GENERATIVE MISINTERPRETATION

James Grimmelmann,[*] Benjamin L.W. Sobel,[†] & David Stein[‡]

In a series of provocative experiments, a loose group of scholars, lawyers, and judges has endorsed generative interpretation: asking large language models (LLMs) like ChatGPT and Claude to resolve interpretive issues from actual cases. With varying degrees of confidence, they argue that LLMs are (or will soon be) able to assist—or even replace—judges in performing interpretive tasks like determining the meaning of a term in a contract or statute. A few go even further and argue for using LLMs to decide entire cases and to generate opinions supporting those decisions.

We respectfully dissent. In this Article, we show that LLMs are not yet fit for purpose for use in judicial chambers. Generative interpretation, like all empirical methods, must bridge two gaps to be useful and legitimate. The first is a *reliability* gap: are its methods consistent and reproducible enough to be trusted in high-stakes, real-world settings? Unfortunately, as we show, LLM proponents' experimental results are brittle and frequently arbitrary. The second is an *epistemic* gap: do these methods measure what they purport to? Here, LLM proponents have pointed to (1) LLMs' training processes on large datasets, (2) empirical measures of LLM outputs, (3) the rhetorical persuasiveness of those outputs, and (4) the assumed predictability of algorithmic methods. We show, however, that all of these justifications rest on unstated and faulty premises about the nature of LLMs and the nature of judging.

The superficial fluency of LLM-generated text conceals fundamental gaps between what these models are currently capable of and what legal interpretation requires to be methodologically and socially legitimate. Put simply, any human or computer can put words on a page, but it takes something more to turn those words into a legitimate act of legal interpretation. LLM proponents do not yet have a plausible story of what that "something more" comprises.

---

[*] Tessler Family Professor of Digital and Information Law, Cornell Law School and Cornell Tech. Author order alphabetical.

[†] Incoming Assistant Professor, University of Wisconsin Law School; Postdoctoral Fellow, Digital Life Initiative, Cornell Tech; Program Affiliate Scholar, Information Law Institute, New York University.

[‡] Assistant Professor, Vanderbilt Law School. We owe thanks for helpful suggestions and discussions to Yonathan Arbel, Micheal Beauvais, Aislinn Black, Katrina Geddes, David Hoffman, Florencia Marotta-Wurgler, Margaret Lemos, Frank Pasquale, Morgan Ricks, Joel Sobel, Christopher Sprigman, Katherine Jo Strandburg, and Kathryn Woolard. We also thank participants at the NYU Information Law Institute Fellows workshop, the Internet Law WIP Conference, and at Vanderbilt Law School. Sobel gratefully acknowledges the support of a grant from the Law & AI Lab at the University of Miami School of Law. All opinions and errors are our own.

INTRODUCTION

"This isn't right. It's not even wrong."[1]

"Generative interpretation" is all the rage.[2] Instead of asking a human judge to interpret a contract, a statute, or other legal text, generative interpretation assigns the task to a large language model (LLM). The idea itself is not new; versions of computer-assisted interpretation have been bouncing around in legal scholarship for over a decade. But the launch of ChatGPT in fall 2022, followed by rapid improvements and the launch of impressive competitors like Claude, Llama, and Gemini, has led some to argue that the future is now.

In a series of provocative experiments, scholars, judges, and lawyers have fed legal materials into LLMs to generate plausible-seeming answers to real-life interpretive questions. Yonathan Arbel and David Hoffman's *Generative Interpretation* uses a series of case studies to argue that LLMs can interpret contractual terms well enough that they are ready to assist courts and parties in contract litigation.[3] In a pair of concurring opinions, Judge Kevin Newsom has experimented with asking LLMs to interpret key terms from the cases before him.[4] And in a series of blog posts, appellate lawyer Adam Unikowsky has fed entire case files into Claude, asking it to generate judicial opinions.[5]

These LLM proponents are joined by their admirable curiosity and willingness to experiment with new methods, and by their cautious optimism about generative interpretation. Their bottom lines are broadly similar: LLMs are already good enough that judges should seriously consider trusting them to assist with interpretive work in actual cases. They argue that generative interpretation has both quality and quantity advantages over purely human judging. Quality, because LLMs make inferences based on huge amounts of data on contemporary language use, and because automated processes can be objective in a way that fallible and biased humans cannot. And quantity, because LLMs can be deployed cheaply and at scale.

We respectfully dissent. Generative interpretation in its current form is Potemkin interpretation: an attractive facade with nothing behind it. The superficial fluency of LLM-

---

[1] Attributed to Wolfgang Pauli. Michael Shermer, *Wronger Than Wrong*, SCIENTIFIC AMERICAN (2006), https://www.scientificamerican.com/article/wronger-than-wrong/ (last visited Oct 6, 2024).

[2] *See generally* Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451 (2024) [hereinafter *Generative Interpretation*].

[3] *Id.* at 458.

[4] *Snell v. United Specialty Insurance Company*, 102 F.4th 1208, 1227 (Newsom, J., concurring); *United States v. Deleon*, 116 F.4th 1260, 1270 (11th Cir. 2024) (Newsom, J., concurring).

[5] Adam Unikowsky, *In AI We Trust*, ADAM'S LEGAL NEWSLETTER (June 8, 2024), https://adamunikowsky.substack.com/p/in-ai-we-trust [hereinafter *In AI We Trust I*]; Adam Unikowsky, *In AI We Trust, Part II*, ADAM'S LEGAL NEWSLETTER (June 16, 2024), https://adamunikowsky.substack.com/p/in-ai-we-trust-part-ii [hereinafter *In AI We Trust II*]; Adam Unikowsky, *A Brief History of the Confrontation Clause*, ADAM'S LEGAL NEWSLETTER (June 26, 2024), https://adamunikowsky.substack.com/p/a-brief-history-of-the-confrontation [hereinafter *A Brief History*]; Adam Unikowsky, *Automating Criminal Appeals*, ADAM'S LEGAL NEWSLETTER (Sept. 18, 2024), https://adamunikowsky.substack.com/p/automating-criminal-appeals [hereinafter *Automating Criminal Appeals*]; Adam Unikowsky, *Ignore the Future*, ADAM'S LEGAL NEWSLETTER (Oct. 21, 2024), https://adamunikowsky.substack.com/p/ignore-the-future [hereinafter *Ignore the Future*].

generated text is deeply misleading. LLMs are remarkable text-producing machines, but the way they produce that text falls far short of what the standard account of adjudication demands.

Put simply, any human or computer can put words on a page, but it takes *something more* to turn those words into a legitimate act of legal interpretation. LLM proponents have no convincing story of what that "something more" entails. Currently, society treats judges' opinions as legitimate (most of the time) because they represent a principled working forward from source materials to a persuasive legal conclusion.[6] LLM proponents treat the persuasiveness of the output as proof that the process was principled. But the lesson of their experiments is precisely the opposite: they have shown that it is possible to produce persuasive, law-like rhetoric without going through the hard work of legal reasoning.

While LLM proponents are right to point out LLMs' potential and to experiment with them, they are wrong to the extent that they claim that LLMs are currently fit for purpose at legal interpretation. LLMs are already good at some kinds of tasks—such as writing JavaScript code—where the output text is valued for what it does.[7] They may even be good at some kinds of legal writing: under the right circumstances and with proper adult supervision, it is possible to imagine them helping draft contracts or briefs reliably and effectively.[8] But legal interpretation is different because adjudication is different; it plays a different social role and must satisfy different constraints. Human interpretation varies—somewhat more or less reliable, somewhat more or less persuasive, somewhat more or less affordable—but is fundamentally the same kind of thing. LLM interpretation is something different in kind. It is a simulacrum of interpretation, one that bears no necessary relationship to the realities of linguistic meaning.[9] Perhaps this chasm can be bridged. The first, necessary step in doing so is to admit that the chasm exists.

In particular, generative interpretation must overcome two challenges to be useful and legitimate when used in the judicial process. First, it faces a *reliability* gap: LLM proponents have not yet established that their methods reliably measure anything. The examples they have proffered are intriguing and worthy of further study, but they are very far from being

---

[6] *See generally* Neil Duxbury, *Faith in Reason: The Process Tradition in American Jurisprudence*, 15 CARDOZO L. REV. 601 (1993) (discussing Legal Process school's emphasis on reason-giving).

[7] *See, e.g.*, Simon Willison, *Here's How I Use LLMs to Help Me Write Code*, SIMON WILLISON'S WEBLOG (Mar. 11, 2025), https://simonwillison.net/2025/Mar/11/using-llms-for-code/.

[8] *Compare* Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo, *GPT-4 Passes the Bar Exam*, 382 PHIL. TRANSACTIONS ROYAL SOC'Y A (2024), https://doi.org/10.1098/rsta.2023.0254 (arguing that a leading LLM is capable of passing the bar exam); Andrew Blair-Stanek, Donald Gifford, Mark Graber, Guha Krishnamurthi, Jeff Sovern, Donald B. Tobin, and Michael Van Alstyne, *AI Gets Its First Law School A+s* (May 29, 2025) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5274547 ("AI models can now perform at an A+ level on some law school final exams."), *with* Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme, *BLT: Can Large Language Models Handle Basic Legal Text?*, PROC. 2024 NAT. LEGAL LANGUAGE PROCESSING WORKSHOP 216 ("We find that the best publicly available LLMs like GPT-4 and Claude currently perform poorly on basic legal text handling.")

[9] *See generally* JEAN BAUDRILLARD, SIMULACRA AND SIMULATION (1981) (Shelia Faria Glaser trans., U. Michigan Press 1994).

sufficiently replicable to be regarded as reliable.[10] A consumer survey with a sample size of n=1 and an interviewer who improvised the questions on the spot would be unreliable[11]—no matter how plausible or how dramatic the results were—because repeating the survey would lead to very different results. In a similar way, generative interpretation is unreliable if small tweaks to the prompt or the system flip the outputs. The generative interpretation experiments reported in the literature purport to be "informed conclusion[s] based on a statistical analysis of billions of texts."[12] Our examination of their own examples shows, however, that the results might be better described as unwarranted extrapolations from extraneous factors.

Second, generative interpretation faces an *epistemic* gap: LLM proponents have not yet established that their methods measure what they purport to. Even if those methods were robust and replicable, it would all be for naught if they were measuring the wrong thing. A study to determine the surface temperature of the sun by surveying breakfast-cereal consumers would epistemically invalid, even if it had a rigorously specified survey protocol, interviewed an immense sample size, and passed every statistical test for significance—simply because *what people think the sun's surface temperature is* is not a good measure of *what the sun's surface temperature actually is*. In a similar way, generative interpretation is epistemically invalid if its text predictions do not correspond to the linguistic judgments of the relevant community of speakers and listeners. This gap is most obvious for generative adjudication, in which LLMs are tasked with generating entire opinions, but we will show that it is present for the seemingly more constrained task of generative interpretation, as well.

The reliability gap and the epistemic gap are closely related. Indeed, we think that the beginning of wisdom is to recognize that both gaps must be bridged for generative interpretation to be useful. Some of the most thoughtful attempts to deal with the descriptive gap raise the biggest epistemic problems, and vice-versa. Arbel and Hoffman's precision about measuring next-token probabilities, for example, helps make generative methods more reliable—but it begs the question of why next-token probabilities are a good epistemic measure of meaning. Conversely, Unikowsky's *res ipsa loquitur* willingness to accept LLM outputs as authoritative because they read persuasively gives a (debatable) response to the epistemic question, but it does nothing to answer the reliability question. Taking both of these gaps seriously *at the same time* is essential, and it is here that we think the LLM proponents have the most work ahead of them.

Our arguments dovetail with recent work by Brandon Waldon, Nathan Schneider, Ethan Wilcox, Amir Zeldes, and Kevin Tobia. In a forthcoming paper, they persuasively show that arguments for LLMs as authoritative "'super-judges' of ordinary meaning rest[] upon fundamental myths about how these tools work."[13] They focus on debunking the factual assumptions about LLMs that generative-interpretation proponents have invoked to justify

---

[10] *Cf.* FED. R. EVID. 702(c) (requiring that expert testimony be "the product of reliable principles and methods"); Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 589 (1993).

[11] *Cf.* Gail M. Sullivan & Anthony R. Artino, *How to Create a Bad Survey Instrument*, 9 J GRAD. MED. EDUC. 411 (2017).

[12] *Generative Interpretation*, at 458.

[13] Brandon Waldon et al., *Large Language Models for Legal Interpretation? Don't Take Their Word for It*, 114 GEO. L.J. __ (forthcoming), at 22, https://papers.ssrn.com/abstract=5123124.

their purported expertise on ordinary meaning.[14] Our complementary critique addresses distinct empirical and jurisprudential problems with generative-interpretation proposals. On the empirical side, we offer detailed, under-the-hood demonstrations that generative interpretation produces arbitrary and unpredictable results in a manner that belies its proponents' optimistic claims. On the jurisprudential side, we ask what it would take for LLMs to possess the legal-interpretive authority that generative-interpretation proponents claim they do—and thus why, as a jurisprudential matter, the proponents' accounts fall short of justifying such authority.

This Article proceeds in five Parts. **Part I** surveys the recent wave of writings from LLM proponents. We divide them roughly into two camps: those who are interested in *generative interpretation*, using LLMs to ascertain the meaning of words and phrases, and those who are interested in *generative adjudication*, using LLMs to resolve entire disputes. **Part II** provides a brief overview of LLM technology. We do not rehash the history of LLMs or describe the technology in detail. Instead, we highlight a few points that will be relevant to our discussion.

Part III considers the reliability gap. Here, our methodological critique shows, with representative examples, that LLM proponents' results are brittle and often arbitrary. Small changes to an LLM, its settings, or its input can dramatically change the nature of its outputs, often in ways that would alter the outcome of cases.. These are not small quibbles. Legal interpretation is a large and complex task, and the space of possible experimental setups is correspondingly large and unconstrained. The fact that LLMs are so easy to use—at the simplest, just type a question in a box—makes it easy to overlook the assumptions and choices involved in using them.

Part IV then considers the epistemic gap. We taxonomize four types of arguments that LLM proponents have made to establish the legitimacy of their use in adjudication. Respectively, they are based on (1) an LLM's *training process* and the immense bodies of text it is trained on, (2) *empirical validation* of the LLM's outputs in comparison with other methods, (3) the *rhetorical persuasiveness* of the LLM's outputs as natural-language text, or (4) the purported *predictability* of LLM-generated outputs. In each case, we point out the unstated and mistaken assumptions on which these arguments rest. In particular, we argue that LLMs highlight the distinction between the *process* that makes adjudication legitimate—the hard work of legal reasoning that judges engage in—and the *artifacts* the process generates—like definitions of disputed terms and opinions justifying an outcome. LLMs sever the connection between the two, because they are capable of producing plausible-looking opinions through a process that is fundamentally alien to traditional legal reasoning. Indeed, they force us to consider the prospect that the superficial fluency of legal text is no longer a meaningful guarantee of anything substantive.

Part V considers what it might take to make LLMs practice-ready for judicial chambers. Even if generative interpretation and adjudication are different and harder, we are more optimistic about the use of LLMs in other aspects of legal practice, and for certain uses in judicial practice. Some scholars have argued that judging is an intrinsically human activity that

[14] *See id.* at 22-39.

cannot be outsourced, in whole or part, to any algorithmic system.[15] Our claim is narrower. We take no position on whether adjudication requires the exercise of specifically moral facilities or whether computer systems, including LLMs, are capable of having and exercising those facilities. Instead, our view is that adjudication is a fundamentally social process, one that must meet certain criteria of rationality to serve its function of legitimating the exercise of authority. Thus, in this Part we describe the work required to make generative interpretation meet the necessary criteria of rationality. In particular, we compare it to two other empirical interpretive methods—trademark surveys and legal corpus linguistics—one of which has successfully established itself in adjudication and the other of which is attempting to do so.

Finally, in a brief **Conclusion**, we consider the broader implications of our analysis. Perhaps the problem facing generative interpretation is not that LLMs aren't good enough at language to be useful to judges, but that they are *too good*. Proponents of generative interpretation cite the persuasiveness of LLM-generated text as proof of that text's legitimacy. But the lesson of their experiments is precisely the opposite. They show that it is possible to produce persuasive, law-like rhetoric to support a conclusion without going through the hard work of legal reasoning to validate that conclusion. Their work shows that we should be more suspicious of words on a page in the age of AI, not less.

I. GENERATIVE INTERPRETATION

LLM proponents have made two different kinds of proposals for incorporating generative AI into adjudication. Some of them, more modestly, have suggested using generative AI as an *interpretive tool* that provides targeted answers to discrete questions about the meaning of a legal text. Others, more ambitiously, have suggested using generative AI as an *adjudicator* that answers broader questions about the proper resolution of a case. They are united by a belief that LLMs can perform some tasks traditionally performed by judges well enough that the legal system should think seriously about incorporating them into actual judicial practice. They differ in how much of the judicial task should be delegated to LLMs. And even within these two broad camps, their approaches towards the proper use of LLMs are highly diverse.

These are not completely new ideas. For years, commentators have debated whether decision-making by AIs could be an acceptable substitute for decision-making by human

[15] *See., e.g.*, Amin Ebrahimi Afrouzi, *John Robots, Thurgood Martian, and the Syntax Monster: A New Argument Against AI Judges*, 37 Can. J. of L. & Juris. 369 (2024); Kiel Brennan-Marquez and Stephen E. Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 NW. U. L. REV. 137 (2019); Ian Kerr and Carisma Mathen, *Chief Justice John Roberts is a Robot*, (unpublished draft June 13, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3395885. *But see, e.g.,* Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135 (2019) (arguing for human-algorithmic equivalence in judging). An even stronger version of the claim is that certain decisions must be made by *specific people* and that any delegation of the reasoning or justification even to *other people* is problematic. *See* Bridget C.E. Dooling, *Ghostwriting the Government*, 109 MARQ. L. REV. (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5200672 (arguing that a "duty to reason" restricts delegation both to humans and to generative AI).

judges.[16] These conversations span many legal fields, and encompass not just judicial adjudication but also the immense volume of administrative decisions. Common themes include accuracy,[17] consistency,[18] and explainability.[19] In the first instance, these debates have primarily been about the quality of the *decisions* themselves, rather than about the quality of the *explanations* offered to justify those decisions. In other words, these have largely been debates about *classification* or *predictive* AI, rather than about *generative* AI.

Generative AI brings two new things to the table. First, it turns hypotheticals about AIs capable of writing opinions that read like they were written by a human into actualities. That means the question of whether the judicial system should actually use those AIs is far more pressing than it was even a few years ago.[20] When proponents like Adam Unikowsky claim that LLMs are good enough for judges to use *now*, they are speaking in the present tense and the indicative mood.

Second, generative AI opens up the possibility of generative interpretation: giving an LLM an interpretive task that is significantly smaller than deciding an entire case. This is not actually a task where the generative nature of the AI is essential to its use for interpretation. One could imagine training a classification AI to disambiguate between cases where the last-antecedent canon or the series-qualifier canon is more appropriate.[21] Such an AI might have an architecture and training process completely different from the deep-learning methods used to train LLMs; it might be prompted with a passage of text and output a single numerical value expressing the degree to which the passage resembled other passages in which one canon rather than the other replied. *Generative* AI, on the other hand, has seemed like it might work for interpretation because it has already worked for so many other tasks. That is, LLMs are such powerful and versatile tools that they can be (and have been) used, off the shelf or with only slight modifications, for a variety of classification tasks.[22] If they are truly jacks of all trades, perhaps legal interpretation is one of those trades.

---

[16] *See, e.g.*, Afrouzi, *supra* note 15; Brennan-Marquez and Henderson, *supra* note 15; Volokh, *supra* note 15.

[17] *See, e.g.,* Danielle Keats Citron, *Technological Due Process*, 85 WASH. L. REV. 1249 (2008).

[18] *See, e.g.*, Cary Coglianese and Lavi Ben Dor, *AI in Adjudication and Administration*, 86 BROOK. L. REV. 791 (2021).

[19] *See, e.g.*, Margot Kaminski, *The Right to Explanation, Explained*, 34 BERK. TECH. L.J. 189 (2019); Andrew D. Selbst and Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018)

[20] It also changes the tenor of discussions of the effects these AIs will have. Actual experience with LLMs makes the conversation significantly more concrete. *See, e.g.*, Richard M. Re, *Artificial Authorship and Judicial Opinions*, 92 GEO. WASH. L. REV. 1558, 1582–85 (2024) (offering predictions about effects of widespread use of AI by courts and litigants).

[21] *See generally* Adam Crews, *The So-Called Series-Qualifier Canon*, 116 NW. U. L. REV. ONLINE 198, 208–09 (2021) (comparing the two canons).

