

ARTIFICIAL INTELLIGENCE AND CONSTITUTIONAL INTERPRETATION

ANDREW COAN* AND HARRY SURDEN†‡

This Article examines the potential use of large language models (LLMs) in constitutional interpretation. LLMs are extremely powerful tools, with significant potential to improve the quality and efficiency of constitutional analysis. But their outputs are highly sensitive to variations in prompts and counterarguments, illustrating the importance of human framing choices. As a result, using LLMs for constitutional interpretation implicates substantially the same theoretical issues that confront human interpreters. Two key implications emerge: First, it is crucial to attend carefully to particular use cases and institutional contexts. Relatedly, judges and lawyers must develop “AI literacy” to use LLMs responsibly. Second, there is no avoiding the burdens of judgment. For any given task, LLMs may be better or worse than humans, but the choice of whether and how to use them is itself a judgment requiring normative justification.

INTRODUCTION	415
I. WHAT IS ARTIFICIAL INTELLIGENCE? WHAT IS CONSTITUTIONAL INTERPRETATION?	422
A. What is Artificial Intelligence?.....	422
1. Advances in Artificial Intelligence	423
2. A Simple Illustration	430
B. Constitutional Interpretation.....	432

*Milton O. Riepe Chair in Constitutional Law, University of Arizona, James E. Rogers College of Law.

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

‡ For helpful discussion and comments, we thank Yonathan Arbel, Anuj Desai, and Andrew Woods, and the participants of the Silicon Flatirons and Byron White Center, Ira C. Rothgerber Jr. symposium on AI and Constitutional Law. The *Colorado Law Review* team provided exceptional editorial assistance.

1.	What Makes Constitutional Interpretation Distinctive.....	433
	a. An Ancient Text and the Dead Hand of the Past.....	433
	b. Difficulty of Amendment, Glittering Generalities, and Judicial Review	435
2.	Constitutional Formalism vs. Constitutional Realism	438
	a. Constitutional Formalism.....	438
	b. Constitutional Realism	441
3.	The Importance of Institutional Context	443
II.	USE CASES FOR LLMs IN CONSTITUTIONAL INTERPRETATION	444
	A. Research Assistant, Critic, Troubleshooter, and Editor	445
	B. Drafting Assistant.....	449
	C. Focused Legal Queries.....	452
	1. Focused Constitutional Queries	453
	2. LLMs as the Ultimate Decider	458
	a. Ultimate and Comprehensive.....	458
	b. Ultimate but Limited	462
III.	POTENTIAL BENEFITS AND DRAWBACKS OF ARTIFICIAL INTELLIGENCE IN CONSTITUTIONAL INTERPRETATION	464
	A. A Naïve View and a Motivating Example.....	465
	B. Technical Limitations	468
	1. Sensitivity to Training and Architecture.....	468
	2. Prompt Sensitivity	471
	3. Randomness.....	473
	C. Legal Theoretical Limitations	475
	D. Benefits.....	482
IV.	A SIMPLE SIMULATION	485
	A. Design and Procedure	486
	B. Results	489
	1. Abortion	489
	2. Affirmative Action.....	490
V.	TAKEAWAYS.....	493
	A. The Law of Conservation of Judgment	494
	B. Artificial Intelligence Literacy and Best Practices	496
	CONCLUSION.....	498
	APPENDIX A	500
	A. Comprehensive – Question 1	500
	B. Limited – Question 1.....	500

C. Limited – Question 2.....	500
D. Limited – Question 3.....	501

INTRODUCTION

In recent years, there have been significant improvements in the ability of artificial intelligence (AI) systems to interpret, understand, and generate language and text. This is notable because until about 2022 AI models struggled with ordinary language processing. Earlier AI systems were unable to produce sensible, relevant, and useful responses to most written inputs.¹ However, today’s AI large language models (LLMs) can analyze a wide array of texts, including complex legal documents, in ways that were, until recently, completely out of reach.² One can, for example, give an LLM the text of the U.S. Constitution and the text or summaries of relevant precedential cases, then ask for its analysis on issues such as the constitutionality of abortion regulation or affirmative action.³ Most advanced “frontier” LLM systems, like ChatGPT, will respond with well-reasoned, coherent, and persuasive text that responsively analyzes just about any question posed.⁴

But just because we now *can* ask AI LLM systems to perform legal and constitutional analysis, *should* we? That is the focus of this Article: cautioning that we must proceed carefully in this area because the technology’s appearance of objectivity is potentially deceptive for judges or other officials who might not understand its subtleties. Such AI LLM systems will indeed provide anyone with confident, often factual, and seemingly objective analyses of constitutional and other legal issues. But the straightforward and convincing nature of the textual output can mask a wide array of subtle decisions, value choices, and policy determinations implicitly made by the AI LLM systems that could easily go unnoticed by decision-makers who use them.

The seeming neutrality and objectivity of the legal analysis produced by LLMs recalls earlier themes in legal thought. At least since Jeremy Bentham, legal formalists in the

1. *See infra* Part I.

2. *See infra* Part I.

3. *See, e.g.*, Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451 (2024); Harry Surden, *Chatgpt, AI Large Language Models, and Law*, 92 FORDHAM L. REV. 1941 (2024).

4. *See infra* Part I.

Anglo-American tradition have dreamed of making the law clearer, more precise, and more predictable in its application, with the ultimate goal of limiting or eliminating the human subjectivity of judging.⁵ The strongest versions of this “legal formalist” project envisioned a system of laws that could be applied by a machine, with perfect reliability and no trace of personal, political, or other bias—and no exercise of the fallible faculty of human judgment.

American legal realists and their forebears subsequently derided this vision of legal formalism as unrealistic “mechanical jurisprudence.”⁶ Such critics have convincingly argued that eliminating value judgments from the legal dispute resolution process is not only impossible but also undesirable.⁷ While the strong formalist view posits that much of judging can and should be turned over to interpretation machines,⁸ nearly every modern scholar who has considered the actual implications of such an approach has concluded that it is both simplistic and unrealistic given the complex and central role of law in managing contested societal values.⁹

Still, the formalist vision persists today. It has retained special allure in the realms of statutory and constitutional interpretation, where textualism and originalism are both routinely advocated as methods for increasing the determinacy of law, while also purportedly reducing the role of subjective

5. See, e.g., JEREMY BENTHAM, AN INTRODUCTION TO THE PRINCIPLES OF MORALS AND LEGISLATION (London, T. Payne & Son 1789); C. C. LANGDELL, SUMMARY OF THE LAW OF CONTRACTS 20–21 (Boston, Little, Brown, and Co. 2d ed. 1880).

6. Roscoe Pound, *Mechanical Jurisprudence*, 8 COLUM. L. REV. 605 (1908). We use the terms “formalism” and “realism” as rough shorthand terms for broad tendencies that cut across many areas of law. By “formalists,” we mean those who emphasize legal determinacy, either descriptively or normatively. By “realists,” we mean those who emphasize legal indeterminacy and the role of moral or political judgment in law. This is obviously a spectrum, rather than a stark dichotomy. The body of the paper will get much more specific about particular variants of these tendencies that are most relevant to our discussion.

7. See, e.g., *id.*; Harry Surden, *The Variable Determinacy Thesis*, 12 COLUM. SCI. & TECH. L. REV. 1 (2011) (summarizing the literature, and arguing that while many legal contexts are indeterminate others are comparatively more determinate and therefore more amenable to computational analysis). More determinate contexts have been largely overlooked by scholars. *Id.*

8. See, e.g., Anthony D’Amato, *Can/Should Computers Replace Judges?*, 11 GA. L. REV. 1277 (1977); see also Surden, *supra* note 7 (summarizing the literature).

9. See Surden, *supra* note 7.

judgment in the interpretive process. This is not the only ground on which these approaches are defended,¹⁰ but it remains an important one—both in the academic literature¹¹ and the broader public discourse.¹² The point looms especially large in constitutional interpretation, where the decisions of life-tenured judges interpreting a Delphic and ancient text can be reversed only by subsequent judicial decision or the extraordinarily onerous amendment process outlined in Article V.¹³

In recent years, formalists have been buoyed by the development of big data techniques for analyzing patterns in the historical usage of words. According to proponents, this sort of “corpus linguistics” analysis can make constitutional originalism “more accurate and credible. It can be more rigorously empirical and transparent. It can fully enter the twenty-first century to enable us to better reach back to the past.”¹⁴ But corpus linguistics, while interesting and sometimes informative, does not materially reduce the normative choices inherent to adjudication. Rather, like the use of dictionaries or historical surveys today, such processes still necessarily present legal officials with a wide range of decisions among distinct selections of sources, definitions, contextual frameworks, interpretations, contemporary translations, and applications. And of course, the choice to employ corpus linguistics is itself a judgment requiring normative justification. Thus, this approach is not, in any robust sense, the kind of objective, value-free

10. See, e.g., William Baude, *Originalism as a Constraint on Judges*, 84 U. CHI. L. REV. 2213, 2214, 2221 (2017) (discussing the historical centrality of constraint as a justification for originalism and textualism and the shift of modern originalism away from this view).

11. See, e.g., Lawrence B. Solum, *The Constraint Principle: Original Meaning and Constitutional Practice* (Apr. 3, 2019) (unpublished manuscript), <https://ssrn.com/abstract=2940215> [<https://perma.cc/XUH5-P4YB>].

12. See, e.g., Carl Hulse, *Architects of the Trump Supreme Court See Culmination of Conservative Push*, N.Y. TIMES (July 3, 2024), <https://www.nytimes.com/2024/07/03/us/politics/trump-supreme-court-conservative-push.html> [<https://perma.cc/6HGX-CQET>] (quoting former White House Counsel Don McGahn describing his goal “to get judges in place who were actually going to read the law as it was written”); Joel Alicea, *Originalism and the Rule of the Dead*, 23 NAT’L AFFS. 149 (2015) (offering a popular defense of originalism emphasizing constraint).

13. U.S. CONST. art. V.

14. James C. Phillips et al., *Corpus Linguistics & Original Public Meaning: A New Tool to Make Originalism More Empirical*, 126 YALE L.J. F. 20, 31 (2016).

interpretation machine that could realize the age-old formalist dream of automating judging.¹⁵

Enter ChatGPT and other LLM systems—a new form of generative AI that has received tremendous attention since the public launch of ChatGPT, an interface that used the GPT-3.5 model, in late 2022.¹⁶ In just two short years, these models have improved and proliferated at an astonishing pace. With some important caveats, they are now capable of outperforming most humans at many complex cognitive tasks, including the bar exam and medical licensing exams.¹⁷ Unlike corpus linguistics or earlier legal informatics tools, such as TurboTax, which were capable of performing computational analysis of basic legal questions in narrow domains, modern LLMs can now perform a variety of complex interpretive functions across a variety of contexts. They can answer constitutional questions posed in ordinary language. They can write fluent and plausible-sounding opinions, and they can explain what a particular constitutional or statutory provision would have communicated to a particular audience in a particular communicative context. This is in stark contrast to corpus linguistics, which largely provided statistics about language usage patterns without deep interpretive insight.¹⁸ With textualists and originalists increasingly recognizing the essential contribution of context to linguistic meaning, these advancements in LLM technology seem to be enormously significant.¹⁹

15. See, e.g., Arbel & Hoffman, *supra* note 3; Jonathan H. Choi, *Measuring Clarity in Legal Text*, 91 U. CHI. L. REV. 1, 1–2 (2024).

16. *Introducing ChatGPT*, OPENAI (Nov. 30, 2022), <https://openai.com/index/chatgpt> [<https://perma.cc/6K95-6E7N>].

17. See Joseph J. Avery et al., *ChatGPT, Esq.: Recasting Unauthorized Practice of Law in the Era of Generative AI*, 26 YALE J.L. & TECH. 64, 69–70 (2023); Ali Abbas et al., *Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions*, 16 CUREUS J. MED. SCI., no. 3, Mar. 11, 2024, <https://www.cureus.com/articles/203719-comparing-the-performance-of-popular-large-language-models-on-the-national-board-of-medical-examiners-sample-questions#!> [<https://perma.cc/85UC-KFTD>].

18. See Arbel & Hoffman, *supra* note 3; see also Choi, *supra* note 15 (explaining the inability of corpus linguistics to account for context).

19. See, e.g., Lawrence B. Solum, *Pragmatics and Textualism* (July 1, 2024) [hereinafter Solum, *Pragmatics*] (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4881344 [<https://perma.cc/82SZ-6EV9>]; Lawrence B. Solum, *The Public Meaning Thesis: An Originalist Theory of Constitutional Meaning*, 101 B.U. L. REV. 1953 (2021) [hereinafter Solum, *Public Meaning Thesis*].

Is ChatGPT the interpretation machine that formalists have been dreaming of for two hundred years? Will it finally bring to fruition a mechanical jurisprudence, untainted by human subjectivity, from an obvious impossibility into a practical reality? In a polarized age, will LLMs allow us to transfer authority over divisive questions of abortion rights, free speech, and federal regulatory power from partisan-identified judges to dispassionate algorithms? While a burgeoning literature examines the relevance and transformative potential of LLMs across different areas of law, there has been no sustained examination of their use in constitutional interpretation.²⁰ This Article takes up that task.

Part I offers a brief overview of AI, with a particular focus on LLMs, and constitutional interpretation. We explain what LLMs are, how they work, and how this new technology is likely to evolve going forward. We then summarize the debate between constitutional formalists and realists over the appropriate place of value judgment and judicial subjectivity in constitutional interpretation.

Part II surveys the potential uses of LLMs in constitutional interpretation—ranging from research and drafting assistant to critic and troubleshooter to ultimate decider of cases. The main takeaway is that the pros and cons of LLMs in this context cannot be intelligently evaluated without careful consideration of how they will be used. Different use cases raise different issues.

Part III explores the potential benefits and drawbacks of using LLMs in constitutional interpretation. Potential benefits include speed, efficiency, and cost; superior research, writing, and analytic capabilities relative to human or other technological alternatives; and greater objectivity in the sense of freedom from personal, political, or other bias. Potential drawbacks include opacity; ingrained bias; manipulability and prompt sensitivity; the difficulty of standardizing LLM use across cases and judges; the variety of high-quality LLMs and consequent need to choose among them; and the stochastic or random character of LLM outputs, which can vary from day to day or even minute to minute and with different parameter settings. The balance of these costs and benefits will vary across the use cases identified in Part II and with institutional context.

20. See Surden, *supra* note 3.

LLMs hold the greatest potential as a quick reference in contexts where speed and efficiency are at a premium, such as busy lower courts with huge caseloads, and for relatively straightforward constitutional questions. But they are not capable of eliminating human subjectivity or value judgment from difficult constitutional cases; at most they will shift the location of that bias from one stage of the decision process—and perhaps from one human decision-maker—to another.

Part IV presents the results of a simple simulation in which we posed the questions presented in *Dobbs v. Jackson Women's Health Organization*²¹ and *Students for Fair Admissions v. President & Fellows of Harvard College*²² to ChatGPT using GPT-4 and Claude 3 Opus, another high-quality LLM.²³ Results were impressively consistent across models but highly sensitive to variation in prompts, illustrating the importance of question-framing in determining outputs. Both Claude 3 Opus and ChatGPT using GPT-4 were also highly sensitive to counterarguments, reversing themselves in every case based on standard arguments that any first-year law student could formulate. Experts refer to this phenomenon of LLMs tailoring their outputs to match user preferences as “AI sycophancy,”²⁴ and it raises serious questions about the reliability and malleability of LLMs as constitutional interpreters. More generally, the extent to which human inputs drive LLM outputs suggests that the use of LLMs for constitutional interpretation will implicate analogous theoretical issues that today confront human constitutional interpreters.

Part V assesses the implications of our analysis for the use of LLMs in constitutional interpretation and future research. Two implications stand out: The first is the importance of attending carefully to particular use cases and institutional contexts. LLMs hold more potential for relatively modest uses, such as research and editorial assistance, and for institutional

21. *Dobbs v. Jackson Women's Health Org.*, 597 U.S. 215 (2022).

22. *Students for Fair Admissions v. President & Fellows of Harvard Coll.*, 600 U.S. 181 (2023).

23. As an illustration of how rapidly LLMs are advancing, updated versions of both ChatGPT and Claude were released as this Article was going to press—ChatGPT using GPT-4o and Claude 3.5 Sonnet.

24. See, e.g., Carson Denison et al., *Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models*, ARXIV (June 29, 2024), <https://doi.org/10.48550/arXiv.2406.10162> [<https://perma.cc/U67M-HRR7>].

contexts with significant resource constraints where decisions have limited impact. The second is that there is no avoiding the burdens of judgment. For any given task, LLMs may be better or worse than humans, but the choice to use them is itself a value judgment. Even more importantly, the outputs that LLMs produce are highly sensitive to the way that interpretive questions are framed. This naturally raises the question of how prompts *should* be framed, which in turn raises most, if not all, of the normative questions that have long bedeviled constitutional theory.

Three prefatory notes are in order before we begin:

First, LLMs are a powerful and useful tool. Our measured skepticism of their potential to transform constitutional interpretation should not be understood as pathologizing LLMs or, conversely, as romanticizing human decision-makers, who have plenty of limitations of their own. The choice between LLMs and humans, at any given margin, is—like any choice—comparative.²⁵ For some tasks, humans will be better at maximizing a particular normative end. For others, LLMs will be better. Which tasks fall into each of these categories will almost certainly change over time.

Second, our main focus is not whether humans or LLMs are better constitutional interpreters. Our focus is whether LLMs offer a plausible hope of avoiding the burdens of normative judgment. They do not. In defending this position, we propose a “law of conservation of judgment.” Like matter or energy, normative judgment in constitutional interpretation can be shifted around, dispersed, or concentrated. It might be transferred from one decision-maker or one stage in the decision-making process to another. But when it is squeezed out of one part of the interpretive process, it inevitably pops up somewhere else.²⁶

Third, there is an ever-present temptation when writing about a revolutionary technology like LLMs to drift into highly

25. See generally Andrew K. Woods, *Robophobia*, 93 U. COLO. L. REV. 51 (2022) (observing that much apprehension about AI, robots, and algorithms is driven by failure to compare these technologies with human agents, whose performance is often far worse).

26. The analogy to the laws of conservation of matter and energy is obviously rough and suggestive rather than precise. Among other differences, the normative judgment exercised by different actors or entities within a legal system cannot readily be quantified.

speculative analysis bordering on science-fiction. It is distinctly possible that LLMs will evolve in ways that make the discussion of this Article—and perhaps the very concepts of judges, courts, and the other law-making and law-enforcing institutions created by the U.S. Constitution—seem quaint and outmoded. But we suspect that possibility is quite a few years off. In any case, it is very difficult to say anything concrete or helpful about a future whose contours we can imagine so hazily. Our analysis will therefore stick fairly close to the ground, focusing on LLMs in their current form while keeping in mind the likelihood of rapid, if still incremental, near-term improvement in their capabilities.

Our overarching goal is to initiate a conversation between experts in constitutional interpretation and experts in AI. As such, we have endeavored to make our discussion accessible to specialists across these fields, as well as merely curious observers from outside them. The downside of this approach is that it requires us to include a certain amount of basic information that will be obvious to some readers. We think this a price worth paying to make our Article accessible to the broadest possible readership.

I. WHAT IS ARTIFICIAL INTELLIGENCE? WHAT IS CONSTITUTIONAL INTERPRETATION?

This Part sets up the remainder of the Article by providing brief overviews of artificial intelligence (AI) and constitutional interpretation. It situates LLMs in the broader historical development of AI and explains how they work and what makes them special. It also explains what is distinctive about constitutional interpretation and the debates between particular variants of formalism and realism that have dominated the field.

A. What is Artificial Intelligence?

There is probably no single definition of AI that most scholars would agree to. However, one practically useful definition of AI is “[u]sing computers to solve problems, make predictions, answer questions, [generate creative output,] or make automated decisions or actions, on tasks that when done

by people, typically require ‘intelligence.’”²⁷ In this view, we can think of AI in terms of particular tasks that we associate with human intelligence, and whether we are able to fully or partially automate these tasks using computers.

The word “intelligence” itself is similarly difficult to define but is usually associated with one or more high-order human cognitive skills, such as abstract reasoning, problem solving, learning, visual comprehension, language understanding or creation, creativity, planning, and critical thinking.²⁸ Activities that people do routinely, like reading a book, playing chess, solving a math problem, or driving a car, require one or more of such advanced cognitive processes. Thus, if we are able to get a computer to partially (or fully) automate such an activity that normally requires advanced cognitive processes when people do them, we can consider that activity an “artificial intelligence task.”²⁹

An important point is that, although computers today can perform various tasks normally associated with human cognition, the underlying techniques that computers use are very different from biological, human cognitive processes. Rather, such AI approaches typically involve statistics, rules, or heuristics to produce useful and intelligent-seeming results through mechanisms that are quite distinct from human intelligence. This is important to emphasize, because there is a tendency to anthropomorphize modern AI systems when they can perform certain tasks at levels that meet or exceed human performance.³⁰ The point is especially relevant in societal domains, like law, that have historically been the province of human judgment.

