the big issue is that many lawyers who use GPT-4 or Claude 3.5 Sonnet are not using the enterprise-grade versions. Rather, many people are using the free or lower-tier "plus" versions. For sensitive legal problems, this can be problematic. For one, at these free or lower levels, OpenAI and Anthropic do not guarantee that they will not train their future AI models on the information that you enter in your prompts. So, for example, if you enter a scenario into a prompt that includes privileged or sensitive information about one of your actual clients, it is possible that OpenAI will use this information as part of the large dataset of documents that it uses to train its next system. While the actual risk of sensitive information leaking into these future AI systems through training is probably low as a practical matter, it is a consideration. More broadly, there are issues of losing lawyer-client privilege by exposing privileged issues to third parties.

However, for comparatively low-stakes analyses that do not involve private client or sensitive information, the use of GPT-4 or similar AI models is probably fine. Still, lawyers have to be extremely careful to double-check the outputs for accuracy and to manually provide the systems with the appropriate laws in their prompts. Many lawyers who use these tools may not understand the nuances of the privacy and security guarantees of the AI system they are using. Thus, there are small, but probably real, risks if one is not using an OpenAI enterprise-grade program that one could accidentally give up attorney-client privilege or expose private information. In general, it is probably better for any sensitive client information to use the law-specific systems that already provide known guarantees.

A. Caveats and Limitations

That said, provided you don't use confidential information, GPT-4 can actually be a quite reliable analyzer for *basic* legal problems. Thus, it can still be very useful as there are many legal scenarios that do not necessarily involve private information, are not too complex, and for which AI large language models can nudge an attorney in the right direction. Here is GPT-4, for example, analyzing a torts fact pattern that I made up, and it does as well as, if not better than, most of my torts students. Provided one uses it correctly, these systems can give you a well-reasoned legal conclusion for just about any basic legal issue. Below is a slip and fall case with facts I created (fig. 11). I've entered this into GPT-4, and it gives a reasonably good analysis and then a prediction about the outcome, which is in line with my own legal prediction. It estimated that the plaintiff has about a 30 percent chance of winning, which is a pretty good assessment. The fact that GPT-4 can do a reasonable legal analysis like this is quite remarkable. Still, it takes care to use it well. One must provide it with the relevant law and carefully analyze its outputs.

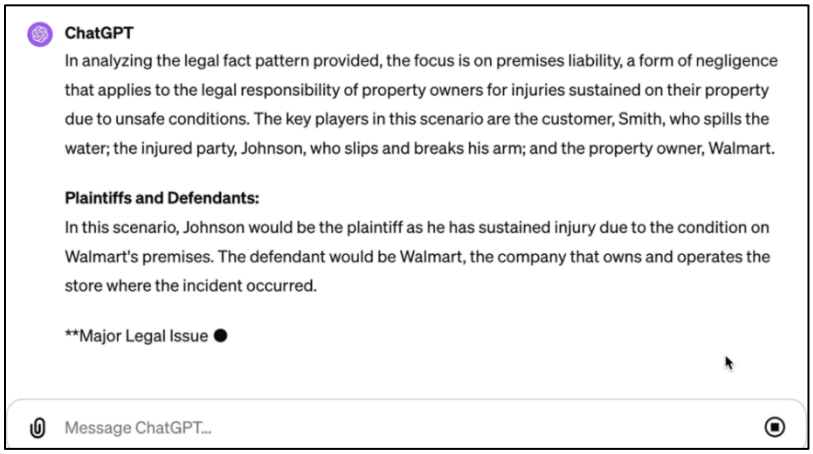

Figure 11. ChatGPT using GPT-4 performs legal analysis

Tools like GPT-4 can also read contracts. One of my research areas is a topic known as "computable contracts." One can ask GPT-4 questions about existing contracts, and it's pretty reliable, provided again that one uses it carefully. Here is an example of GPT-4 answering scenarios about a homeowners insurance policy and doing a reasonably good job citing the

actual provisions about whether something is covered or not (fig. 12).