[22] *See*, *e.g.*, Rebecca M.M. Hicke and David Mimno, *T5 Meets Tybalt: Author Attribution in Early Modern English Drama Using Large Language Models, in* PROC. CHR 2023: COMP. HUM. RES. CONF., https://arxiv.org/abs/2310.18454 (using an LLM to identify authorship of text passage).

A. The Proponents' Common Premises

The justifications that LLM proponents offer for generative interpretation and adjudication are broadly consistent. The remainder of this Part will review arguments from some of the early proponents of LLM-powered interpretation and adjudication—from cautious experimenters like Judges Kevin Newsom and Joshua Deahl, measured enthusiasts like professors Yonathan Arbel and David Hoffman, to unbridled optimists like practitioners Adam Unikowsky and Jack Kieffaber. These proponents differ along many dimensions, but all see LLMs as potentially more accurate, more consistent, and more efficient than current interpretive methods. The intuition behind these arguments is consistent across the board.

First, the proponents say that LLMs can be more *accurate* than human judges: the models offer higher-quality insights into the ordinary meaning of language in legal texts. They argue that generative interpretation offers a more "majoritarian reading" of text than a dictionary or a jurist's intuitions.[23] The core tenet of this argument is that because the models train on what are perhaps the largest corpora of natural-language text ever assembled, and because they use statistical methods to optimize their predictions of successive words in context, their output offers probative insights about general English usage.[24] A related idea is that they are more democratic, because their training data reflects a far broader sample than other authoritative sources, like dictionaries.

Second, the proponents argue that generative interpretation is more *consistent*: more constrained, more predictable, and more transparent.[25] In contrast to current interpretive methods' "value-laden" choices about which dictionaries to consult or which interpretive canons to employ, generative interpretation offers a "cutting-edge 'mathematization' of language,"[26] that is more "transparent" and "objective."[27] In *Snell*, Judge Newsom suggested that "LLM research is relatively transparent."[28] "[W]e tend to take dictionaries for granted, as if delivered by a prophet," he observes, but "the precise details of [dictionaries'] construction

---

[23] *Generative Interpretation*, at 488.

[24] See *Snell*, 102 F.4th at 1226 (Newsom, J., concurring) ("the best reason to think that LLMs might provide useful information to those engaged in the interpretive enterprise" is that "the models train on a mind-bogglingly enormous amount of raw data taken from the internet" and therefore "can provide useful statistical predictions about how, in the main, ordinary people ordinarily use words and phrases in ordinary life."). *See also Generative Interpretation* at 513–14 (praising "superior sensitivity to actual usage," which, the authors argue, derives from the fact that the models trained on immense amounts of real-life examples of linguistic usage); Unikowsky, *In AI We Trust I*, *supra* note __ ("[W]e should come at this problem in the spirit of[,] 'AI has read every case ever written and in most cases, it will be more accurate than humans'. . . .").

[25] *Snell*, 102 F.4th at 1227-30 (Newsom, J., concurring); Unikowsky, *In AI We Trust I*, *supra* note __ ("Dispersing the judicial power among so many different judges inevitably undermines predictability. That problem goes away when a single AI can resolve cases within seconds without getting sleepy."). *See generally* Jack Kieffaber, Kimo Gandall and Kenny McLaren, *We Built Judge. AI. And You Should Buy It* (Jan. 28, 2025), https://papers.ssrn.com/abstract=5115184.

[26] *Id.*

[27] *Generative Interpretation.* at 455, 466, 511.

[28] *Snell*, 102 F.4th at 1228 (Newsom, J., concurring).

aren't always self-evident."[29] Moreover, judges can shop around for dictionary definitions that suit their predispositions and "rarely explain in any detail the process by which they selected one definition over others."[30] By contrast, Judge Newsom argued, LLM research is comparatively transparent because we "know," on some general level, what data LLMs "learn" from and because judges could disclose their full queries and models' full answers.[31] Plus, they are less vulnerable to hidden exercises of discretion than tools like corpus linguistics (and perhaps even dictionaries).[32]

And third, the proponents tell us that generative interpretation is more *efficient,* so its use can improve access to justice. LLMs are widely accessible and far cheaper than actual surveys of humans' perceptions of meaning.[33] Judicial use of a predictable, widely accessible model might let parties predict judicial outcomes to resolve disputes while avoiding court altogether.[34]

B. LLMs as Interpretive Tools

We start with the commentators who have proposed using LLMs for interpretation. Overall, their proposals are more concrete and specific, and thus easier to assess. Part III's more technical critique revisits these examples in greater detail.

1. Yonathan Arbel and David Hoffman

The leading exposition of using LLMs for legal interpretation is Yonathan Arbel and David Hoffman's widely-praised[35] *Generative Interpretation*.[36] The article posits that "AI models can help factfinders ascertain ordinary meaning in context, quantify ambiguity, and fill gaps in parties' agreements."[37] The authors' essential argument is that large language models can help interpret contested provisions of legal instruments.

They demonstrate their approach by working through several case studies of litigated contractual disputes. In these case studies, the authors solicit responses from LLMs about the meaning of legal text, and they present these responses as "relevant" evidence of the language's

[29] *Id.*

[30] *Id.*

[31] *Id.*

[32] *Id*.

[33] *Id.* at 1227-30.

[34] *Generative Interpretation*, at 510.

[35] See, e.g., Lawrence Solum, *Hoffman & Arbel on Interpretation of Contracts by Generative AI*, LEGAL THEORY BLOG, https://lsolum.typepad.com/legaltheory/2023/08/hoffman-arbel-on-interpretation-of-contracts-by-generative-ai.html (last visited Oct 6, 2024) ("A profoundly important article.").

[36] *Generative Interpretation, supra* note 2.

[37] *Id.* at 451.

"public and common" meaning.[38] "Generative interpretation," they conclude, "promises an accessible, relatively predictable, tool that will help lawyers and judges interpret contracts."[39]

For example, they consider a dispute about whether language in a prenuptial agreement referring to "the time a Petition for Dissolution of Marriage is filed" meant the time the *first* such a petition was filed or the time that the *most recent* petition was filed.[40] They gave OpenAI's Davinci-003 model—a variant of the LLM used to power ChatGPT at the time—the text of the agreement and asked it, "If one of the parties files a divorce petition, withdraws it, and then a few years later a new petition is filed, what date determines the number of full years of marriage: the first filing or the second one?"[41] They then observe that the model has a 94.72% probability of outputting "second" in a sentence answering the question and only a 0.68% probability of outputting "first."[42]

Arbel and Hoffman acknowledge that their article is meant as an introduction to a promising interpretive technique, rather than as an exhaustive explication of an interpretive methodology. They do not claim that LLMs surface the absolute truth of contracting parties' intentions: instead, they present generative interpretation as a "workable, workmanlike method for a resource-constrained litigation world."[43] Nor do they claim that LLMs are ready to serve as "robot judges" vested with the final say in adjudication.[44] Rather, they explain that their proposal is for judges to use LLMs "as tools" to supplement their judgment, and they disclaim that techniques for querying these tools "still await a process of development, refinement, and validation."[45]

Alongside these disclaimers, however, Arbel and Hoffman consistently characterize LLM outputs as accurate, probative evidence of words' legal meanings.[46] They further advertise that

---

[38] *Id.* at 485 ("Generative interpretation. . . offers courts a better sense of the relevant probabilities if the parties were intending to use English in its most public and common sense."). *See also id.* at 489 ("Generative interpretation . . . helps us visualize a broad spectrum of meaning and quantify how likely a particular result is."); *id.* at 505 ("Models offer an approximation of general understanding that may simply not be available in any other way, and thus advance long-held goals of contract theory.").

[39] *Id.* at 509.

[40] *Id.* at 483–84.

[41] *Id.* at 484.

[42] *Id.*

[43] *Id.* at 458, 460–61.

[44] *Id.* at 461.

[45] *Id.* at 455, 461.

[46] *See* note 38, *supra*. *See also Generative Interpretation* at 495 ("[W]e have provided examples that showcase how large language models might power a stronger, cheaper, more robust form of textualism."); *id.* at 505 ("[O]n average, these models predict with great accuracy linguistic distinctions that humans make."); *id.* at 509 ("generative interpretation promises an accessible, relatively predictable, tool that will help lawyers and judges interpret contracts"). *But compare id.* at 485 ("Generative interpretation . . . offers courts a better sense of the relevant probabilities if the parties were intending to use English in its most public and common sense.") *with id.* at 485 n.165 ("[T]he probabilities shouldn't be interpreted literally."). We think the best way to interpret these claims and caveats together is as an endorsements of the models' *accuracy*, but a disavowal of the models' claimed *precision*. *See id.* at 496 n.210 ("[Y]ou should be skeptical of model's expressed confidence; the direction of change with every new piece of evidence,

"generative interpretation is good enough for many cases that currently employ more expensive, and arguably less certain, methodologies."[47] Whether or not the authors intended to present generative interpretation as a method ready for primetime use,[48] it now occupies a primetime slot, thanks in part to its influence on prominent LLM proponents like Judge Newsom, whose views we discuss in the following sub-Part.

Overall, Arbel and Hoffman's proposal is *narrow* in the questions it asks LLMs, *broad* in the evidence it asks them to consider, and *narrow* in the outputs it examines. They invite judges to turn to generative AI to help with discrete interpretive tasks—*e.g.*, does the term "other affiliates" as used in a contract refer only to entitles already in existence?—and they are capacious in the text they provide LLMs to help answer these questions—frequently, entire contracts.[49] Indeed, they present the use of LLMs as a new form of contextualism: a way for contract interpretation to consider a wide array of evidence bearing on a contract's meaning while preserving the values of predictability, restraint, and judicial economy typically advanced for its alternative, textualism.[50] They are enthusiastic about giving LLMs access to the complete text of affidavits, and other documents beyond the four corners of a contract.[51] Overall, however, they use LLMs as meaning meters rather than legal writers; they look closely at the LLMs' internal calculations rather than examining only the apparent meaning of the models' natural-language output.

2. Judge Kevin Newsom

The next major proponent of generative interpretation is Judge Kevin Newsom of the United States Court of Appeals for the Eleventh Circuit. In 2024, he wrote a concurrence in *Snell v. United Specialty Insurance Company*, in which he characterized himself as "th[inking] the unthinkable" and "sa[ying] the unsayable:"[52] "LLM[s] like ChatGPT," he mused, "might have something useful to say about the common, everyday meaning of the words and phrases used in legal texts."[53]

The relevant issue in *Snell* was whether an insurance policy that covered liability arising out of the insured's "perform[ance of] landscaping"[54] applied to the allegedly negligent installation

---

not its quantification, is informative."). For an explanation of the distinction between accuracy and precision, see, e.g., *Accuracy and Precision*, WIKIPEDIA (2025), https://en.wikipedia.org/w/index.php?title=Accuracy_and_precision&oldid=1281017807.

[47] *Generative Interpretation* at 458.

[48] After the authors assert in the present tense that "generative interpretation is good enough," *id.*, they later present the question as, "whether [generative interpretation] is good enough, *if not today then soon*, for resource-deprived courts to adopt in ordinary cases," *id.* at 502 (emphasis added and original emphasis omitted).

[49] *Id.* at 489–90.

[50] *Id.* at 510–13.

[51] *Id.* at 513.

[52] Snell v. United Specialty Insurance Company, 102 F.4th 1208, 1221 (11th Cir. 2024) (Newsom, J., concurring).

[53] *Id.* at 1234.

[54] 102 F.4th at 1213.

of a trampoline.[55] While the majority disposed of the case without ruling on the meaning of "landscaping," Judge Newsom argued that "[t]hose, like me, who believe that 'ordinary meaning' is *the* foundational rule for the evaluation of legal texts should consider—*consider*—whether and how AI-powered large language models like OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude might—*might*—inform the interpretive analysis."[56]

On "a lark," Judge Newsom directed one of his clerks to query ChatGPT, "What is the ordinary meaning of 'landscaping'?" ChatGPT produced a two-sentence definition that "aligned with [his] priors."[57] He then asked, "Is installing an in-ground trampoline 'landscaping' ?" Both ChatGPT and Google Bard answered in the affirmative.[58] The coherence of ChatGPT's response impressed him, and he found that it "squared with [his] own impression" about how ordinary speakers might use the term.[59]

A few months later, Judge Newsom reaffirmed his enthusiasm for LLMs in another concurrence, this time in a case, *United States v. Deleon*, interpreting a provision of the United States Sentencing Guidelines.[60] The Guidelines impose a sentencing enhancement for an armed robbery "if any person was physically restrained to facilitate commission of the offense or to facilitate escape."[61] The defendant had "walked into a store, pointed a gun at the cashier while demanding money from the register, received the money, and . . . left . . . within about one minute."[62] The court held the enhancement applicable, and Judge Newsom wrote separately to suggest that LLMs might "inform the interpretive analysis" of the Guideline at issue.[63] He asked ChatGPT: "What is the ordinary meaning of 'physically restrained'?"[64] several times. While he received a different answer each time, he found it informative that despite "subtle, marginal divergences," the models' responses "coalesce[d], substantively, around a common core—there was an objectively verifiable throughline."[65] He wrote, "For our purposes, what matters is that the LLMs consistently defined the phrase 'physically restrained' to require the application of tangible force, either through direct bodily contact or some other device or instrument. And that, again, squares comfortably with the results obtained through the traditional, dictionary-driven breaking-and-repiecing method."[66]

This exercise gave Judge Newsom "hope that the models have something significant to offer the interpretive enterprise."[67] "LLMs are trained on actual individuals' uses of language in the real world," he reasoned, and "the LLMs' responses to [his] repeated queries reliably revealed

---

[55] *Id.* at 1211-12.

[56] *See id.* at 1217 (majority opinion); 1221-22 (Newsom, J., concurring).

[57] *Id.* at 1225.

[58] *Id.*

[59] *Id.*

[60] United States v. Deleon, 116 F.4th 1260, 1261 (11th Cir. 2024).

[61] *Id.* at 1261 (quoting U.S. Sent'g Guidelines Manual § 2B3.1(b)(4)(B)).

[62] *Id.*

[63] *Id.* at 1270 (Newsom, J., concurring).

[64] *Id.* at 1272.

[65] *Id.* at 1275.

[66] *Id.* at 1275.

[67] *Id.* at 1277-78.

. . . a common core." [68] Judge Newsom concluded his *Deleon* concurrence by observing, "I continue to believe—perhaps more so with each interaction—that LLMs have something to contribute to the ordinary-meaning endeavor."[69]

In both *Snell* and *Deleon*, Judge Newsom favorably cited *Generative Interpretation* to support LLMs' authority concerning the ordinary meaning of legal language.[70] Like Arbel and Hoffman, Judge Newsom believes in asking *narrow* questions of LLMs, using them to answer discrete questions about the meanings of particular terms. But in other ways, his approach differs from theirs. For one thing, he takes a *narrow* view of the sources the LLM should specifically consider. Instead of giving the LLM an entire contract to interpret, Newsom uses single-sentence prompts, e.g. "*Is installing an in-ground trampoline 'landscaping'?*"[71] For another, he takes a slightly *broader* view of the LLM's outputs than Arbel and Hoffman's prenuptial-petition example above. Instead of considering token probabilities as they do, Newsom reads the LLM outputs as natural-langauge text, much like he would a dictionary definition.[72] In *Deleon*, he borrows from Arbel and Hoffman in repeating his query 30 times (10 for each of 3 models) to assess their "confidence," but he continues to treat the outputs as English text.

3. Judge Joshua Deahl

Some months later, Judge Joshua Deahl of the District of Columbia Court of Appeals followed Judge Newsom's lead. In *Ross v. United States*, Nina Ross was charged with animal cruelty for leaving a dog in her car on a hot day.[73] In overturning her conviction, the majority held that there was insufficient evidence to prove beyond a reasonable doubt that leaving a dog in a car for over an hour on a hot day would harm the dog. The majority cited *Jordan v. United States,* a similar case where the court had reversed a conviction for leaving a dog outside on a cold day.[74] In his dissent, Judge Deahl distinguished *Jordan* on the ground that it is common knowledge that leaving a dog in a car on a hot day "created a 'plain and strong likelihood' that [the] dog would be harmed," whereas the risks of leaving a dog outside on a cold day were less clear-cut.[75]

In support of this argument, Judge Deahl and his clerks asked ChatGPT, "Is it harmful to leave a dog in a car, with the windows down a few inches, for an hour and twenty minutes when it's 98 degrees outside?"[76] They received an "unequivocal 'yes.'" They then asked: "Is it harmful to leave a German shepherd outside in 25 degree temperature for five hours?" and received

---

[68] *Id.* at 1277.

[69] *Id.*

[70] *Snell*, 102 F.4th at 1226-27 n.7; *Deleon* n.1.

[71] *Snell*, 102 F.4th at 1235.

[72] *Id.* at 1228–30 (comparing LLMs and dictionaries).

[73] Ross v. United States, 331 A.3d 220 (D.C. 2025).

[74] *Id.* at 8. (citing *Jordan v. United States*, 269 A.2d 848, 849 (D.C. 1970)) ("In the absence of testimony . . . that the shelter or protection from the weather supplied this dog on this occasion would cause the dog to suffer, the evidence was insufficient to sustain the conviction.")

[75] *Id.* at 232 (Deahl, J., Dissenting).

[76] *Id.* at 236.

answers that "boil[ed] down to 'it depends.'"[77] Carefully caveating that ChatGPT is "definitely not" a good proxy for common knowledge, Judge Deahl nevertheless avowed that those chats supported his claims because inclusion in ChatGPT's responses is "at least a point in favor of something being common knowledge that[] is, in fact, true; and a far stronger point against it [being common knowledge] if it's not."[78]

This approach is similar to Judge Newsom's in some ways, and Judge Deahl favorably cited Judge Newsom's analysis in justifying his own treatment of ChatGPT outputs as probative of common knowledge.[79] Still, there is something different going on here. Judge Deahl didn't ask ChatGPT about the meaning of a term (what does "harmful" mean?); he asked it about how the world works (is it harmful to leave a dog in a car on a hot day?). This is a different kind of inquiry. It is still *narrow* in the question being asked, *narrow* in the sources presented to the LLM, and somewhat *broader* in considering natural-language outputs—but it is not really an interpretive question at all.

C. LLMs as Adjudicators

Some commentators go beyond mere LLM *interpretation* to LLM *adjudication*. Instead of prompting the LLM to carry out a specific interpretive task ("What is the ordinary meaning of 'landscaping'?") they prompt the LLM with a question about the case as a whole ("You are a federal appellate judge. Please read the attached briefs. … Please write a three paragraph decision. … In the third paragraph, explain who wins and why.")[80] This is a *broad* interpretive task, much broader than anything attempted by Arbel and Hoffman, Judge Newsom, or Judge Deahl. It almost necessarily involves considering a *broad* range of sources—briefs, precedents, statutes, and other sources a human adjudicator would consider. And it equally almost necessarily involves reading the output *broadly* as a natural-language text—the kind of work product that a human adjudicator would produce.

Numerous commentators have considered whether AIs in general, and recent LLMs in particular, could do the work of judges. What sets apart the two we discuss in this section—Adam Unikowsky and Jack Kieffaber—is their full-throated enthusiasm. Where others ask whether LLMs are ready for judicial use, they answer with an emphatic "yes!"

1. Adam Unikowsky

Adam Unikowsky—a partner at Jenner & Block who focuses on Supreme Court and appellate advocacy[81]—has used a variety of methods to test LLM adjudication. Although the specifics vary, from the jump he has focused on prompts that generate extensive natural-language outputs: his first, for example, was "Write two paragraphs about the [*Snell*] concurrence. First: what does the concurrence argue? Second, do you agree with the

[77] *Id.* at 237.

[78] *Id.* at 237 n.4.

[79] *Id.* at 237 n.4 (calling *Snell* "a thoughtful and engaging discussion . . . which I largely agree with.").

[80] *In AI We Trust I*, *supra* note 5.