1. Advances in Artificial Intelligence

The remarkable progress in AI since about 2022 deserves particular attention. Notably, there have been improvements in

27. Surden, *supra* note 3, at 1944 (quoting 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 26).

28. *Id.*

29. *Id.*

30. This is, in fact, almost unavoidable when talking about LLMs, and we use a fair amount of anthropomorphic language in this Article. Such language should always be understood figuratively, not literally.

the ability of AI systems to understand and create language and, more broadly, to reason and problem solve in areas associated with human knowledge, surpassing earlier AI iterations.³¹ Because AI has a long history extending back to at least the 1950s,³² it is helpful to survey how AI techniques have evolved to understand why AI previously struggled with abstract areas rooted in language, such as law, and see how recent changes have led to the modern era of more broadly capable LLMs, such as GPT-3.5 or newer, that are the focus of this Article.

Starting in the 1950s and continuing through the 1980s, AI was largely focused upon “computer rules” and “knowledge representation.”³³ The goal of such knowledge representation AI systems was to represent or *model* different aspects of the world, using expert knowledge manually encoded in formal programming languages that computers could easily process.³⁴ For example, in medicine, such systems aimed to codify the diagnostic knowledge and processes of doctors into formal computer rules, allowing computers to sometimes deduce nonobvious diagnoses. Although this early symbolic AI approach achieved some successes, its limitations became quickly apparent: Hand-coded expert rules about law, medicine, or other phenomena were often “brittle” in the sense that they couldn’t handle exceptions, nonstandard “hybrid” scenarios, discretion, or nuances. Many social or natural phenomena were just too complex to be modeled manually in terms of lists of general rules and exceptions.

Spurred in part by these limitations, a new approach to AI became popular starting in the 1990s: machine learning.³⁵ The machine-learning approach contrasted starkly with knowledge representation. Rather than relying on experts to encode rules, machine-learning approaches use algorithms to infer rules automatically from patterns identified in large datasets.³⁶ For

31. Louis Castricato et al., *Suppressing Pink Elephants with Direct Principle Feedback*, ARXIV 1–2 (Feb. 13, 2024), <http://arxiv.org/abs/2402.07896> [<https://perma.cc/CE7M-L6U8>]; Yangfan Hu et al., *Toward Large-Scale Spiking Neural Networks: A Comprehensive Survey and Future Directions*, ARXIV (Aug. 19, 2024), <http://arxiv.org/abs/2409.02111> [<https://perma.cc/N46N-5UPL>].

32. STUART J. RUSSELL & PETER NORVIG, *ARTIFICIAL INTELLIGENCE: A MODERN APPROACH* 17–18 (Stuart Russell & Peter Norvig eds., 4th ed. 2021).

33. *Id.* at 22–23.

34. *Id.* at 22–24.

35. *Id.* at 24–25.

36. *Id.* at 24–26.

example, instead of coming up with a list of words that experts think are likely to occur in spam, machine-learning systems detect spam by analyzing large datasets of user-marked spam emails. This enables them to identify phrase patterns highly associated with spam emails through statistical inference.³⁷ Such data-oriented, statistical approaches turned out to be much more effective for many real-world tasks because naturally occurring patterns in data better capture the complexities of changing phenomena than expert-crafted lists of computer rules.³⁸

Around 2010, one particular machine-learning technique began to show remarkable results: deep “neural networks.”³⁹ Neural networks were originally developed in the 1940s and 1950s, and the approach received its name because it was very loosely inspired by a simple model of how human brain neurons work.⁴⁰ Neural networks are computer algorithms that process information through layers of connected nodes, learning from examples to recognize patterns and solve complex problems. By the 1980s, neural networks had largely fallen out of favor with the AI research community because of some perceived limitations.⁴¹ By 2010, however, researchers were able to revive this long dormant technique due to subsequent theoretical research advances, the increased availability of training data due largely to the Internet, and steadily improving hardware.⁴² The result was an approach known as “deep learning,” which took neural networks and scaled them up to sizes previously unattainable in terms of computation and data. This scaling made deep neural networks much more effective for a wide range of tasks than earlier machine-learning techniques had been. From 2010 to 2020, deep learning neural network systems were able to engage in tasks such as automated language translation, autonomous driving of cars, prediction, image recognition, and playing games like chess, at previously unprecedented levels.

37. *Id.* at 24–25, 826 (using email spam as an applied example of the N-gram model).

38. *Id.* at 25–26.

39. *Id.* at 26.

40. *See id.* at 17.

41. *See id.* at 23–24.

42. *See id.* at 26.

Notwithstanding these major advances, deep learning during this period continued to struggle with human language.⁴³ The subfield of AI that focuses on understanding written and spoken language is known as “natural language processing” (NLP).⁴⁴ The term “natural” in “natural language” is meant to refer to the ordinary languages that people use to communicate, such as English, Spanish, or Japanese, as opposed to formal “languages,” such as Python, C, Java, or Prolog, that are used to encode information and program computers.⁴⁵ Creating robust AI systems that could reliably read and “understand” any human-written document or answer complicated written questions eluded many prior AI machine-learning and knowledge-representation attempts since the 1950s. Similarly, despite the success of deep learning in other domains, most deep learning NLP systems of this era were still unable to understand and sensibly respond to most natural-language questions.⁴⁶

A new AI era began in November 2022 with OpenAI’s release of ChatGPT: a user-friendly interface built on the GPT-3.5 model.⁴⁷ Much to the surprise of most AI researchers, this was the first NLP AI system that could sensibly react to and analyze just about any textual input or document.⁴⁸ ChatGPT was an example of an accessible way to interface with a *large language model* (LLM), a type of NLP AI system that was designed to generate coherent, seemingly human-written text.⁴⁹ LLMs like GPT-3.5 were created using new deep learning architectures, which enabled them to analyze much larger amounts of text within which to detect and learn the patterns of human language. These AI models learned to understand and generate language in a way that closely simulated human writing through “training” on billions of pages of previously

43. *Id.* at 26–27.

44. See Prakash M. Nadkarni et al., *Natural Language Processing: An Introduction*, 18 J. AM. MED. INFORMATICS ASS’N 544, 544–45 (2011).

45. RUSSELL & NORVIG, *supra* note 32, at 252.

46. Jianyang Deng & Yijia Lin, *The Benefits and Challenges of ChatGPT: An Overview*, 2 FRONTIERS COMPUTING & INTELLIGENT SYS. 81 (2023).

47. See *Introducing ChatGPT*, *supra* note 16.

48. Katikapalli Subramanyam Kalyan, *A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4*, ARXIV (Feb. 27, 2023), <https://arxiv.labs.arxiv.org/html/2310.12321> [<https://perma.cc/B3WX-6M53>].

49. Sheng Lu et al., *Are Emergent Abilities in Large Language Models Just In-Context Learning?*, ARXIV (July 15, 2024), <https://arxiv.org/abs/2309.01809v2> [<https://perma.cc/48CT-58DS>].

human-written pages available on the Internet and elsewhere, such as Wikipedia, news websites, books, federal and state statutes, court decisions, contracts on sites like EDGAR, and legal motions. For the first time, ChatGPT using the GPT-3.5 model was an AI system able to produce intelligent and pertinent responses to nearly any query or instruction. Remarkably, it did so merely by predicting and producing one word at a time as it incrementally created its answers based on the prompt given and the words the system had already generated in the response. It employed the intricate patterns of human language and knowledge about the world that it had learned during its earlier training.

To be clear, GPT-3.5 was not always accurate in its responses or analysis—it suffered from well-known accuracy problems and a tendency to make up facts—a process known as “hallucination.”⁵⁰ But factual accuracy was not even the biggest technical hurdle for such AI systems prior to that time.⁵¹ Rather, going back to November 2022, LLMs prior to GPT-3.5 had much more severe limitations. These systems could not even respond sensibly to misleading or mistaken user inputs such as “What is the capital of Paris?” that were too far outside of their training.⁵² So, even though GPT-3.5 made factual and reasoning errors, what astonished AI researchers was that it could analyze and respond to arbitrary text of any kind—texts, contracts, computer code—sensibly *at all*.⁵³

GPT-3.5 could also perform basic reasoning and problem solving.⁵⁴ This was extremely surprising, as such reasoning and problem solving appeared to be an “emergent” property. GPT-3.5 was not initially trained to perform these tasks; rather, these capabilities emerged as a byproduct of its use of deep learning to detect patterns in millions of existing, human-written web pages

50. See generally Hongbin Ye et al., *Cognitive Mirage: A Review of Hallucinations in Large Language Models*, ARXIV 1–2 (Sept. 13, 2023), <http://arxiv.org/abs/2309.06794> [<https://perma.cc/3LRQ-MVMX>] (discussing hallucinations in LLMs).

51. *Id.*

52. Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, PROC. OF THE 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 610 (2021), <https://dl.acm.org/doi/10.1145/3442188.3445922> [<https://perma.cc/QZE6-GC54>].

53. *Id.*

54. Lu et al., *supra* note 49.

and documents.⁵⁵ ChatGPT using GPT-3.5 thus appeared to be the first AI LLM system to “understand” the underlying meaning of most articles, documents, questions, or other texts with any reliability.

The unexpected natural language capabilities of GPT-3.5 were the result of several earlier engineering advances from 2017 to 2022.⁵⁶ One of the most significant was the “transformer architecture”—a neural-network-based breakthrough developed by Google in 2017—which allowed AI systems to understand the context of what was being asked.⁵⁷ In language, meaning is often heavily influenced by the surrounding words, and people rely upon context words for nuances of meaning. For example, if one were to ask, “What is a crane?” the answer would depend upon whether the preceding context words had alluded to construction sites, in which case the question would refer to the construction machine. By contrast, had the earlier context referred to wetlands and flying, a human reader would understand the word “crane” to refer to the bird.

Understanding the context of language—the surrounding words and linguistic setting—is key to natural language comprehension. But, for various technical reasons, it had previously proved challenging for AI systems to incorporate context, particularly for longer texts.⁵⁸ This limitation inhibited the ability of earlier AI systems to understand ordinary text. Google’s transformer architecture enabled AI systems to “see” the entirety of a text and what was being asked and to mathematically incorporate the meaning of surrounding context words.⁵⁹ This innovation dramatically improved automated language understanding. As we will discuss, the design of the transformer architecture means that the presence of particular context words used in a prompt or question can influence the model’s response to a question in subtle and surprising ways. In creating GPT-3.5, OpenAI built upon Google’s transformer architecture and added several engineering advances of its own.

55. *Id.*

56. *Id.*

57. Ashish Vaswani et al., *Attention Is All You Need*, ARXIV, <http://arxiv.org/abs/1706.03762> [<https://perma.cc/SVJ4-66KB>] (last updated Aug. 2, 2023).

58. *Id.*

59. *Id.*

As a result, GPT-3.5 exhibited unprecedented generality and usefulness compared to earlier NLP AI systems.⁶⁰

Surprisingly, just four months later, in March of 2023, OpenAI released an even more capable model to run ChatGPT: GPT-4.⁶¹ GPT-4 could not only do everything that GPT-3.5 could do, it exceeded the earlier model's ability to analyze text, reason, perform analysis of documents and create text, at levels never before seen in AI research.⁶² As of the writing of this Article in mid-2024, GPT-4 and its variants remain the state of the art, and the model has steadily improved in capability since then. Capable competitors such as Claude 3.5 from Anthropic, Gemini from Google, and Llama 3 from Meta have emerged in the interim.⁶³ These competitor, GPT-4 class models are similarly able to “understand” and respond appropriately to a nearly infinite variety of human text.⁶⁴

While technologically unfeasible just two years ago, today judges and others can use AI systems to seek legal analysis and answers about constitutional, statutory, case law, and nearly any other legal question. Modern AI systems usually respond with coherent, well-reasoned, and persuasive text. But *should* judges and other legal actors be doing this? The answer to this question is complex, but this Article strongly counsels caution. One obvious limitation of LLMs is accuracy. As of mid-2024, systems occasionally hallucinate.⁶⁵ There are, however, reasons to believe that technical advances will significantly reduce this problem in the near future.⁶⁶ But putting accuracy aside, there are more significant jurisprudential issues that warrant

60. *Introducing ChatGPT*, *supra* note 16; Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, ARXIV (Mar. 4, 2022), <http://arxiv.org/abs/2203.02155> [<https://perma.cc/9QEY-G85U>]; Walid Hariri, *Unlocking the Potential of ChatGPT: A Comprehensive Exploration of Its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing*, ARXIV (Apr. 17, 2024), <https://arxiv.org/abs/2304.02017> [<https://perma.cc/RB9U-YE6L>].

61. *GPT-4 Is OpenAI's Most Advanced System, Producing Safer and More Useful Responses*, OPENAI, <https://openai.com/product/gpt-4> [<https://perma.cc/TV7S-QVM7>].

62. OpenAI, *GPT-4 Technical Report*, ARXIV (Mar. 4, 2024), <http://arxiv.org/abs/2303.08774> [<https://perma.cc/7WBU-SCZK>].

63. See Hung Phan et al., *RAG vs. Long Context: Examining Frontier Large Language Models for Environmental Review Document Comprehension*, ARXIV (July 10, 2024), <http://arxiv.org/abs/2407.07321> [<https://perma.cc/JZ5H-F7F4>].

64. *Id.*

65. Ye et al., *supra* note 50.

66. *Id.*

prudence. LLMs *appear* to be providing confident, well-reasoned, and objectively neutral answers to legal questions. However, behind the scenes, the responses that AI systems produce are in fact masking important policy choices for society.

2. A Simple Illustration

To illustrate what this might look like in practice, consider a simple example. A judge deciding a case might ask an advanced LLM to interpret the infrequently invoked, and relatively noncontroversial, Third Amendment, which states: “No Soldier shall, in time of peace be quartered in any house, without the consent of the Owner, nor in time of war, but in a manner to be prescribed by law.”⁶⁷ To play this out, we posed the following prompt to ChatGPT using the GPT-4o model: “Under the 3rd Amendment, is the Governor of the State of Colorado barred from quartering themselves in my house without my permission?”

ChatGPT using GPT-4o responded:

The Third Amendment specifically prohibits the quartering of soldiers in private homes without the owner’s consent during peacetime and outlines the conditions for such actions during wartime. Since the governor is not a soldier, this scenario does not fall under the scope of the Third Amendment. However, even though this scenario is not covered on the 3rd Amendment, there are likely other state, local, or Federal regulations that would prevent the Governor from quartering themselves in your house without your permission.⁶⁸

As the reader can see, GPT-4o provides a well-reasoned, straightforward and apparently thoughtful response, probably along the lines of responses that some similarly situated lawyers would offer. According to this response, the Third Amendment does not cover the governor of Colorado. (This is a remarkably coherent response compared to what could have been produced

67. U.S. Const. amend. III.

68. Andrew Coan & Harry Surden, *Quartering the Governor Under the Third Amendment, ChatGPT*, OPENAI (unpublished prompt) (on file with authors).

by LLMs only two years earlier.) We could imagine this being persuasive to an inquisitive judge attempting this same query in their chambers. But an important nuance is revealed when we pose this same exact query to an equally advanced but different LLM, Claude 3 Opus from Anthropic.

Claude 3 Opus produces a comparably well-reasoned, but directionally different response:

Yes, the Third Amendment to the U.S. Constitution would prohibit the Governor of Colorado (or any other government official) from quartering themselves in your private home without your consent, even during peacetime. While the amendment specifically refers to “soldiers,” it has been interpreted to apply to any agent of the government, not just military personnel. The Third Amendment was adopted in response to the Quartering Acts passed by the British Parliament during the colonial era, which allowed British soldiers to take residence in private homes. Although rarely invoked, the amendment still stands as a protection against government intrusion into private homes and a recognition of the fundamental right to privacy in one’s own residence.⁶⁹

This simple example illustrates some of the risks of using LLMs in constitutional interpretation, or in law more broadly, without proper AI literacy and a strong understanding of the models’ limitations. Both AI models gave well-reasoned, plausible, and *confident* constitutional interpretations to the same legal issues but came to different legal conclusions. How did this happen? During the underlying computations, each AI model implicitly made many subtle, substantive, and value-laden interpretive choices, and their internal probabilistic computational mechanisms guided them in different directions.

Both responses focused on the meaning of the word “soldier.” GPT-4o implicitly followed a more *textualist* interpretative approach, adopting a more literal (and perhaps historical) definition of the word “soldier” as limited to members of the armed forces. Presumably, this definition led the model to the conclusion that a state governor, while an agent of a government, was not *literally* a member of the armed forces, and

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

therefore not a “soldier” under a literal interpretation of the Third Amendment. By contrast, Claude 3’s algorithm followed a different computational path, adopting a broader interpretation of the word “solider” as any agent of the government. This nonliteral or purposive interpretation reflected the Amendment’s original goal of safeguarding private homes against forced government intrusion.

Of course, this example is not meant to illuminate the meaning of the Third Amendment. Rather, it illustrates how highly capable LLMs can make normatively significant legal judgments that may not be obvious to most users. Without a sophisticated view of the mechanics and limits of the technology, judges and other legal actors might be easily lulled into thinking that LLMs are providing “correct” and “objective” answers to their questions, when in fact those answers often reflect important interpretive choices that have traditionally been the province of judges. These choices are subtle and not easy to detect, even by those trained in AI, let alone nontechnical judges and regulators who may just be casually using AI systems in their daily work.

Today, it is easy for judges and lawyers to access a modern LLM like ChatGPT powered by GPT-4 or Claude 3 powered by the Opus model to perform constitutional interpretation. In some circumstances, and in perhaps many, it will also be attractive for judges and lawyers to use LLMs in this way, given their ability to produce informative, confident, and apparently well-reasoned responses almost instantly. But is this an appropriate path? To answer this question intelligently, we need to understand what constitutional interpretation entails and what distinguishes it from other forms of interpretation and legal reasoning.

B. Constitutional Interpretation

Constitutional interpretation is the method by which courts—paradigmatically, the U.S. Supreme Court—decide constitutional questions.⁷⁰ It is distinct from other forms of

70. Some scholars use interpretation in a narrow, technical sense, to refer to the search for the Constitution’s original public meaning or communicative content, as opposed to its legal content. *See, e.g.*, Lawrence B. Solum, *The Interpretation-Construction Distinction*, 27 CONST. COMMENT. 95 (2010). However, we adopt the broader usage which encompasses all techniques for discovering,

interpretation and judicial decision-making in a number of important ways, most notably the age and vagueness of the Constitution, the difficulty of amending it (or overriding a judicial decision interpreting it), and the public salience of many of the issues it implicates. These distinctive features have given the broader debate between formalist and realist theorists a greater intensity and a somewhat different shape in the context of constitutional interpretation than it has elsewhere. This debate largely revolves around the kind of difficult constitutional questions that make their way to the U.S. Supreme Court. It has a different complexion in the lower courts and in constitutional disputes that never make it to court, which often have clearer and more determinate answers.

1. What Makes Constitutional Interpretation Distinctive

The Constitution is old, vague, difficult to formally amend, and has been interpreted to address many of the most controversial issues in American society. On top of that, judicial review of legislative action gives the final authority to interpret the Constitution to life-tenured federal judges, culminating in the U.S. Supreme Court, which has declared itself “the ultimate expositor”⁷¹ of the Constitution, which under the Supremacy Clause of Article VI is “the supreme Law of the Land.”⁷² This combination of factors is a recipe for more intense controversy over how the Constitution is interpreted than we see in any other legal domain. That controversy, in turn, gives the dream of a formalist interpretation machine that can resolve legal disputes objectively without the exercise of normative judgment, an especially strong allure.

a. An Ancient Text and the Dead Hand of the Past

The Constitution’s age creates powerful countervailing pressures well captured by two landmark U.S. Supreme Court

assigning, and resolving the meaning of the Constitution, linguistic, legal, or otherwise, in the course of deciding a constitutional case.

71. See, e.g., *United States v. Morrison*, 529 U.S. 598, 616 n.7 (2000) (explaining that “ever since *Marbury* this Court has remained the ultimate expositor of the constitutional text”).