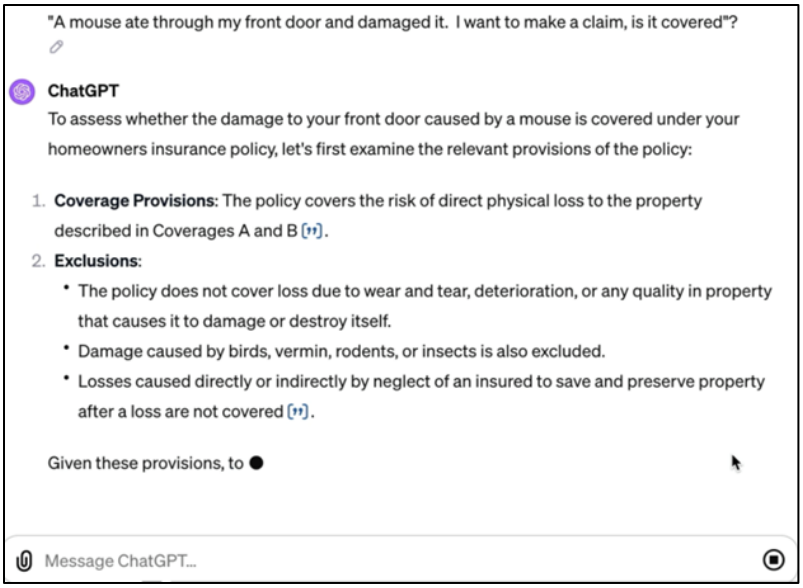

Figure 12. ChatGPT using GPT-4 analyzes a home insurance contract for a hypothetical policy claim

Thus, advanced AI large language models can be very useful as a tool, as long as one is careful and double-checks the work. One way to think about it is to use the following mental model: Think of GPT-4 as if it had the capability of the best, top-of-the-class third-year law student intern. On the one hand, if a lawyer had a top-notch third-year law student as an intern, this could be incredibly useful. The student intern could excel at producing initial drafts of documents or doing basic research that could be extremely helpful as a starting point for a lawyer to build upon. On the other hand, a lawyer likely would not turn in the work-product from a student intern directly to a judge or a client. Rather, a lawyer would take the intern's draft and double-check it and change it substantially into a polished, finished product. This is because even the smartest student lacks experience and can make basic errors in law or reasoning. But, just because the top student intern sometimes makes basic errors does not mean that their initial drafts are not useful. To the contrary, a strong initial draft can be extremely useful for an experienced attorney to more productively and quickly turn it into a finished product. Similarly, frontier large language models can produce excellent initial drafts of documents,

particularly when you provide them with examples of earlier model documents in the prompt. But they will occasionally make errors about the law or have other details wrong, and need their work supervised.

The key is to have a strong familiarity and literacy of the capabilities and limits of the AI tools themselves, and not to conceive of them as error-free, final-product-producing, or final-answer-producing machines. Rather, they should be understood as tools that can produce initial drafts or starting-point answers upon which lawyers can and should add their expertise.

B. Best Use Cases and Reliability

Let's explore in more depth which use cases in law systems like GPT-4 are fairly reliable today. The most reliable use case is to use advanced large language models to summarize legal or other documents or ask questions about a document's contents. For example, I can take a complex motion from a court case, copy its text and paste the entire text into ChatGPT prompt interface and then below that, ask it a question about the contents of the text I just copied. Even though ChatGPT's interface gives me the option to upload a PDF of the document, at the current moment for technical reasons I will not go into, it is often more effective to copy and paste the full text of the document from the PDF prompt interface as I have done here, rather than upload a PDF of that document, even though it seems like those two actions should result in the same thing.

What is interesting about this use case is that one can interact with one's documents to, for example, ask clarifying questions about a document's contents or ask the large language model for clarifying examples. The ability to interact with one's documents in a dialogue is quite a new capability in the history of AI. Normally, documents are static objects that AI can merely read or search. But large language models now give us the ability to "talk" to one's documents through the large language model and to ask clarifying questions or request examples or counterexamples of concepts described in the document. These documents can range from motions to court opinions to statutes. For example, if one encounters difficult-to-understand statutory language and pastes the text into ChatGPT using GPT-4, the AI model can generally do a good job of helping one understand the basic meaning of the statute. Again, a lawyer would then ensure

that they have a fuller picture that includes any court-based interpretations of the plain meaning, but the model can initially point you in the right direction. Today, summarizing documents and asking clarifying questions is a very effective and quite reliable use case in law.

Other reliable use cases involve using AI to draft basic documents by providing it with examples of previous, similar documents in the prompt. Generally speaking, this is more reliable when the documents are only a few pages long rather than long, complex documents.