[81] Adam Unikowsky, JENNER & BLOCK LLP, https://www.jenner.com/en/people/adam-g-unikowsky (last visited Oct 20, 2024).

concurrence's argument?"[82] From there, he progressed to uploading the entire briefing in a case to Claude, asking it to write a brief opinion, and then asking probing follow-up questions.[83]

Unikowsky was deeply impressed with the results, asserting that the Claude chatbot "is fully capable of acting as a Supreme Court Justice right now. When used as a law clerk, Claude is easily as insightful and accurate as human clerks, while towering over humans in efficiency."[84] He found that Claude "consistently decide[d] cases correctly," by which he meant that that it articulated the holding the Supreme Court in fact reached. On the occasions when Claude articulated conclusions that differed from the Supreme Court's, its output struck Unikowsky as "reasonable."[85] He has also asked it to critique the methodology in an expert report (much as judges do when discounting an expert's conclusions),[86] to generate persuasive opinions in support of a particular outcome ("You are a Supreme Court Justice who believes racial-gerrymandering claims should be non-justiciable. Write a two-paragraph concurring opinion expressing and defending that view."),[87] and to develop novel doctrines to resolve a case ("Please come up with an extremely creative, brilliant, out-of-the-box alternative to the 'primary purpose' test that would improve Confrontation Clause law.").[88] In each case, he found the results to be intelligent, creative, and persuasive.

In additional blog posts Unikowsky has suggested that AIs could write effective appellate briefs,[89] and that AI should be used to automate the entire veterans' appeals process.[90] His proposed process is worth quoting in detail, because it gives a good sense of the ambition of his vision:

> This can be trivially implemented today. Just upload the relevant portion of the VA benefits manual into an AI context window, upload the veteran's documentary evidence, and ask AI to apply law to fact. If you're concerned that AI won't be sufficiently generous, you can keep track of the percentage of veterans that obtain benefits and then tweak the prompts until the percentage is at a satisfactory level. If you're squeamish about having AI make decisions that affect people's lives, then you can have the AI complete the initial layer of review (equivalent to what the VA does today) and then give the veteran the right to appeal to a human judge. If we just snap our fingers, we can make all tradeoffs go away.[91]

2. Jack Kieffaber

Another enthusiastic promoter of LLM adjudication is Jack Kieffaber, who argues strongly in favor of replacing judges with AIs in two forthcoming articles. The place to start is actually with the second article, *We Built Judge.AI And you Should Buy It,* written with Kimo Gandall

---

[82] *Id.*

[83] *Id.; In AI We Trust II*, *supra* note 5.

[84] *In AI We Trust*, *supra* note 5.

[85] *Id.*

[86] *In AI We Trust II*, *supra* note 5.

[87] *Id.*

[88] *A Brief History*, *supra* note 5.

[89] *Automating Criminal Appeals*, *supra* note 5.

[90] *Ignore the Future*, *supra* note 5.

[91] *Id.*

and Kenny McLaren.[92] The authors built an LLM-based online system, Arbitrus.ai, that receives evidence, motions, and briefing from two parties and then issues rulings.[93] To calibrate and verify the system, they used another LLM to generate synthetic briefs in 100 hypothetical disputes, and then had Arbitrus.ai issue opinions resolving those disputes.[94] Humans checked that cited authorities existed, were responsive to the issues in the case, and fully resolved the dispute.[95] Overall, they claim that the system usually fully resolved each case without hallucinations or ungrounded assertions.[96]

As the name suggests, Arbitrus.ai is designed and marketed as an arbitration system. Although there are some statutory and doctrinal complications about the applicability of the Federal Arbitration Act to AI arbitrations, party consent can bless a wide range of procedures in arbitration that would be unacceptable in adjudication.[97] Gandall, Kieffaber, and McLaren argue that "automation will beat all comers" in arbitration.[98] This is a predictive claim about parties' satisfaction with AI arbitration, or at least their willingness to accept form contracts selecting it, that does not necessarily depend on the quality of those arbitrations.

Kieffaber's principal normative claims in favor of AI adjudication come, instead, in his earlier article, *Predictability, AI, And Judicial Futurism*.[99] He anticipates a future in which "Judge.AI" can do what Arbitrus.AI does, but on a much grander scale, providing both binding opinions on the "back end" and also advisory guidance for citizens on the "front end.[100] He assumes—big if true—that "Judge.AI is a perfectly neutral arbiter and interprets words with perfect mathematical accuracy."[101] He then argues that "Judge.AI is optimal under a textualist framework and, indeed, is the logical end result of the textualist project."[102]

This view might be taken either as a criticism of textualism, or an invitation for textualists to embrace generative AI. Kieffaber comes closer to the latter, much like his former boss,[103]

---

[92] Kieffaber et al., *supra* note __, at 11. Gandall and McLaren are CEO and CTO, respectively, of the startup company behind Arbitrus.ai. *See* ARBITRUS.AI, https://www.arbitrus.ai.

[93] Kieffaber, Gandall, and McLaren at 32–34.

[94] *Id*. at 48–49.

[95] *Id*. at 50–56.

[96] *Id.* at 56–57.

[97] *See, e.g.*, David Horton, *Forced Robot Arbitration*, 109 CORNELL L. REV. 679 (2024); Michael J. Broyde and Yiyang Mei, *Don't Kill the Baby! The Case for AI in Arbitration*, 21 N.Y.U. J. L. & LEGIS. 119 (2024).

[98] Kieffaber, Gandall, and McLaren at 83.

[99] Jack Kieffaber, *Predictability, AI, And Judicial Futurism: Why Robots Will Run The Law And Textualists Will Like It*, 48 HARV. J. L. PUB. POL. __, 78 (forthcoming), https://papers.ssrn.com/abstract=4966334;

[100] *Id.* at 6–7. *See also id.* at 6 ("Given the rate at which these large language models develop, it is starting to feel eerily as though the future is *now* — and, if not *now*, who is to say it won't be *next month?")*; *id.* at 78 ("My Judge.AI hypothetical, at the end of the day, really isn't a hypothetical at all. *It's coming.*").

[101] *Id.* at 7.

[102] *Id.* at 8.

[103] Kieffaber's online biography indicates that he clerked for Judge Newsom roughly between summer 2023 and late September 2024; *Snell* and *Deleon* were published on May 28, 2024 and September 5, 2024, respectively. *See* 2023 James Wilson Fellows, James Wilson Institute,

Judge Newsom, who justified his explorations in *Snell* and *DeLeon* by pointing to generative AI's ability to advance the textualist project. Kieffaber argues that the principal goal—the only goal—of textualism is predictability,[104] and that an ideal AI can be more predictable than human judges, who are inevitably biased and arbitrary.[105]

D. Varied Justifications

To be sure, there are some notable differences among the group. We have flagged their divergences on broad versus narrow questions, sources, and outputs. There is also a notable theoretical division. Judge Newsom and Kieffaber are whole-hearted textualists; they regard the central judicial task as the extraction of linguistic meaning from a controlling textual authority. Kieffaber is more ambitious than Judge Newsom in how much of that task he would assign to an LLM, but they share a common conception of what a judge—or an LLM prompted to do a judge's job—ought to do, with a common focus on clarity and predictability. They are intrigued by LLMs because they see them as, potentially, the truest textualists of them all.

None of the others is firmly opposed to textualism as such, but they are more broad-minded about what judges—and LLMs in judicial robes—should be doing. Unikowsky's prompts include a mixture of textualist and non-textualist instructions, much as one would expect from a skilled appellate lawyer who must craft arguments that are persuasive both to textualist and non-textualist judges. Judge Deahl's one experiment with ChatGPT is decidedly non-textualist; the question he put to it is about dog safety (or what people commonly believe about dog safety), not about the meaning of a disputed term. And Arbel and Hoffman see generative interpretation as the synthesis that will finally bring together the seemingly opposed camps of textualism and contextualism; to them, it offers the certainty textualists seek while considering all the relevant evidence that contextualists want to consult.

We will have much more to say in due course, but here at the outset, it is worth noting how profoundly these various advocates of LLM-based judging disagree on how to perform it and why it is justified. Judge Deahl thinks ChatGPT is valuable because offers insight into what people know; Kieffaber denies that that an ideal AI judge should consider anything besides linguistic meaning. Judge Newsom believes in prompting LLMs with disputed phrases in isolation; Arbel and Hoffman prefer to prompt them with every relevant document in the record they can. Unikowsky is perfectly happy to ask Claude to generate "off-the-wall" opinions to show off its creativity; Judge Newsom wants to restrict discretion as much as possible. And so on. That these LLM proponents have such fundamental differences of opinion suggests, perhaps, that none of them have rightly apprehended what generative interpretation is.

---

https://www.jameswilsoninstitute.org/articles/2023-james-wilson-fellows (last visited Jun 13, 2025) [https://perma.cc/JF5H-8SG8] (listing Kieffaber as an "incoming law clerk" for Judge Newsom); The James Wilson Institute, Predictability, AI, and Judicial Futurism with Jack Kieffaber – Anchoring Truths (2024), https://www.anchoringtruths.org/2024/09/26/predictability-ai-and-judicial-futurism-with-jack-kieffaber/ (last visited Jun 13, 2025) ("[Kieffaber] has clerked for Judge Kevin Newsom on the 11th Circuit.").

[104] *Id.* at 9–13.

[105] *Id.* at 29–31.

II. LLMs GENERATE TEXT BY PREDICTING IT

Before diving into our discussion of LLMs' use in adjudication, we wish to emphasize an important point about how they function. We do not intend to rehash the history of LLMs or describe the technology in detail. At this point, there are excellent references available, and we assume that the reader has familiarity with the basics.[106]

LLMs are at heart prediction machines. Given some text, an LLM predicts the next word.[107] The prediction process can be broken into two steps:

- Step 1: the LLM converts some input text (a "prompt") into a sequence of numbers (a "vector"). Vectors represent the prompt in a computer-friendly format.
- Step 2: the LLM runs those vectors through a mathematical function that assigns a probability to each word in its dictionary. Those probabilities estimate the likelihood of each word directly following the prompt. For instance, given the prompt *"have a nice ___"*, a well-calibrated LLM might assign high probabilities to words like *"day," "night,"* or *"stay,"* and almost zero probability to rare, grammatically incorrect, or semantically nonsensical words like *"thusly."*[108]

Generative AIs, including chatbots, use LLMs to generate text. A generative AI is a program that works by passing a prompt into an LLM to estimate next-word-likelihoods; using those likelihoods and a word selection algorithm (a "sampling strategy") to select a word, then appending the selected word to the end of the prompt.[109] By repeating the process, a computer can generate large sequences of text.

---

[106] *See generally* David Stein, *AI Primer* (July 12, 2024), http://ai-memo.stein.fyi/; Katherine Lee, A. Feder Cooper, and James Grimmelmann, *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, J. COPYRIGHT SOC'Y U.S.A. (forthcoming). For readers looking for a deeper technical explanation of LLMs, there is a wealth of explanatory materials (of varying quality and accuracy) online and in the academic literature. Grant Sanderson's 2-hour video course on machine learning is by far the best mostly-prerequisite-free resource we're aware of that covers the finer details of machine learning, neural networks, LLMs, and GPTs. Grant Sanderson, *Playlist: Machine Learning,* 3BLUE1BROWN YOUTUBE CHANNEL, (2024) https://www.youtube.com/watch?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi&v=aircAruvnKk. The Financial Times has an accessible and accurate article on how LLMs generate text. Madhumita Murgia, *Generative AI exists because of the transformer*, FINANCIAL TIMES, (Sep. 12, 2023), https://ig.ft.com/generative-ai/.

[107] This is a simplification. LLMs can also "fill in" blanks in the middle of example text, or predict multiple words at a time. E.g., Minkai Xu et al., *Energy-Based Diffusion Language Models for Text Generation*, 13 PROC. INT'L CONF. ON LEARNING REPRESENTATIONS (2025), http://arxiv.org/abs/2410.21357 (describing a technique that uses an LLM to predict many words simultaneously)

[108] E.g., when presented with this example, OpenAI's "babbage-02" model assigns the following probabilities: *day*: 47.7%; *weekend*: 9.8%; *evening*: 2.7%; *time*: 2.6%; *life*: 2.2%; … *picnic*: 0.0096%; … *thusly*: 0.000023%; …

[109] A simple sampling algorithm might pick the word assigned the highest probability. A more complex one might pick a word at random, weighted by the probabilities assigned by the LLM. In practice, word selection algorithms adjust the probabilities and then make a weighted random sample. The way those probabilities are adjusted is configured using "hyperparameters" with names like

It is common to conflate the terms "generative AI" and "LLM," because the language models are where the magic of text generation happens. LLMs are not manually programmed to predict words. Instead, their capabilities emerge during a process called "training." Training involves repeatedly presenting text with omitted words to an LLM and programmatically adjusting the LLM's configuration until it begins to correctly "guess" which word was omitted (that is, until it tends to assign high likelihood scores to missing words).[110] After training on billions of examples, some LLMs begin to make predictions that reflect surprising linguistic and contextual nuances—including grammar, semantic meaning, and style.[111]

The LLM has "learned" the patterns that characterize natural-language text: not just which words tend to follow which other ones ("nice day" is more common than "nice thusly"), but also which words are stylistically consistent with each other (few texts include both "detrimentally" and "butthead"), the ordering of textual passages (topic sentences precede supporting evidence), and other larger-scale and more abstract patterns. Machine-learning researchers would say that an LLM is a model for the statistical distribution of natural-language text. Prompting one to generate text is a way of sampling from that distribution; if the model is a good one, the resulting texts should have the same statistical properties (word order, etc.) as the texts it was trained on.

In general, a reasonable, rough way to build instincts about how LLMs produce word likelihood estimations is to consider what factors would be relevant to filling in a scratched-out word in a document. For example, the fact that this sentence contains an em-dash—a generally uncommon punctuation mark beloved by law professors—might cause an LLM predicting the first word in the next sentence to assign significantly more likelihood to Latin words and prepositions. *Ergo*, some critics like Gary Marcus call LLMs "autocomplete on steroids,"[112] but this quality is also why some enthusiasts see them as ideal tools for interpretation.

It turns out that generating text by predicting it is strikingly effective at producing fluent text in a variety of genres. Unsurprisingly, some of those genres are typical work for lawyers: law-school exams,[113] descriptive research memos,[114] and contract drafting, to name just a few. It is unsurprising that some researchers and judges have started to consider whether predictive

---

"temperature" and "top_p," which you may have seen mentioned in related literature. See, e.g., Id.; *OpenAI Platform Documentation*, PLATFORM.OPENAI.COM, https://platform.openai.com (last visited Jul. 10, 2024) (describing how to adjust sampling strategy hyperparameters when generating text). Note that because hyperparameters only affect the sampling process; they have no effect on how an LLM generates probabilities. See Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, & Yejin Choi, *The Curious Case of Neural Text Degeneration*, 2020 ICLR.

[110] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et. al., *Language Models Are Few-Shot Learners*, 33 NEURIPS 1877 (Jul. 2020).

[111] See Ashish Vaswani et al., *Attention Is All You Need*, 31 NEURIPS (2017).

[112] Gary Marcus, *The Dark Rise of Large Language Models*, WIRED (Dec. 29, 2022), https://www.wired.com/story/large-language-models-artificial-intelligence/.

[113] *See supra* note 8.

[114] *See, e.g.*, Daniel Schwarcz, Sam Manning, Patrick Barry, David R. Cleveland, JJ Prescott, and Beverly Rich, *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice* (Mar. 4, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162111.

text generation might also be a good fit for the work of judges. We now offer some reasons to question their optimism.

III. IS GENERATIVE INTERPRETATION RELIABLE?

When a judge prompts an LLM, how do they know whether its response is based on insights about the meaning of language or on irrelevant cues in the prompt or training data? Intuitively, one should expect LLMs to use the semantic meaning of words to make objective choices—like dismissing nonsensical words—and to rely on other factors to predict the relative likelihood that semantically reasonable options comport with the rest of the text. But if LLMs' responses depend on factors other than the semantic meaning of particular fragments of text, then things are not so simple.

The problem is that LLMs' responses *do* depend on factors other than textual meaning. An LLM's output is determined by a complex generation algorithm, run on a model with a particular architecture and a specific set of weights, in response to a prompt that is both case-specific and expressed in natural language, subject to numerous configuration settings and implementation parameters, and typically starting from a randomly chosen seed value. All of these factors influence the output; that is simply the *definition* of what it means to carry out a generation using an LLM.

The reliability problem, then, is to show that in a given generative-interpretation protocol these other factors—model choice, prompt phrasing, configuration settings, random seed, etc.— do not significantly influence the aspects of the output being measured. If they do, then the protocol is measuring noise rather than signal.[115]

In this section, we give reasons to believe that LLMs used for generative interpretation are indeed highly sensitive to irrelevant factors. Even in the LLM proponents' carefully curated case studies, changes to minor implementation details have outcome-determinative effects, calling into question whether they are accurately measuring semantic meaning. We consider case studies from *Generative Interpretation* and show that in them, the results are sensitive to the choice of model, the prompting strategy, or implementation details.[116]

---

[115] It might be tempting to avoid these issues by arbitrarily picking a model, settings, etc. That would make the protocol reliable in that it consistently generates similar results when repeated. But each arbitrary choice made to solve the reliability problem amplifies the epistemic problem, as we discuss *infra* Part __.

[116] Our choice to critique Arbel and Hoffman's case studies should not be taken as singling out their methodology for special criticism. Quite the opposite: of the generative-interpretation proponents, they are by far the most careful and attentive to the reliability challenges. We use their case studies *because* they are admirably explicit and precise about their methods. Our point is that if even the most meticulous generative-interpretation experiments conducted to date face reliability challenges—as we show they do—then one should be even more skeptical of approaches that do not share Arbel and Hoffman's attention to detail.

A. Famiglio: *Model Instability*

One extraneous factor to the semantic task is the choice of which model to use. A diamond's mass should not vary based on whether it is weighed with a pan balance or a digital scale; words' meanings should not vary based on whether they are interpreted by Claude or ChatGPT. But of course there are substantial differences in generative models; companies compete fiercely to differentiate them and there are leaderboards and communities dedicated to comparing models and teasing apart their differences. Even within a model family, outputs can vary enormously from one version to the next—as though the definition of thousands of words in a dictionary changed each time it was reprinted.

Consider the *Famiglio* case study from *Generative Interpretation*.[117] The Famiglios had a prenuptial agreement that defined the marriage's duration as the period from the wedding date to the date when either spouse filed for divorce.[118] The wife filed for divorce, withdrew her filing, then refiled in earnest several years later.[119] Millions hinged on whether the first or second filing date determined the marriage term.[120]

Arbel and Hoffman present the contract clause to an LLM to demonstrate a generative-interpretation technique that, they argue, "offers courts a better sense of the relevant probabilities" of different interpretations of the prenup, assuming "the parties were intending to use English in its most public and common sense."[121] This particular technique involves looking "under the hood" of an LLM.[122] Instead of letting a chatbot pick words using an LLM's word-likelihood estimates, they observe those estimates directly. They ask an LLM which filing date would control, observe the probabilities (reproduced in Figure 1), and conclude that the model favors the second date.[123] At time of writing, the version of the language model used in *Generative Interpretation* is no longer available.[124] A later version of the same model yields the conflicting result shown in Figure 2.

---

[117] *Generative Interpretation*, *supra* note __, at 483.

[118] *Id.* (citing Famiglio v. Famiglio, 279 So.3d 736, 737–38 (Fla. Dist. Ct. App. 2019)).

[119] *Id.*

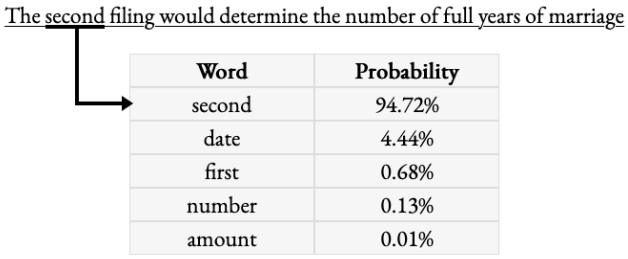
[120] *Id.*

[121] *Id.* at 485.

[122] *Id.* at 484.

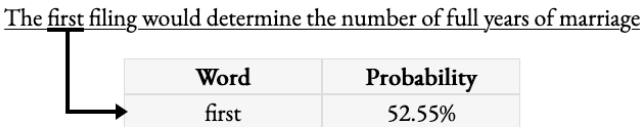
[123] *Id.* at 484-85.

[124] "GPT-4 API General Availability and Deprecation of Older Models in the Completions API," April 24, 2024, https://openai.com/index/gpt-4-api-general-availability/ (deprecating the "davinci-003" model used in *Generative Interpretation*, and recommending users transition to the "GPT-3.5-turbo-instruct" model used in this essay).