72. U.S. CONST. art. VI, cl. 2.

decisions: *McCulloch v. Maryland*⁷³ and *Trop v. Dulles*.⁷⁴ In upholding Congress's power to charter a national bank in *McCulloch*, Chief Justice Marshall famously observed that a Constitution "intended to endure for ages to come" needed to be interpreted flexibly to ensure that the political branches can respond to "the various crises of human affairs."⁷⁵ This imperative of legislative flexibility to deal with unanticipated problems has been the major engine behind the Supreme Court's expansive interpretation of the federal spending and commerce powers, as well as some important decisions permitting experimentation in the realm of separation of powers.⁷⁶

On the other hand, the Constitution's age has also created a persistent felt need to keep the Constitution's individual rights protections in tune with contemporary values. The point of constitutionalizing rights, after all, is generally understood to be placing exceptionally important interests beyond the power of democratic majorities that control the government. But the understanding of which interests and values require this protection has changed greatly over time. In holding that a punishment of denationalization for military desertion violated the Eighth Amendment, *Trop v. Dulles* famously held the meaning of "cruel and unusual punishment" as "not static. The Amendment must draw its meaning from the evolving standards of decency that mark the progress of a maturing society."⁷⁷ This evolutionary approach to individual rights has been the major engine behind landmark decisions on race and sex discrimination, same-sex intimacy and marriage, freedom of speech, and more;⁷⁸ though the U.S. Supreme Court's recent turn to "history and tradition" on abortion and gun rights appears to represent a major shift away from this approach.⁷⁹

73. *McCulloch v. Maryland*, 17 U.S. 316 (1819).

74. *Trop v. Dulles*, 356 U.S. 86 (1958).

75. *McCulloch*, 17 U.S. at 415.

76. See, e.g., Andrew Coan & David S. Schwartz, *The Original Meaning of Enumerated Powers*, 109 IOWA L. REV. 971 (2024); DAVID S. SCHWARTZ, *THE SPIRIT OF THE CONSTITUTION: JOHN MARSHALL AND THE 200-YEAR ODYSSEY OF MCCULLOCH V. MARYLAND* (2019).

77. 356 U.S. at 101.

78. See, e.g., DAVID A. STRAUSS, *THE LIVING CONSTITUTION* (2010); *Obergefell v. Hodges*, 576 U.S. 644 (2015).

79. *United States v. Rahimi*, 144 S. Ct. 1889 (2024); *Dobbs v. Jackson Women's Health Org.*, 597 U.S. 215 (2022); *Bruen v. N.Y. State Rifle & Pistol Ass'n*, 597 U.S. 1, 26 (2022).

The U.S. Supreme Court's recent Second Amendment decisions illustrate a final way in which the Constitution's great age influences debates over its interpretation. In *Bruen v. New York State Rifle & Pistol Ass'n*, the U.S. Supreme Court relied on the history and tradition surrounding the adoption of the Second Amendment to limit the legislative power of a contemporary state government to regulate handguns.⁸⁰ This is the inverse of the flexibility emphasized in *McCulloch* and amounts to a highly controversial acquiescence to the dead hand of a distant past, whose values, circumstances, and assumptions about the world differed vastly from those of present-day Americans.⁸¹ In contrast to decisions like *Dobbs*, which rely on the history and original meaning of an ancient text to roll back judicial limitations on legislative power, decisions like *Bruen* invoke that history to impose judicial limitations on legislative power. This muscular originalist approach—sometimes called “judicial engagement”⁸² or, more pejoratively, “conservative judicial activism”⁸³—is the major engine behind the recent series of decisions limiting congressional restrictions on the president's removal power over the heads of independent agencies.⁸⁴ It is also the engine most likely to drive a renewed conservative assault on the federal commerce power and the delegation of rulemaking power to administrative agencies, should such an assault materialize in the near future.⁸⁵

*b. Difficulty of Amendment, Glittering Generalities,
and Judicial Review*

The Constitution's age might attract less attention if it were easier to amend. But Article V requires a two-thirds majority of both Congressional houses just to propose an amendment, and the approval of three-quarters of state legislatures to actually

80. *Bruen*, 597 U.S. at 26.

81. See, e.g., Andrew Coan, *The Dead Hand Revisited*, 70 EMORY L.J. ONLINE 1 (2020); David A. Strauss, *Common Law, Common Ground, and Jefferson's Principle*, 112 YALE L.J. 1717 (2003).

82. See, e.g., Clark Neily, *Litigation Without Adjudication: Why the Modern Rational Basis Test Is Unconstitutional*, 14 GEO. J.L. & PUB. POL'Y 537 (2016).

83. See, e.g., ERIC J. SEGALL, ORIGINALISM AS FAITH (2018).

84. *Seila Law LLC v. Consumer Fin. Prot. Bureau*, 591 U.S. 197 (2020); *Free Enter. Fund v. Pub. Co. Acct. Oversight Bd.*, 561 U.S. 477 (2010).

85. See, e.g., Coan & Schwartz, *supra* note 76; Andrew Coan, *Eight Futures of the Nondelegation Doctrine*, 2020 WIS. L. REV. 141.

ratify it.⁸⁶ Empirical research suggests that few, if any, national constitutions are more resistant to change.⁸⁷ This rigidity has been much criticized by American constitutional theorists. One has gone so far as to identify Article V as the single worst provision of the original constitutional text.⁸⁸

The conventional view is that the extreme difficulty of this amendment process makes originalism, which fixes constitutional meaning at the time the Constitution was written, less attractive, while making living-constitutionalist approaches, which permit constitutional meaning to evolve, more attractive. The purpose of amendment is to fix constitutional problems. As the stringency of amendment procedures increases, the need for substitute mechanisms for fixing problems with the Constitution's original meaning would seem also to increase. One obvious substitute is the process of interpretation, which, exercised creatively, holds the potential to radically reshape the Constitution's practical operation, within capacious outer limits. The more stringent the amendment procedure, the greater the apparent need for this sort of alternative mechanism for addressing constitutional problems. This also makes the choice of *how* to interpret the Constitution more significant. In the limiting case of a perfectly frictionless amendment procedure, interpretive choice would not matter at all, since any decision—however arrived at—could and would be costlessly reversed whenever the amending authority disapproved.⁸⁹

The vagueness of much of the constitutional text—sometimes described as its “glittering generalities”⁹⁰—increases the plausibility of using constitutional interpretation as a substitute mechanism for addressing constitutional problems. It probably also increases the demand for such informal or nontextual amendments by interpretation. It is much easier to square a living-constitutionalist approach to constitutional

86. U.S. CONST. art. V.

87. See, e.g., Andrew Coan & Anuj Desai, *Difficulty of Amendment and Interpretive Choice*, 1 J. INST. STUD. 6 (2015) (collecting sources).

88. Stephen M. Griffin, *The Nominee Is . . . Article V*, 12 CONST. COMMENT. 171 (1995).

89. This paragraph and portions of the previous one are adapted from Coan & Desai, *supra* note 87, where the authors also argue these points and include supporting citations.

90. CHRISTOPHER G. TIEDEMAN, *THE UNWRITTEN CONSTITUTION OF THE UNITED STATES* 79 (New York, G.P. Putnam's Sons 1890).

interpretation with respect for textualism or originalism when the text is open-ended. Examples include the First Amendment's Freedom of Speech Clause, the Fourteenth Amendment's Due Process Clause and Equal Protection Clause, and the Eighth Amendment's Cruel and Unusual Punishment Clause. This was David Strauss's point when he observed that "[t]he genius of the Constitution is that it is specific where specificity is valuable, general where generality is valuable—and that it does not put us in unacceptable situations that we can't plausibly interpret our way out of."⁹¹

The institution of judicial review cuts in a different, arguably opposite, direction. The constitutional text nowhere expressly grants federal courts the power to review the constitutionality of legislative and executive action, but this power is now well established. Indeed, not only do the courts routinely exercise the power of judicial review but they are also customarily understood to have the final and supreme authority to decide constitutional questions.⁹² If this were simply a matter of ministerially executing a clear and uncontroversial set of constitutional commands, ascertainable through objective professional expertise, that would be one thing. But of course, that is not the case. Due to the vagueness and age of the Constitution and the many contentious issues it implicates, constitutional questions often have a range—sometimes, an extremely broad range—of plausible answers. This gives rise to the famous “counter-majoritarian difficulty.”⁹³ Federal judges are appointed rather than elected and serve for life, subject only to removal by impeachment in extreme cases. In a democratic society, can it possibly be legitimate for them to set aside the acts of elected officials on the basis of vague constitutional text?⁹⁴

The well-settled character of judicial review and even judicial supremacy in constitutional interpretation can be

91. Strauss, *supra* note 81, at 1737, 1744.

92. See *Cooper v. Aaron*, 358 U.S. 1, 18 (1958); LARRY D. KRAMER, *THE PEOPLE THEMSELVES: POPULAR CONSTITUTIONALISM AND JUDICIAL REVIEW* 109 (2005) (describing and lamenting this widespread conventional wisdom).

93. *ArtIII.S2.C1.10.3 Counter-Majoritarian Difficulty*, CONSTITUTION ANNOTATED, https://constitution.congress.gov/browse/essay/artIII-S2-C1-10-3/ALDE_00013155 [<https://perma.cc/2XNR-MXQY>].

94. For the classic statement of these familiar difficulties, see ALEXANDER M. BICKEL, *THE LEAST DANGEROUS BRANCH: THE SUPREME COURT AT THE BAR OF POLITICS* 16–19 (1962).

misleading in this regard. While few today dispute courts' power to conclusively decide constitutional questions, there is intense dispute about how they should exercise this power and to what extent, if any, judges should understand themselves as empowered to employ their own faculties of moral and political judgment in deciding constitutional cases.⁹⁵ On the one hand, as Alexander Bickel famously asked, is it not in tension with our democratic commitments for unelected and unaccountable judges to exercise this kind of de facto law-making power over the most hotly contested social and political questions, on which the U.S. Supreme Court Justices disagree along essentially the same lines as the rest of us?⁹⁶ On the other hand, does constitutional democracy not require that certain values and interests—paradigmatically, those of historically disadvantaged minorities—be placed beyond the powers of transient political majorities?⁹⁷ This brings us back to the formalist-realist debate.

2. Constitutional Formalism vs. Constitutional Realism

The version of the debate between formalists and realists that prevails in constitutional law is rooted in the combination of factors discussed in the preceding Section. That debate is far too messy, long-running, and sprawling to describe in detail here. But much of it turns on the importance and practicability of constraining the exercise of moral and political judgment by judges in constitutional cases where their decisions are extremely difficult to reverse. Formalists of various stripes think such constraint is both practicable and important. Realists, also of various stripes, doubt whether it is practicable and deny its desirability.

a. Constitutional Formalism

The interpretive approach most commonly associated with formalism in constitutional law is originalism. Originalism comes in many flavors and is defended on many grounds. But its

95. See, e.g., Andrew Coan, *What is the Matter with Dobbs?*, 26 U. PA. J. CONST. L. 282 (2024) (discussing morality in judicial decisions).

96. BICKEL, *supra* note 94, at 130–31.

97. For the classic statement, see JOHN HART ELY, *DEMOCRACY AND DISTRUST: A THEORY OF JUDICIAL REVIEW* 135–80 (1980).

most common current variant holds that judges should seek to ascertain and follow the original public meaning of the Constitution—the meaning that members of the American public would have understood the constitutional text to communicate in its original context, which is to say at the time of its drafting and ratification.⁹⁸ The most historically influential argument for originalism—and still probably the most important—is rooted in the importance of constraining judges. The idea is that judges have no legal or democratic warrant for imposing their own moral or political values on the American people. Originalism attempts to guard against this risk by limiting the role of judges to implementing the commands of the sovereign people as embodied in the Constitution’s original public meaning.⁹⁹

Of course, originalism developed as a theory of interpretation to guide the decisions of human judges. Originalists acknowledge that humans are susceptible to personal and political bias and the temptations of motivated reasoning. At least some judges, some of the time, probably also act in bad faith, knowingly making decisions on partisan or political grounds and dressing them up in seemingly plausible legalistic rhetoric. But originalists generally believe the best hope of reining in these inevitable shortcomings is for judges to act *to the maximum extent possible* as passive and objective enforcers of the Constitution’s original public meaning, without regard to their own personal moral and political views.¹⁰⁰

The italicized caveat above is important. Originalists acknowledge that the historical evidence of original meaning will sometimes be unclear. Other times, the historical evidence will clearly establish that the meaning of the constitutional text was vague or ambiguous at the time of its adoption. And even where original meanings are clear in the abstract, their application to particular facts—including modern technologies that the founders could never have imagined—may raise difficult questions that require the exercise of normative

98. See, e.g., Solum, *Public Meaning Thesis*, *supra* note 19, at 1990; Solum, *supra* note 11, at 3.

99. See, e.g., Antonin Scalia, *Originalism: The Lesser Evil*, 57 U. CIN. L. REV. 849, 864 (1989); Robert H. Bork, *Neutral Principles and Some First Amendment Problems*, 47 IND. L.J. 1, 4 (1971).

100. See, e.g., Baude, *supra* note 10, at 2228; Scalia, *supra* note 99, at 853.

judgment.¹⁰¹ Nevertheless, constraint-focused originalists believe the ideal is for judges to exercise as little moral and political judgment as possible and for constitutional cases to be resolved as much as possible by the application of original public meaning.

Richard Posner describes this as the “transmission belt” model of judging.¹⁰² The mechanical metaphor is suggestive. While sophisticated originalists recognize that applying originalism is not, and cannot be, mechanical, originalism’s commitment to constraining judges¹⁰³ implies that an interpretation machine ascertaining the Constitution’s original public meaning with perfect accuracy and applying that meaning with perfect objectivity, regularity, and predictability is, in fact, the ideal to which human judges should aspire. If such a machine were ever invented, constraint-oriented originalists would presumably find it very attractive. At a minimum, they might be expected to endorse its use as an adjunct to human judging—a kind of android law clerk. Conceivably, they might even be persuaded to turn over the task of constitutional interpretation entirely to the machine.

Originalism is not the only formalist approach to constitutional interpretation. The strong judicial-restraint approach associated with James Bradley Thayer also emphasizes the importance of constraining the role of moral and political judgment in constitutional decision-making. Instead of enforcing original public meaning, Thayer’s approach calls for judges to defer to legislative judgments except in cases where it is indisputably clear that the Constitution has been violated.¹⁰⁴ Some common-law approaches to constitutional interpretation might also be characterized as formalist, to the extent that they call for judges to adhere strictly to precedent rather than exercising their own independent moral and political judgment.¹⁰⁵ But because of originalism’s great contemporary

101. See, e.g., Solum, *Public Meaning Thesis*, *supra* note 19, at 1992; Solum, *supra* note 11, at 23.

102. RICHARD A. POSNER, *HOW JUDGES THINK* 252 (2010).

103. We do not mean to imply that all originalists share this commitment. As noted earlier, they do not. But here, and throughout, our focus is on constraint-oriented originalists, unless otherwise noted.

104. See James B. Thayer, *The Origin and Scope of the American Doctrine of Constitutional Law*, 7 *HARV. L. REV.* 129, 144 (1893).

105. See, e.g., STRAUSS, *supra* note 78; Strauss, *supra* note 81.

prominence, we shall confine our discussion of constitutional formalism to constraint-oriented originalist approaches.

b. Constitutional Realism

Most non-originalist approaches to constitutional interpretation, which are often grouped under the banner of “living constitutionalism,” qualify as realist in the sense we are using that term. Like originalism, these approaches come in many flavors, but they share two central commitments in common. First, living constitutionalists are skeptical that judges can avoid moral and political judgment when deciding constitutional cases. Second, they view the exercise of moral and political judgment by judges as, on balance, salutary, though many living constitutionalists emphasize the importance of judicial humility and sensitivity to questions of comparative institutional competence.¹⁰⁶ In both of these respects, living constitutionalists are the mirror image of originalists and constitutional realism is the mirror image of constitutional formalism.

Begin with the practicability of avoiding moral and political judgment in constitutional decision-making. The skepticism of living constitutionalists on this score rests on three distinct foundations. First, living constitutionalists think originalists significantly understate the fallibility of human judges and particularly the susceptibility of human judges to motivated reasoning. To substantiate this view, they point to a voluminous political science literature demonstrating that judicial votes are highly correlated with political ideology, especially at the U.S. Supreme Court.¹⁰⁷ Second, living constitutionalists are skeptical that original public meaning provides clear or determinate answers to many hard constitutional cases, either because of the limits of historical evidence, the vagueness of the constitutional text, or both.¹⁰⁸ Third, and perhaps most

106. See, e.g., Coan, *supra* note 95 (making these points and collecting sources); see also STRAUSS, *supra* note 78.

107. See, e.g., NEAL DEVINS & LAWRENCE BAUM, *THE COMPANY THEY KEEP: HOW PARTISAN DIVISIONS CAME TO THE SUPREME COURT* (2019); LEE EPSTEIN ET AL., *THE BEHAVIOR OF FEDERAL JUDGES: A THEORETICAL AND EMPIRICAL STUDY OF RATIONAL CHOICE* (2013); JEFFREY A. SEGAL & HAROLD J. SPAETH, *THE SUPREME COURT AND THE ATTITUDINAL MODEL REVISITED* (2002).

108. See, e.g., Coan & Schwartz, *supra* note 76.

important, living constitutionalists insist that choosing an approach to constitutional interpretation is itself a moral and political choice, requiring normative justification.¹⁰⁹

This brings us to the second central commitment of living constitutionalism. Even if it were possible for judges to avoid moral and political judgment when deciding constitutional cases, living constitutionalists are skeptical that this would be normatively desirable. In particular, they are skeptical that it would be normatively desirable for contemporary judges, appointed by relatively recent presidents, to defer to the moral and political judgments of long-dead framers and ratifiers, whose values and worldviews were starkly different from those of present-day Americans. As David Strauss observes, Americans would never allow the voters of contemporary Norway or Canada to decide our most important constitutional debates. But those voters have much more in common with present-day Americans than the people who framed and ratified most of the American constitutional text.¹¹⁰

Most living constitutionalists are also skeptical that the constitutional judgments of elected officials are always more trustworthy than those of life-tenured judges, operating within the institutional culture and structures of the federal court system. The upshot is that living constitutionalists believe that it is normatively desirable for judges to exercise moral and political (though not partisan) judgment when deciding constitutional cases.¹¹¹ At the very minimum, living constitutionalists insist that judges deciding constitutional cases cannot avoid making moral and political judgments about when to exercise their own judgment and when to adhere to original public meaning or defer to the judgments of other institutional actors.¹¹² Although few living constitutionalists have directly considered the possibility of replacing human judges with an interpretation machine, the logic of their position implies that this decision, too, requires a normative judgment—and a normative defense. The same would be true when deciding how to frame questions for such an interpretation machine to

109. See, e.g., Coan, *supra* note 95 (making this point and collecting sources).

110. Strauss, *supra* note 81; see also Coan, *supra* note 81.

111. See Coan, *supra* note 95 (collecting sources).

112. *Id.*

answer. We shall have much more to say on these matters in subsequent Parts.

3. The Importance of Institutional Context

The debate between constitutional formalists and realists has largely focused on the kinds of controversial questions that come before the U.S. Supreme Court. In such cases, there are nearly always plausible legal arguments on both sides.¹¹³ Constitutional formalists think judges should decide between those arguments on the basis of original public meaning—or, at any rate, on the basis of some criterion other than their own moral and political judgment.¹¹⁴ Constitutional realists doubt this is possible and, at any rate, think the moral and political judgment of judges is at least some of the time normatively superior to the various criteria defended by formalists.¹¹⁵ The stakes of this debate are very high because when the U.S. Supreme Court resolves such cases it shapes public policy on vitally important questions for the entire country.

In all of these respects, the constitutional questions that come before the U.S. Supreme Court are exceptional, rather than normal. The kinds of constitutional questions most often posed in the federal district courts—and quite often in the federal courts of appeals—generally have clear or fairly clear answers on which most or all judges applying any mainstream interpretive approach would agree. The same is true for many, if not most, of the constitutional questions that never make their way to court. Questions arising at the lower levels of the federal judicial system—and completely outside it—also tend to have lower stakes. Decisions of federal district courts have no precedential effect, and the decisions of federal courts of appeals govern particular geographic regions, rather than the whole country.¹¹⁶ These courts also have far larger caseloads and far

113. *Id.*

114. *See supra* Section I.B.1.

115. *See supra* Section I.B.2.

116. For classic discussions of these widely understood institutional points, see RICHARD A. POSNER, *THE FEDERAL COURTS: CHALLENGE AND REFORM* (1996); Frederick Schauer, *Easy Cases*, 58 S. CALIF. L. REV. 399 (1985). For a more recent treatment, see Aaron-Andrew P. Bruhl, *Statutory Interpretation and the Rest of the Iceberg: Divergences Between the Lower Federal Courts and the Supreme Court*, 68 DUKE L.J. 1 (2018).

fewer resources to devote to each decision than the U.S. Supreme Court does.¹¹⁷ The same is generally true for government officials and government institutions grappling with constitutional questions outside of court.¹¹⁸ For all of these reasons, the plausibility and attractiveness of the formalist program of clear legal answers generated with maximum speed and efficiency is significantly greater and less controversial outside the rarefied realm of the U.S. Supreme Court. For the same reasons, the case for employing a fast, low-cost, and relatively reliable interpretation machine seems likely to be stronger in these contexts. Again, we shall have more to say about this in subsequent Parts.