C. AI and Constitutional Law

In my last couple of minutes, let us turn to AI and constitutional law. We have seen the frontier models like GPT-4 that can read arbitrary legal texts and return reasonable-sounding legal analyses. Is this a good idea to use GPT-4 or the future GPT-5 to interpret constitutional documents? Is this reliable? This is a complicated question. While the models can generally provide rough summaries of areas of law and a constitutional amendment's plain text, at the moment we should not rely upon them in this context for a number of reasons. For one, such models rely heavily upon the "context" of what we provide them in the prompt, and in many cases the necessary context—such as earlier court opinions interpreting the meaning of the Constitution—are not available to the model at that moment. More importantly, however, such models' outputs can be highly sensitive to various factors that might appear to a lay user not to make a difference, such as the particular words the user chose in her prompt, or the particular AI model that was used.

For example, remember earlier I asked ChatGPT using GPT-4 a question about the Third Amendment: "Can the Governor of Colorado stay in my house under the Third Amendment?" (fig. 5). The model responded with a confident-sounding, well-reasoned answer (fig. 6), indicating that the Governor can't stay there, but not because of the Third Amendment. Rather, according to GPT-4's response, the Third Amendment applies to only to soldiers under the plain-meaning of the Amendment's words.

By contrast, we can ask Claude, another equally capable frontier AI system from a different company, the same exact, seemingly straightforward constitutional question. However, as

one can see, Claude responds equally confidentially but with the opposite answer. According to Claude, "Even though the Third Amendment contains the word 'soldiers,' it should be interpreted to apply to all government officials. Thus, Third Amendment means the government cannot force you to house soldiers or other government officials, including the Governor of Colorado."[24] We apparently have our first AI circuit split. One AI system confidentially gives one answer to a seemingly straightforward constitutional query, while another AI system gives you the exact opposite, but equally confident answer.

What this example highlights is the problematic nuances of using AI in important legal settings without proper AI literacy and an understanding of the technical and legal theoretical limitations of these systems.

These systems can give the superficial appearance that they can confidently and authoritatively "answer" legal or constitutional questions. Moreover, undoubtedly, some judges out there are currently, either openly or secretly, typing such queries into ChatGPT or Claude in their chambers. And these systems will definitely provide one with a confident and generally well-reasoned answer. Moreover, it is entirely possible that such judges will see such a reasonable-seeming answer and defer to it. But what such a judge may be missing is that, in the background, such systems are implicitly doing value interpretations that today are normally done explicitly by judges.

For example, under the Third Amendment, does the word "soldier" mean that the Amendment literally only applies to members of the miliary, as GPT-4's answer claimed? Or is it more broadly supposed to apply to forcible housing of government officials who are not technically soldiers, as Claude's answer asserted? That's an interpretive question, and the interpretation of this word to mean one thing or another could be determinative to the outcome of the case. What's the "right" answer? That's the point: There is no right answer—it's the job of judges to determine the best interpretation of the answer given precedent, policies, facts, and other relevant information. However, current AI systems can non-transparently make such value "choices," behind the scenes in ways that may not be obvious to the judges and other legal

24. Andrew Coan & Harry Surden, *Quartering the Governor Under the Third Amendment*, *Claude*, ANTHROPIC (unpublished prompt) (on file with authors).

officials who are using them. Moreover, the answers that these systems give are often sensitive and can change, depending upon the particular words given in the prompt. For example, just to use a word like "originalism" in the prompt in many cases may "nudge" the AI model to produce a more originalist response.

Thus, while I do think that such AI systems can ultimately be valuable for judges and lawyers in providing more information upon which these experts can make more informed decisions, it is important that the technology is understood as an information-providing *tool*, and *not* as a neutral, question-and-answer legal oracle, the way it may superficially appear to be today.

We are here today to have a conversation about these topics to ensure that lawyers understand that when they ask legal questions of these AI systems, these value and interpretive choices are often implicitly occurring in the background in a non-transparent manner. AI systems can be very valuable in law, but they require a reasonable degree of literacy as to their strengths and weaknesses, and awareness of the different interpretive and value choice-points that are invisibly occurring as these systems produce answers.

VIII. PREDICTING THE FUTURE OF AI

As I end here, I am going to peer a bit into the future of AI. Before I do that, however, let me give you an invitation to be skeptical of those who predict the future of technology, including me. I invite you to be very skeptical about those who purport to predict the future of AI and other technologies, particularly on longer time frames, and particularly about the details. Today, many in the media or on social media try to predict the future of AI in five, ten, or twenty years. Many also make predictions about what is going to happen with respect to society, jobs, or art, and sometimes these predicted effects are good for society, and sometimes bad. The important reality for you to take away: These people don't actually know the answer to what they are predicting.