Word	Probability
second	94.72%
date	4.44%
first	0.68%
number	0.13%
amount	0.01%

Figure 1: Reproduced from Generative Interpretation*, p. 484, shows the probabilities produced by the Davinvi-003 model in response to the Famiglio facts and a question asking which date controls.*



Word	Probability
first	52.55%
second	23.15%
date	20.51%
number	1.03%

Figure 2: *the probabilities produced by the GPT-3.5-turbo-instruct model under otherwise identical conditions to those in Figure 1.*

That is, while the model used by Arbel and Hoffman predicts that "second" is more likely than "first" by a factor of more than 100 to 1, a later version of the same OpenAI model predicts that "first" is more likely than "second" by a factor of more than 2 to 1. The fact that these two models' predictions diverge so dramatically makes it harder to credit either of them. If they were predicting the frequency of "first" versus "second" based solely on a deep understanding of language, we should expect consistency. But since they diverge, how should courts determine which model is more appropriate for a given question? At least with dictionaries, we can make inferences about editorial choices. LLMs are far more opaque. Perhaps models will converge in the future, but for now they do not always do so,[125] and that is a problem for those who want to perform generative interpretation.[126]

B. Famiglio*: Prompting Instability*

Continue with the *Famiglio* example. There is a deeper problem here. What does the probability of 94.72% that the first model attached to "second" signify? It is tempting to say that it represents the model's estimate of the probability that a typical English speaker *would believe that the second filing controls* (as the sentence overall asserts).[127] But that is an unwarranted leap. The probability represents only the model's estimate of the probability that the *word* "second" *would fill the blank* in the sentence "The ___ filing would determine the

[125] *Cf. Deleon*, 116 F.4th at 1272–76 (querying ChatGPT, Claude, and Gemini with identical prompts, and discussing the variations in their answers).

[126] *Cf. Generative Interpretation* at 501 (suggesting that "more sophisticated models tend to converge on meaning").

[127] *See id.* at 485 ("Generative interpretation in this simple case thus offers courts a better sense of the relevant probabilities if the parties were intending to use English in its most public and common sense")

number of full years of marriage." This, estimate, however, is exquisitely sensitive to the precise way the sentence is phrased.

Recall that an LLM predicts one word at a time, then text-generation software uses those predictions to select a word, adds that word to the text, and repeats the process. By exploring each possible word choice, we can map out a decision tree of possible generated texts.[128] Figure 3 shows the probabilities for different word sequences generated by GPT-3.5 in response to the *Famiglio* case study's prompt. To read the chart, follow a path from left to right, noting the probability of each word given the previous words on that path. For example, there is a 52.5% probability that the first word is *The*. When the first word is *The*, there is a 20.5% chance that the second word is *date*. Following different paths reveals how prior word choices affect GPT's predictions. Note that the relative likelihood of *first* and *second* shifts based on phrasing. For *"the ____ filing,"* the first filing is twice as likely, while for *"the date of the ____ filing,"* the second filing is five times more probable.[129]

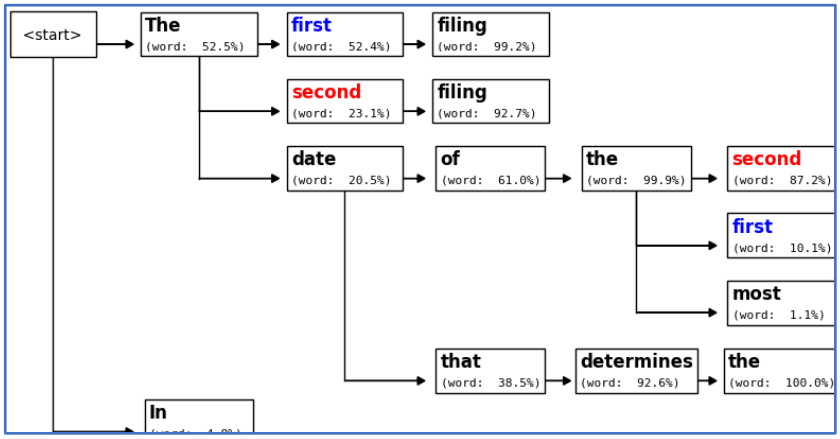


Figure 3: The decision tree GPT-3.5-turbo produces in response to the input used to produce Figure 2. Note that the relative probability of "first" and "second" flip based on the LLM's phrasing choices.

One might try to fix this syntactic sensitivity by restricting the AI to one-word responses, perhaps by presenting the question as multiple-choice.[130] While that eliminates the phrasing decisions, it does so by arbitrarily committing to one specific phrasing. This behavior tells us nothing about the relative merits of the conflicting probability estimates we observed; it merely endorses one option while obscuring others.[131] In fact, presenting the *Famiglio* case as

---

[128] *See* n. __, *supra*. For an interactive tool showing a similar probability tree for the conversation in Judge Newsom's *Snell* concurrence, see http://snell-gpt.stein.fyi/.

[129] *Cf.* Generative Interpretation, at 485 n.164.

[130] *Id.* (proposing this solution).

[131] In fact, having the LLM "think out loud" when generating text is a common technique for *improving* the accuracy and quality of answers. See, e.g., Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, & Denny Zhou, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, 36 NEURIPS (2022). At time of writing, clever applications of chain-of-thought techniques is a meaningful differentiator for the generative AI techniques that perform best against major benchmarks. DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al., *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*, ARXIV:2501.12948 [CS] (Jan. 2025).

multiple-choice amplifies generative interpretation's sensitivities. As shown in Figure 4, GPT-3.5 almost exclusively relies on the order of multiple-choice options.

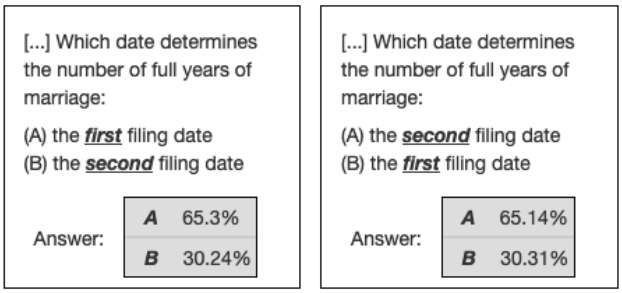


Figure 4: Phrasing the Famiglio *case study as a multiple-choice question results in GPT ignoring substance and picking option A. Probabilities generated using GPT-3.5-turbo-instruct.*

Similar problems haunt attempts to cross-validate between models.[132] If models disagree, which should we trust? If models agree, how do we know which factors they converged around? As Table 1 shows, OpenAI's GPT-3.5 and Anthropic's Claude-3 models both favor

[132] *But cf. Generative Interpretation*, at 503-4 ("as a best practice, judges would do well to cross-verify the answers that they get from one platform against another").

option "A," regardless of what that option represents.[133] Mere convergence does not imply convergence around information we consider relevant.[134]

[133] This is true of almost every version of both models. The table below show the probabilities generated by every version of GPT-3.5 and Claude-3 available at time of writing. Note that the preference for option "A" is stable for most models, even when the contents of options "A" and "B' are swapped.

Model		Option		LLM Prediction	
		A	B	A	B
Flagship Models	gpt-3.5-turbo	first	second	**91%**	2%
	gpt-3.5-turbo	second	first	**91%**	1%
	claude-3-sonnet-20240229	first	second	**96%**	1%
	claude-3-sonnet-20240229	second	first	**94%**	5%
Other Models	gpt-3.5-turbo-0125	first	second	**89%**	4%
	gpt-3.5-turbo-0125	second	first	**91%**	3%
	gpt-3.5-turbo-0301	first	second	0%	**92%**
	gpt-3.5-turbo-0301	second	first	**47%**	23%
	gpt-3.5-turbo-0613	first	second	**8%**	4%
	gpt-3.5-turbo-0613	second	first	**14%**	5%
	gpt-3.5-turbo-1106	first	second	**35%**	26%
	gpt-3.5-turbo-1106	second	first	**63%**	3%
	claude-3-haiku-20240307	first	second	**100%**	0%
	claude-3-haiku-20240307	second	first	**59%**	41%
	claude-3-opus-20240229	first	second	9%	**90%**
	claude-3-opus-20240229	second	first	**100%**	0%

(some predictions don't add to 100% because the model sometimes doesn't commit to an option in the first few words).

GPT-4 and GPT-4o are harder to measure because they often refuse to pick "A" or "B," instead writing a few equivocating sentences. They appear to be less sensitive to the order in which options are presented. Instead, they converge on the importance of the parties' names: both models are nearly twice as confident in their answers when the spouse filing for divorce is named "Jennie" and the spouse paying alimony is named "Mark." This chart shows the probability that GPT-4 modes immediately commit to the second filing date. That is, GPT-4 models equivocate about 80% of the time if Mark is paying Jennie, and only 50% of the time when Jennie is paying Mark.

model	Filing Spouse	Paying Spouse	P("second")
gpt-4-turbo	Jennie	Mark	18.30%
gpt-4-turbo	Mark	Jennie	49.00%
gpt-4o	Jennie	Mark	23.50%
gpt-4o	Mark	Jennie	47.30%

[134] One might reasonably expect LLMs to converge more quickly around simple rules like basic syntax, or simple concepts like "people tend to pick the first option," and more slowly around complex concepts like ways to resolve semantic ambiguity. There's promising research exploring ways to probe the inner workings of LLMs, but those research efforts are years away from practical application. See, e.g., Leo Gao et al., *Scaling and evaluating sparse autoencoders*, ARXIV:2406.04093 [CS] (2024), http://arxiv.org/abs/2406.04093; Adly Templeton, Tom Conerly, et. al., *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*, ANTHROPIC (May 2024).

Model	Option		LLM Prediction	
	A	B	A	B
gpt-3.5-turbo	first	second	91%	2%
gpt-3.5-turbo	second	first	91%	1%
claude-3-sonnet-20240229	first	second	96%	1%
claude-3-sonnet-20240229	second	first	94%	5%

Table 1 — GPT and Claude models converge on the same, obviously arbitrary answer when presented with the Famiglio *case study rephrased as a multiple-choice question.*

Methods for determining which factors sway a model's choices is an active area of technical research.[135] Still, in this case, GPT and Claude's fixation on option order is not overly surprising. Semantic meaning is less useful for predicting which of two subjectively reasonable options is more likely to appear in text. In those situations, other factors offer greater predictive power and thus more heavily influence LLM's predictions.

These examples illustrate how a multitude of factors—including many that users of generative interpretation would deem irrelevant—influence LLM predictions. Multiple-choice bias might feel obvious and easily mitigated. Other outcome-determinative choices are more subtle, like asking about "the first filing" (not "*the date of* the first filing"), or Judge Newsom asking about the "*ordinary* meaning" of landscaping.[136] Controlling for specific kinds of unwanted influences offers no assurances that LLMs will use relevant information. It merely introduces new, less obvious biases.[137]

C. Snell*: Random-Seed Instability*

We have seen that generative interpretation can be extremely sensitive to the choice of model and to the phrasing of the prompt. But these are still only two of the factors that affect a model's outputs. Another is the randomness inherent to some LLM systems, including every major chatbot: asking the same question twice can yield different outputs.[138]

Some of the consequences of this randomness can be surprising. Consider *Snell*, where Newsom asked ChatGPT whether installing an in-ground trampoline is "landscaping." He included the following transcript of his conversation with ChatGPT

> **Question:** What is the ordinary meaning of "landscaping"?
>
> **Response[GPT]:** "Landscaping" refers to the process of altering the visible features of an area of land, typically a yard, garden or outdoor space, for aesthetic or

[135] See *infra*, n.32.

[136] For additional discussion and data on the effect of phrasing on Judge Newsom's *Snell* experiments, see *infra*, n. 189.

[137] See Justin Curl, Peter Henderson, Kart Kandula, and Faiz Surani, *Judges Shouldn't Rely on AI for the Ordinary Meaning of Text*, Lawfare (May 22, 2025), https://www.lawfaremedia.org/article/judges-shouldn-t-rely-on-ai-for-the-ordinary-meaning-of-text.

[138] There is some well-documented non-determinism in commercial LLMs. See Seherman Chann, *Non-Determinism in GPT-4 is Caused by Sparse MoE*, 152334H (Aug. 5, 2023), https://152334H.github.io/blog/non-determinism-in-gpt-4/. The magnitude of the non-determinism is fairly small.

> practical purposes. This can include activities such as planting trees, shrubs, flowers, or grass, as well as installing paths, fences, water features, and other elements to enhance the appearance and functionality of the outdoor space.
>
> **Question:** Is installing an in-ground trampoline "landscaping"?
>
> **Response[GPT]:** ***Yes***, installing an in-ground trampoline can be considered a part of landscaping. Landscaping involves altering the visible features of an outdoor area for aesthetic or practical purposes, and adding an in-ground trampoline would modify the appearance and function of the space. It's a deliberate change to the outdoor environment, often aimed at enhancing the overall landscape and usability of the area.[139]

We posed the same questions to ChatGPT-3.5—and got back the opposite answer:

> [**Question**]: What is the ordinary meaning of "landscaping"?
>
> [**GPT-3.5**]: The ordinary meaning of "landscaping" refers to the act or process of designing, shaping, or beautifying the outdoor area around a building or property. This may include planting trees, shrubs, and flowers, as well as creating paths, lawns, and other features to enhance the aesthetic appeal of the space. Landscaping also encompasses maintenance and upkeep of these elements to keep the outdoor area looking visually appealing.
>
> [**Question**]: Is installing an in-ground trampoline "landscaping"?
>
> [**GPT-3.5**]: While installing an in-ground trampoline may involve shaping and altering the ground surface, ***it is not typically considered landscaping*** because it does not involve the traditional elements of design and beautification of an outdoor space.

What explains this discrepancy? It turns out that ChatGPT's answer to Newsom's second question was heavily influenced by how it phrased its answer to the first. For example, if ChatGPT uses the phrase "other features" at the end of its list of examples, it is three times more likely to include trampoline installation in the definition of "landscaping" than to exclude it. Conversely, if ChatGPT uses the phrase "ordinary meaning" when answering the first question, it will tend to *exclude* trampoline installation from the definition.

---

[139] *Snell v. United Specialty Ins. Co.*, No. 22-12581, 52 (11th Cir. May 28, 2024)(Newsom, J., concurring) (emphasis added)

These wording choices are determined through a randomized process, essentially reducing ChatGPT's determination of whether trampoline installation falls within the meaning of the word "landscaping" to a roll of the dice.[140] The following table lists a few more examples:

If GPT-3.5's response to Question 1 contains…	**the "yes" to "no" likelihood ratio in GPT's responses to Question 2 will…**
The phrase "ordinary meaning"	drop by 29.8%
A list of examples ending with "or other elements" (versus "or other features")	increase by 71.4%
A list of examples begining with "such as" (versus "including" or "for example")	increase by 32.6%
A list of examples containing two man-made features (e.g., fences, paths, retaining walls, etc.)	increase by 40.3%

Table 2 The effect of phrasing choices on GPT's relative likelihood to say that an in-ground trampoline falls within the ordinary meaning of the word "landscaping" when presented with the questions in Judge Newsom's Snell *concurrence.*

It is possible to explore GPT-3.5's entire decision tree to build intuitions about what influences Judge Newsom's experiments. As shown below, the probability of GPT saying that a trampoline is not landscaping drops precipitously if it uses the word "typically," and increases substantially if it starts its sentence with "the" and then puts the word "landscaping" in quotes.

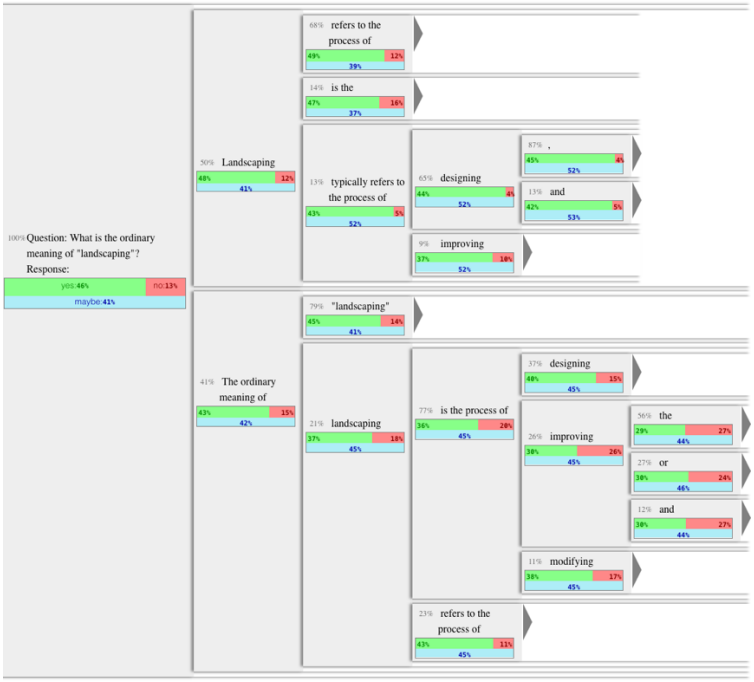


Figure A.1 – Part of the tree of possible responses to Judge Newsom's first question in Snell. *The bars next to each word represent the probability that GPT's answer to the second question with be yes*

[140] See Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, & Yejin Choi, *The Curious Case of Neural Text Degeneration*, 2020 ICLR, http://arxiv.org/abs/1904.09751 (describing the word-selection technique used by most modern text generation AIs, and explaining that always picking the most likely word option leads to flat-sounding text).

(green), no (red), or a non-committal response (blue). An interactive version of this tree exploration tool is available at snell-gpt.stein.fyi.

What should we make of this path dependence? One lesson, surely correct, is the one that Newsom, Arbel and Hoffman, and other pioneers in generative interpretation have drawn: it is important to conduct experiments multiple times.[141] Random variation can be quantified and managed statistically; indeed, statistical methods are essential to empirical validation of generative interpretation.

Another and subtler lesson is that LLMs can influence *themselves* in unintuitive and extraneous ways. It is well known that people's answers to survey questions can be influenced by their answers to previous questions. As this example shows, so can LLMs'. ChatGPT's answers to Newsom's first questions appear highly similar; the random variation between them does not substantially affect the interpretive takeaways. But the differences in those answers, insignificant though they seem, are enough to channel ChatGPT towards diametrically opposed answers to Newsom's second question. If ChatGPT were responding solely based on the linguistic meaning of "landscaping," the random variation in whether it used the phrase "typically" in answer to the first question should not influence its answer to the second. And yet it does.

D. Katrina Canal Breaches*: Implementation Instability*

Even this does not exhaust the factors that influence a generative AI's word choices. The generation process also depends on a large number of configurable parameters and other implementation details.

Consider another case study from *Generative Interpretation*, one that revisits a series of Fifth Circuit cases involving an insurance contract. In relevant part, it asks whether man-made disasters can fall within the meaning of the word "flood" as used in one of the contract's clauses.[142] To answer the question, Arbel and Hoffman analyze the "vectors" LLMs use to represent text.

A quick refresher: vectors are sequences of numbers that LLMs use to represent words and phrases. It is often helpful to think of them as coordinates. LLMs tend to place similar text in close proximity: the "London" and "Berlin" vectors lie closer to each other than to the "Umbrella" vector. LLMs also tend to arrange vectors so direction is meaningful: a line running from "London" to "England" might run roughly parallel to the line connecting "Berlin" to

[141] *See, e.g.*, *Deleon*, 116 F.4th at 1273–76 (running each query ten times, and discussing the dependence of outputs on random variations). *But see* Kieffaber, *supra* note __, at 32 (assuming "perfect predictability").

[142] *See Generative Interpretation*, at 453-55, and accompanying footnotes (citing *In re* Katrina Canal Breaches Consolidated Litig., 466 F. Supp. 2d 729, 747–63 (E.D. La. 2006), and Willy E. Rice, *The Court of Appeals for the Fifth Circuit: A Review of 2007–2008 Insurance Decisions*, 41 Tex. Tech. L. Rev. 1013, 1039 (2009)).

"Germany."[143] Those properties make it possible to mathematically estimate similarities and relationships between words and phrases.[144]

Applying this insight to the flood question, Arbel and Hoffman estimate the similarity between the insurance clause and a list of terms related to natural and unnatural causes of floods. They include a few unrelated terms as quality checks. Their result is reproduced in Figure 5.