II. USE CASES FOR LLMs IN CONSTITUTIONAL INTERPRETATION

What would it mean for LLMs to interpret the Constitution? The first essential point to recognize is that LLMs are not machines that “go of themselves.”¹¹⁹ They are tools which require direction by a human user. Different kinds of human users interested in interpreting the Constitution—judges, law clerks, lawyers, government officials, private citizens, and so forth—could use LLMs in a wide range of ways for a wide range of purposes. Indeed, many are undoubtedly doing so already. Some of these uses are relatively modest. Others would de facto transfer the coercive power of the state, in substantial measure, to an algorithmic model that few members of the relevant community of users truly understand.

This Part describes an inexhaustive continuum of these use cases, arranged in roughly ascending order of ambition. Those at the lower end of this spectrum are not specific to

117. See, e.g., Peter S. Menell & Ryan Vacca, *Revisiting and Confronting the Federal Judiciary Capacity “Crisis”: Charting a Path for Federal Judiciary Reform*, 108 CALIF. L. REV. 789 (2020); Merritt E. McAlister, *“Downright Indifference”: Examining Unpublished Decisions in the Federal Courts of Appeals*, 118 MICH. L. REV. 533 (2020).

118. See, e.g., Jonah B. Gelbach & David Marcus, *Rethinking Judicial Review of High Volume Agency Adjudication*, 96 TEX. L. REV. 1097 (2018) (describing enormous caseloads of many administrative agencies).

119. Cf. James Russell Lowell, *The Place of the Independent in Politics*, in LITERARY AND POLITICAL ADDRESSES 190, 207 (1890) (emphasis added) (“After our Constitution got fairly into working order it really seemed as if we had invented a machine that *would go of itself*. . .”).

constitutional interpretation and are broadly on par with familiar technological tools like Google, Westlaw, Wikipedia, Grammarly, document automation software, and compliance management platforms, though LLMs operate using a very different underlying mechanism.¹²⁰ Use cases at the upper end of the spectrum are more tailored to constitutional law and would represent a substantially more radical shift in American constitutional practice. As such, these uses raise more difficult questions of constitutional theory.

A. Research Assistant, Critic, Troubleshooter, and Editor

Users of all kinds, from ordinary private citizens to U.S. Supreme Court Justices, are likely to find LLMs highly useful as research assistants. For starters, LLMs are fast, tireless, available at all hours, and comparatively cheap.¹²¹ They have also been trained on enormous quantities of data, likely including almost every reported decision ever issued by any U.S. court at any level and nearly all academic articles and books ever written on constitutional law. Unlike Google or Westlaw, LLMs are not especially useful for document retrieval, and they have a widely recognized—albeit improving—tendency to hallucinate when asked for direct quotes or specific citations. But LLMs are highly capable at descriptive summary, which they can provide at varying levels of depth, generality, detail, and length. If a private citizen wants to understand the broad contours of free speech law or a specific corner of it, such as the speech rights of high school students or newspapers or advertisers, the best LLMs will provide lucid and informative answers with a high degree of reliability.¹²² Those answers will be expressed in fluent and lucid prose, easily accessible to a general reader. And, of course, LLMs always stand ready to answer a limitless number of follow up questions. Moreover, although earlier generations of LLMs such as GPT-3.5 suffered from hallucinations and inaccuracies, more current models have significantly improved in this area. There are reasons to believe

120. See *supra* Section I.A.

121. The highest quality existing models charge modest monthly subscription fees, but their marginal cost rounds to zero.

122. See, e.g., Surden, *supra* note 3; Arbel & Hoffman, *supra* note 3.

that accuracy will dramatically improve in upcoming models, due to technological innovations.¹²³

As an example, we posed the following question to two leading LLMs, ChatGPT's GPT-4o and Claude 3 Opus: "You are a Supreme Court Justice and constitutional law professor. In 150 words, summarize the free speech rights of public employees under the U.S. Constitution." Both provided clear, accurate, and informative answers that heavily overlapped with one another. Here is ChatGPT using GPT-4o's answer:

Public employees in the United States, while retaining their free speech rights under the First Amendment, are subject to certain limitations based on their employment status. The Supreme Court established a balancing test in the landmark case *Pickering v. Board of Education* (1968) to determine when a public employee's speech is protected. According to this test, the speech must be on a matter of public concern, and the employee's interest in expressing themselves must outweigh the government's interest in promoting efficiency and discipline in public services. Further refined by *Garcetti v. Ceballos* (2006), the Court ruled that when public employees speak as part of their official duties, their speech is not protected under the First Amendment. This framework aims to protect public discourse while ensuring that governmental functions are not disrupted by internal dissent.¹²⁴

This is not the most elegant prose we have ever read, and it glosses over several important nuances. But it is pretty good. When we posed the same question without a word limit, the answers of both ChatGPT and Claude were better: more comprehensive, and more nuanced. The models can also be relatively straightforwardly coaxed into improving their prose or

123. See, e.g., Yuyan Chen et al., *Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models*, ARXIV 1–2 (July 4, 2024), <https://arxiv.org/abs/2407.04121> [<https://perma.cc/NQ6P-9T6E>].

124. Sample GPT-4o, Response to: "You are a Supreme Court Justice and constitutional law professor. In 150 words, summarize the free speech rights of public employees under the U.S. Constitution." *Summarize Free Speech Rights, ChatGPT*, OPENAI (June 3, 2024), <https://www.openai.com/chatgpt> [<https://perma.cc/R6TD-YCGX>].

elaborating on a nuance garbled or omitted in their original answers.

For example, when we asked ChatGPT about “public employees whose duties include significant policymaking authority or political judgment,” it correctly responded that such employees “are generally afforded less protection under the First Amendment for their speech related to their employment.”¹²⁵ Its explanation for this distinction was a bit vague but still serviceable and a few follow-up questions were sufficient to quickly clarify the basic outlines of the law in this area. Of course, it takes some pre-existing knowledge of First Amendment law to know what questions to ask. Most judges and lawyers will have at least some of that knowledge. Ordinary citizens and nonlawyer government officials generally will not have it—nor will they know what they do not know. For all of these reasons, the models powering Claude and ChatGPT are not (yet) good substitutes for consulting an attorney on complex and high-stakes questions. But they can be highly useful resources for general information and relatively simple, low-stakes questions. For these purposes, LLMs’ ability to answer highly specific and detailed questions, to provide additional clarification, and to answer endless follow-up questions makes them far more powerful than traditional search engines like Google or standard reference texts accessible to a general readership.

In addition to answering questions based on their training data, LLMs can also respond to questions about documents uploaded by the user, such as briefs, judicial opinions, or law review articles. For example, a judge or law clerk could ask an LLM to summarize the key arguments of several dozen amicus briefs in concise bullet points. An attorney could ask the LLM to do the same for a collection of law review articles. LLMs are not perfect at these tasks, but they are very good and will continue to get better.¹²⁶

On the basis of the LLM’s answers, judges and attorneys can zero in on the most relevant briefs or articles for closer reading—or ask the LLMs for further elaboration, summary, and so

125. *Id.*

126. *See, e.g.,* Surden, *supra* note 3; Arbel & Hoffman, *supra* note 3.

forth.¹²⁷ Of course, all of these are jobs that a human research assistant, law clerk, or junior associate could do—perhaps marginally better than an LLM, given the current state of the technology. But the cost and time required to achieve a marginal and uncertain improvement in accuracy would be substantial. Especially for private citizens and government officials seeking to understand the broad contours of constitutional law quickly, cheaply, and reliably, LLMs are extremely attractive and useful research assistants.¹²⁸ This is especially true for the large fraction of constitutional questions that have relatively clear and determinate answers.

Finally, judges, lawyers, and private citizens can also upload documents containing their own constitutional analysis and ask LLMs for evaluation, feedback, or criticism. For example, a law clerk drafting an opinion or order might ask ChatGPT or Claude to evaluate her discussion of a particularly thorny constitutional question in a draft opinion or order. Such a request can be quite nuanced, focusing the LLM’s “attention” on the specific doubts or concerns, asking it whether those doubts and concerns are well founded, and, if so, how they might be most effectively addressed. The law clerk might then incorporate the LLM’s suggestions and ask for additional feedback until she is comfortable with the result.¹²⁹ Obviously, processing and acting on the feedback of an LLM requires the exercise of judgment. In that sense, it is not fundamentally different than evaluating and acting upon the editorial suggestions of a human co-clerk or supervising judge. At present, LLMs are probably not better editors of constitutional analysis than the modal federal law clerk or judge. But these colleagues are not always available and, even when they are, their time is a scarce and valuable resource. For many questions, especially those that are more straightforward and involve

127. We express no view on any ethical rules or norms that might limit a judge’s or attorney’s discretion to rely on an LLM for these tasks under current law. Our point is simply that, as a technical matter, LLMs can already perform these tasks well enough to be quite useful. We assume that ethics rules will eventually catch up to the technology if they have not already, but nothing in our discussion turns on that one way or the other.

128. See, e.g., Adam Unikowsky, *In AI We Trust*, ADAM’S LEGAL NEWSL. (June 8, 2024), <https://adamunikowsky.substack.com/p/in-ai-we-trust> [<https://perma.cc/6PP2-6ZNT>] (highlighting these capabilities and going so far as to suggest that LLMs can and should currently decide cases).

129. See, e.g., Surden, *supra* note 3; Arbel & Hoffman, *supra* note 3.

relatively low stakes, the editorial feedback of LLMs will often be quite helpful. It might even be formally incorporated into the workflow of law offices and judicial chambers before any substantial written work product is circulated to human colleagues.¹³⁰

All of the use cases described in this Section are general, rather than specific, to constitutional law. But we suspect that many and perhaps most of the uses of LLMs to answer questions of constitutional interpretation will fall into this category. We therefore thought it important to explain these uses in some detail and to explain why they do not raise any particularly interesting or difficult theoretical questions. Like any other research or editing tool, LLMs should be used with care and good judgment appropriate to the significance of the task at hand. The same, of course, is true of human research assistants and editors. We shall therefore put these use cases to one side for the remainder of this Article.

B. Drafting Assistant

Closer to the midpoint of our continuum of ambition, LLMs might be used to draft formal legal documents addressing questions of constitutional law, including party and amicus briefs, judicial opinions, and judicial orders. For present purposes, we are talking about drafting *assistance* only, based on detailed instructions from a human lawyer, law clerk, or judge, with robust ex post review and editing by those same humans. Like the use cases we have already discussed, this use of LLMs is not confined to constitutional law. Nor does it actually transfer the responsibility or authority to interpret the Constitution from humans to machines.

But there is also a potentially significant difference. Unlike research memos or editorial suggestions, judicial opinions and orders carry authoritative weight both for the parties to a case and, in the case of appellate decisions, as a matter of *stare decisis*. Frequently, the precise wording of these documents has practical consequences. A brief does not carry this kind of authority but does represent a solemn submission to a court and carries an implied affirmation by the signing attorneys that the

130. See, e.g., Unikowsky, *supra* note 128; Surden, *supra* note 3; Arbel & Hoffman, *supra* note 3.

brief's contents are truthful and legally plausible—or at least nonfrivolous.¹³¹

This potential use of LLMs raises two major questions. First, are LLMs capable of drafting formal legal documents at a high (or even a minimally competent) level? Second, does this use raise any significant practical or theoretical questions beyond those of more modest uses of LLMs? The answer to the first is emphatically, yes. The answer to the second is probably not.

Readers of this Article have almost certainly seen some—and perhaps many—comically bad examples of LLM-drafted prose. Many people have an unduly low estimation of LLMs' drafting capabilities as a result. But the best models have improved significantly in this area since the launch of ChatGPT built on GPT-3.5 in late 2022.¹³² Numerous studies show that human readers are not able to reliably distinguish LLM prose from prose written by humans.¹³³ One informal study found that blind-graded GPT-4 essays earned mostly As in undergraduate classes at Harvard University.¹³⁴ Another law-specific study found that GPT-3, which is primitive by current standards, passed four law school exams at a highly-ranked public law school, albeit near the bottom of the class.¹³⁵ Almost as important, this study simply presented GPT-3 with a single prompt, rather than working with it iteratively to produce the best possible exam responses.¹³⁶ As such, it almost certainly dramatically underestimates the drafting capacity of the current top LLMs in the hands of reasonably sophisticated users.

We have ourselves used LLMs for numerous kinds of drafting, including fundraising appeals, grant applications, and letters of recommendation (with appropriate privacy safeguards), and we found the results produced by the best current models remarkably impressive. Anthropic's

131. FED. R. CIV. P. 11.

132. See *supra* Section I.A.

133. See, e.g., Balazs Kovacs, *The Turing Test of Online Reviews: Can We Tell the Difference Between Human-Written and GPT-4-Written Online Reviews?*, 24 MKTG. LETTERS: J. RSCH. MKTG. 1 (2024).

134. Maya Bodnick, *ChatGPT Goes to Harvard*, SLOW BORING (July 18, 2023), <https://www.slowboring.com/p/chatgpt-goes-to-harvard> [<https://perma.cc/W6LQ-K7XA>].

135. Jonathan H. Choi et al., *ChatGPT Goes to Law School*, 71 J. LEGAL EDUC. 387, 391 (2022).

136. *Id.* at 389.

Claude 3 Opus produces especially polished prose and is capable of passably mimicking the prose style of individual human writers when presented with samples of their work. As an experiment, we used Claude 3 Opus to produce a 10,000-word draft law review essay based on a human-drafted introduction and a high-level outline. In roughly six hours of work, the model produced a complete and highly sophisticated draft essay that we strongly suspect would have been accepted for publication by a well-respected law review, albeit with a few more hours of editing work and the addition and verification of citations by a student research assistant.¹³⁷ This same essay would probably have taken the authors of this Article two solid weeks of work to complete.

Law review essays and published appellate judicial opinions involve roughly comparable levels of analytical intricacy and sophistication. If anything, published appellate opinions are probably more regimented and formulaic. That is certainly true of most routine judicial orders and unpublished decisions. If sophisticated users can get today's best LLMs to produce high-quality grant applications, fundraising letters, and law review essays, they can almost certainly do the same for many formal legal documents.¹³⁸ Those documents would need serious ex post review to check for accuracy and fidelity to the prompt provided. Given the current state of the art, direct quotations and citations would either need to be provided in the initial prompt or scrupulously checked after the fact. But today's best LLMs are capable of drafting not just minimally competent but genuinely high-quality prose of the kind that appears in many, if not most, formal legal documents, including those involving constitutional questions.¹³⁹ We strongly suspect LLMs are already in widespread use for this purpose, especially by practicing lawyers but increasingly by law clerks and judges.¹⁴⁰

137. *Anthropic*, ANTHROPIC (Mar. 9, 2024), <https://www.anthropic.com> [<https://perma.cc/6FYC-J4PH>]. Ethical norms surrounding “co-authorship” of academic papers with LLMs are still evolving and the draft essay described in the text was not submitted to law reviews for that reason. All of the text in this Article was produced by its human authors.

138. *See, e.g.*, Unikowsky, *supra* note 128 (producing very high quality results with Claude 3 Opus).

139. *Id.*

140. There have been numerous widely publicized cases of attorneys filing formal submissions containing hallucinatory AI work product. *See, e.g.*, Bob Ambrogi, *Not Again! Two More Cases, Just this Week, of Hallucinated Citations in*

Should this prospect concern us? We do not think so, provided the human users in question exercise reasonable caution and appropriately independent judgment. Indeed, the use of LLMs as drafting assistants in the manner we have been discussing strikes us merely as a new form of the delegated responsibility that has long been rampant in legal practice. Partners and senior associates sign briefs drafted by junior associates. Judges sign opinions drafted by law clerks—often with far less detailed instructions than a responsible judge or law clerk would typically supply to an LLM. It is certainly possible to cross the line in this respect. Courts of appeals have occasionally faced criticism and controversy over the limited involvement of judges in the various forms of summary disposition developed to manage their overwhelming caseloads.¹⁴¹ But we do not see any consequential distinction between delegating legal drafting responsibilities to LLMs and delegating these same drafting responsibilities to human subordinates, even for authoritative legal documents like judicial orders and appellate decisions. In both cases, what matters is that the authoritative decision-maker in question is exercising genuinely independent judgment over the instructions provided and the ultimate work product. We shall therefore put this use case to one side for the remainder of this Article.

C. *Focused Legal Queries*

We now come to the use case that most people probably have in mind when they think about AI and constitutional interpretation. Rather than relying on LLMs for background research and editorial assistance or using them as ghost writers with detailed *ex ante* instructions and robust *ex post* review, judges and their law clerks might actually ask LLMs to interpret the Constitution and rely on their interpretations to decide

Court Filings Leading to Sanctions, LAWSITES (Feb. 22, 2024), <https://www.lawnext.com/2024/02/not-again-two-more-cases-just-this-week-of-hallucinated-citations-in-court-filings-leading-to-sanctions.html> [https://perma.cc/RP9L-JTA6]. This is almost certainly just the tip of the iceberg.

141. See, e.g., McAlister, *supra* note 117; WILLIAM M. RICHMAN & WILLIAM L. REYNOLDS, *INJUSTICE ON APPEAL: THE UNITED STATES COURTS OF APPEALS IN CRISIS* (2013).

cases, motions, and so forth.¹⁴² This use case can be helpfully sub-divided into three sub-cases. First, LLMs might be used to answer focused constitutional queries. Second, LLMs might be asked to actually decide cases under a specified interpretive approach. Third, LLMs might be asked to decide cases without specifying an interpretive approach. In each of these cases, judges might treat the LLM's interpretation as advisory or conclusive. The remainder of this Section briefly explains what each of these uses would entail. Subsequent Parts will then provide more in-depth analysis of their potential benefits and drawbacks, with a particular focus on the use of LLMs to actually decide cases.

1. Focused Constitutional Queries

Constitutional decision-making involves several steps. Those steps might be broken down and sequenced in any number of ways. But one plausible breakdown and sequence is as follows: (1) choose an interpretive method; (2) consult all interpretive sources, authorities, and modalities relevant under that interpretive method; (3) apply the method to those sources, authorities, and modalities to arrive at an interpretation of the applicable constitutional provisions; and (4) apply that interpretation to the facts of the case to arrive at a decision. Each of these steps could be further subdivided and each is susceptible to various objections.¹⁴³ But for present purposes, the important—and uncontroversial—point is simply that constitutional decision-making has multiple steps. Any one of these might be delegated to an LLM, while the others are performed by a human judge.

Of the four steps we just laid out, step two is the one that we think judges are most likely to delegate to LLMs. For example, nearly all mainstream approaches accord some interpretive weight to the Constitution's original public meaning. Many approaches also accord some weight to the original intentions, purposes, and expectations of the

142. Cf. *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208 (11th Cir. 2024) (Newsom, J., concurring) (proposing, very tentatively, to use LLMs as one tool among many for ascertaining the “ordinary meaning” of contracts).

143. For an alternative breakdown, see Solum, *Public Meaning Thesis*, *supra* note 19; Solum, *supra* note 70.

Constitution's drafters and ratifiers.¹⁴⁴ All lower-court judges and nearly all U.S. Supreme Court Justices accord substantial weight to the constitutional interpretations embraced by previous U.S. Supreme Court opinions.¹⁴⁵ And some interpretive approaches also accord significant weight to the ordinary contemporary understanding of the Constitution—at least on some constitutional questions. The “evolving standards of decency” test for defining “cruel and unusual punishment” under the Eighth Amendment¹⁴⁶ is the best, but not the only, example in an important U.S. Supreme Court decision.¹⁴⁷

A judge or justice might plausibly ask an LLM for the answer to any of these questions and treat the answer as authoritative or highly persuasive while exercising her own independent judgment about the other steps of the constitutional decision-making process.¹⁴⁸ We will refer to this use case as a “focused constitutional query.” For example, in a federalism case about Congress’s power to pass the Clean Air Act, a U.S. Supreme Court Justice might ask ChatGPT whether the original public meaning of the Constitution granted Congress a general legislative power to address important national problems or instead strictly limited Congress to those powers expressly enumerated in the constitutional text. If that Justice were an originalist, she might treat ChatGPT’s answer as fully resolving this question of constitutional meaning and apply that meaning to the facts of the case in her decision. If the Justice were instead a pluralist along the lines of Philip Bobbitt¹⁴⁹ or Richard Fallon,¹⁵⁰ she might treat ChatGPT’s answer as resolving the question of original public meaning and

144. See, e.g., Mitchell N. Berman, *Originalism Is Bunk*, 84 N.Y.U. L. REV. 1 (2009).

145. See, e.g., STRAUSS, *supra* note 78.

146. *Trop v. Dulles*, 356 U.S. 86, 101 (1958).

147. See, e.g., *Rochin v. California*, 342 U.S. 165 (1952).

148. Here, of the four steps the Justice would only use (1) choice of interpretive method; (3) applying that method to the LLM’s answers to interpret the applicable constitutional provisions; and (4) apply that interpretation to the facts of the case to arrive at a decision.