They do not know, just as I do not know, the details about the future and society with respect to AI. But we may often be fooled by the confidence with which one commentor predicts a future of AI-induced abundance while a different commentor predicts an AI future of doom and gloom and unemployment. But their confidence does not reflect the reality that these

commentators do not know the details about the future, in the same way that I do not know, and you do not know.

Research shows that humans are really bad at predicting the future, in particular the details about the specific directions from which upcoming technologies will develop; what new, unforeseen technologies will arise and when; and the details about how future technologies will actually affect society in terms of politics, jobs, art, policies, and so forth. There is a long history of people confidently, but quite incorrectly, predicting the future arc of technologies and their detailed impact on day-to-day life ten, twenty, or thirty years in the future—and they are typically wrong. And the reason is the technology, and society itself, is just too complex and uncertain to predict with any specificity in the longer time horizon beyond five years or so. Moreover, the details about these unpredictable technologies actually matter. So even if one can very roughly predict that AI will continue to improve over time or that computing will get better over time—all reasonable assumptions—that is very different from accurately predicting the details about how these continued improvements will affect our day-to-day lives, politics, society, jobs, the arts, and so forth. And it is those unpredictable details that matter the most.

Moreover, there are commentators who just happen to get predictions approximately right many years in advance. But they are usually those who happened to get lucky, and through selection and hindsight bias, appear as if they knew what they were talking about. In other words, they may turn out to be roughly right, but they turn out to be right due to luck and not using reliable methods. And we can see this if we look at the many others who also made confident predictions about the impact of technology that later turned out to be wrong—that's selection bias. When it comes to predicting the detailed impacts of technology such as AI, history has shown that nobody can confidently predict what the future will hold with any specificity, whether good or bad, including me.

What I have just discussed are confident, detailed predictions about technology on a long time-scale. However, there are some very modest predictions that we *can* make at a high level of generality and on limited time scales that *can* be helpful. Generally speaking, the best that one can do reliably is to make predictions on a narrow time frame, say one to two years in the future, based upon known, current technologies and known trends. For example, we can modestly project out existing

AI trends in the next year or two, and generally understand that the technology will tend to get faster, less expensive, smaller, and more capable.

In my view, taking modest time frames and making very general predictions rooted in known technologies and trends—as opposed to speculating about future technological developments that have not yet been invented or proven, and avoiding speculating about specific societal changes brought about by technology—is the most reliable way to predict a very narrow slice of the future. But that's about it—beyond that, history shows that we are not good at reliably predicting how fast these technologies will develop or accelerate; what new, unknown technological developments will occur; how these technologies will affect jobs or society ten years from now; or what the technology will look like in detail in twenty years. And to that end, I very much invite you to be skeptical of those who purport to predict the broad impacts of AI, or other technologies, particularly those who predict in timeframes beyond two or three years, and particularly those who espouse extreme visions of prosperity or harm or who purport to know the details about how the technology will ultimately develop or how it will impact society. There is no shortage of those making these strong predictions confidently, but that does not mean that their predictions are actually reliable.

A. Near-Future Trends in AI

What is reasonable to believe will develop in the short term with respect to AI based upon known, current technology and trends; taking our limited predictive view; and projecting out modestly in the next one or two years?

For one, "frontier" large language model systems are likely to get larger and more powerful. OpenAI is reportedly currently training GPT-5, its next generation AI model, which is likely to be more capable than GPT-4. We don't know how much better it will actually be but these large language model systems tend to get better as they grow data. This is known as "scaling laws"—large language models seem to reliably, at least so far, improve in capabilities the bigger they are in terms of computation and data. So, for example, GPT-3, the large language model from 2022, which had moderate performance, was about 175 billion "parameters" in size. It was big but small compared to GPT-4, which was estimated to be about ten times bigger, at one trillion

parameters. And GPT-5 might be as large as ten trillion parameters. Each one of those parameters means it has the ability to detect more patterns and absorb more high-level abstract concepts, and so forth. So, these systems will generally get better as their size (in terms of the level of computation that they use and the amount of data that they are trained on) increases, compared to previous generation systems. This is the reason to be skeptical: We just don't know how much better these new systems will be compared to today, nor what the actual impact on society improved AI systems will have.

There are some other modest AI trends that we can project-out: The earlier GPT-3 system trained on web pages that had a lot of unreliable data. There's a lot of good information on the Internet, such as Wikipedia, but there is also lot of unreliable data, such as many low-quality comments on social media. In their early days, researchers tended to train systems such as GPT on just about any data on the Internet that was available, without filtering for quality. Today, there is a lot more emphasis on training on higher quality data, which research has shown to dramatically improve the quality of the systems. So future large language models will increasingly be trained on *high-quality data*, such as textbooks and research papers, rather than just being trained on all data that happens to be available, whether high-quality research or low-quality social media comments.