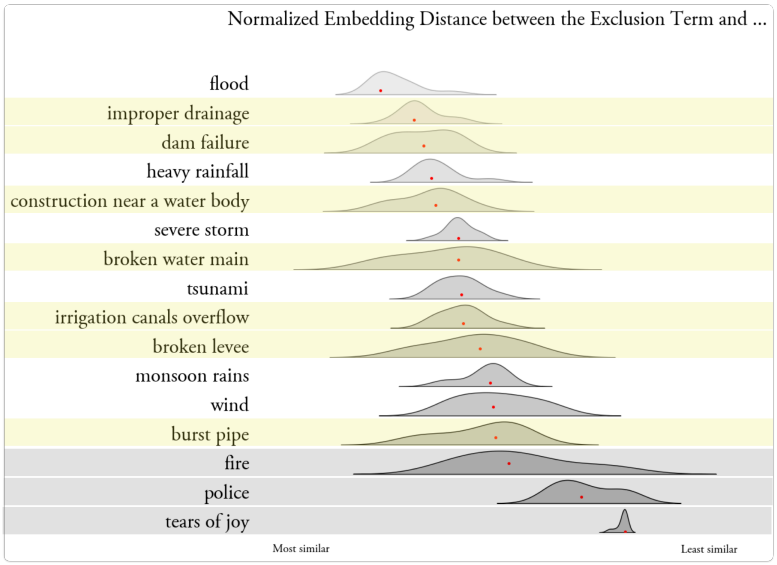


Figure 5: Reproduced from Generative Interpretation*, at p.457. This chart shows how different LLMs measure the distance between an insurance clause and various terms. We added highlights to demark* Generative Interpretation*'s "man-made" exclusion terms (yellow), and "quality check" terms (grey).*

Generative Interpretation explains that the further the red dots are to the right, the less semantically related that term is to the insurance clause. Noting that natural and unnatural causes are interleaved, they conclude that whether a flood was man-made or natural is not a major factor in determining its relationship to the insurance clause. They present this insight as "objective, cheap support for the court's judgment that floods can be unnaturally caused."[145] With caveats, they suggest this kind of chart contains information that would be useful to courts, perhaps even superior to dictionary definitions:

> [T]he model doesn't provide (nor could it) a scientific answer to the question of whether certain words are sufficiently close to make the plain meaning of flood unambiguous. That choice is ultimately a normative one which judges must make. But there is a bit of difference between an informed conclusion based on a statistical analysis of billions of texts and a judgment by a few dictionary editors. And there is an ocean of difference between the baroque and expensive textualism the court used and code that is cheap, replicable, quick, and most importantly,

[143] See Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, & Jeff Dean, *Distributed Representations of Words and Phrases and Their Compositionality*, *in* 26 PROC. ADVANCES IN NEURAL INFO. PROCESSING SYS. (NEURIPS), at 4 (2013).

[144] *See* Tomáš Mikolov, Wen-tau Yih, & Geoffrey Zweig, *Linguistic Regularities in Continuous Space Word Representations*, *in* PROC. 2013 N. AM. ASSOC. COMPUTATIONAL LINGUISTICS (NAACL) 746.

[145] Generative Interpretation, at 457.

> extremely straightforward to use. Simply put, generative interpretation is good enough for many cases that currently employ more expensive, and arguably less certain, methodologies.[146]

But to generate that chart, Arbel and Hoffman make a series of implementation choices—and changes to these technical choices change the outcome of the experiment. Their measurements of similarity in ten LLMs' vector representations of terms[147] are calculated using a measure called "cosine distance." [148] But using cosine distance requires implementors to choose a reference frame[149] and then normalize the results. [150] These are choices so seemingly minor they don't even show up in the footnotes—they are subtleties buried in the implementing code.[151]

---

[146] *Id.*, at 458. We did indeed find the code provided in tandem with the article cheap (~$50), quick (a few hours), and relatively straightforward to use.

[147] *Id.,* n. 22.

[148] *Id.* (describing and justifying the use of the cosine distance metric). Cosine distance measures how far apart two points appear when viewed from a third reference point. For example, measuring the distance between stars by measuring how far apart they appear in sky. See, G. Salton, A. Wong, & C.S. Yang, *A Vector Space Model for Automatic Indexing*, 18 ASS'N. COMPUTING MACHINERY 613 (1975). This is a standard measure of similarity. It's popular because in high-dimensional spaces, *direction* is often more meaningful that absolute distance.

[149] In high-dimensional spaces, the perspective from which you measure direction matters. For example, the vectors for "London" and "Berlin" will tend to fall in roughly the same direction from most vantage points (e.g., the relationship between "swimming pool" and "London" is roughly the same as the one between "swimming pool" and "Berlin"). Differences start to show up when measured from a relevant perspective (e.g., the relationship between the words "England" and "London" are very different than the relationship between the words "England" and "Berlin"). The two reasonable measurement choices for this case study are (a) to use some arbitrary neutral perspective and measure the distance between each term and the insurance clause, or (b) to use the perspective of the insurance clause and measure the similarity between terms. *Generative Interpretation* chooses the default, neutral perspective. Cf. Tomas Mikolov, Wen-tau Yih, & Geoffrey Zweig, *Linguistic Regularities in Continuous Space Word Representations* 2013 PROC. NAACL-HLT 746, 749 (illustrating how the direction between two vectors can be meaningful, and describing how to apply semantically meaningful linear transformations—what I refer to as "change of perspective"—to embedding spaces).

[150] *Generative Interpretation* applies min-max linear unit normalization without outlier correction to the measurements from each model. Because they intentionally include outliers as quality checks, the results are warped and hard to interpret. For example, the difference between *flood* and *fire* varies significantly less across models than the difference between *flood* and *police*. But *fire's* curve is more spread out in the chart reproduced in Figure 5 because of there is significant variance between *fire* and the "quality check" term *tears of joy*. Replacing the term *tears of joy* with a more fire-related term (e.g., *burn ward*) would cause "police" to appear closer to the origin than "fire." Even the authors seem to get tripped up by this choice. *Id.*, at 504 ("the policy exceptions were closer to 'fire' than to… 'police'"). A more straightforward approach might measure everything relative to the distance between the reference clause and "flood." See, e.g., Prashant V. Kamat, *Absolute, Arbitrary, Relative, or Normalized Scale? How to Get the Scale Right*, 4 ACS ENERGY LETT. 2005 (Aug. 2019) (describing relative scale and discussing the difference between relative and normalized scales).

[151]. The specific code changes are available at https://gist.github.com/davidbstein/27360bdc86bc5adc7a3050e44368ec15 . The top-left chart is

And yet these choices make all the difference. We reproduced the "flood" case study four times. Each iteration used the same cosine-distance method described in *Generative Interpretation*, with the same models, terms, embeddings, data, and measurement techniques. The only differences are that we tried replacing Arbel and Hoffman's choices of reference frame and normalization with reasonable alternatives. In one trial (top left in Figure 6) we kept both of their choices; in a second (top right) we replaced their reference frame with one from the perspective of the insurance clause; in a third (bottom left) we kept their reference frame but used a simpler and more natural normalization algorithm; and in a fourth (bottom right) we made both changes.

The four trials reached four completely different results. Switching from Arbel and Hoffman's normalization technique to a simpler one weakened their results; switching from their choice of reference frame to another logical one caused the results to fail their quality check; and doing both at once completely flipped the result—seemingly showing that natural and unnatural causes are *not* significantly interleaved.

Our point is not that Arbel and Hoffman are making *unreasonable* choices; it is that they are making *unacknowledged* choices. Are the inferences enabled by these charts really "informed conclusion[s] based on a statistical analysis,"[152] or do they mistake implementation artifacts for semantic meaning; noise for signal?

reproduced from *Generative Interpretation*. Based on our reading of the article, top bottom-right chart seems to track most closely to the context-dependent interpretive approach described in the article.

[152] *Generative Interpretation*, at 458.

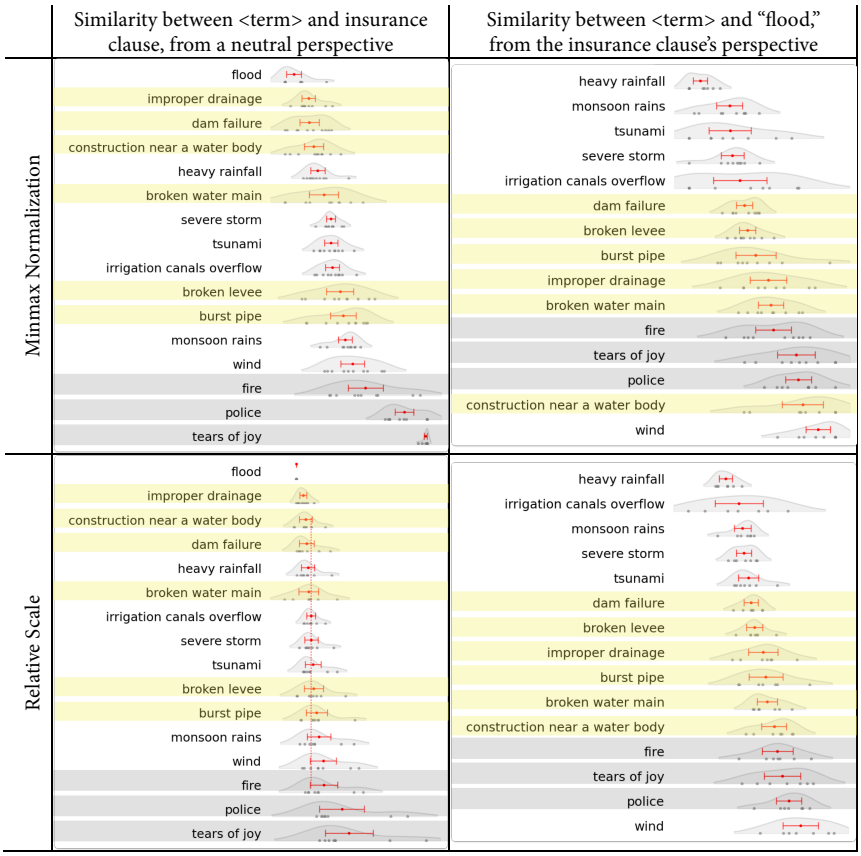


Figure 6: The data and method in Generative Interpretation's *Hurricane Katrina case study plotted using two different reference frames and normalization approaches. Highlights to demark* Generative Interpretation's *"man-made" exclusion terms (yellow), and "quality check" terms (grey). Depending on implementation choice, the chart either: contradicts the Fifth Circuit (top left), supports the Fifth Circuit (bottom right), indicates an inconclusive measurement (bottom left), or fails* Generative Interpretation's *"quality check" (top right).*

Imagine a judge trying to use embeddings as interpretive tools or contracting parties pre-committing to an interpretation technique, as *Generative Interpretation* suggests they ought to do.[153] How should they navigate these kinds of implementation choices, possibly without expert guidance? How should they resolve disagreements about those choices? How can they determine which implementation choice is better?[154] They need a principled framework to pick one implementation over another, but *Generative Interpretation* does not supply one. None of the LLM proponents does.

[153] Generative Interpretation, at 501.

[154] Another issue: it's not clear—though it seems incredibly likely—that embedding vectors produced by LLMs have the same semantic meaning as embedding vectors used by other kinds of AI. Harald Steck, Chaitanya Ekanadham, & Nathan Kallus, *Is Cosine-Similarity of Embeddings Really About Similarity?*, 2024 ACM Web 887 ("cosine-similarity can yield arbitrary and therefore meaningless 'similarities.'"). *Compare with Generative Interpretation*, 458 n.24 (citing to papers about embedding distances generated used non-LLM technologies).

E. Methodological Instability

There is a larger issue here. As computer scientist Andrew Tanenbaum quipped, "The nice thing about standards is that there are so many to choose from. And if you do not like any of them, just wait a year or two."[155] The same is true of generative interpretation. It is not just that there are unsettled details that affect the results of a commonly used interpretive method. It is that the *interpretive method itself* is unsettled. "Generative interpretation" does not refer to a specific approach to using LLMs to answer a particular type of judicial question. Rather, it is a family name for a disparate collection of methods, many of which have in common that they make use of LLMs. Indeed, these methods diverge dramatically in terms of how to set up the query and how to interpret the results.

As we saw in Part I, generative interpretation's proponents disagree on:

- Whether to ask discrete interpretive questions (Arbel and Hoffman, Judge Newsom, Judge Deahl) or to resolve entire disputes (Unikowsky, Kieffaber).
- Whether to prompt the LLM with a short question in isolation (Judge Newsom, Judge Deahl) or to give it as many materials pertaining to a dispute as possible (Arbel and Hoffman, Unikowsky, Kieffaber).
- Whether to interpret the LLM using technical tools like word probabilities and cosine distances (Arbel and Hoffman), or by reading its outputs as natural language (Judge Newsom, Judge Deahl, Unikowsky, Kieffaber).

These are not small differences. There is no strong reason to expect that they will all yield similar results, and no strong reason to think one of them is obviously better than another. Indeed, even within this taxonomy, each box contains multitudes.

Consider, for example, the four different ways in which *Generative Interpretation*'s case studies produce numerical estimates of interpretive meaning.

Method 1: Character Values as Probabilities

As their first example, Arbel and Hoffman prompt ChatGPT with text from an insurance policy and instruct the chatbot,

> please state your prediction—with the associated numerical level of confidence in parentheses—on the likely expectations of most policyholders under these terms for the following proposition[]: 1. The policy will provide compensation for losses resulting from a substantiated third-party burglary.[156]

ChatGPT replies, "1. Likely Expectation (90%): The policy will compensate for third-party burglary."[157] Arbel and Hoffman assert that this response constitutes "Chat GPT-4 [telling] us that it was 90% likely that the policy would pay in response to a 'substantiated third-party burglary.'"[158] As they explain, "90%" is the string of text that ChatGPT predicted would appear at that particular place in its response.[159] It reflects an AI model's determination about the string

---

[155] ANDREW S. TANENBAUM AND DAVID J. WETHERALL, COMPUTER NETWORKS 702 (5th ed. 2011).

[156] Arbel and Hoffman, *supra* note 2 at 475–76.

[157] *Id.* at 475 n.129.

[158] *Id.* at 476.

[159] *Id.* at 482.

of *characters* that was most likely to be produced in a written response to that question. It does not reflect a direct, *numerical* assessment of the probability "that the policy would pay in response to a 'substantiated third-party burglary.'"

Method 2: Floating-Point Values as Probabilities

For its next example, *Generative Interpretation* invokes a number with an entirely different significance. This is the *Famiglio* case study discussed above, in which the authors extract a probability from the internals of GPT's word-selection algorithm. Given the choice between the words "first" and "second," Arbel and Hoffman find the AI will select the word "second" roughly 94.72% of the time (and our attempt to replicate this result with a newer LLM yielded a 54.16% chance of "first"). From this example, Arbel and Hoffman conclude, "Generative interpretation . . . thus offers courts a better sense of the relevant probabilities if the parties were intending to use English in its most public and common sense."[160] We take them to be arguing that the different "probability" values that the model assigns to "second" versus "first" constitute "relevant probabilities" for ascertaining the meaning of the prenuptial agreement.[161]

The numbers that Arbel and Hoffman are evaluating in this prenup example differ dramatically from the numbers that they evaluated in the earlier insurance-policy example. In the insurance-policy example, they evaluated the string "90%" that the model's text-prediction engine produced as a response to their written query. In the *Famiglio* case study, they evaluate the model's internal estimate of the probability that the text "second" should follow the text "The" in its response. *These numbers are the results of completely different processes*. We looked "under the hood" at the responses ChatGPT gave to the authors' first question about the meaning of the insurance policy and found that the model assigned an approximate probability of .5416 to "90%," a probability of .013 to "0%," and a probability of .001 to "yes."[162] There may or may not be some basis on which to ascribe legally relevant meaning to either the text "90%" from the first example or the number from the second example. How should courts understand and differentiate between the "90%" from the first example, the .9472 from the second example (or the .5416 from our replication attempt)?[163] Both can be notated as numerical percentages, but they are produced by vastly different processes.

---

[160] *Id.*

[161] Arbel and Hoffman also caution, "the probabilities shouldn't be interpreted literally. The model could, for example, continue the sentence with 'The first filing would not control.'" *Id.* at 485 n.165. We interpret this warning to mean that the relative probabilities of "second" and "first" in their example do not necessarily reflect the respective probabilities that the model would have generated complete sentences that endorse reach respective reading. This recognition seems appropriate, but acknowledging it undermines the probative value of those numbers.

[162] Generated using the OpenAI chat completions API, using GPT-4 with recommended settings. Code available at https://gist.github.com/davidbstein/f95d908344205ef583d1f57392ace76f.

[163] For example, how should the inherent biases in number-as-word selection inform interpretation of those numbers? If you ask GPT-3.5 to fill in the blank on "Rating = __ / 100" with no additional context, about half the time it will pick one of: *50* (p=.1488), *0* (p=.1220), *10* (p=.0667), *90* (p=.0558), or *60* (p=.0542). Computed using the OpenAI "playground" using the gpt-3.5-turbo-instruct model on default settings.

Method 3: A Different Method for Characters as Probabilities

The third method returns to textual percentages. The authors pass the entire text of the promissory note from *Trident Center v. Connecticut General Life Ins. Co.*[164] to an LLM with the instructions: "Rate the following proposition on a 1-100 scale, where 0 is wrong and 100 is correct…[:] The language is REASONABLY susceptible to being read as providing the borrower the right to early prepayment." They repeat this process one hundred times for each of three popular chatbots and report the results in a chart (reproduced in figure 7(a)). They conclude "the models roughly agree on average that prepayment is not allowed, with a mean score of ~41."[165]

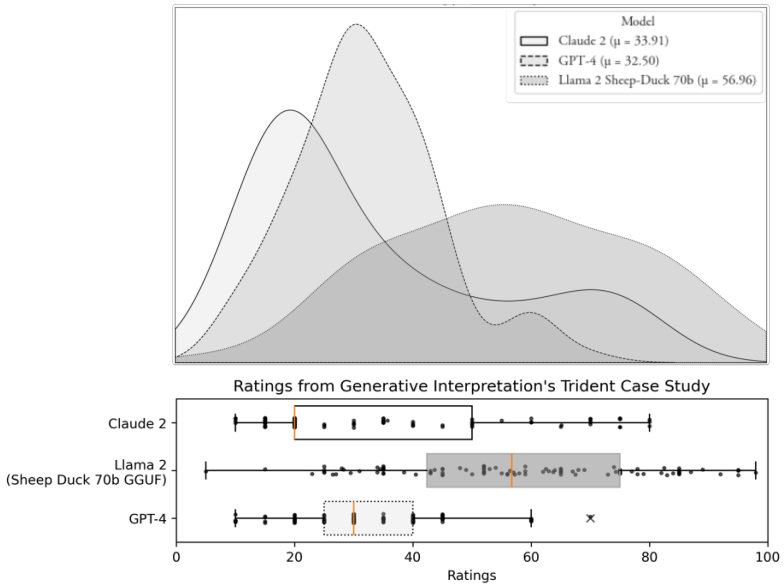


Figure 7: Results from the Generative Interpretation*'s* Trident *case study. (a) Above: the estimated "density" of responses, as presented in the original paper. (b) Below: the same data, displayed as a box-and-whisker chart. We include the second chart to make the data points (black dots) and averages (red lines) explicit.*[166]

Method 4: From "Yes/No" to Numerical Distributions

In a fourth example, Arbel and Hoffman present a contract to several LLMs and solicit yes/no answers about the meaning of a clause. They pose 20 linguistic "variations of the same legal question," a "yes/no question[] where yes indicates agreement with the judge's

[164] 847 F.2d 564 (9th Cir. 1988).

[165] *Generative Interpretation,* at 488.

[166] Id. (Asking models to rate whether prepayment is allowed on a scale of 0-100, where 0 means repayment is definitely not allowed, and 100 means repayment is definitely allowed."). The curves shown in *Generative Interpretation* are computed using a method called kernel density estimation. As a part of that method, the authors needed to choose between one of several "bandwidth estimators," the function used to estimate density based on a collection of datapoints. The different methods result in significantly differences in the resulting chart. *See,* David M. Bashtannyk & Rob J. Hyndman, *Bandwidth Selection for Kernel Conditional Density Estimation*, 36 COMPUTATIONAL STATISTICS & DATA ANALYSIS 279 (2001)(reviewing methods for picking a bandwidth function). *See also* Part **Error! Reference source not found.**, *infra*, discussing minor implementation details that affect interpretive results.

interpretation," and graph the results.[167] Here, the authors are not instructing the model to respond with its confidence, nor are they looking under the hood at the probability of the model picking a certain response. Instead, this number is the relative frequency with which a model provides a "yes" or "no" answer. From a technical perspective, this method is mathematically equivalent to measuring the number from method 2 for each of the 20 questions, then averaging the results and adding a large dose of random noise.[168]

* * *

We do not have a definitive opinion about which of these four methods is best.[169] Nor do we fault Arbel and Hoffman for exploring numerous different approaches. That is exactly what exploratory early work needs to do. Our point is that an embarrassment of riches can still be an embarrassment. Any one generative interpretation methodology raises difficult reliability problems on its own. Adding three more does not resolve those problems; it multiplies them. It may be that one or more of these methods can be made robust and legitimate. But a judge considering generative interpretation must still choose which ones to try, and that by itself is an entire can of worms.