149. PHILIP BOBBITT, *CONSTITUTIONAL INTERPRETATION* (1991); PHILIP BOBBITT, *CONSTITUTIONAL FATE: THEORY OF THE CONSTITUTION* (1982) (articulating and defending a pluralist approach consisting of several constitutional modalities, such as text, history, structure and prudential reasoning).

150. Richard H. Fallon, Jr., *A Constructivist Coherence Theory of Constitutional Interpretation*, 100 HARV. L. REV. 1189 (1987) (advocating a somewhat different brand of pluralism that strives to bring different modalities into coherence).

proceed to ask it (or conduct her own inquiry) about other sources of interpretive authority like historical practice, U.S. Supreme Court precedent, and constitutional structure.

There are four points worth noting about this potential use of LLMs:

First, in principle, it is compatible with any interpretive approach. Judges and justices of different interpretive persuasions would often pose different questions to the LLM. But all interpretive approaches turn on forms of authority or argument—such as original meaning, purpose, structure, or precedent—that LLMs could be empowered to authoritatively or presumptively decide.

Second, it is a difficult question whether LLMs can provide better answers to all—or any—of these questions than human judges and law clerks relying on party and amicus submissions and their own independent legal research. The answer depends not just on the relative competence and biases of the LLMs and alternative human decision-makers in the abstract.¹⁵¹ It also crucially depends on the time and attention a human decision-maker can afford to devote to the issue in question and the sophistication of the LLM and its user. A U.S. Supreme Court Justice who decides fewer than seventy cases and writes perhaps fifteen opinions per year with the assistance of four highly capable law clerks can very likely outperform the best current LLMs on most plausible, focused queries. The same is not necessarily true for a busy federal district judge or state trial court judge deciding a novel constitutional question posed by one of two dozen motions in limine in a complex criminal case or a busy court of appeals processing tens of thousands of cases per year.¹⁵² Of special note for originalists, LLMs are not currently very good at confining their answers to specific historical points or periods, such as the meaning of the commerce power in 1789. They are even worse at recreating the thick and evolving historical, political, economic, and cultural context which shapes the content communicated by particular words and phrases. But the capability of LLMs will certainly improve over time, and special-purpose, hybrid tools may be developed that turn the

151. The proper criteria for evaluating competence and bias are obviously normative questions—and contested ones. Our point is that it is difficult to assess the relative competence and bias of LLMs and human decision-makers in the domain of constitutional law by any plausible set of criteria.

152. See, e.g., Menell & Vacca, *supra* note 117; McAlister, *supra* note 117.

power of LLMs on large data sets such as those used in corpus linguistics.¹⁵³

Third, however powerful LLMs become, the way judges frame their prompts is likely to remain extremely significant. This should hardly be surprising. As every first-year law student learns, any legal issue can plausibly be framed in a number of different ways, and this framing often decisively determines the result.¹⁵⁴ Consider, again, a hypothetical federalism case involving the constitutionality of the Clean Air Act. A judge might ask ChatGPT whether the original public meaning of the Constitution granted Congress a general legislative power to address important national problems or instead strictly limited Congress to those powers expressly enumerated in the constitutional text. This framing arguably rests on a false dichotomy. The original public meaning of the Constitution may have limited Congress to those powers expressly or *impliedly* granted by the constitutional text, which might have been understood in context as subject to *broad* (or *liberal* or *reasonable*) rather than *strict* construction, so as to ensure the national government was able to accomplish the important purposes for which it was established.

Of course, this game can be played by both sides. An alternative framing of the question might load the dice in the opposite direction, asking whether “the original public meaning of the Constitution granted Congress those powers normally inherent in sovereign governments, even absent express enumeration, as Alexander Hamilton argued as Treasury Secretary and many of the Constitution’s prominent Anti-Federalist opponents argued during the ratification debates.”¹⁵⁵ Judges might employ a variety of techniques to get around this problem, asking LLMs to frame the question for themselves based on the parties’ briefs or based on an ostensibly neutral presentation of the facts of the case.¹⁵⁶ But these, too,

153. See, e.g., Surden, *supra* note 3; Arbel & Hoffman, *supra* note 3.

154. For this reason, lawyers practicing before the U.S. Supreme Court invest substantial time and energy arguing over the question presented, and the Justices often draft their own questions presented—rather than granting certiorari on the questions proposed by the parties—for their own reasons. See Benjamin B. Johnson, *The Origins of Supreme Court Question Selection*, 122 COLUM. L. REV. 793 (2022).

155. See Andrew Coan & David S. Schwartz, *Interpreting Ratification*, 2 J. AM. CONST. HIST. 91 (2023).

156. See, e.g., Unikowsky, *supra* note 128; Arbel & Hoffman, *supra* note 3.

are judicial choices, which could themselves be framed in a range of different ways, with a foreseeable impact on the answers provided by the LLM. If the LLM's answer is ambiguous or equivocal, the same question of framing arises again with respect to any additional prompts or queries the judge might make to clarify the LLM's original response.¹⁵⁷

Fourth, a special—and especially significant—case of the issue-framing problem arises with respect to the burden of persuasion, the standard of proof, and what we will call “the standard of determinacy.” Lawyers are accustomed to thinking about these matters in connection with questions of factual or evidentiary proof. But as Gary Lawson has persuasively argued, they are equally applicable to questions of law.¹⁵⁸ Indeed, they are unavoidable. Should legislative and executive action enjoy a presumption of validity such that every constitutional challenger should bear the burden of persuasion? Or should courts instead employ a presumption of liberty, placing the burden of persuasion on the government? Whatever the answer, should the standard of proof be a preponderance of the evidence? Clear and convincing evidence? Something else?

Constitutional theorists of all stripes acknowledge the possibility that constitutional meaning is simply indeterminate on some questions—whether we are talking about the original public meaning or communicative content of the document, the contemporary public meaning, the meaning implicit in historical practice, or something else. But just how clear does a constitutional meaning have to be to qualify as determinate? There is nothing close to a consensus on this neglected question even within—much less *across*—different interpretive approaches.¹⁵⁹ Yet LLMs cannot answer focused constitutional queries without applying some standard of determinacy, burden of persuasion, and standard of proof. Which standards and burden they choose can either be directed by the judge posing the query or supplied by the black box of the LLM's algorithm. Either way, this is a choice judges unavoidably make when posing focused queries to LLMs.

157. See generally Surden, *supra* note 3.

158. GARY LAWSON, EVIDENCE OF THE LAW: PROVING LEGAL CLAIMS (2017); see also Coan & Schwartz, *supra* note 76.

159. LAWSON *supra* note 158; see also Solum, *Public Meaning Thesis*, *supra* note 19.

2. LLMs as the Ultimate Decider

At first blush, it may seem more straightforward and less fraught for judges simply to delegate ultimate decision-making authority to LLMs. But this even more ambitious use of LLMs in constitutional interpretation raises a host of issues, many of them overlapping with the issues raised by focused constitutional queries. We envision two main ways that judges might delegate ultimate decision-making authority over constitutional interpretation to LLMs: one more comprehensive and one more limited.¹⁶⁰ We doubt judges are currently engaged in either of these to any meaningful extent, but as LLMs continue to become more commonplace and their capabilities improve, it is highly probable that judges will do so in some form or fashion.¹⁶¹ To prepare for that eventuality, this Section describes what each of these use cases would entail and explains the different issues each of them would raise.

a. Ultimate and Comprehensive

First, and most comprehensively, judges might simply instruct an LLM to decide a case as if it were a judge or justice of the relevant court. For example, in a constitutional case before the U.S. Supreme Court, an individual Justice—or the Court as a whole—might ask the LLM to decide the question presented and provide a written opinion explaining its reasoning. For a more robust response, the Justice or Justices could, and likely would, upload the party and amicus briefs. In cases with more than a handful of amici, this might overwhelm the LLM's capacity. But this could be addressed in a number of ways. Most obviously, the Justices' law clerks could ask the LLM to prepare one-page summaries of the individual amicus briefs for each side, compile them in supplementary documents, and upload

160. These are two ideal types, and each could obviously take various forms or be employed in various permutations (with each other and other use cases).

161. This may seem unthinkable today, but we expect the "Overton window" to shift quickly as the capabilities and understanding of LLMs expand. See *A Brief Explanation of the Overton Window*, MACKINAC CTR. FOR PUB. POL'Y (2019), <https://www.mackinac.org/OvertonWindow> [https://perma.cc/TMT8-4R5P] (defining the Overton window). The enthusiastic advocacy of LLM judging by Adam Unikowsky, a sophisticated and successful appellate lawyer and former U.S. Supreme Court clerk, is likely to appear as a harbinger in retrospect. See Unikowsky, *supra* note 128.

them along with the party briefs. This current capacity limitation, however, is unlikely to persist for long as the technology improves.¹⁶²

Of course, as the U.S. Supreme Court currently operates, the question presented is drafted either by one of the parties or by the Court itself acting on the basis of the parties' submissions.¹⁶³ A Justice or group of Justices committed to delegating this important framing decision could upload the parties' petition for certiorari and opposition briefs and task the LLM with formulating the question presented. This request itself requires framing. It could be posed in terse, minimalist terms, or it could provide expansive guidance—perhaps rooted in the Supreme Court's rules of practice—for formulating the question presented and, in the process, the scope of the Court's review. If torn between these approaches, the Justices could even ask the LLM for advice between them and to draft the language for a more detailed prompt if that is the direction the LLM recommends or the Justices decide. As a technical matter, all of this is perfectly possible today, though the capability of LLMs to outperform U.S. Supreme Court Justices—according to any plausible set of normative criteria—is doubtful.¹⁶⁴

That may well change as the technology improves. But there are two points worth noting about this potential use of LLMs that seem likely to apply so long as the technology exists in roughly its current form.

First, to state the obvious, the U.S. Supreme Court Justices have strong views about how the Constitution should be interpreted. They also have a strong self-interest in retaining the U.S. Supreme Court's power to shape the future of the country—and, presumably in most cases, a strong sense of personal responsibility to do so according to their own best judgment. To comprehensively delegate ultimate decision-making power to LLMs would undermine all of these, transferring the power long vested in the Justices to a nontransparent, probabilistic, and proprietary algorithm.¹⁶⁵ Even in a world where LLMs are much more advanced than they are today, it is hard to imagine the Justices voluntarily

162. See *supra* Section I.A.

163. Johnson, *supra* note 154.

164. *But cf.* Unikowsky, *supra* note 128 (enthusiastically advocating the superiority of current LLMs to human judges).

165. See *supra* Section I.A.

surrendering this much control without a very strong reason to expect that LLMs would produce better results by the Justices' own lights.¹⁶⁶ This decision would almost certainly be influenced by the Justices' views of the sorts of results typically generated by LLMs. Justices who generally find those results congenial would presumably be more inclined to delegate more authority to LLMs, with the inverse being true for Justices who find the results disagreeable. This decision would almost certainly be influenced by the Justices' view of the sorts of results typically generated by LLMs. Justices who generally find those results congenial would presumably be more inclined to delegate more authority to LLMs, with the inverse being true for Justices who find the LLMs' results disagreeable. If this *ex post* evaluation is a major driver of LLM use, then the Justices' normative priors with LLMs will play a large—and quite possibly a dispositive—role even in decisions that they (hypothetically) formally delegate to LLMs. Even Justices who find the results produced by LLMs congenial might see little practical value in using them given their undemanding caseload and the personal prestige that all of the Justices derive from making (or being seen as making) constitutional decisions for themselves.

Second, the institutional positions and working conditions of lower-court judges in the state and federal systems are very different from those of U.S. Supreme Court Justices. These differences might make them considerably more open to the use of LLMs as comprehensive ultimate deciders. For starters, these judges are much busier than U.S. Supreme Court Justices, and they have much less staff support relative to their workloads. The modal lower-court case is also much more routine than the modal U.S. Supreme Court case, and unlike the Justices, lower-court judges are strictly bound by U.S. Supreme Court precedents. This dramatically reduces the room for interpretive disagreement in most lower-court cases. A routine case, by definition, is one in which nearly all reasonable judges will reach the same conclusion, regardless of their methodological,

166. This need not, of course, be a crudely consequentialist or ideological judgment. A Justice might be persuaded to delegate decision-making authority to an LLM because she believes it likely to apply her own preferred interpretive approach more accurately and reliably than she can. In theory, a Justice might also conclude that the LLM's judgment of what constitutes better performance is better and more trustworthy than her own. But as a matter of basic psychology this seems unlikely.

philosophical, or ideological inclinations.¹⁶⁷ Out of practical necessity, the federal courts of appeals already shunt a huge fraction of their cases into various forms of summary disposition that receive very little focused attention from the judges who nominally decide them.¹⁶⁸ Many federal district judges lean heavily on law clerks and magistrate judges for a large share of their work.¹⁶⁹

For now, the symbolic significance of formally and publicly delegating comprehensive ultimate decision-making authority to LLMs is probably a bridge too far. But the practical value of such delegation for resource-constrained courts deciding mostly routine constitutional (and other) questions under binding U.S. Supreme Court precedent could be very substantial. The process by which LLMs answer routine constitutional questions is no more transparent or human-like than the process by which they answer more controversial constitutional questions. But precisely because these cases generate greater consensus among human judges, LLMs are likely to more consistently arrive at the same outcome as human decision-makers and to raise fewer controversial normative questions.¹⁷⁰ They are also likely to produce better written justifications of their decisions relative to human decision-makers as measured by any plausible set of criteria. As LLMs improve in their capabilities and diffuse through society in general and workplaces of all kinds, we suspect that the temptation to use them as comprehensive ultimate deciders in at least some routine matters will become irresistible.¹⁷¹

167. See notes 116–117 and accompanying text.

168. See McAlister, *supra* note 117.

169. See, e.g., Menell & Vacca, *supra* note 117; McAlister, *supra* note 117; RICHMAN & REYNOLDS, *supra* note 141.

170. For purposes of this paper, we bracket whether there is something morally unique or deontologically obligatory about human, judicial decision-making. See, e.g., Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137 (2019).

171. This may take a while to become formal and public and is likely to be masked behind a fig leaf of independent human judgment whenever that finally happens. But behind-the-scenes we expect LLMs to be used in something like this fashion with increasing frequency in the near future.

b. Ultimate but Limited

Much more than the typical lower-court judge, U.S. Supreme Court Justices care about how questions of constitutional interpretation are decided. They also decide far more cases (and questions within cases) in which the choice of interpretive approach is plausibly dispositive. For these reasons, the Justices seem less likely than lower-court judges to delegate ultimate and comprehensive decision-making authority to LLMs. But there is another intriguing possibility: Justices might ask LLMs, presumptively or conclusively, to decide constitutional questions under the Justices' own preferred approaches to constitutional interpretation.¹⁷² For example, Justice Neil Gorsuch or Justice Clarence Thomas might ask an LLM to decide whether a state criminal conviction by an eight-member jury is unconstitutional under an original public meaning approach. Justice Breyer, when he was still on the Court, might have asked whether this reading of the Seventh Amendment was a common-law or "active liberty" approach to constitutional interpretation.¹⁷³

Like the delegation of comprehensive and ultimate decision-making authority, this more limited approach would presumably involve uploading party and amicus briefs and asking the LLM to decide the question presented. It would therefore involve all of the same issue-framing problems as a more comprehensive delegation. In a more limited form, it would also face many of the same obstacles to implementation. U.S. Supreme Court Justices seem unlikely to voluntarily surrender any substantial portion of the decision-making authority from which the power, prestige, and solemn ethical responsibilities of their office derive. Still, it is at least conceivable that some Justices might be more open to delegating decision-making

172. For the same reason the Justices might be more interested in this approach, lower-court judges seem likely to be less interested. Their more routine cases are much more likely to be controlled—to some substantial extent—by binding U.S. Supreme Court precedent. They are correspondingly less likely to turn on the choice of interpretive method. Because lower courts are also much more resource-constrained relative to their caseloads, they are also more likely to be attracted to the high speed, low cost, and simplicity of comprehensive delegation.

173. See STEPHEN BREYER, *ACTIVE LIBERTY: INTERPRETING OUR DEMOCRATIC CONSTITUTION 5* (Alfred A. Knopf ed., 2005) (advocating an interpretive approach focused on "the people's right to 'an active and constant participation in collective power'").

authority to an LLM instructed to apply their own preferred interpretive method. This might be especially true if LLMs improve to the point where their ability to synthesize historical evidence, linguistic usage, and judicial case law plausibly exceeds that of even the best human analysts (according to the normative criteria supplied by a Justice's preferred interpretive approach). Justices who sincerely embrace originalism or other constraint-focused approaches might be especially drawn to this possibility.

The banal point to make about this use of LLMs is that the human Justice's choice among theories of interpretation is likely to be a major driver of the result. The more subtle point is that any given interpretive approach might be described in different ways and at greatly differing levels of detail. Sticking with the example of originalism, a Justice who wished to delegate ultimate but limited decision-making authority to an LLM might simply ask—as we suggested earlier—what the original public meaning of the Constitution has to say about the question before the Court. But this minimalist approach leaves many questions to be resolved by the LLM. A Justice committed to a particular vision of originalism might want to provide substantially more detail, for example, about the distinction between original meaning and intentions, expectations, and political preferences; the role of social, political, economic, cultural, and legal context in shaping the communicative content of linguistic utterances; the distinction between public and private meanings and the special case of terms of art; and so on.¹⁷⁴ Such a Justice might also wish to expressly instruct the LLM on the burden of persuasion, the standard of proof, and the standard of determinacy (as we discussed earlier in connection with focused constitutional inquiries). An even more fastidious originalist Justice might choose to upload an article or book explaining their preferred version of originalism—Justice Antonin Scalia's *A Matter of Interpretation*,¹⁷⁵ for example, or the latest article by Lawrence Solum or William Baude.¹⁷⁶

174. See, e.g., Solum, *Public Meaning Thesis*, *supra* note 19; Solum, *supra* note 11; see also Coan & Schwartz, *supra* note 76.

175. ANTONIN SCALIA, *A MATTER OF INTERPRETATION: FEDERAL COURTS AND THE LAW* (Amy Gutmann ed., 1997).

176. As a technical matter, we doubt that the best current LLMs are currently capable of adhering to such dense and technical instructions. But this does not eliminate the issue. It simply means that many important nuances of interpretive

All of these human choices about how and at what level of detail to instruct the LLM can significantly affect its ultimate output. As with delegations of comprehensive and ultimate decision-making authority, a judge or justice might attempt to get around this problem by asking the LLM to decide which level of detail and which specific form of originalism to apply. But this workaround is itself a human choice—one that sacrifices much of the control that differentiates limited from comprehensive delegation. Both this choice and the choice to delegate limited authority to an LLM in the first place seem highly likely to turn on the typical results of such delegation. As with comprehensive delegations, Justices who approve of those results are much more likely to use LLMs in this way, and vice versa. If this correlation is high enough, these ex post normative judgments will amount to a kind of human tail wagging the AI dog.

III. POTENTIAL BENEFITS AND DRAWBACKS OF ARTIFICIAL INTELLIGENCE IN CONSTITUTIONAL INTERPRETATION

This Part examines the potential benefits and risks of employing advanced LLMs in constitutional interpretation. In terms of risk, the primary concern is that LLMs will be used in legally consequential situations by actors who do not properly understand their limitations. To illustrate this concern, we will focus on the use of LLMs by legal officials—such as judges and government administrators—who possess the authority to officially interpret and apply the law (as opposed to lawyers or members of the public whose interpretive decisions are largely predictive or persuasive). The use of LLMs by judges and other legal officials is particularly important because their interpretations and applications of law have direct real-world impacts on individual rights, government powers, societal norms, and case outcomes.

We are emphatically not suggesting that advanced LLMs have no place in official legal decision-making if used appropriately. On the contrary, these models can be quite beneficial in various roles, such as summarizing documents or providing background information to inform judicial or other official judgments. It is crucial not to view the risks of LLMs in

method will, at present, be resolved by the LLM in nontransparent and basically arbitrary (or random) fashion.

isolation or to ignore the fact that human officials also make mistakes and are subject to cognitive errors, biases, and the temptations of motivated reasoning. As a result, humans may occasionally—or even often—make under-informed, confused, personally-motivated, or haphazard decisions. But in our view, the more immediate risk is that LLMs will be used in an unsophisticated manner by inexperienced legal officials without a full appreciation of the technology’s strengths and weaknesses.