There are also new AI architectures and designs that are coming out. There has been a lot of innovation. The transformer—the invention mentioned earlier upon which most of the modern AI era has been built—was invented several years ago in 2017. Since that time, there have been a lot of inventions on top of similar architectures that are incrementally improving how fast and how accurately AI large language models operate. Thus, we are slowly seeing the adoption of these more recent software and algorithm inventions entering the mainstream.

There have also been improvements on the hardware side. Much of current AI is limited by immense hardware requirements to train and run sophisticated systems like GPT-4. Importantly, the underlying hardware that powers these AI systems seems to reliably get faster and more powerful each year. For instance, Nvidia, the maker of the graphic processing units that models like GPT-4 often use to process user questions (what is known as "performing inference"), just released a new graphics processing unit, which has 200 billion transistors on a single chip. As AI hardware becomes more powerful, which most

expect it to do, AI becomes less expensive and faster and can likely do more computation and reasoning than today's systems.

Another emerging, near-term trend in AI is that we are starting to see increases in AI agents or autonomous systems. An AI "agent" or "autonomous system" is one that gives the AI system high-level goals, rather than specific instructions, and the system itself is tasked with figuring out the underlying details about how to make it work. For instance, you might ask an AI agent to purchase something on your behalf or to do research on a topic for you. Unlike most AI systems today where a human user is controlling each step of the AI interaction, with an AI agent the AI system itself is making a plan about the steps needed to reach a certain goal, making a series of decisions and performing actions on its own to get there. In this regard, we are seeing AI algorithms for better planning, and other software improvements.

Another AI technological trend of the future that we are seeing is longer context windows, meaning you can put more helpful contextual text in the prompt to guide AI models, such as GPT-4, to produce more accurate answers. Right now, we're seeing up to one million tokens—long enough to put in several books. For example, if this trend continues, one could put entire legal treatises into the next generation of GPT and ask detailed and useful questions about the law. Today, most large language models limit us to about fifty to one hundred pages of text that we can reliably include in a prompt.

Finally, another trend, in law in particular, we are seeing is more reliability compared to the large language models of the past and more verification. We are seeing a trend in composite systems that, before they immediately provide an answer, will make a plan, will check resources, and perform verifications to help improve the reliability of AI systems and reduce hallucinations.[25] We are already seeing this with OpenAI's "o1-preview" model, released in October 2024, which spends time planning, thinking, and verifying before it provides an answer to a user, which dramatically increases its accuracy and capabilities on advanced math and logic problems.[26]

Thus, these are all trends in AI that I expect in 2025 and onwards that will likely create more reliable and capable

25. *See, e.g.*, *Introducing ChatGPT Search*, https://openai.com/index/introducing-chatgpt-search [https://perma.cc/WB6G-VSYF].

26. *See, e.g., Introducing OpenAI o1*, https://openai.com/index/introducing-openai-o1-preview [https://perma.cc/W9LT-GLQM].

systems than we have today. These predictions are all based upon current known trends and currently available technologies. However, just how reliable and how much more intelligent these AI systems of the next two years will be is hard to know with any degree of certainty.

IX. THE NEED FOR UNDERSTANDING AND IMPROVEMENT

However, there are still limits to these systems, and while some of these limits will likely be improved upon in future iterations, many will still persist into the near future. As with any tool, we need to understand the nuances of these limits in order to use these tools correctly. If we do not understand these limits, we will think that we are getting legally authoritative, objectively correct answers to legal questions, when, in fact, an AI model is giving us contingent output, based upon subtle, internal interpretation choices it has made.

To end on an optimistic note: I believe that AI can bring many benefits to the law if we, as a society, use it thoughtfully. I believe that we can use it to improve the quality of the law, and that lawyers can use it to produce higher caliber legal work for clients. But we should take this as a broader opportunity to improve the law and the legal system more generally. We should use this moment with AI to make the law more transparent than it has been. I believe that AI, if used with care, can make the law more equitable and less biased; and I believe that we can use it to improve access to justice in many contexts. This is a moment where we can choose to use this incredible new set of technological tools to *create* the future of the law and the legal system that embodies what we as a society want it to be—fairer and more accessible for all of us.

Thank you very much.

I appreciate your time.