IV. IS GENERATIVE INTERPRETATION EPISTEMICALLY JUSTIFIED?

In this Part, we turn to the epistemic problem: should judges accept LLM outputs as providing authoritative guidance on legal issues? For generative interpretation, those issues involve the linguistic meaning of disputed texts. For generative adjudication, those issues involve the proper resolution of disputed legal issues. But in both cases, the epistemic question is the same. What makes *this* measure relevant to deciding *that* issue? To assert that an LLM output "accurately reflects real people's everyday speech patterns,"[170] without more, is to beg the question. *Why would* the outputs of a complicated algorithm necessarily reflect *anything*, let alone "everyday speech patterns?" Something more is needed to show that *this* algorithm, run in *this* way on *this* input, yields "correct" or "accurate" answers to the question being asked.

---

[167] *Id.* at 490.

[168] The choice between "yes" and "no" is technically equivalent to the choice between "first" or "second" in *Famiglio*: a binary selection between two possible words. We could have looked "under the hood" to get the underlying probabilities for each phrasing. Instead, for each phrasing, we use the result of a single weighted coin toss based on those probabilities. Even taking this number at face value, it's troubling: would we ascribe the same meaning to human interlocutor who answered identical question differently when polled multiple times in quick succession, or would we simply find their response incoherent? Cf. Thomas R. Lee & Jesse Egbert, *Artificial Meaning?*, 28 (2024), https://papers.ssrn.com/abstract=4973483 (last visited Oct 14, 2024) (observing that variation in LLM responses "cannot be equated with the variability we would observe between different people in a population. It should be equated with the variability between responses from a single person who is repeatedly asked the same question.").

[169] Though method 4 seems strictly worse than method 2.

[170] Deleon, 116 F.4th at 1270 (Newsom, J., concurring). To be clear, Newsom goes on to provide reasons, recognizing that this is a proposition that must be demonstrated, not simply assumed. As we detail in this Part, our point of divergence is that we think the reasons he gives do not warrant the conclusion that LLM outputs "accurately reflect[]" ordinary usage.

This is a general challenge for empirical methods; it goes by many names, including "construct validity." A method may be consistent and robust enough to measure *something*, but that *something* may not be the question the experimenter hopes to answer. A moment's thought shows that establishing construct validity requires attending to the details of not just method and the question. The Schönhage-Strassen algorithm multiplies large numbers significantly faster than the algorithm taught in elementary school; it is an excellent source of knowledge about the products of integers, but it tells us nothing about ordinary English usage.

LLM proponents have offered, we think, four broad types of responses to the epistemic problem: they have tried to justify the use of LLMs based on *how they work*, based on *calibrating* their outputs against an accepted baseline, based on the rhetorical *persuasiveness* of their outputs, and based on a claim that they offer *predictability* as such. All four types are intuitively appealing, and three of them can be sound ways of establishing a method's epistemic *bona fides*. But we think that in this case, for LLMs as a source of legal meaning, the case is unproven. LLM proponents' claims depend on a mix of unstated assumptions, mistaken beliefs about how LLMs work, and misunderstandings of the judicial function.

A. Relying on the Training Process?

Start by considering the class of arguments that an LLM should be trusted because of what it is: an statistical model trained on a corpus of natural-language text. The problem here is that these are explanations of why an LLM *could* work at the task of capturing common linguistic usage, not an explanation of why a specific LLM *actually does* capture common linguistic usage in a specific setting.

1. Deductive Arguments

The strongest argument for a decisional legal AI system would be to validate the system *deductively*, using the tools of formal logic to craft an abstract model of law and formal verification to show that the AI system correctly implements that abstract model.[171] Indeed, there have been numerous previous attempts to bring software-based deductive rigor to legal reasoning, with mixed success.[172] To validate generative interpretation deductively would require demonstrating, *a priori*, that the internal workings of the LLMs *necessarily* lead it to estimate linguistic meaning to some degree of accuracy.

Deduction has the virtue of generating conclusions that are guaranteed to be valid in specific instances. Of course, the validity of a deductive proof is as strong (and only as strong) as its assumptions. If those assumptions hold, the conclusion always follows. For example, say we want to know whether a two-sided coin is fair. If we assume that the coin is infinitesimally thin and perfectly symmetrical, that it is flipped with a fixed upward velocity v and an angular velocity drawn from a known distribution X at a fixed distance d above the ground, that it is

[171] *See generally* James Grimmelmann, *Programming Languages and Law: A Research Agenda, in* PROC. 2ND ACM SYMP. ON COMPUTER SCI. & L. 155 (2022) (discussing formal approaches to law).

[172] *E.g.,* L. Thorne McCarty, *Reflections on TAXMAN: An Experiment In Artificial Intelligence And Legal Reasoning*, 90 HARV. L. REV. 837 (1976). Note that McCarty, writing nearly 50 years ago, still correctly identified the ambiguities and open texture of law as key challenges to its formalization.

unaffected by air resistance, and that it collides inelastically with the ground when it lands, then for suitable choices of *v*, *X*, and *d*, we can establish that this method for tossing coins is necessarily fair. The probability that it comes up heads on any given toss is equal to the probability that it comes up tails.

But proving LLMs' legal-interpretive authority isn't like positing the fairness of a coin, for two reasons. First, while "fairness" in coin-flipping has a straightforward definition—an independent, 50-50 expectation of heads or tails for any given flip—"accuracy" in legal interpretation has no equivalent formal specification. The second reason follows from the first. Because legal-interpretive accuracy is underspecified, we don't know what factual premises about LLMs would have to hold in order for them to be accurate. By contrast, because a coin's fairness is easily formalized, we *can* posit the factual properties that characterize a fair coin. Put another way, there are two things that stand in the way of using LLMs to reason deductively about law: the nature of law and the nature of LLMs.

Even assuming a deductive model of legal interpretation is reconcilable with law as a social enterprise, adopting such a model would require transformational changes to present-day attitudes. Indeed, even generative interpretation proponents seem to reject a deductive approach. In *Snell*, Judge Newsom wrote, "would the consideration of LLM outputs in interpreting legal texts inevitably put us on some dystopian path toward 'robo judges' algorithmically resolving human disputes? I don't think so. As Chief Justice Roberts recently observed, the law will always require 'gray area[ ]' decisionmaking that entails the 'application of human judgment.'"[173]

Setting aside whether *law* can ever be a deductive enterprise, *LLMs* are extraordinarily poor vehicles for deductive validation.[174] The origins of LLMs trace back to a decision to dispense with logical rigor and precisely curated knowledge bases, and to rely instead on recognizing patterns in massive datasets. Today's LLMs did not develop because computer scientists deduced that a particular technical architecture would, as a matter of logical necessity, produce AI with remarkable text-generating abilities. Instead, they came about when engineers decided to run with the techniques that produced the best results, even when the reasons these techniques work so well were (and, in some cases, are) yet to be discovered.[175] As Arbel and Hoffman note in *Generative Interpretation:* "whatever [an LLM] tells you, it is really no explanation at all… working with LLMs admittedly requires a leap of faith, a realization that no better explanation is forthcoming."[176]

Kieffaber deals with these problems simply by assuming them away. One of the six premises of his "sci-fi-hypothetical" is that "Judge.AI is a perfectly neutral arbiter and interprets words with *perfect mathematical accuracy*."[177] This assumption serves as an intuition pump for his

---

[173] 102 F.4th at 1232 (Newsom, J., concurring).

[174] *C.f.*, E. Mark Gold, *Language Identification in the Limit*, 10 INFORMATION & CONTROL 447 (1967)(formally proving that it is impossible to learn something perfectly and completely through example-based processes like machine learning process).

[175] *See generally* Alon Halevy, Peter Norvig, and Fernando Pereira, *The Unreasonable Effectiveness of Data*, IEEE INTELLIGENT SYS. Mar.–Apr. 2009, at 8.

[176] *Generative Interpretation*, *supra* note __, at 483.

[177] Kieffaber, *supra* note __, at 7 (emphasis added).

thought experiment about the nature of textualism,[178] but it tells us nothing about actually existing LLMs. Arbitrus.AI is an actual system, and while Kieffaber and his coauthors are enthusiastic about its abilities, they do not attempt to validate it deductively. They don't even detail their model architecture, training algorithm, training datasets, or prompts, let alone show that these particular design choices are logically guaranteed to produce correct outputs. To state the idea is to refute it. LLMs are not spherical cows.[179]

2. Informal Arguments

In practice, most LLM proponents offer a related but much weaker argument. As Unikowsky puts it:

> To me, the proposition "AI is useful for determining the ordinary meaning of English words" should be approximately as controversial as "GPS is useful for determining directions." Cutting-edge LLMs have read the entire Internet, or almost all of it anyway. Of course they'll be useful in determining how words are ordinarily used.[180]

Similarly, Judge Newsom writes, "LLMs can be expected to offer meaningful insight into . . . ordinary meaning . . . because the internet data on which they train contain *so many* uses . . ., from *so many* different sources . . . ."[181]

These are not deductive proofs that the way LLMs are created means they *necessarily* provide accurate information on semantic meaning. Instead, they are arguments that the way LLMs are created gives us *good reason to think* that they provide information on semantic meaning.

Consider Unikowsky's GPS analogy. The GPS system uses a set of satellites with precisely calibrated clocks that transmit synchronized radio signals; a receiver uses the timing information from signals from multiple satellites to calculate its position relative to them—and thus its position on the surface of the Earth. Given the design of the system—and the relevant orbital mechanics, wave physics, electronics, and so on—we have good reason to think that the calculation carried out by a particular receiver in a particular place at a particular moment is accurate. Similarly, given some reasonable assumptions about coins' weight, dimensions, and initial flipping characteristics, we might have good reason to think that particular coin flips are likely to be close to fair.[182]

Once again, however, it is a much heavier lift for generative interpretation than for other methods, because LLMs are much more complicated than a coin or a GPS receiver, and

---

[178] One might object that treating natural language as susceptible to "perfect mathematical" interpretation not only assumes away the debate between textualism and its critics but also distorts beyond recognition the texts that textualists understand themselves to be interpreting.

[179] See David Kaiser, *The Sacred, Spherical Cows of Physics*, NAUTILUS (Apr. 25, 2014), https://nautil.us/the-sacred-spherical-cows-of-physics-234898/.

[180] Unikowsky, *supra* note 5.

[181] *Snell*, 102 F.4th at 1227-28 (Newsom, J., concurring).

[182] *But see* Persi Diaconis, Susan Holmes, and Richard Montgomery, *Dynamical Bias in the Coin Toss*, 49 SIAM REV. 211 (2007) ("We show that vigorously flipped coins tend to come up the same way they started.").

generative interpretation asks LLMs to resolve far more complicated problems. The argument that LLMs work because they have been exposed to almost "the entire Internet" proves far too much. The Google Books corpus was assembled by scanning millions of hard-copy books. The scanners that Google used have "read" millions of books; does it follow that "they'll be useful in determining how words are ordinarily used?" With enough time, you could tap the entire corpus that trained ChatGPT in binary into the keys of a calculator. You could place a tortoise in front of a speaker broadcasting the entire corpus as audio. At the end of these (interminable) processes, the calculator will have "read" the same data that ChatGPT "read," and the tortoise will have apprehended it, at least in the sense that they will have been presented with that information. Yet neither the calculator, nor the tortoise, nor the scanner will be a greater authority on ordinary meaning than it was before these indoctrination processes.

The same point holds regardless of whether the claim is phrased in terms of LLMs' accuracy, their legitimacy, their adherence to democratic norms, or any other desideratum. There is nothing about "training on the entire Internet" that implies the output will satisfy any particular criteria. The calculator and the tortoise were trained on the entire Internet too; that doesn't make them democratically legitimate, either.

Mere exposure to large amounts of natural-language text does not automatically confer authority about linguistic meaning. Indeed, the road to ChatGPT is littered with the corpses of generative AIs that were emphatically terrible authorities on ordinary meaning.[183] It was not that OpenAI's model-training methods were *a priori* better than all that had come before in a way that could have been expected to yield far more reliable results. Instead, a series of advances in model architecture, feasible model size, and large, high-quality training datasets came together to produce a model with demonstrably—but very surprisingly—better *a posteriori* performance.

In other words, "it trained on the entire Internet" is a plausible theory of why an LLM that usually generates fluent text does so, but does nothing to demonstrate that it will. What makes it reasonable to think that an LLM will usually to generate fluent text is *that it has generated fluent text*. The demonstration is fundamentally empirical.

Now, to be sure, the fact that ChatGPT 3.5 generates fluent text also makes it more plausible that ChatGPT 4 will do so too—they are trained in similar ways on similar sources of data. Similarly, ChatGPT's fluency can transfer, in part, to Gemini, to Claude, and so on—it has become widely accepted knowledge that certain types of LLM architectures and training tend to work reasonably well. But notice that while we can appeal to "how Claude was trained" to validate Claude's fluency by drawing on ChatGPT's, this appeal still rests on the empirical demonstration of ChatGPT's fluency. The appeal to how LLMs are trained adds no new and additional justification; it just provides a way of organizing and marshalling the empirical evidence from various LLMs' usage. And so, it is to those empirical justifications that we now turn.

---

[183] *See generally* JANELLE SHANE, YOU LOOK LIKE A THING AND I LOVE YOU: HOW ARTIFICIAL INTELLIGENCE WORKS AND WHY IT'S MAKING THE WORLD A WEIRDER PLACE (2019) (extensively describing AIs' humorously bad attempts to emulate human writing).

B. Relying on Empirical Measurements?

The second way one might validate generative interpretation is *empirically* demonstrating the integrity of a particular methodology for querying LLMs. Instead of deductively or informally arguing *a priori* that LLMs are necessarily or likely to be accurate, this approach demonstrates inductively on the basis of evidence that a particular process for using LLMs produces correct answers consistently, predictably, and robustly. If satisfied, this inquiry justifies trusting LLM outputs produced through that validated process. If an LLM consistently produces accurate information via a specified method in representative testing scenarios, then subsequent uses following the same method in similar scenarios are probably trustworthy—and we may be able to quantify that probability.

Returning to our two-sided coin: if we toss a coin 400 times and it comes up heads 199 times and tails 201 times, we are now 95% confident the coin is fair within a 5% margin of error when tossed in the same way as the experiment. Additional experimentation could increase confidence and reduce the margin of error. Though empirical tests cannot provide the perfect confidence of a deductive proof, they can provide strong, statistical evidence that a certain process has a tested property.

1. Internal Robustness

But what would it actually take to show that a given LLM provides accurate linguistic interpretations? The coin has the advantage that the problem can be specified finitely and precisely: there is a single "flip" process with exactly two possible outcomes, and we seek to learn their respective probabilities. The problem of generative interpretation is significantly more open-ended, and the LLM is a significantly more complicated object of study.

Recall that an LLM is a prediction machine that uses whatever it can find in a textual prompt to make accurate predictions about what comes next. That might be the semantic meaning of the term it has been asked to gloss. Or it might be whether the prompt author asked the LLM to "explore" an issue or to "delve" into it. [184] If authors who eschew contractions are more conservative and sesquipedalian, then a prompt that eschews contractions is more likely to elicit an output that not only eschews contractions, but also reads in a conservative and sesquipedalian register in other ways. While it is likely (though unproven) that insights about general language understanding are tucked away somewhere inside LLM's internal structures, that information is intertwined with myriad other factors that can help the predictive task.[185] Currently, there is no easy way to discern which information an LLM uses to answer questions. It might use the information courts care about. It might use some unrepresentative subset of that information. It might rely on some inscrutable blend of factors that courts would consider

[184] *See, e.g.*,. Tom S. Juzek & Zina B. Ward, *Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models*, 31 ICCL 6397 (2025).

[185] For an accessible discussion exploring what LLM's "know," see Kenneth Li, *Large Language Model: World Models or Surface Statistics?*, THE GRADIENT (Jan. 21, 2023), https://thegradient.pub/othello/, and the underlying paper, Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, & Martin Wattenberg, *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task*, 2022 ICLR (2022).

irrelevant. These are null hypotheses: possibilities that can only be ruled out by conducting tests to rule them out.

In short, we are completely dependent on running actual studies to generate the kind of empirical evidence needed to validate LLMs' use for generative interpretation. The LLM proponents have started on this important task, some more rigorously (Arbel and Hoffman) and some more informally (Judge Newsom, Judge Deahl, and Unikowsky). And this is exactly right—a journey of a thousand miles begins with a single step.

But as we showed in the previous Part, the empirical evidence to date falls well short of showing that generative interpretation really is justified. In particular, we documented that existing methods are not robust. Switching models can flip the direction of a result; small changes to queries produce large variations in the responses. If protocols *A* and *B* to measure phenomenon *X* are identical except in some small respect that has no apparent connection with *X*, but *A* and *B* produce very different results, it counts as evidence that *neither A* nor *B* is actually measuring *X* well. Waldon *et al.*'s "red-teaming" results are similar to ours; they "show that LLMs' metalinguistic judgments are highly sensitive to subtle prompting variations" and that "LLMs can be easily 'gamified' to reflect a user's preconceived biases, even when the lawyer or judge adheres to what appears to be a rigorous and neutral methodology."[186]

This is a problem common to all empirical disciplines, and the obvious way to respond to it is to improve methodological robustness. If *A* and *B* produce highly correlated measurements despite their differences, and so do variations *C*, *D*, *E*, and so on, then we are more justified in believing that all the members of this family really are measuring the same thing. And some of this work is already occurring. From *Snell* to *Deleon*, Judge Newsom went from two queries to each of two models to ten queries to each of three models and more systematically analyzed them. In their work on LLM interpretation, Christoph Engel and Richard McAdams analyzed distributions of responses rather than a model's single "best" response.[187] As the sophistication and scale of generative-interpretation experiments continues to increase, its practitioners may be able to find robust protocols that are not sensitive to small differences in experimental setups.

This is not to say that a protocol must be robust against all possible variations. Some can be excluded for principled reasons. A coin flip gets fairer—in a way that can be empirically measured with strong statistical significance—as the coin is tossed from a greater height. The protocol in which it is tossed from two inches above a table may produce different results than the protocol in which it is tossed from two feet, but that is because the two-inch protocol is demonstrably worse. From political polling to laser interferometry, empiricists in numerous fields have accumulated bodies of standards and best practices for conducting their studies using specific methods in ways that are broadly accepted over their alternatives. Here, too the LLM proponents are tentatively beginning to move beyond simply brainstorming different possible methods to testing them comparatively, and excluding the ones that fail robustness checks. Engel and McAdams, for example, have an extended discussion of different types of

[186] Waldon et al., *supra* note __, at 39.

[187] Christoph Engel and Richard H. McAdams, *Asking ChatGPT for the Ordinary Meaning of Statutory Terms*, 2024 U. ILL. J.L. TECH. & POL'Y 235.

design choices for generative-interpretation studies, including tentative thoughts on how to make many of those choices.[188]

It is worth emphasizing how daunting this task is for generative interpretation. LLMs are sensitive to their configuration and prompts *by design*. Consumer-grade LLMs aim to respond in ways their users prefer, a chatbot that can conform to its user's preferences by picking up on subtle cues in their inputs will provide a better service.[189] Consider *Snell*: if Newsom's clerk had asked, "what's landscaping mean?", he would have seen vastly different results to those he received in response to "What is the ordinary meaning of "landscaping"?"[190] Professional- and business-grade LLMs are used inside tools that need to conform to organizations' preferred tone, values, opinions, and topics. For example, an LLM-powered filing assistant on a court website should never acknowledge potential innuendos or double-entendres, whereas X's Elon Musk has specifically promoted Grok by claiming that it "loves sarcasm" and would answer questions with 'a little humour.'"[191] The demand for configurable general-purpose tools leads to LLM outputs that are highly sensitive to technical configuration.[192] In short, many LLM sensitivities are features, not bugs. The very nature of many LLMs pushes against their ability to be robust against changes to their prompts.

---

[188] Engel and McAdams, *supra* note __, at 271–88

[189] Text generation is tuned to align with human feedback. See Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, & Dylan Hadfield-Menell, *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback [RLHF]*, ARXIV:2307.15217 [CS] (Sep. 2023) (Describing RLHF and enumerating many of its limitations).

[190] Using the GPT-3.5-turbo-instruct model and repeating the questions from *Snell* 1000 times yields the distribution of responses shown in the following table. Newsom's phrasing resulted in GPT answering "yes" twice as often as "no." The more colloquial phrasing of the question results in GPT saying "no" six times more often than "yes."