A. A Naïve View and a Motivating Example

To explore this issue, it is helpful to begin with a simplistic, perhaps even naïve, perspective on using LLMs in constitutional interpretation. This view asks: Why not simply input constitutional law, relevant precedents, and particular factual scenarios into the most advanced LLMs, like GPT-4 or Claude 3.5 Sonnet, and rely on their seemingly objective, consistent, and neutral resolutions of constitutional controversies? On this view, one could improve upon the frail constitutional judgments of human judges by leveraging the computational power, immense knowledge base, and seemingly detached “neutrality” of LLMs to provide more objective and “correct” legal answers. This perspective posits that LLMs could serve as more informed, neutral, and fair decision-makers in constitutional contexts than today’s more ideologically motivated and cognitively limited human judges. Let us call this the “naïve view” of LLMs as objective legal oracles. While views this simplistic may not be widely held among legal officials, we have encountered several judges (lacking technical background or AI experience) who expressed opinions not far removed from this perspective.¹⁷⁷ This naïve view is therefore useful to consider as a point of contrast to illustrate some of the concerns.

AI systems are now widely accessible to the public, including within judges’ chambers, providing unprecedented

177. Compare Unikowsky, *supra* note 128 (advocating something approximating this view), with Snell v. United Specialty Ins. Co., 102 F.4th 1208 (11th Cir. 2024) (Newsom, J., concurring) (advocating a much more tentative and limited use of LLMs by judges). Unikowsky is not naïve. He understands the capabilities of LLMs far better than most lawyers and judges today, and his writings are valuable illustrations of the impressive power and usefulness of LLMs in legal and constitutional analysis. But, for reasons elaborated in this Part, we think his enthusiasm requires significant tempering and qualification.

opportunities for explicit or undisclosed use in legal decision-making. However, most judges and legal officials who now have ready access to such systems are unlikely to be experienced users with full awareness of LLMs' capabilities and shortcomings. Moreover, these AI models do not clearly and effectively communicate their own weaknesses to such users. Rather, today's LLMs exhibit unprecedented fluency and coherence, and will nearly always provide apparently well-reasoned, well-supported, and confident answers—even in contexts in which there are socially contested values and for which there are not actually any objectively “correct” answers that all reasonable interpreters would agree on. Such outputs can be deceptively compelling to those who do not fully understand the technical subtleties of LLMs, such as their inherent randomness, sensitivity to framing and word choice, and the fact that seemingly inconsequential prompt choices can lead to diametrically opposite substantive responses. By contrast, the AI systems of the past were much less capable and therefore made their own weaknesses much more readily apparent even to nontechnical users.

The point is not that the answers that LLMs produce to legal questions are necessarily “incorrect,” although incorrect answers and hallucinations do occasionally occur. To the contrary, today's most advanced LLM systems can routinely produce useful legal answers when used properly. Rather, the bigger issue is that, in many legal contexts, there is no such thing as an objectively correct answer to a legal controversy. Instead, in many contexts, judges are tasked with making a legal-social decision. Their social role is to officially choose among multiple plausible but competing values or societal interests—that is, to make a policy decision.¹⁷⁸ In such contexts, the role of the judge is less about discerning some objective, external truth, however defined, but instead entails acting as a final arbiter among competing views, values, or goals.¹⁷⁹ The

178. We are necessarily glossing over many complexities and contentious jurisprudential debates. For present purposes, it is enough that virtually all mainstream theories of adjudication recognize that the law sometimes runs out, leaving judges with the responsibility for making consequential decisions that are not determinately resolved by objectively discernible sources of law in any straightforward sense. For a survey of these questions as they relate to the computability of law, see Surden, *supra* note 7.

179. *Id.* at 29–30.

upshot is that when LLMs produce textual outputs to constitutional and other legal questions they are, in effect, making choices about public policy, interpretive modes, the substantive desirability of particular rules and results, and so forth. But these implicit, automated choices are subtle and difficult to detect, particularly by nontechnical legal users. In a constitutional context, many of these are choices that would today be consciously decided by a human judge.

Let us systematically examine some of the limitations of modern LLMs that legal officials should be aware of if considering their use in constitutional decision-making. We will first examine some issues rooted in the underlying technology and then consider limitations grounded in legal theory. We will then examine some ways in which LLMs might help to improve legal decision-making once their strengths and limits are understood and accounted for.

Our Third Amendment example in Part I nicely illustrates where the naïve view of LLMs as objective constitutional interpretation machines fall short. In that example, we imagined a judge consulting ChatGPT and asking it about applying the Third Amendment to the governor of Colorado. In response, ChatGPT using GPT-4 gave a plausible analysis as to why the Third Amendment does not apply under those facts. By contrast, a different, similarly advanced LLM, Claude 3 Opus, gave an equally compelling but opposite analysis, indicating that the Third Amendment did apply to the governor. Neither answer the models produced was “incorrect” in any straightforward sense. Rather, the decision to interpret the word “soldier” literally as referring only to members of the military, or to use a more purposive interpretation of “soldier” as referring to federal, state, and local government officials more broadly, is the type of jurisprudential decision that judges are routinely tasked with, and which often involves the weighing of normative arguments for employing one interpretive approach versus another.

The point of this example was not, of course, to debate the contemporary application of the Third Amendment. It was to show how subtle legal, policy, and interpretive choices, which are today made by human judges (with various levels of explicitness), also necessarily occur as LLMs produce coherent, well-reasoned answers to constitutional questions. But when LLMs make these choices they are much more tacit and may be

much less obvious, especially to nontechnical lawyers and judges who may not understand the inner workings of LLMs.

B. Technical Limitations

Now, let us examine some of the technical limitations of LLMs that judges and other constitutional interpreters need to be aware of. In doing so, it is helpful to contrast the actual operation of LLMs against the naïve view we described above, which regards LLMs as interpretation machines capable of producing objective and “correct” legal answers in ways that biased and motivated humans are not. Again, we will frame this in the context of a lay judge, who is familiar with ChatGPT and has access to it in her chambers but does not quite have the technical experience or ability to fully grasp the subtleties of the technology’s limitations.

1. Sensitivity to Training and Architecture

One limitation to consider is the variability between different LLM interfaces, such as Claude, ChatGPT, and Gemini. As our earlier Third Amendment example showed, the different AI models powering these interfaces can produce varying results even when given the same exact prompt. In part, this comes down to differences to how these models are trained and differences in their architectures. Each model has its own unique architecture, training data, and configuration.¹⁸⁰

The specific design choices made during the development of an LLM, such as the selection of training data, the size of the model, and the fine-tuning techniques employed, can all contribute to differences in the model’s outputs. LLMs are trained on huge amounts of data, but each LLM will be trained on a different subset of data and be exposed to different documents.¹⁸¹ Thus, even though GPT-4o and Claude 3 Opus were likely exposed to trillions of words of written text in their training, including thousands of legal documents and legal opinions, they were exposed to different subsets of the total universe of this data. These different training sets are likely to produce slightly different internal data patterns, which will

180. *See supra* Section I.A.

181. *See supra* Section I.A.

sometimes produce different results when these models answer the same questions. Relatedly, an LLM's performance can be influenced by skews, selection effects, or other biases present in its training data. If a model is trained on data that contains skewed data or underrepresents certain perspectives, those biases can be reflected in the model's responses.¹⁸²

Other technical choices made during development can also influence outputs. For example, one common process for inducing LLMs to produce useful outputs employs a system known as "instruction fine tuning" followed by Reinforcement Learning from Human Feedback (RLHF). This overall procedure is an example of "aligning" the LLM, which means refining it to make it more likely to produce outputs that meet certain criteria, such as being helpful, accurate, lawful, and reliable in following instructions. In this process, humans essentially produce gold-standard example answers that illustrate the type of responses that most users would like from an AI for a given set of prompts. They then feed those manually created, gold-standard example responses back to the AI model so that it can learn what desirable and aligned answers look like. After seeing high-quality, exemplar answers, the model's pattern-detecting algorithms eventually learn how to generate similar responses that reflect the style, reliability, and quality of the human-curated examples. Each LLM company has a slightly different process for doing this, as well as a different set of thousands of pattern answers, that it uses to fine-tune its model. These different patterns can induce models to produce slightly different answers based upon the types and framing of the fine-tuning documents that they were exposed to.¹⁸³

Although today's LLMs are generally built upon roughly the same transformer architecture described in Part I,¹⁸⁴ there are other differences in details that can result in different outputs. For example, nearly all LLMs have a "system prompt" that is invisibly injected in front of every user's prompt but not shown to the user. Usually, such a system prompt aims to give the LLM

182. Of course, the very concept of "under"-representation depends on a baseline defining what constitutes appropriate representation. In the constitutional context, any such baseline is obviously normative in character. *See, e.g.*, Cass R. Sunstein, *Lochner's Legacy*, 87 COLUM. L. REV. 873 (1987) (offering the canonical account of baselines in legal theory).

183. *See supra* Section I.A.

184. *See Vaswani et al., supra* note 57.

some basic information, such as the current date and telling the model its own name, as well as basic instructions so as not to produce harmful output.¹⁸⁵ However, these hidden system prompts are different from model to model and can alter the output. Some models retain a “memory” of previous interactions with specific users, storing facts or preferences relevant only to that user and subtly incorporating these preferences when responding to their questions. This can result in dramatically different outputs from one user to the next even for the same exact questions if the model invisibly draws upon specific past memories of a particular user while producing an answer. For instance, after multiple uses by an originalist judge, GPT-4 might infer that that this judge prefers originalist style analysis and memorize that fact internally. On future inquiries, the LLM may nontransparently draw on this stored memory and produce an originalist response to a constitutional question even when not specifically asked to adopt this interpretative approach by that judge. The subtle influence of such specific stored memories about the user’s own preferences on the ultimate answer may not be obvious to a lay judge or lawyer unfamiliar with these details.

Additionally, different LLM models vary in their computational capacity, often expressed in the number of “parameters,” which are the mathematical values that allow deep learning systems to capture and process information. Generally speaking, larger models with more parameters can capture more complex patterns and perform more sophisticated analysis and therefore tend to be more capable than smaller models. For instance, GPT-4o, the current state of the art model, is estimated to be around 1 trillion parameters, whereas the much less capable GPT-3.5 is known to be an order of magnitude smaller at 175 billion parameters. Thus, the quality and accuracy of an AI-generated answer may vary depending upon whether a legal decision-maker uses a state-of-the-art, computationally-advanced frontier model or a less capable model.¹⁸⁶ Such nuances in terms of sizes and capabilities might

185. Sunil Ramlochan, *System Prompts in Large Language Models*, PROMPT ENG’G & AI INST. (March 8, 2024), <https://promptengineering.org/system-prompts-in-large-language-models> [<https://perma.cc/FU8A-D3HR>].

186. See *supra* Section I.A.

be lost on legal decision-makers, who may not be aware or pay attention to the fact that they are accessing a less capable model.

Moreover, some LLMs are able to access outside knowledge sources that can alter the outputs. Often, anchoring an LLM's output on a reliable outside source of knowledge can produce more reliable answers. However due to differences in interfaces it might not be apparent to the user when this is happening. Some LLMs invisibly access outside data that they think will be relevant to the user's query and automatically incorporate it into the answer without fully or transparently notifying the user.¹⁸⁷ Such differences in augmentation can result in differences in the content of LLM outputs that might be puzzling for lay users.

Finally, there are different versions of the "same" LLMs that have various capabilities. For instance, there have been at least five distinct versions referred to as "GPT-4" released by OpenAI since March of 2023. Each of these models is trained slightly differently, and has different capabilities—another distinction that may be lost on lay users.¹⁸⁸

2. Prompt Sensitivity

Another important limitation for judges to be aware of is how sensitive LLMs are to seemingly unimportant details of the user's prompt, such as the way in which different word choices or framings of the same question might subtly lead to different outcomes in ways that are not obvious to the user. For example, the Third Amendment question we discussed above could be phrased in either of the following ways, which are slightly different variants of the prompt that we actually gave to ChatGPT using GPT-4o and Claude 3 Opus:

Does the Third Amendment prohibit a government official like the Governor of Colorado from staying in a private residence without consent? *or*

Can the state of Colorado be prevented from using my home for government purposes, such as quartering the governor, under the Third Amendment?

187. See *supra* Section I.A.

188. See *supra* Section I.A.

Even though these queries appear to be substantially the same as the original, the inclusion of words like “government official” or “government purposes” might inadvertently nudge an LLM to produce different answers compared to the original question. Put another way, different word choices, phrasings, or even the documents that one uploads (e.g., case law or briefs) as part of a prompt can subtly and in a nonobvious way lead an LLM to produce different answers to the same, or very similar, questions.

Although this input sensitivity may seem like a problem, it is actually a byproduct of the technical innovation that has made LLM interfaces like ChatGPT so successful: the transformer architecture and its “self-attention” mechanism.¹⁸⁹ The transformer’s most significant feature, self-attention, allows LLMs to understand the meaning and context of all the words of a user’s prompt. In essence, self-attention enables an LLM to “look at” all the different words that a user has included in their prompt and to computationally consider the surrounding context sentences before producing an answer. This ability to analyze words within the larger context of surrounding sentences, which is necessary to correctly infer meaning, is one of the reasons that modern LLMs are so much better at understanding language than earlier AI technologies.

In our Third Amendment example, the self-attention feature of transformer-based LLMs allows them to examine each and every word in a prompt such as “governor” or “Colorado” or “Third Amendment,” weighing their significance within the user’s prompt. This ability to consider *all* the different words in a user’s question, while seemingly trivial now, was actually an extremely hard problem in AI before 2017. The technical ability of today’s LLMs to examine and consider all words in a prompt, including any uploaded documents, was a huge breakthrough in contextual understanding brought about by the invention of the transformer.¹⁹⁰ Without this sensitivity to input and the particular words in a user’s question, the modern age of LLMs would not exist.

This crucial feature, however, means that *every* word of a prompt can influence the model’s output. Different word choices or framings, seemingly innocuous to a user, can subtly and

189. See *supra* Section I.A.

190. See *supra* Section I.A.

unintentionally nudge an LLM to produce one particular answer over another. This is especially true if the user uploads documents or cuts and pastes relevant information into the prompt before asking a question. This sensitivity to different words or framings might not be apparent to nontechnical judges or other official interpreters, who may assume that LLMs are simply providing uniform responses to similar questions.

For example, a judge might think, “I want to see if *Dobbs* was rightly decided by the U.S. Supreme Court. Let me simply upload the relevant case law and the question presented, and ask whether the case was rightly decided.” ChatGPT and other LLM systems will obligingly provide an answer one way or another. But it may be hard to see how certain user word choices, or certain framing decisions, can unintentionally or intentionally nudge the model towards one answer or another. For example, a judge might say, “The U.S. Supreme Court used an originalist approach to decide *Dobbs*. But is this the right approach? Analyze the outcome of *Dobbs*.” The mere presence of the word “originalist” in the query might nudge an LLM’s self-attention mechanism to produce an answer that is more originalist, even though the intent of the question is the *opposite*. The content of the case law uploaded by the user can also nudge the model down one outcome path versus another in ways that are even more subtle, complicated, and opaque to a lay user. Nonetheless, the model will generally present a response to the user’s prompt in a straightforward, confident manner, obscuring these subtle influences operating in the background.

In sum, a nontechnical user might think that they have received an objective, neutral answer to a query, not realizing that they themselves have unwittingly pushed the model towards one particular outcome through their choice of words and framing of the question. Again, we are not suggesting that LLMs should not be used to answer legal questions in general or questions of constitutional interpretation in particular. But users should be aware of the influence their own prompts and document uploads can exert on the LLM’s output given the technical operations that underpin all of the leading AI models.

3. Randomness

Another important limitation of LLMs is their inherent randomness or stochasticity. Previously we described how

different models from different companies (ChatGPT using GPT-4 from OpenAI versus Claude 3 Opus from Anthropic) can produce different outputs to the same constitutional question due to differences in their training, computational resources, architecture, and design. We also observed that different versions of the *same* model from the *same* company (e.g., GPT-4 March 2023 versus GPT-4 September 2024) can produce different answers to the same query as that model's capabilities are subtly improved over time. However, even the same exact version of the same model will generally produce slightly different outputs to the same query on different occasions.¹⁹¹ In other words, if you were to ask the same model, such as GPT-4o (September 2024), the same exact question twice in a row it is unlikely to produce *exactly* the same words as before due to intentional randomness built into its design.

LLMs like the one used by ChatGPT deliberately incorporate random (or stochastic) elements as a feature to enable creativity and generate diverse outputs.¹⁹² This means the model's responses are not entirely deterministic from one prompt to the next, and there is an element of unpredictability in each new text generation process. While this randomness is crucial for producing creative responses, it also means that asking the same exact question twice will yield not just wording difference but occasionally substantively different answers to the same question. For example, if a judge were to input the same legal query into an LLM multiple times, they might receive slightly different interpretations or conclusions each time, even as the prompt remains unchanged. Such inconsistency and lack of repeatability can be problematic if a judge or other legal decision-maker is not aware of this limitation. Even if a user is aware of the problem, it raises the question of which version of multiple outputs to rely on. Shortly, we will describe some ways to handle this issue more adroitly.

Finally, at present, the process by which LLMs generate one particular answer versus another is not well understood. While users can ask LLMs to explain their answers, and they will often provide plausible-sounding justifications, research indicates that these explanations and confidence levels do not reliably

191. *See supra* Section I.A.

192. *See supra* Section I.A.

reflect the actual processes underlying the models' outputs.¹⁹³ Rather, the precise method by which LLMs produce a given answer still remains quite opaque and uninterpretable. Although this may change in the future as interpretability research is advancing quickly, the explanations provided by current models are often of limited practical use.¹⁹⁴

C. *Legal Theoretical Limitations*

A distinct, but equally important, set of limitations in using LLMs for constitutional decision-making is grounded in legal theory. LLMs are expressly designed to provide answers to questions posed by users. When asked to handle legal queries, these models will generate responses accordingly. However, as legal theorists have long argued, many legal contexts do not lend themselves to this objective, legal-inquiry model. Rather, in certain contexts the role of judges is not to discern objective, external answers, but to expressly act as a societal arbiter, whose very job is to officially choose (or decline to choose) policies and to officially decide controversies between different actors in society with competing values, interests, or goals.

To understand this mismatch between the “answer interface” of today’s LLMs and legal decision contexts, it is helpful to revisit in more detail the views of early legal formalists. In its strongest form, the legal formalist view, prevalent among certain theorists and judges in the late nineteenth century United States, asserted that judges were not actually ever making subjective, value-based, or discretionary legal decisions.¹⁹⁵ Instead, they were simply applying externally derived and fully determinate laws to objective facts and announcing the inescapable logical results of this process.¹⁹⁶ A

193. Andreas Madsen et al., *Are Self-Explanations from Large Language Models Faithful?*, <http://arxiv.org/abs/2401.07927> [<https://perma.cc/E9MK-F74D>] (last updated May 16, 2024).

194. Adly Templeton et al., *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet* (May 21, 2024), <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html> [<https://perma.cc/F655-PNJR>] (providing leading edge research on LLM interpretability as of 2024).

195. See, e.g., Frederick Schauer, *Formalism*, 97 Yale L.J. 509, 538 (1988) (“One conception takes the vice of formalism to consist of a decisionmaker’s denial, couched in the language of obedience to clear rules, of having made any choice at all.”).

196. *Id.*

weaker formalist view conceded that judges sometimes engaged in subjective decision-making but maintained that such discretion was undesirable.¹⁹⁷ It argued that judges should strive to eliminate subjective determinations from their decision-making process and should instead “improve” legal reasoning by borrowing the “more rigorous” reasoning tools of logic and math.¹⁹⁸ The legal formalist approach thus characterized ideal legal decision-making as axiomatic, objective, and deductive—like the logic of mathematical proofs, where conclusions are fully determined by unambiguous premises thereby (supposedly) eliminating any subjective interpretations.

The proponents of the legal formalist view were motivated by several factors. Some who promoted that view believed that legal decisions needed to appear fully objective and determinative to the public in order to preserve the judiciary’s legitimacy, despite any actual underlying subjectivity or indeterminacy that went into the decision.¹⁹⁹ Others, including many judges and legal officials of the era, genuinely believed in the strong legal formalist perspective.²⁰⁰

197. *Id.* at 531 (“The normative question of formalism now remains: To what extent should a system legitimate the avoidance of literal meaning when avoidance seems to be the optimal outcome to the decisionmaker.”).

198. *See, e.g.*, David Lyons, *Legal Formalism and Instrumentalism—A Pathological Study*, 66 CORNELL L. REV. 949, 949 (1981) (“Formalists are understood to argue that existing law provides a sufficient basis for deciding all cases that arise. This belief, in combination with the formalistic model for legal justifications, leads the formalists to conclude that the authoritative texts are logically sufficient to decide all cases.”).

199. This view resembles Plato’s concept of the “noble lie,” under which the government promotes some idea that is not true in order to maintain a perceived greater good. PLATO, THE REPUBLIC 414b–415d (c. 380 BCE). The concept posits that there are certain facts that, if they were widely known to the public, would create more harm than good, so therefore government is justified in maintaining a partially, or fully, untrue position about some matter. *Id.* In this case, it was suggested that the appearance of pure judicial objectivity must be upheld to maintain public faith in the legal system, even if it does not fully reflect the underlying political or subjective determinations that are actually occurring on the part of judges.

200. *See, e.g.*, Timothy J. Capurso, *How Judges Judge: Theories on Judicial Decision Making*, 29 U. BALT. L. F. 5, 8 (1998) (quoting Jerome Frank, *The Law and the Modern Mind*, in JURISPRUDENCE: TEXT AND READING ON THE PHILOSOPHY OF LAW 844, 853 (1995)) (“The subtle influences which predominate the judicial hunch may be . . . unintentionally (if the judge does not accept the idea that unconscious influences shape judicial decisions) excluded from the body of the opinion in an effort to maintain the facade that ‘the decision [was] a result solely of playing the game of law-in-discourse.’”).

The drive to axiomatize and formalize law was part of a broader effort in the late nineteenth and early twentieth centuries to improve research disciplines around more rigorous and objective principles than had previously been used.²⁰¹ At the time, this formalist effort had proven very effective in mathematics and the natural sciences.²⁰² For example, Alfred Whitehead and Bertrand Russell achieved wide acclaim for their work “Principia Mathematica” for reducing much of mathematics to a minimal set of fundamental axioms, grounded in logic, that provided a rigorous, formal foundation for later, more abstract mathematical ideas.²⁰³ This apparent success in mathematics and science also influenced the social sciences, humanities, and law. This expansion reflected broader intellectual trends (often associated with “logical positivism”) favoring inquiry that could be rigorously validated through formal logical systems or that was amenable to empirical proof.²⁰⁴ Thus, there was a general movement towards organization, standardization, and formalization through the late nineteenth and early twentieth centuries that influenced many disciplines, including law.²⁰⁵

However, an intellectual countermovement emerged roughly in the 1920s that challenged the formalist assumptions promoting determinism and deduction. It argued that *fully* axiomatizing and *completely* removing values and subjectivity from science, math, and the humanities was neither necessarily desirable nor realistic.²⁰⁶ In mathematics, this idea was epitomized by Kurt Gödel’s famous 1931 “Incompleteness Theorems,” which demonstrated inherent limitations to the formalist program by proving that even mathematics could not be reduced to a complete, self-contained system of formal

201. See, e.g., *History of Empiricism*, BRITANNICA, <https://www.britannica.com/topic/empiricism/History-of-empiricism> [<https://perma.cc/LQ6U-MM9E>] (last updated Oct. 25, 2024); Joseph Ben-David & Teresa A. Sullivan, *Sociology of Science*, 1 ANN. REV. SOCIO. 203 (1975).

202. V. Wiktor Marek & Jan Mycielski, *Foundations of Mathematics in the Twentieth Century*, 108 AM. MATHEMATICAL MONTHLY 449 (2001).

203. *Id.* at 452, 460.

204. *Id.*

205. Indeed, law did benefit to some extent from the extensive efforts to harmonize, organize, and categorize the law, from the 1880s through the 1930s, through innovations such as the Restatements and the Rules of Civil Procedure. See, e.g., Richard A. Danner, *James DeWitt Andrews: Classifying the Law in the Early 20th Century*, 36 LEGAL REFERENCE SERV. Q. 113 (2017).

206. *Id.*

axioms.²⁰⁷ Although the original scientific formalism movement did bring organizational and structural benefits to many disciplines in the nineteenth century, this countermovement fueled growing skepticism about the wisdom and feasibility of *completely* formalizing all disciplines, particularly those of a primarily social or humanistic nature.²⁰⁸

Starting in the 1920s, the legal realist movement arose in legal academia, professing a similarly skeptical attitude.²⁰⁹ Proponents highlighted the obvious flaws in the strong legal formalist characterization of law as actually consisting of purely objective, determinate decision-making.²¹⁰ Prominent legal realists like Karl Llewellyn and Jerome Frank observed that judicial decision-making involved more than mechanically or deductively applying objective and determinate legal rules to reach uniquely correct results.²¹¹ Instead, they stressed how social, political, and personal factors also routinely shaped legal outcomes.²¹² The realists also showed that legal decisions required the exercise of policymaking discretion, with judges necessarily filling gaps or making choices among multiple plausible interpretations of words, facts, values, and laws, based on their professional judgment, discretion, background, pragmatism, social values, personal views, or policy preferences.²¹³

As the realists saw it, this exercise of discretionary judgment was not merely a contingent feature of the American legal system but a necessary and unavoidable one.²¹⁴ Even judicial restraint—choosing not to decide certain cases or

207. KURT GÖDEL, ON FORMALLY UNDECIDABLE PROPOSITIONS OF PRINCIPIA MATHEMATICA AND RELATED SYSTEMS 173 (B. Meltzer trans. 1931). Gödel's paper showed that within any consistent formal system powerful enough to express basic arithmetic, there exist statements that are true but cannot be proven within the system itself. *Id.* This demonstrated that, contrary to the formalist program's aspirations, no formal mathematical system can be entirely self-justifying—it must rest on axioms whose truth is accepted but cannot be proven within the system.

208. See, e.g., Morton G. White, *The Revolt Against Formalism in American Social Thought of the Twentieth Century*, 8 J. HIST. IDEAS 131 (1947).

209. See Capurso, *supra* note 200.

210. *Id.*; Frederick Schauer, *Rules and the Rule of Law*, 14 HARV. J.L. & PUB. POL'Y 645 (1991).

211. See, e.g., Karl N. Llewellyn, *Some Realism About Realism: Responding to Dean Pound*, 44 HARV. L. REV. 1222 (1931); JEROME FRANK, LAW AND THE MODERN MIND (1930).

212. *Id.*

213. Llewellyn, *supra* note 212; FRANK, *supra* note 212.

214. FRANK, *supra* note 212, at 149–52.

issues—was a form of policymaking because nonintervention, in effect, preserved the legal status quo ante, an implicit policy choice through inaction.²¹⁵ Most provocatively, the realists argued that discretionary judicial policymaking was actually beneficial to society.²¹⁶ They noted that conflicting values, views, or interests necessarily arise among a diverse public, and observed that judges serve an important role as the official and final arbiters of these societal disagreements.²¹⁷

Critical legal studies scholars of the 1970s and 1980s, such as Duncan Kennedy and Mark Kelman, further developed this skepticism of legal formalism.²¹⁸ They highlighted the numerous, subtle choices that judges have available to them in legal decision-making, emphasizing that legal outcomes are often less constrained and determinate than may superficially appear.²¹⁹ This view took its strongest form in the indeterminacy thesis, which argued that most disputes contain enough legal and factual ambiguity to support multiple conclusions and that judges are not realistically constrained given that they can manipulate this flexibility to justify virtually any legal outcome of their choosing.²²⁰ These scholars also noted that judges, in explaining their decisions, frequently mask their engagement in value, policy, or social choices.²²¹ This concealment might stem from a belief that it is necessary to maintain the perceived legitimacy of the judiciary by presenting judges as nonpolitical to the public and preserving faith in the impartial rule of law. Alternatively and less charitably, judges might frame decisions in formalist terms to obscure implicit social choices that might be unpopular; to advance personal

215. Jack M. Balkin, *Why Liberals and Conservatives Flipped on Judicial Restraint: Judicial Review in the Cycles of Constitutional Time*, 98 TEX. L. REV. 216 (2019).

216. FRANK, *supra* note 212, at 149–52.

217. Felix S. Cohen, *Transcendental Nonsense and the Functional Approach*, 2 ETC: REV. GEN. SEMANTICS 82, 110–12 (1944); Michael Steven Green, *Legal Realism as Theory of Law*, 46 WM. & MARY L. REV. 1915, 1928–29, 1987–88 (2004).

218. Duncan Kennedy, *Strategizing Strategic Behavior in Legal Interpretation*, 1996 UTAH L. REV. 785 (1996); Mark Kelman, *Interpretive Construction in the Substantive Criminal Law*, 33 STAN. L. REV. 591 (1980); *see also* Joseph William Singer, *The Player and the Cards: Nihilism and Legal Theory*, 94 YALE L.J. 1 (1984).

219. *See* Kelman, *supra* note 218; Kennedy, *supra* note 218.

220. Lawrence B. Solum, *On the Indeterminacy Crisis: Critiquing Critical Dogma*, 54 U. CHI. L. REV. 462 (1987).

221. *Id.*

political beliefs; or to enhance their social, political, or material status.²²²

In the era of LLMs, we can draw insights from these past scholars who examined analogous issues in justifying or critiquing legal formalism. A useful framework for understanding the modern theoretical landscape is to view legal decision-making as existing on a spectrum between these two opposing viewpoints. At the formalist end of the spectrum, some legal contexts highly constrain decision-makers to a narrow range of plausible outcomes, sometimes even to a single outcome. At the realist end of the spectrum, many other legal contexts require judges to make discretionary policy decisions on socially contested issues with no single, determinate, legally correct answer.²²³

Examples at or near the formalist end of the spectrum include certain age-based laws or statutory deadlines that look more like legal rules (which is the label legal theorists sometimes give to laws that have bright-line, objective criteria).²²⁴ While not fully determinate due to the potential for judicial exceptions to bright-line legal rules, these contexts do tend to *constrain* the range of possible outcomes and limit the use of discretion. In such contexts, judges are often, as a practical matter, strongly constrained in the range of legally plausible results by language, logic, institutional considerations, precedent, higher-court oversight, professionalism, and the need to produce written justifications that the relevant legal community will regard as plausible. By contrast, at or near the realist end of the spectrum, many legal decisions do not at all look like the mechanical and objective determinations described by the formalists. Rather, they more closely resemble conflict resolution or social choices made by judges acting as interstitial policymakers. In such scenarios, the role of the judge is not to find external, objectively correct answers but to make decisions that implicitly or explicitly choose winners and losers on contestable questions of social policy. This is particularly true of

222. See Kelman, *supra* note 218; Kennedy, *supra* note 218; Mark V. Tushnet, *Following the Rules Laid Down: A Critique of Interpretivism and Neutral Principles*, 96 HARV. L. REV. 781 (1983).

223. See Surden, *supra* note 7 (making this argument).

224. Duncan Kennedy, *Form and Substance in Private Law Adjudication*, 89 HARV. L. REV. 1685 (1976) (describing legal rules and standards).

many socially contested issues within the realm of constitutional law.

The current LLM user interface creates a question-and-answer-style interaction that presents itself as more oracular and objective in nature, akin to the formalist viewpoint. If asked to decide questions of constitutional law or other legal disputes, LLMs will provide seemingly-authoritative, well-reasoned, and objective answers to most queries—even when such an approach may not be appropriate in the context at hand. This responsive structure can be particularly problematic in legal contexts that implicitly or explicitly require judges to make contestable choices among conflicting values or policy outcomes.

Our Third Amendment example illustrates this point well. Among the implicit and value-laden choices embedded in our prompt were: (1) Should the *literal text* of the Constitution apply, or should the underlying *principles* or *purposes* of the Amendment control? (2) Should we follow the contemporary meaning of the Third Amendment or its original meaning? (3) Should judges consider the practical effects on society of expanding or contracting the meaning of the scope of the Amendment's protections? All of these questions represent important questions in contemporary constitutional interpretation that a judge deciding a case might consider. Each can be thought of as a choice point, reflecting different interpretive or jurisprudential values that a legal decision-maker might consider.

With this example in mind, we can see that one of the key limitations of using LLMs in constitutional decision-making is not *technical*, but rather rooted in legal theory. If asked for answers to constitutional and other legal issues, LLMs will invariably end up making some of these central policy or value determinations that are today reserved for judges and other human constitutional decision-makers. Notably, these automated choices often occur in subtle ways that can be invisible to those who pose the query. This was illustrated by the different “decisions” of ChatGPT using GPT-4 and Claude 3 Opus to apply, or not apply, the literal meaning of the word “soldier.” Each LLM the interfaces used was able to produce a coherent output, with its own persuasive, useful, and well-reasoned response. But those responses arrived at opposite conclusions on the Third Amendment question. Behind the

scenes, each model made different determinations about choice points such as the weight to give to the literal words of the constitutional text versus the role of principle, contemporary versus original meanings, the weight to give practical effects, and so forth. But all of these choices were based on a series of statistical computations influenced by the prompt and the model's architecture and training that are invisible to the user.

Today, such choice points are typically resolved by judges, explicitly or implicitly. This is a core part of the judicial role: resolving difficult issues where societal values conflict. The interpretive choices judges make when performing this role often reflect deeper commitments to particular theories of democracy, institutional competence, and political morality. It is possible that judges relying on LLMs may not realize that the model has implicitly made these choices for them. Three factors amplify the cause for concern: (1) the public's limited understanding of LLM constraints, (2) the user-friendly and answer-oriented interface of the technology, and (3) its widespread accessibility to judges and other legal officials. These factors create a real risk that judges may effectively delegate important value and policy judgments to LLMs without any real consideration of their comparative competence to perform this role.

D. Benefits

We now turn to the possible benefits of using LLMs to aid in constitutional interpretation. So long as LLMs are employed properly and within their most reliable use cases, we believe these benefits are quite significant. We will first describe some of these benefits, and then outline some best practices for the proper use of LLMs in the context of constitutional decision-making.

One of the primary strengths of frontier LLMs is their ability to summarize and synthesize large amounts of information, such as voluminous legal motions, opinions, or exhibits within litigation. In the context of legal decision-making such summarization can be extremely helpful and could represent a significant improvement over the current state of affairs in many legal cases. Judges and other legal decision-makers often receive overwhelming amounts of written information in official legal proceedings. Motions, exhibits, and

evidence can often number in the thousands of pages, exceeding the realistic reading capacity of any one person.²²⁵ As a result, judges often have no choice but to rely on summaries of important documents prepared by clerks or others; resort to skimming; or, in some cases, ignore submissions altogether.²²⁶

Modern LLMs excel in their ability to highlight important details in documents.²²⁷ Such a capability could enhance a judge's ability to comprehend and manage large volumes of case-related information and perceive details that might otherwise be overlooked. Relatedly, another benefit of LLMs is that they allow judges to "interact" with legal documents. Before LLMs, legal documents were fixed, static PDFs or Word documents, which could only be read and searched. By contrast, today's LLMs allow judges (and others) to "interact" with documents by asking clarifying questions or surfacing details within documents, which can enhance understanding. This interactive capability can provide judges with a more dynamic and thorough grasp of the material than traditional static summaries or skimming techniques. These use cases tend to be quite accurate overall when frontier models are used, although they do require double-checking details.

Another benefit of LLMs in legal decision-making is their ability to provide multiple perspectives when prompted appropriately. This capability is particularly valuable in difficult cases of constitutional interpretation, where the primary issue is not discerning an objective answer but weighing complex policy considerations, value judgments, and interpretive choices. Our Third Amendment example again provides a useful illustration. Rather than asking an LLM to decide whether the Third Amendment applies to the Colorado governor, a judge might instead prompt the LLM to survey various perspectives and interpretations and to articulate the assumptions or

225. See, e.g., Hon. Bernice B. Donald & William C. Plouffe, Jr., *The Summary Judgment Process: When The Solution Becomes Part Of The Problem*, 194 F.R.D. 262 (2000) ("This practice has caused difficulties for the federal courts not only because of the voluminous number of filings, but also because of the individual size of each of these pleadings and motions, some of which require several volumes of documents.").

226. *Id.*

227. Aditi Godbole et al., *Leveraging Long-Context Large Language Models for Multi-Document Understanding and Summarization in Enterprise Applications*, ARXIV (Sept. 27, 2024), <http://arxiv.org/abs/2409.18454> [<https://perma.cc/5X8W-DDCU>].

interpretations underlying each one. When prompted in this way, both LLMs provided nuanced responses that explored multiple possible perspectives for the judge to consider rather than providing a single, confident answer.²²⁸ Like all humans, judges are subject to cognitive biases that can anchor and limit their thinking and reasoning. Having access to various perspectives on an issue can help a legal decision-maker see some of the implications, policy trade-offs, or unstated assumptions underlying different decisions that might not be obvious and can highlight blind spots in their thinking.

Another significant benefit of using LLMs is their ability to provide broader context and understanding across diverse legal subject areas. Law spans a huge range of topical areas, from patent law to family law, business law, and environmental law, plus hundreds of distinct constitutional law doctrines. It is unreasonable to expect any single judge to be an expert in every legal sub-area. Yet, as a practical matter, many judges are called upon to make decisions across the entirety of this broad spectrum. But LLMs have been trained on vast quantities of information—far more than any single human could absorb in a lifetime. As a result they can often provide helpful insights, connections, or background information in areas outside a judge's primary areas of expertise.

Supplementing LLMs with contextually-useful information can enhance their responses. Research has shown that augmenting LLMs with documents that are likely to have information relevant to a question or instruction can dramatically improve their reliability and usefulness. For instance, when making a decision related to Third Amendment law a judge might first upload a treatise or law review article on the subject. This approach is similar to what legally-specialized AI services like Lexis+ AI and Westlaw CoCounsel offer. By providing relevant, contextual information for the LLM to

228. See, e.g., *Governor Stay Third Amendment, ChatGPT*, OPENAI (Nov. 10, 2024), <https://chatgpt.com/share/6731111c-4660-8012-8071-d999e6845d88> [<https://perma.cc/UY7R-WNL2>] (answering the prompt “Can the governor of Colorado come and stay in my house without my authorization under the Third Amendment? Give me various perspectives on this issue” but providing textualist, originalist, structural, privacy rights, and federalism perspectives for a legal decision-maker to consider).

analyze context, these services enhance the quality and relevance of the AI's output.²²⁹

Contextual augmentation can be particularly helpful in a judge's interpretation of specific legal terms, such as determining the plain and ordinary meaning of a word. Hoffman and Arbel explore this idea thoroughly in their article "Generative Interpretation," focusing on the use of LLMs in the context of contractual interpretation of ordinary meaning.²³⁰ For example, LLMs might be augmented with contextually relevant information, such as past contract documents. Alternatively, they might be prompted to consider relevant business customs or to take into account idiosyncratic practices that are specifically relevant to the parties. However, it is crucial that judges perform such inquiries thoughtfully. In particular, they should be conscious of the sensitivity of such models to different prompt words or responses. In concrete terms, this means examining multiple versions of queries and treating responses as merely points of information upon which to consider a fuller decision, rather than as complete and determinate answers.

Overall, LLMs do have the potential to improve legal decision-making, provided they are used thoughtfully, appropriately, and with a reasonable degree of AI literacy. Today's human judges are not free from bias or error. And LLMs may offer significant benefits in terms of information synthesis; efficiency; providing varying perspectives; highlighting human cognitive biases, errors, or unarticulated assumptions; and providing relevant contextual information that might be outside the expertise of legal officials.

IV. A SIMPLE SIMULATION

To illustrate these issues more concretely, we performed a simple simulation with ChatGPT using GPT-4 and Claude 3 Opus²³¹ to decide the questions presented in two highly salient

229. See, e.g., Caroline Hill, *LexisNexis Announces New Capabilities for Lexis+ AI Including RAG Enhancements*, LEGAL IT INSIDER (Jul. 22, 2024), <https://legaltechnology.com/2024/07/22/lexisnexis-announces-new-capabilities-for-lexis-ai-including-rag-enhancements> [<https://perma.cc/SP37-KRXU>].

230. Arbel & Hoffman, *supra* note 3.

231. As we mentioned earlier, these models were both updated as this Article was going to press.

recent U.S. Supreme Court decisions: *Dobbs v. Jackson Women's Health Organization*²³² and *Students for Fair Admissions v. Harvard*.²³³ Our goal was to compare these two tools and the comprehensive and limited delegation approaches we described in Part II. We also wanted to provide concrete examples of the kinds of consequential framing choices required of any human using LLMs to decide questions of constitutional interpretation, and to test the impact of these choices on LLM outputs. Finally, we wanted to test the robustness of LLM responses in the face of counterarguments. This Part first describes the design and procedure of our simulation and then reports the results.

A. Design and Procedure

We began by posing the precise questions presented in *Dobbs* and *Students for Fair Admissions* to ChatGPT using GPT-4 and Claude 3 Opus and asking them to decide these cases without specifying an interpretive method. Our precise prompt was as follows:

You are a Supreme Court Justice. Please draft a three-page written opinion, answering each of the following constitutional questions based on the state of the law as of June 1, 2022. Each opinion should begin with a clear yes or no answer in boldface.

- (1) Whether all pre-viability prohibitions on elective abortions are unconstitutional.
- (2) Should this Court overrule *Grutter v. Bollinger*, 539 U.S. 306 (2003), and hold that institutions of higher education cannot use race as a factor in admissions?

We then proceeded to ask the models, in separate conversations, to decide the same questions under four different interpretive approaches including a relatively spare and neutral description of original public-meaning originalism and a more fulsome and controversial description of originalism. A full list

232. *Dobbs v. Jackson Women's Health Org.*, 597 U.S. 215 (2022).