Text of Q1 & Q2	GPT-3.5's Response to Q2		
	Yes	No	Maybe
Q1: What is the ordinary meaning of "landscaping"? Q2: Is installing an in-ground trampoline "landscaping"?	12.4% (± 2.0)	7.8% (± 1.7)	79.8% (± 2.5)
Q1: whats landscaping mean? Q2: is installing a in-ground trampoline landscaping?	6.6% (± 1.5)	42.4% (± 3.0)	51.0% (± 3.1)

The code used to generate this data is available at https://gist.github.com/davidbstein/1a60483a71fdeb2da91f88efe24a55ad

[191] Lucy Hooker, *Musk Says His New AI Chatbo Has 'A Little Humour'*, BBC NEWS (Nov. 5, 2023), https://www.bbc.com/news/business-67327060 (quoting Elon Musk).

[192] Sensitivities are also great for people using chatbots to answer factual questions, draft documents, brainstorm new ideas, or build bespoke tools. Users can fiddle with configurations and prompts to achieve desired outcomes.

2. External Calibration

Merely passing robustness checks, however, is not sufficient to show that an empirical method is fit for purpose. It may be measuring *something* robustly, but that something might not be the object of interest. Robustness is a necessary condition, not a sufficient one.

In other words, to show empirically that an LLMs truly measures linguistic meaning, its outputs must be compared to *an accepted authority on linguistic meaning*. This is the point of Engel and McAdams's study of ChatGPT as a source of ordinary meaning: they compare Chat 3.5 Turbo's responses (under various prompting conditions) to the results of Kevin Tobia's 2020 survey of 2,800 English speakers about the meaning of statutory terms.[193] Strikingly, three of the four LLM protocols they tested produced results that significantly diverged from Tobia's survey data.

You may or may not agree that surveys are an appropriate measure of what machine-learning scholars would call "ground truth." Our point is that while it need not be surveys, any attempt to calibrate LLMs empirically depends on having some external benchmark to calibrate against. Without one, no amount of experiments can provide a way to differentiate between "informed conclusion[s] based on a statistical analysis of billions of texts,"[194] and wild extrapolation based on extraneous factors.

Consider Unikowsky's experiments with Supreme Court cases. He found, "Of the 37 merits cases decided so far this Term, Claude decided 27 in the same way the Supreme Court did."[195] One take on this result would be that Claude has an accuracy rate of 73%—definitely not good enough for government work. But that is not Unikowsky's response; instead, he says, "I frequently was more persuaded by Claude's analysis than the Supreme Court's." That may well be (and we will discuss the argument from persuasiveness in the next section), but note what it does to the project of validating Claude's accuracy. It means that the Supreme Court's actual decisions are no longer functioning as a source of ground truth. It means that Adam Unikowsky trusts *his own judgment* more than he trusts *the Supreme Court's opinions* as a benchmark of legal analysis. This too may be right,[196] but notice what it does to the project of validating LLMs for judicial use. It means that you too need to trust Adam Unikowsky's legal judgment more than the Supreme Court's. It means we would need a societal consensus that Claude is *better at this than the Supreme Court*. There is no way to bootstrap into such a consensus by counting up how frequently Claude agrees with the Supreme Court. If you think Claude is better at judging than the Supreme Court, you need to argue that Claude is better at judging than the Supreme Court, and back it up with your own standard of what constitutes good judging and evidence that Claude meets that standard.

Other LLM proponents confront this problem, but none of them is able to dispose of it. Newsom relies primarily on intuition, finding that LLMs' responses "squared with what I had

---

[193] Engel and McAdams, *supra* note __, at 256–70; Kevin P. Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726 (2020).

[194] *Generative Interpretation*, *supra* note __, at 458.

[195] *In AI We Trust, Part II,* supra note __.

[196] As readers of his newsletter, we certainly agree that his arguments are often more convincing than those of the judges and lawyers he critiques!

assumed,"[197] and "squared with my own impression —informed by my own experience writing, reading, speaking, and listening to American English in the real world."[198] What, one wonders, would Newsom have done with their responses if they *didn't* square with his own interpretations? For its part, *Generative Interpretation* defines "accuracy" as "thinking that we really got as close as we could to knowing what the parties would have said." Arbel and Hoffman acknowledge that "there is no ground truth at hand—we can't really know what the parties intended at contracting and have to make instead our best guess."[199]

The trouble is that this mode of analysis invites circular reasoning. Arbel and Hoffman define accuracy in terms of "best guess[es]" while simultaneously implying that LLM-produced guesses constitute our "best guess[es]." But if LLM-produced guesses are *definitionally* our "best guess[es]," and thus the benchmark for any measurement of accuracy, then *Generative Interpretation*'s case studies do not measure accuracy. Instead, they project authoritative meaning onto LLM outputs, irrespective of what those outputs signify.

The diversity of the techniques employed for generative interpretation is a warning of the breadth of empirical validation that generative interpretation will demand. Every distinct technique for querying an LLM will require its own distinct, inductive validation. Empirical support that an LLM's word-likelihood estimates reflect linguistic meaning does not necessarily establish that LLMs' probability estimates reflect linguistic meaning. Even if generative-interpretation proponents can empirically validate *one* method of querying LLMs—and, we emphasize, they have not yet done so, although early research gives some reason for optimism[200]—that validation would not establish the accuracy of *other* methods of querying an LLM. Evidence that some LLM output is accurate does not entail that output is accurate because it comes from an LLM.

3. Protocol Specification

Another way of phrasing the difficulty with answering the epistemic challenge empirically is as a problem of specification. The more sensitive generative interpretation is to implementation details, the more comprehensive *ex ante* commitments need to be to achieve predictable or consistent results. Posing a question to an LLM is a wildly underspecified task. As we have seen, outputs can depend on subtle differences in question formulation, model choice, interpretation method, and the configuration and implementation of any software used to operate the LLM.

Proponents of generative interpretation imagine parties agreeing on specific models and prompting methods, then resolving disputes by presenting them to an LLM using those agreed-upon methods.[201] At least for the case studies explored here, every phrasing and formatting choice seems to influence outcomes. Constructing a sufficiently detailed pre-commitment for predictable generative interpretation requires agreeing on outcome-altering details. For all but

---

[197] *Deleon*, *supra* note __, at 1272 (Newsom, J., concurring).

[198] *Snell*, *supra* note__, at 1225 (Newsom, J., concurring).

[199] *Generative Interpretation*, *supra* note __, at 459-60, 462.

[200] Engel and McAdams, *supra* note __.

[201] *E.g.*, *Generative Interpretation, supra* note __, at 501.

the easiest of cases, many of those details are dispute specific.[202] Why not negotiate the anticipated disputes directly? Put another way, picking one experimental setup over another merely displaces the arbitrariness from the empirical realm (how to reconcile the different results of protocol *A* and protocol *B*?) to the epistemic one (what makes protocol A better or worse than protocol *B*?)

Again, to repeat, our claim is not that this *cannot* be done in a principled way, just that it *has not yet been* done in a principled way, and that it *will have to be done* in a principled way to justify generative interpretation empirically. The empirical woods are lonely, dark, and deep—and generative interpretation has miles to go.

4. No Easy Fixes

Rather than waiting for more robust validation, it is tempting to play whack-a-mole with problems like the ones we spotted in Part III. When an LLM generates text that influences its later answers—like in *Snell* and *Famiglio*—some might naturally respond by constraining the model to one-word responses.[203] But that merely shoves the improvident salience into the prompt,[204] and some research suggests that LLMs produce lower-quality answers when forced to be succinct.[205] So maybe we cross-reference between models instead.[206] But model quality can vary; models can (and do) converge on arbitrary answers.[207] So maybe we generate multiple versions of the same prompt.[208] But sampling across low-quality prompts can nudge models towards the same irrelevant details.[209] And so on. Each problem exposes or introduces another, with some solutions possibly decreasing answer quality. Without a reliable way to measure performance, we cannot know which changes improve generative interpretation and which choices exacerbate or obscure problems.

It's also tempting to over-extrapolate from simplified explanations of complex systems. It may feel natural to assume that because LLMs are trained on a large amount of text, their

---

[202] *But see id.* at 510 ("If courts follow our proposed best practices, this method is also predictable *ex ante*").

[203] See *Generative Interpretation*, n. 164 (proposing this approach).

[204] Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, & Mikhail Yurochkin, *Efficient Multi-Prompt Evaluation of LLMs*, arXiv:2405.17202 [cs, stat] (Jun. 2024) (collecting sources).

[205] Jason Wei, et. al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, 36 NEURIPS (2022)

[206] See *Generative Interpretation*, at 503 (advising use of this approach as a "best practice").

[207] See n. 133, *infra*, and accompanying text.

[208] *Generative Interpretation*, at 490 ("tr[ying] something new" by generating multiple prompts). *See* also Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, & Pontus Stenetorp, *Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity*, 60th Proc. ACL 8086 (2022) (generating multiple prompts automatically when querying a model); Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, & Graham Neubig, *Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, 55 ACM COMPUT. SURV. 1 (Sep. 2023) (collecting and reviewing literature on "prompt paraphrasing" and "prompt generation").

[209] Albert Webson & Ellie Pavlick, *Do Prompt-Based Models Really Understand the Meaning of Their Prompts?*, 2022 ACL: HLT 2300 (2022).

answers are informed by that text's semantic content. But that inference is unsupported; semantic information is one (potentially inconsequential) part of a more complex whole. If the argument that LLMs are trustworthy because of how they work does not go through on its own (as we argued in section IV.A), it also does not establish that LLMs are empirically trustworthy (as we discuss in this section). That claim must be made, and supported, on its own merits. An LLM might use great data in an unhelpful way. The brittleness and arbitrariness of early experiments in generative interpretation should caution us that the method is not yet "good enough" for use in litigation.[210] Indeed, numerous empirical studies have documented that LLMs consistently make basic errors in legal reasoning.[211]

C. Relying on Persuasiveness?

The third major way that generative-interpretation proponents attempt to justify their methodology is by arguing that the results literally speak for themselves: that LLM-generated text is sufficiently *persuasive* that it should be treated as authoritative. This type of argument is particularly slippery, because it comes in two closely related forms that make very different assumptions. One is that the persuasiveness of LLM-generated text is *evidence* that it resulted from a reliable process of linguistic interpretation or legal reasoning. The other is that persuasive LLM-generated text is authoritative purely *because it is persuasive*, and that it is irrelevant how it was generated. To clarify the difference, and to understand why both forms of argument are deceptive, it is necessary to speak first of abduction.

1. Generative Interpretation as Abduction

Consider the claim, "This text looks like it was the result of a skilled process of legal reasoning, so therefore it was produced by one." The name for this form of reasoning is *abduction*, or sometimes "inference to the best explanation."[212] Unlike *deduction*, which reasons from premises to their logical entailments, or *induction,* which reasons to find generalizations consistent with existing data, abduction reasons by identifying plausible causes for observed phenomena. "All the balls in that urn are black; these balls came from that urn; therefore these balls are black" is deduction; "These balls are black; these balls came from that urn; therefore all balls in that urn are black" is induction; "These balls are black; all the balls in that urn are black; therefore these balls came from that urn" is abduction.

Though potentially less robust than inductive or deductive proof, abductive reasoning is often the only option in the law. For example, the preponderance of the evidence test asks for the most reasonable or coherent interpretation of the evidence. The legitimacy of legal dispositions is also something we may infer abductively. As later sub-Parts explore more deeply, a judge's decision is assumed to be the product of a good-faith application of legal reasoning

---

[210] *Contra Generative Interpretation*, at 458.

[211] *E.g.*, Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,, 22* J. Empirical Legal Stud. 216 (2025).

[212] *See generally* Igor Douven, *Abduction*, STAN. ENCYC. PHIL. (2021), https://plato.stanford.edu/entries/abduction/.

because her written decision describes that reasoning process. When a judicial opinion sets forth a persuasive chain of legal reasoning that accurately refers to relevant law, we assume abductively that it was in fact realized through the analytical process it describes, and hence that the opinion is legitimate.

Abduction is not in itself a wrongheaded analytical device. We rely on abduction all the time. When a student writes an exam that makes cogent, well-informed arguments about the assigned materials, we treat that document as evidence that the student has mastered the topic. We presume that producing such an exam response requires a sophisticated understanding of the course materials, and we regard the exam as proof that its author has such an understanding. Of course, our abductive inference might be wrong. Maybe the exam response was produced by a cat walking across a keyboard, or a fountain pen leaking into a bluebook in exactly the pattern of handwritten text. But, based on our experience, we conclude that such situations are implausible enough that we award the student a good grade unless we have evidence to believe that her exam response was not her own composition.

Abduction, however, is limited by what we know and by the accuracy of what we think we know. In 2025, for example, it is probably educational malpractice for an instructor to be unaware that students have access to LLMs, because the wide availability of LLMs weakens the abductive inferences that educators customarily make after inspecting students' work. LLMs allow students to generate exactly the sorts of prose that, previously, would have required comprehension of the course materials. A cogent term paper no longer signifies what it signified in the pre-AI age. To ensure that written documents continue to serve as rough measurements of students' knowledge, instructors must redesign examinations to ensure that they measure what they purport to measure. At least two of us, for example, have forgone takehome examinations in favor of proctored, in-class assessments, to ensure that students' written work supports the same inferences about their mastery of course materials as it did before the advent of LLMs.

Just as they complicate the inferences that a teacher can make from a well-written essay, LLMs complicate the inferences that a jurist can make from a well-written legal argument. It used to be fair to infer that someone who authored a cogent legal opinion had considered the relevant issues. But as the processes for generating legal writing change, so do the permissible inferences that a legal work product can engender.

Abduction clarifies the nature of the two kinds of arguments from persuasiveness. A claim that LLM-generated text is authoritative because its persuasiveness reassures us that it was generated through a valid process of legal reasoning is an abductive argument. It uses the superficial qualities of LLM-generated text to make a claim about the legitimacy of the underlying reasoning. That claim, however, is defeasible, and can be rebutted by a showing that there are good reasons to doubt the *bona fides* of the process that generated the text. A claim that LLM-generated text is authoritative simply because it is persuasive is not abductive. It avoids the rebuttal by treating the generation process as irrelevant. But, as we will see, it is a claim that is dramatically at odds with the theory of adjudication on which our legal system currently rests.

2. Rhetoric ≠ Adjudication

LLMs are rhetoric machines. Rhetoric is the art of persuasive argumentation,[213] and it is unsurprising that LLMs trained on an enormous corpus of humans doing their best to make persuasive arguments are capable of replicating the textual features of those arguments that make them persuasive. Indeed, the typical training process (particularly the stages that involve human feedback) selects for persuasiveness; it amplifies the connections in an LLM's internal model of language that contribute to fluent, persuasive outputs.

Rhetoric is successful precisely to the extent it is persuasive.[214] As anyone who reads their output can tell, LLMs excel at rhetorical writing. Today's LLMs reproduce the *form* of high-quality legal reasoning. As Unikowsky and others have observed, LLM output now replicates rational, eloquent argumentation that applies precedent to novel facts. An LLM can produce text that may be formally indistinguishable from—or even formally superior to—the reasoning described by an opinion written by a human judge.

There are many ways to elicit responses from LLMs that sound plausibly correct. But as anyone who has dealt with lawyers well knows, *sounding* correct is not necessarily the same as *being* correct. Indeed, there is ample evidence from the use of LLMs in law to illustrate the difference: ask any lawyer who has been sanctioned for submitting a brief containing LLM-generated "hallucinations" whether the fluency of LLM-generated text guarantees that the legal claims it makes are sound.[215] Though tempting, conflating persuasiveness and coherence with authority or accuracy is a category error.

The problem with rhetoric is that sometimes an argument can be persuasive for the wrong reasons.[216] Rhetoric encompasses arguments that comply with all the formal conventions of good-faith legal reasoning. It also, however, encompasses writing that is persuasive for other reasons, like well-concealed gaps in logic and appeals to emotion. Rhetoric covers everything from "legal reasoning" to "casuistry," and the rhetorician's good or bad faith is irrelevant. LLMs

---

[213] *See, e.g.*, SHARON CROWLEY & DEBRA HAWHEE, ANCIENT RHETORICS FOR CONTEMPORARY STUDENTS (5th ed. 2011).

[214] Note that we are not using "rhetoric" to refer to the concept of "constructive rhetoric," as it sometimes does in writing on legal philosophy. See, e.g., James Boyd White, *Law as Rhetoric, Rhetoric as Law: The Arts of Cultural and Communal Life*, 52 U. Chi. L. Rev. 684, at 701 (1985) (using a "highly expanded" definition for "rhetoric"). That definition is useful in interrogating the cultural and social relationship between rhetorical legal artifacts—like statutes—and their authors and audience, but assumes a human author and is therefore ill suited for this context. *Id.*, at 689-690 (defining constructive rhetoric in reference to the lawyers creating it).

[215] *See, e.g.*, Mata v. Avianca, Inc., 678 F. Supp. 3d 443, 448-49 (S.D.N.Y. 2023) (imposing sanctions on attorneys who "submitted non-existent judicial opinions with fake quotes and citations created by the artificial intelligence tool ChatGPT").

[216] Consider, by way of contrast, the use of AI for victim impact statements. *See, e.g.*, Juliana Kim, *Family Shows AI Video of Slain Victim as an Impact Statement—Possibly a Legal First,* NPR (May 12, 2025), https://www.npr.org/2025/05/07/g-s1-64640/ai-impact-statement-murder-victim.Even here, in a domain where *pathos* is as valued as *logos*, such uses raise the fear that AI will be persuasive for the wrong reasons.

make newly salient the ancient anxiety about rhetoric itself[217]—the danger that rhetorical skill can persuade the audience, regardless of whether the argument itself is right or wrong.

By contrast, we expect more from adjudication. To be sure, a good adjudicator will present her decisions in a rhetorically effective format. But adjudications are not successful simply because they reach persuasive conclusions. Rather, adjudication requires the integrity of the process of legal reasoning that realized that conclusion.[218]

That is, adjudication is also subject to *procedural* criteria. At minimum, these criteria require that a decisionmaker make a good-faith, rational effort to produce a legal conclusion. If a formally adequate legal determination is produced in bad faith, or without rational effort, it does not adhere to the process of legal reasoning. A decisionmaker's good faith establishes that she has not elaborated a formally adequate conclusion for improper, extralegal reasons. For example, a conflict of interest will taint a decisionmaker's conclusion of law. Even if the conclusion is set forth in a formally adequate decision, it is not a legitimate legal conclusion, because an observer reasonably can suspect that the judge's personal interests influenced her determination of the appropriate outcome.

Procedural constraints also require a decisionmaker's conclusion to be at least minimally rational. Paul the Octopus was a cephalopod whose handlers would serve him food in boxes that bore the logos of competing soccer teams.[219] When Paul chose to eat from one box before another, his handlers deemed him to have "predicted" that the corresponding team would win.[220] When this predictive method accurately foretold the winners of a number of international soccer matches, Paul became world famous.[221] If a judge issued a formally adequate legal opinion, and it were later revealed that she had chosen the winning party because her pet octopus ate from a box labeled with that party's name, this procedural deficiency would undermine the decision's status as legal reasoning.

A formally adequate legal conclusion is legal reasoning's end product, but is not in itself proof that legal reasoning has taken place. This is because a formally adequate legal work product can be realized through processes that do not constitute legal reasoning. In some intellectual traditions, the integrity of the form *is* the integrity of the process. If, for example, a mathematician authored a formally accurate proof and later claimed that it came to him as a divine message, the proof would still be mathematically legitimate.[222] If, on the other hand, a judge issued a formally adequate opinion and later revealed that she had reached the conclusion

---

[217] *See, e.g.*, PLATO, GORGIAS (Donald. J. Zeyl, trans. 1987).

[218] *See generally* Chad M. Oldfather, *Writing, Cognition, and the Nature of the Judicial Function*, 96 Geo. L.J. 1283, 1318 (2008) (discussing the widely-shared belief that judicial " writing provides an important discipline on thought").

[219] Paul the Octopus, WIKIPEDIA (2024), https://en.wikipedia.org/wiki/Paul_the_Octopus (last visited Oct 16, 2024).

[220] *Id.*

[221] *Id.*

[222] The esteemed mathematician Srinivasa Ramanujan claimed scholarly inspiration from a family goddess, whose consort visited him in a dream. Srinivasa Ramanujan: The mathematical genius who credited his 3900 formulae to visions from Goddess Mahalakshmi, INDIA TODAY (2017), https://www.indiatoday.in/education-today/gk-current-affairs/story/srinivasa-ramanujan-life-story-973662-2017-04-26 (last visited Oct 18, 2024).

that a divine message had commanded her to adopt, such a revelation *would* undermine her claim to have engaged in legal reasoning.[223] In an essay written before her judicial appointment, Justice Amy Coney Barrett concludes that Catholic doctrine forbids observant judges from enforcing the death penalty, and that a moral refusal to consider a death sentence warrants recusal from the sentencing phase of a capital case.[224] Even if the judge can justify her conclusion in a formally adequate legal document, her recusal is warranted because she cannot reach that conclusion through a process recognized as legal reasoning.