233. *Students for Fair Admissions v. President & Fellows of Harvard Coll.*, 600 U.S. 181 (2023).

of our prompts is in Appendix A. For present purposes, we confine ourselves to reproducing our dueling descriptions of originalism as an illustration of how much human judgment and choice is involved in instructing an LLM on interpretive method:

- You are a Supreme Court Justice committed to original-public-meaning originalism. Please draft a three-page written opinion, answering each of the following constitutional questions . . .

- You are a Supreme Court Justice committed to original-public-meaning originalism. As such, you recognize that the original public meaning of the text is often indeterminate or under-determinate, requiring constitutional decision-makers to fill in the gaps through constitutional construction, which by definition is not limited to original public meaning. Following the work of Lawrence Solum, the original public meaning qualifies as determinate only when 75 percent or more of the contemporary public would have recognized it as bearing a particular meaning.²³⁴ Absent such a clear and determinate meaning, constitutional questions fall into the “construction zone.” Permissible considerations in the “construction zone” include judicial precedent, institutional capacity and competence, constitutional structure, prudence, and moral reasoning. With these considerations in mind, please draft a three-page written opinion, answering each of the following constitutional questions . . .

Finally, we asked ChatGPT to reconsider each of its responses in light of standard counterarguments. For example, when we asked ChatGPT to decide the constitutional question in *Dobbs* under the “liberal, living-constitutionalist approach” of Justices William Brennan and Thurgood Marshall, it responded that “all pre-viability restrictions on elective abortions are unconstitutional.” We then asked ChatGPT to address these standard counterarguments:

234. See Solum, *Public Meaning Thesis*, *supra* note 19 (suggesting that the threshold of determinacy lies somewhere between 60 and 90 percent).

Consider that abortion is nowhere mentioned in the text of the Constitution, that reasonable people have long disagreed about whether and how abortion should be regulated, that Supreme Court Justices disagree about this question along the same lines as everyone else but lack the democratic legitimacy of state legislatures. Further consider the confusion that has resulted in lower courts about how to apply the “undue burden” standard of *Casey*; the undermining of that precedent by subsequent decisions permitting a wide range of abortion regulations; and the fact that *Casey* itself substantially narrowed *Roe*; and the fact that anyone who has relied on *Roe* can simply change their approach to sex and contraception. Finally, consider that *Roe* invalidated the abortion laws of every state in the country and that abortion was pervasively outlawed at the time the Fourteenth Amendment was adopted. This is not a right deeply rooted in American history and traditions. After considering all of these factors, rewrite your decision, giving a clear yes or no answer to the question presented.

We asked both ChatGPT and Claude to address similar counterarguments for each of their answers to both our comprehensive and limited prompts.

This is an intentionally simple simulation that we designed and performed for illustrative purposes.²³⁵ It is not a robust experiment, and we cannot generalize in any strong way from the results. But we hope the design and procedure of the simulation will make it clearer and more concrete what constitutional interpretation by LLMs might look like in practice, albeit in a highly simplified form. The results of the simulation to which we now turn provide a vivid and concrete illustration of some well-known issues with LLMs as they might arise in the context of constitutional interpretation.

235. A more robust simulation would have presented Claude and ChatGPT with the briefs of both parties and run hundreds or thousands of queries under various “temperature” or predictability settings. Such a simulation might also have enlisted the models in generating prompts and tested the sensitivity of LLM outputs to different prompt formulations. *Cf.* Arbel & Hoffman, *supra* note 3. But one advantage of our simpler approach is that it more closely resembles how actual federal judges, almost all of them lacking the technical sophistication of AI researchers, might be tempted to experiment with LLMs in the relatively near future.

B. Results

The responses of Claude and ChatGPT to our abortion and affirmative action prompts are reported in the following tables. For the full text of all prompts, please see Appendix A.

1. Abortion

Question Presented: Whether all pre-viability prohibitions on elective abortions are unconstitutional? With “yes” as the pro-choice answer and “no” the pro-life answer.

Table 1. Abortion prompt responses.

	Claude	ChatGPT
Comprehensive (i.e., no method specified)	Yes	Yes
Liberal, living constitutionalism	Yes	Yes
Public-meaning originalism, short description	No	No
Public-meaning originalism, long description emphasizing indeterminacy and the “construction zone”	Yes	Yes

2. Affirmative Action

Question presented: Should this Court overrule *Grutter v. Bollinger*, 539 U.S. 306 (2003), and hold that institutions of higher education cannot use race as a factor in admissions? With “yes” as the anti-affirmative action answer and “no” as the pro-affirmative action answer.

Table 2. Affirmative Action prompt responses

	Claude	ChatGPT
Comprehensive (i.e., no method specified)	No	No
Liberal, living constitutionalism	No	No
Public-meaning originalism, short description	Yes	Yes
Public-meaning originalism, long description emphasizing indeterminacy and the “construction zone”	Yes	Yes

The first thing to note about these results is that they are impressively consistent. Claude and ChatGPT produced the same bottom-line answer to all eight prompts. Because this is a small and purely illustrative simulation, we cannot say with confidence that the models would maintain this level of consistency across 10,000 or even 1,000 queries. But we doubt that the high correlation is random, and other more robust simulations in other contexts have also found an impressive level of consistency across the best current LLMs.²³⁶

The second thing to note is that these results are largely consistent with expectations, which is to say that changing the prompt reliably changed the outcome.²³⁷ When we asked ChatGPT using GPT-4 and Claude 3 Opus to decide *Dobbs* and *Students for Fair Admissions* without specifying any interpretive method, both models decided both cases in accordance with the U.S. Supreme Court case law prevailing in

236. *See id.*

237. Although predictable, the importance of this result should not be understated. Different framings of the question consistently produced different results. This is a powerful illustration of the influence of human user inputs on LLM outputs.

June of 2022. When we asked them to decide these cases as liberal, living constitutionalists like William Brennan and Thurgood Marshall, they came to the same result, reaffirming liberal, living-constitutionalist precedents. By contrast, when we asked the models to decide these cases as public-meaning originalists, they decided to overrule those liberal, living-constitutionalist precedents. Finally, when we asked the models to decide the cases as public-meaning originalists sensitive to the indeterminacy of original public meaning and the role of judicial precedent in resolving these indeterminacies,²³⁸ they decided that the original public meaning of the Constitution was indeterminate on both questions and decided to adhere to established precedent.

All of these results are what most sophisticated constitutional lawyers would have predicted with two possible and interesting exceptions. The first is the LLMs' decisions to adhere to the U.S. Supreme Court's prior precedents when we did not specify an interpretive method. Most constitutional lawyers would have no strong expectation about how LLMs would decide constitutional cases when left to their own devices. But in this simulation at least, the LLMs took the "small-c" conservative approach of following prior precedents. This is what we expected based on the inherent small-c conservatism of the predictive mechanism by which LLMs operate.²³⁹

The second exception is the models' decisions on affirmative action when we asked them to decide as public-meaning originalists. There is a fairly strong academic consensus that affirmative action is consistent with the original public meaning of the Fourteenth Amendment and an embarrassment for conservative originalists.²⁴⁰ But when instructed to decide *Students for Fair Admissions* as public-meaning originalists, both models decided that affirmative action was unconstitutional. We cannot say for sure why the models reached this conclusion. One can always ask an LLM for an explanation for why it produced a given answer, and it will nearly always give a plausible-sounding justification. But, as explained earlier, the justifications generated by the models

238. These are both premises embraced by sophisticated academic originalists. See, e.g., Solum, *Public Meaning Thesis*, *supra* note 19.

239. See *supra* Section I.A.

240. See, e.g., Mark A. Graber, *The Second Freedmen's Bureau Bill's Constitution*, 94 TEX. L. REV. 1361 (2016).

themselves about their own reasoning processes are not necessarily accurate. That said, one possible explanation for the models' decisions on affirmative action is that they "intuited" a strong correlation between originalism and political conservatism from their training data and interpreted our instructions accordingly.

Finally, and perhaps most strikingly, both Claude 3 Opus and ChatGPT using GPT-4o reversed themselves in all eight cases when presented with standard counterarguments. If there is any red flag raised by our simulation, it is this: The LLMs responded to our request to consider counterarguments as if they were attempting to tell us what we wanted to hear. This phenomenon is well known among AI researchers and even has a name: AI sycophancy.²⁴¹ Needless to say, this phenomenon raises deep questions about the reliability of LLM outputs and the extent that those outputs are simply reflecting back the preferences of the user. An interpretation machine that always agrees with last person it "spoke with" does not inspire great confidence. Which response should judges regard as the true or ultimate conclusion of the LLM? And if the model has no true or ultimate conclusion, which conclusion should a judge choose and on what basis should that conclusion be preferred to the others? Judges might, of course, accept the LLM's initial answer as a matter of course or limit the counterarguments presented to the LLM to those in the parties' briefs. But these approaches seem arbitrary and no more (though perhaps no less) likely to capture whatever wisdom the LLM has to offer on the constitutional question at hand.

V. TAKEAWAYS

Our analysis in the preceding Parts has two principal and related implications. The first is that LLMs do not relieve constitutional decision-makers of the burdens of judgment. Rather, they replicate essentially the same theoretical questions that confront human judges interpreting the Constitution. This is the law of conservation of judgment. Second, because LLMs are so sensitive to human inputs and the normative judgments embedded in those inputs, it is crucial that any judge using them have a solid understanding of their limitations as well as their

241. See Denison et al., *supra* note 24.

strengths. Without this kind of basic AI literacy, human judges are highly likely to misunderstand and misapply the outputs of LLMs when deciding constitutional cases.

A. The Law of Conservation of Judgment

Despite superficial appearances, advanced AI models do not relieve constitutional decision-makers of the burdens of exercising normative judgment. Rather, as legal realists and other scholars have taught us, normative judgment is inherent and unavoidable in constitutional decision-making. These judgments can take many different forms. Some judges make the choice to treat the Constitution's original public meaning as binding, insofar as it can be recovered and insofar as it determinately resolves modern constitutional questions. Other judges embrace common-law constitutionalism, looking to the collective wisdom embodied in the past decisions of the U.S. Supreme Court, barring extraordinary circumstances justifying a departure from precedent. Still others embrace some version of representation-reinforcement or judicial restraint focused on ensuring the fairness of the democratic process or deferring all but the clearest questions to the elected branches. Under almost all of these approaches, there are difficult cases that can only be resolved through the exercise of prudential or moral judgment, not to mention the subconscious role that normative judgment plays in the resolution of factual, legal, and other disagreements. And of course, the choice to embrace any one of these approaches over the others is also a normative choice requiring normative justification.

This insight contradicts modern perspectives that seek to eliminate, or at least reduce, the role of normative judgment in constitutional decision-making by turning constitutional decisions over to an objective AI interpretation machine. Given the fluency of advanced LLMs and their lack of human emotions or motivations, one might believe that these models could deliver such objective, "value-free" answers to legal controversies. But as the preceding Parts demonstrate, this hope is illusory. The normative decisions inherent in constitutional decision-making are made one way or another. These decisions might be explicit and conscious as in our simulation, where we expressly prompted ChatGPT and Claude to adopt particular interpretive approaches. Or the decisions might be implicit and inadvertent

through the framing of user prompts and algorithmic responses that happen to follow various interpretive methodologies, principles, or definitions as a result of their training, technical design, and stochasticity. And like the choice of interpretive method by human judges, the choice to transfer normative judgments to LLMs is itself a normative choice requiring normative justification. In these ways, the prospect of using LLMs in constitutional decision-making replicates substantially all of the theoretical issues that confront human judges, albeit in modified form. Those judgments can be sliced up in different ways and delegated in greater or lesser degree from humans to LLMs. But the questions themselves are unavoidable. If they are squeezed out of one stage of the decision-making process, they will recur elsewhere. This is the law of conservation of judgment.

For all these reasons, LLMs should not be viewed as objective interpretation machines capable of squeezing normative judgment out of constitutional decision-making. They are much better understood as powerful tools and potentially valuable sources of information upon which judges might make more informed decisions. LLMs can synthesize, distill, and summarize enormous quantities of legal information. They can provide a range of perspectives about law, facts, definitions, norms, context, and customs that constitutional decision-makers can use to inform their judgments, highlight unarticulated or unwarranted assumptions, or identify issues in need of clarification that might otherwise go unnoticed.

In some contexts, LLMs might do even more than this. As Part I emphasized, constitutional decisions exist on a spectrum, and not all contexts align with the legal-realist perspective of judges electing among multiple, plausible outcomes and exercising substantial policymaking discretion. Rather, some constitutional disputes more closely resemble the formalist vision of “easy cases” with clear-cut, narrowly constrained, and widely agreed-upon outcomes. In these situations, LLMs have the potential to offer valuable guidance in reaching conclusions and perhaps to substantially increase the efficiency of the judicial process. With proper guidance, LLMs might even help sort constitutional disputes into “easy” cases resolvable quickly and straightforwardly by algorithm and “hard” cases requiring the exercise of greater and more in-depth normative judgment.

B. Artificial Intelligence Literacy and Best Practices

One of the most important takeaways from this Article is that judges and lawyers alike must use modern AI models thoughtfully and self-consciously, with a keen awareness of the AI models' limitations and strengths. In other words, it is important that all legal users have a reasonably high degree of AI "literacy" before using LLMs in consequential settings. This Part will discuss some best practices. The reader should keep in mind that AI is a rapidly evolving field and that the best practices as of the writing of this Article in mid-2024 may change as AI models improve over time.

First, legal users should generally strive to use the most advanced frontier models available. As we have emphasized throughout, LLMs come in a variety of sizes and capabilities: from the less-capable open-weight Llama models to the free GPT-3.5 models to the (currently) most advanced frontier models, like GPT-4o from OpenAI and Claude 3.5 Sonnet from Anthropic. The capabilities of these models vary significantly, with the smaller, less powerful models often providing inferior or incorrect results. However, lay users may not be fully aware of which model they are using. For example, we have encountered multiple users who have discounted the abilities of LLMs after a negative or unimpressive session, unaware that they were actually using a free and less capable model. Those who use LLMs in law should actively strive to use the most capable models available and should inquire about the version and sophistication of the model that they are choosing.

Second, users should be aware of the differences that adding relevant information to a prompt, such as the actual text of a statute, can make in producing an accurate and useful legal response from an LLM. Information can generally be added to inform an LLM response in one of two ways: (1) explicitly by the user, who might upload relevant law or legal rules; or (2) in the background, by the AI system itself, which might aim to retrieve information that it deems relevant while automatically and invisibly adding that information to the user's prompt. As the technical discussion earlier made clear, this kind of supplemental information serves as "context" for the LLM's response. Adding such context, in the form of party briefs, relevant case law, or statutory text, can dramatically improve the reliability and the accuracy of an LLM's response.

Specialized LLM-based legal systems like Lexis+ AI and Westlaw CoCounsel aim to supplement their LLM responses with relevant and retrieved information from their deep and reliable legal data repositories.

Context is not a panacea, however. In some cases, adding information to a prompt can actually hurt the outcome. And given the statistical machinery underpinning the operation of LLMs, the addition of certain words might unintentionally and invisibly nudge an LLM to produce one type of response versus another. For instance, adding a seemingly innocuous phrase like “originalism is one interpretive methodology, but not the only one” might accidentally nudge the LLM to produce a more originalist response based upon the inclusion of the word “originalist.” Users of LLMs should be aware of the sensitivity of their responses to particular words included in the prompt and to any uploaded materials.

A best practice for dealing with this problem is the following: Legal users of LLMs should always pose a question in multiple ways to assess whether different framings or word choices materially alter the outcome. In this way, the user can understand if the LLM is producing relatively consistent results over time or if it is producing dramatically different results based upon the framing of the question. Examining multiple responses can help to ensure consistency and accuracy. Moreover, it goes without saying that judges and lawyers should double-check the outputs of LLMs before using them in consequential contexts.

Relatedly, judges and other constitutional decision-makers should generally ask LLMs for multiple competing perspectives on any given legal question. Such varied perspectives can be very valuable for a legal decision-maker. However, as we have cautioned, in contexts that involve contestable value or policy choices, users must not mistake LLMs for objective interpretation machines and be aware of the different normative judgments that LLMs might invisibly make. A good practice, in addition to asking the LLM for multiple perspectives on the topic, is to ask the LLM to include in its answer any value, policy, or legal assumptions that it made. These models can generally do a fairly good job of identifying certain assumptions embedded in their own answers even if they are not describing their actual “reasoning” process (because LLMs do not reason the way humans do). Moreover, the interactive nature of LLMs allows

users to ask for not just simple answers but rather to ask follow-up questions asking the LLM to clarify or expand on issues. This ability to interact over time and request clarifications is one of the biggest strengths of modern LLMs compared to static documents and PDFs.

Finally, legal users should employ LLMs in the contexts and for the tasks for which they are best adapted and most reliable, given the state of the technology. Current LLMs, for example, excel at summarizing and answering questions about legal documents and providing multiple perspectives. It is important to pay attention to the particular use cases in law and be aware of employing LLMs in appropriate institutional contexts. This requires that legal users, particularly legal officials, attain some reasonable level of AI literacy on the basic functionality, strengths, and limitations of the technology before employing LLMs in practice. As the technology evolves, some of the limitations described in this Article may be overcome, and the strengths will likely grow and expand. LLM users in law are advised to periodically revisit their understanding of AI capabilities as the technology is likely to evolve significantly from its current state over time.

CONCLUSION

In recent years, generative AI has improved at a dizzying pace and continues to do so. LLMs have great potential to improve the speed and efficiency of the judicial process in constitutional cases and more broadly. But as this new technology suffuses our legal system and society, it is vital to distinguish carefully among the ways in which it is being used in different institutional contexts. Each of those uses involves different trade-offs and each represents a choice requiring normative justification. In many contexts, LLMs have the capacity to improve on human decision-making. They might even have an important role to play in informing—or actually deciding—more routine constitutional cases. But LLMs are not machines that “go of themselves.” They require guidance and instruction from human users and decisions about how, when, and whether to delegate authority from human judges to LLMs. The power to make constitutional decisions can be apportioned between humans and LLMs in any number of ways, many of which simply replicate the major controversies over

constitutional interpretation in slightly different forms. Others involve brand-new issues about when and whether to prefer the decisions of opaque algorithms over fallible human judges. But, for better or worse, there is no escaping the burdens of judgment.

APPENDIX A

A. Comprehensive – Question 1

You are a Supreme Court Justice. Please draft a three-page written opinion, answering each of the following constitutional questions based on the state of the law as of June 1, 2022. Each opinion should begin with a clear yes or no answer in boldface.

- (1) Whether all pre-viability prohibitions on elective abortions are unconstitutional.
- (2) Should this Court overrule *Grutter v. Bollinger*, 539 U.S. 306 (2003), and hold that institutions of higher education cannot use race as a factor in admissions?

B. Limited – Question 1

You are a Supreme Court Justice committed to a liberal, living-constitutionalist approach to constitutional interpretation similar to that of Justice William Brennan and Justice Thurgood Marshall. Please draft a three-page written opinion, answering each of the following constitutional questions based on the state of the law as of June 1, 2022. Each opinion should begin with a clear yes or no answer in boldface.

- (1) Whether all pre-viability prohibitions on elective abortions are unconstitutional.
- (2) Should this Court overrule *Grutter v. Bollinger*, 539 U.S. 306 (2003), and hold that institutions of higher education cannot use race as a factor in admissions?

C. Limited – Question 2

You are a Supreme Court Justice committed to original-public-meaning originalism. Please draft a three-page written opinion, answering each of the following constitutional questions based on the state of the law as of June 1, 2022. Each opinion should begin with a clear yes or no answer in boldface.

- (1) Whether all pre-viability prohibitions on elective abortions are unconstitutional.
- (2) Should this Court overrule *Grutter v. Bollinger*, 539 U.S. 306 (2003), and hold that institutions of higher education cannot use race as a factor in admissions?

D. Limited – Question 3

You are a Supreme Court Justice committed to original-public-meaning originalism. As such, you recognize that the original public meaning of the text is often indeterminate or under-determinate, requiring constitutional decision-makers to fill in the gaps through constitutional construction, which by definition is not limited to original public meaning. Following the work of Lawrence Solum, the original public meaning qualifies as determinate only when 75 percent or more of the contemporary public would have recognized it bearing a particular meaning. Absent such a clear and determinate meaning, constitutional questions fall into the “construction zone.” Permissible considerations in the “construction zone” include judicial precedent, institutional capacity and competence, constitutional structure, prudence, and moral reasoning. With these considerations in mind, please draft a three-page written opinion, answering each of the following constitutional questions based on the state of the law as of June 1, 2022. Each opinion should begin with a clear yes or no answer in boldface.

- (1) Whether all pre-viability prohibitions on elective abortions are unconstitutional.
- (2) Should this Court overrule *Grutter v. Bollinger*, 539 U.S. 306 (2003), and hold that institutions of higher education cannot use race as a factor in admissions?