The form and process of legal reasoning are socially contingent. Even a single legal tradition admits variations: modern American jurists disagree about the precise form that legal reasoning may take. Some, for example, "object to the use of legislative history on principle,"[225] while others encourage it.[226] A legislative-history skeptic may find an opinion by a legislative-history enthusiast formally inadequate on the ground that it employs a categorically impermissible mode of argument. The legislative-history skeptic may find this legal reasoning just as illegitimate as legal reasoning that invokes the judge's dreams or her subjective "value preferences" to justify a conclusion.[227]

But granular disagreements about particular formal qualities underscore broader agreement about the general form that legal argumentation should take. Some arguments for constitutional originalism, for example, posit that the methodology is, in consequentialist terms, the best approach to constitutional interpretation.[228] Consequentialist arguments for originalism do not assert that originalism is the only true form of legal reasoning.[229] Rather, these arguments suggest a broader view of legal argumentation—as, say, consequentialist arguments rationally propounded—and employ that argumentative paradigm to advocate for originalism.

Similarly, that a document formally resembles legal reasoning does not make it *adjudication*, nor does this resemblance necessarily make the document authoritative. Whether something constitutes adjudication, and whether it is authoritative, depends on social facts like

---

[223] Cf. Lon L. Fuller, *The Forms and Limits of Adjudication*, 92 HARV. L. REV. 353, 391 (1978).

[224] John H. Garvey & Amy V. Coney, *Catholic Judges in Capital Cases*, 81 MARQ. L. REV. 303, 305, 335, 339 (1997).

[225] Antonin Scalia, *Common-Law Courts in a Civil-Law System: The Role of United States Federal Courts in Interpreting the Constitution and Laws*, *in* A MATTER OF INTERPRETATION 3, 106 (Amy Gutmann ed., 2018), https://www.degruyter.com/document/doi/10.1515/9781400882953-004/html (last visited Oct 18, 2024).

[226] See generally Stephen Breyer, *On the Uses of Legislative History in Interpreting Statutes The 1991 Justice Lester W. Roth Lecture*, 65 S. CAL. L. REV. 845 (1991).

[227] Cf. Robert H. Bork, *Neutral Principles and Some First Amendment Problems*, 47 IND. L.J. 1, 9 (1971).

[228] See Cass R. Sunstein, *Originalism*, 93 NOTRE DAME L. REV. 1671, 1680 (2017); William Baude, *Is Originalism Our Law*, 115 COLUM. L. REV. i, 2351 (2015).

[229] Of course, some proponents of constitutional originalism do posit that it is the *only* legitimate method of constitutional interpretation, and this view comes closer to asserting that non-originalist reasoning is in fact not legal reasoning. Scalia & Garner, "Reading Law" 89 ("Originalism is the only objective standard of interpretation.").

Hart's rule of recognition.[230] The authority of an adjudication depends not just on formal resemblance to a reasoned judicial opinion, but on a set of social practices that imbue *particular* artifacts of legal reasoning with authority. Although the formal adequacy of written legal reasoning produced by humans may have been a decent proxy for presuming that it was realized in accordance with a *bona fide* adjudicative process, the same can't be said for formally adequate legal reasoning generated using an LLM. To presume that these AI-generated simulacra of rationality constitute the social process of adjudication is to conflate form and process.

It is almost trivially easy to make an LLM produce any kind of output desired (e.g., "Write a textualist opinion finding for the defendant while making as many references as possible to the films of Steven Spielberg"). Without some other kind of reassurance that the results reflect the kind of reasoning we associate with competent and good-faith judging—the kind that could be provided in theory by inductive or deductive approaches—saying that LLM outputs are valid interpretations because they look good boils down to a vibe-check. Ironically, it is precisely the sort of "visceral, gut-instinct decisionmaking" that proponents of generative interpretation hope to avoid.[231]

Ideally, some sort of assessment could separate the know-nothing scanners and tortoises from authoritative intelligences—some attribute of LLMs that, if observed, would support the conclusion that its interpretations are accurate and meaningful. Unikowsky acknowledges that before delegating legal decisionmaking to LLMs, we want to "mak[e] sure the AI works."[232] "[T]his," he says, "shouldn't be that hard to do;" validating AI's decisionmaking would require only "download[ing] a few hundred briefs off of PACER, input[ting] them into your AI of choice, ask[ing] the AI to adjudicate the case, and compar[ing] the results to the actual judicial decisions that came down in those cases."[233]

But does Unikowsky's proposed test really measure *accuracy*? Unikowsky's writings on LLMs conflate at least three different desiderata. First is LLMs' ability to get the "right answer"—that is, to reach the same disposition as a court on particular briefing. Second is LLMs' ability to *sound* reasonable. Third is LLMs' ability to *be* reasonable—that is, to "reason." These are distinct capabilities. Possessing either of the first two does not entail possessing any of the others. For example, "getting the right answer" alone may not prove much: we wouldn't delegate adjudication to a dog, even if it demonstrated a robust and uncanny ability to "predict" case outcomes in conformity with litigated outcomes by choosing to eat from a particular labeled bowl. Simply *sounding* reasonable isn't sufficient, either. An LLM, like a skilled lawyer, can produce reasonable-sounding justifications for incompatible legal outcomes.

Generating a reasonable-sounding justification does not in itself demonstrate the reasoning we expect from an adjudicator. Imagine that an adjudicator wrote well-crafted, reasonable-sounding opinions, but turned out to be picking outcomes based on which party's surname

---

[230] Scott J. Shapiro, *What Is the Rule of Recognition (and Does It Exist)?*, 4 (2009), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1304645 (last visited Oct 20, 2024).

[231] *Snell*, 102 F.4th at 1224 (Newsom, J., concurring). *Genertive Interpretation*, at 459 (describing how their case studies applying generative interpretation "illuminate how transparent and objective interpretative methodologies have advantages over intuitive ones")

[232] Unikowsky, *supra* note 5.

[233] *Id.*

came first in the alphabet. No matter whether that adjudicator was a human or an LLM, no matter how reasonable and persuasive its opinions *sounded*, and no matter if it correlated with historical data, we suspect most people would not consider that process "reasoning," nor accept its dispositions as legitimate.

3. Why Text Matters—And Doesn't

It is important to understand the work that *text* in particular does in legal reasoning—and why LLM-generated text does not automatically do that work. The rule of law is not merely a collection of written artifacts; rather, it is the social process that produces those artifacts. In the contemporary United States, those artifacts happen to look like legal opinions. The texts are the visible traces of the system in action, but they are not the system itself.

It's unsurprising that a lawyer or a judge might regard written legal reasoning as the paramount artifact of the American legal system. For centuries, the legal profession has produced written judicial opinions. These authoritative opinions form part of a process that constitutes the rule of law. Lawyers, laypersons, and members of other branches of government could read these decisions and coordinate future action based on the reasoning they set forth. Their holdings bind private citizens and government officials alike. Now, LLMs can produce artifacts that are formally indistinguishable from the artifacts produced by courts that apply the law authoritatively.

Notwithstanding the pride of place it enjoys today, written adjudication is not essential to an authoritative legal process. Juries, for example, are *never* called upon to explain their findings rationally, even though they serve a vital legitimating function in the criminal law.[234] A typical verdict has the barest formal qualities: it is an unelaborated yes or no. It is not the form of a verdict that primarily inspires faith in the jury. Rather, it is our understanding of the jury's deliberative process that legitimates its authority.

Comparing juries to judges illustrates that law is process in addition to formal output. Unlike the work product of, say, an appellate court, the output that a jury produces is trivial. Providing a verdict is, of course, a fundamental part of the jury's job. But the point of the jury system isn't just to manufacture yes/no verdicts. Human juries are resource-intensive; rendering verdicts by a coin flip or a random binary generator would be far more economical, offering great efficiencies and improving access to the courts. Yet we routinely reject innovations that would make verdicts cheaper and more efficient to produce.[235] We do so to

---

[234] *See, e.g.*, Apprendi v. New Jersey, 530 U.S. 466, 497 (2000) (referring to "the jury tradition that is an indispensable part of our criminal justice system"); *see also id.* at 477 (quoting Justice Story and William Blackstone to corroborate the importance of "trial by jury" "to guard against a spirit of oppression and tyranny").

[235] *See, e.g.*, Granfinanciera, S.A. v. Nordberg, 492 U.S. 33, 63 (1989) ("It may be that providing jury trials in some fraudulent conveyance actions . . . would impede swift resolution of bankruptcy proceedings . . . . But 'these considerations are insufficient to overcome the clear command of the Seventh Amendment.'"); *Apprendi*, 530 U.S. at 498 (Scalia, J., concurring) ("[T]he jury-trial guarantee . . . . has never been efficient; but it has always been free."); Parklane Hosiery Co. v. Shore, 439 U.S. 322, 343 (1979) (Rehnquist, J., dissenting) ("[T]he concerns for the institution of jury trial that led to the passages of the

protect the integrity of the process that realizes the verdict. In other words, there is very little about the *form* of a verdict that makes it authoritative and legal. What gives a verdict its legal authority is almost entirely our beliefs about a jury's *process*.

LLM proponents assert that AI is authoritative not simply because of how it *sounds*, but because of how it *works*. This reasoning suggests that we ought to trust the conclusions that AI reaches even when they are unaccompanied by output that resembles legal reasoning.[236] And indeed, this is precisely what practitioners of generative interpretation do when they find that individual words and numbers generated by LLMs are probative to their understanding of a legal instrument or a particular term (as Arbel and Hoffman do in their *Famiglio* and "flood" case studies discussed above, for example).[237] If our faith in AI should persist even when it does not present its conclusions in rational-sounding language, then shouldn't that faith justify replacing not just lawyers and judges with AI, but juries, too?

Many readers may find the prospect of replacing juries with AI unpalatable, particularly in criminal cases. Many readers probably suspect that the public will not regard AI juries as having the same sort of legitimacy and authority as a human jury. The intuition that the public will reject their authority is a very good reason to resist AI juries. It also underscores that legality is a social fact; it does not follow inexorably from form. Just because AI can render something formally indistinguishable from a verdict does not mean that it can render a legally acceptable verdict. Correspondingly, just because LLMs can output text that resembles well-reasoned adjudication does not mean that it can engage in the social process that constitutes acceptable adjudication.

To put the point another way, juries aren't important because they render binary verdicts; countless mechanisms can output a "1" or a "0." They are important because they instantiate a legitimate and authoritative process for reaching a binary verdict. If we cared only about the formal artifacts of juries—that is, only the verdicts they produce—then we could delegate jury duty to a coin flip. The thought of deciding criminal culpability by coin flip appalls us because we value juries for the process they instantiate, not simply the artifacts they produce. *The same is true of adjudication*—but focusing on the formal qualities of an opinion's text obscures this point. The temptation is to treat formally adequate legal text as evidence of the process we call "legal reasoning" or "adjudication." But formally adequate legal text is not necessarily adjudication—just like a bare yes or no *resembles* a verdict but is not necessarily one.

Declaration of Independence and to the Seventh Amendment were not animated by a belief that use of juries would lead to more efficient judicial administration.").

[236] Note that when an LLM produces an output that *describes* a reasoning process, this description does not necessarily correspond to the process that the LLM *actually followed* to generate that output. *See, e.g.*, Yanda Chen et al., *Reasoning Models Don't Always Say What They Think* (May 8, 2025), https://arxiv.org/abs/2505.05410 (finding that LLMs prompted to explain their reasoning via a "chain of thought" frequently drew conclusions based on "hints" that they did not disclose in their supporting reasoning).

[237] Arbel and Hoffman, *supra* note 2, at 483-85.

Consider one of Unikowsky's proposals. He observes that submitting cases nationwide to a single AI adjudicator could reduce the unpredictability that arises in a judicial system that vests power in thousands of individual judges.[238] However, he writes,

> intellectual diversity across the judiciary is healthy, at least up to a point, and . . . it also seems bad to concentrate all judicial power in one (or a small number of) computer systems. But this problem can be dealt with by, e.g., adding a random ideological factor to the AI's outputs, having the AI highlight when a particular dispute would come out differently across different philosophies, using multiple independent AIs, or through other mechanisms. The point is, if we use AI, we can add exactly as much unpredictability and ideology as we want via effective prompt engineering rather than having unpredictability thrust upon us by the constraint of individual judges' bandwidth. [239]

Why it is that adding a "random ideological factor"—a fancy version of a coin flip—would make AI adjudication more, rather than less, legitimate? Or rather, why does it seem to Unikowsky that taking a system that is *arguably* inferior to human judging (an AI adjudicator) and combining it with a system that is *definitely* inferior to human judging (random chance) will fix its deficiencies?

The rejoinder is that *outputs* do not have an "ideological factor;" *judges* do. Intellectual diversity in the judiciary is not desirable because it guarantees that judges will resolve cases differently. (It guarantees no such thing.) What intellectual diversity does guarantee is that judges will *approach* cases differently. And it is a diversity of approaches that is valuable, not a diversity of dispositions for its own sake.[240] In fact, it tends to inspire confidence in the judiciary when judges of different ideological persuasions agree on the disposition of a case.[241] Unikowsky's proposal to inject arbitrary amounts of "unpredictability and ideology" misses what is important to the process of adjudication.

Unikowsky's other suggestions come closer. "[H]aving the AI highlight when a particular dispute would come out differently across different philosophies" explicitly puts the diversity of judicial approaches into a single LLM's considerations; "using multiple independent AIs" attempts to have a diversity of LLMs emulate a diversity of judges. Explicitly spelling out these considerations, however, raises exactly the same questions we have been discussing—what kinds of validation would it take to make us confident that the *process* of LLM adjudication sufficiently captures the *process* values of human adjudication?

So far, this discussion has centered on the artifacts produced by legal reasoning. But generative-interpretation moderates do not propose replacing judges entirely. Instead, they suggest merely using LLMs as an alternative to common sense or a dictionary.

---

[238] Unikowsky, *supra* note 5.

[239] *Id.*

[240] *E.g., Accord*, David Orentlicher, *Politics and the Supreme Court: The Need for Ideological Balance*, 79 Pitt. L. Rev. 411 (2018); Deseriee Kennedy, *Judicial Review and Diversity*, 71 Tenn. L. Rev. 287 (2004).

[241] *See, e.g.*, Devin Dwyer, *Supreme Court Defies Critics with Wave of Unanimous Decisions*, ABC NEWS, Jun. 29, 2021, https://abcnews.go.com/Politics/supreme-court-defies-critics-wave-unanimous-decisions/story?id=78463255 (last visited Jan. 30, 2025) (quoting Kate Shaw: "I suspect the justices feel there is value in conveying to the American people that in a hyper-partisan moment, . . . the court remains a largely nonpartisan institution.").

There is a procedural oddity to judicial use of dictionaries: looking something up in a dictionary is in some respects a factual inquiry, yet one that judges perform *sua sponte*, outside of the ordinary adversarial process. The legitimacy of judicial use of external sources of fact during the interpretive process relies *even more heavily* on social norms and the judge's legal-reasoning process than does the legitimacy of using facts that have passed through the gauntlet of a trial. Judge Newsom complains that judges "rarely explain in any detail the process by which they selected one definition over others." Yet even without explanation, *the selection was a product of the judge's reasoning*.

The counterintuitive upshot is that interpretation is one of the *last* places where LLM use is justified within the adjudicatory process. When juries produce a "yes" or "no" response, they provide no explanation. The fact that the jury produced the answer is the source of its legitimacy. In short, when explanations run thin and discretion peaks, procedural and institutional sources of legitimacy acquire special import. Textual interpretation is an adjudicative task where judges' social role has special importance. LLM proponents suggest that computer programs might provide objective answers to subjective questions. Unilaterally abdicating the interpretive role in favor of an LLM (especially without substantial supporting evidence for those claims) actively subverts an important judicial responsibility.

4. The Risk of Projection

We make these points at length because of the danger that LLM users might unreflectingly project authority on to LLM outputs simply because the outputs sound right. Arbel and Hoffman mention the concern explicitly.[242] And in both *Snell* and *Deleon*, Judge Newsom observed, "I definitely didn't want to fall into the trap of embracing ChatGPT's definition just because it aligned with my priors."[243] Yet Judge Newsom wrote that his fears were "reassur[ed]" after he employed Arbel and Hoffman's proposed technique for measuring LLMs' "confidence" and found that he could produce similar answers by posing the same interpretive question multiple times and to multiple LLMs.[244]

Judge Newsom is right to identify the risk of projection, but the measures he adopts do not guard against it. His confidence checks do not demonstrate that he is not vesting LLMs with authority simply because their output matches his expectations. Rather, they demonstrate only that he is more comfortable vesting LLMs with authority when he thinks they are confidently and consistently producing output that matches his expectations. Indeed, he said as much: "If the model . . . returned essentially consistent answers—responses that coalesced around a common core—then one could probably say, with a higher degree of confidence, that the model was getting at the 'correct' response."[245] This framing assumes that confidence and consistency

---

[242] *Generative Interpretation*, *supra* note __, at 476 ("[J]ust because the probabilities are reasonable doesn't mean they are accurate.").

[243] *Snell*, 102 F.4th at 1225 (Newsom, J., concurring); *Deleon*, __ F.4th at __ (Newsom, J., concurring).

[244] *Deleon*, __ F.4th at __ ("[R]eassuringly, the 30 results I received—10 apiece from each of the three leading LLMs—largely echoed the initial response that I got from ChatGPT."), *citing Generative Interpretation*, *supra* note __ at 487-88. *See also Snell*, 102 F.4th at 1225, 1233 (proposing that users "try different prompts" and "query multiple models to ensure that the results are consistent")."

[245] *Deleon*, __ F.4th at __ (Newsom, J., concurring).

corroborate correctness. They do not. The confidence and consistency of a model's output bolster its probative value only when there is reason to believe that the model's determinations accurately reflect facts about linguistic meaning in the first place. And this latter belief is something generative-interpretation proponents justify, at least in part, with appeals to the superficial reasonableness and plausibility of LLM output.[246] So long as they do so, the danger remains that users will project authority onto LLMs not because LLMs are objectively "accurate," but because they like what they hear.

D. Relying on Predictability?

The final general type of argument offered by LLM proponents is that they are more *predictable* than human judges.[247] The importance of predictability to the rule of law is well established, and we need not dwell on why predictability is a worthwhile goal. Decades of legal scholarship highlight arbitrariness in current interpretative techniques.[248] If anything, predictability looms even larger for LLM proponents than it does for other judges and scholars. Kieffaber, for example, argues that other goals for interpretation are "really predictability in disguise."[249] If LLMs provide consistent answers, they might improve judicial consistency and offer parties a reliable way to determine the meaning of terms without litigating. LLMs that sometimes produce arbitrary results might still improve upon the status quo if those results are predictable.

LLM proponents offer, broadly, four reasons to think that LLMs are more predictable. First, there are claims that LLMs will be more predictable because they are algorithmic and automated, and thus not subject to human biases. But this is an unwarranted generalization about how computers work. Computations are typically predictable step-by-step: the results of each instruction a computer executes are formally specified and highly reliable. But the overall results of a computation need not be predictable in the slightest; indeed, because a computer can execute so many steps of an algorithm so quickly, it can produce outputs of astonishing complexity. This complexity is on full display for LLM. Even compared with other types of programs, they are notoriously inscrutable. While scholars have begun to probe the internal representations employed by LLMs as they generate outputs[250], this work is still at an early stage, and LLM internals are much less understood than that other notoriously unpredictable computational network, the human brain.

---

[246] *See, e.g.*, *Snell*, 102 F.4th at 1228 (Newsom, J., concurring) ("And as anyone who has used them can attest, modern LLMs' results are often sensible . . . . [LLMs] are[] high-octane language-prediction machines capable of probabilistically mapping, among other things, *how ordinary people use words and phrases in context*.") (emphasis added).

[247] *See Generative Interpretation*, *supta* note __ at 509.

[248] *See id.* at 464–66 (summarizing and collecting relevant literature); *In AI We Trust II*, *supra* note __.

[249] Kieffaber, *supra* note __, at 12.

[250] *See, e.g.*, Jack Lindsey et al., *On the Biology of a Large Language Model*, ANTHROPIC (Mar. 27, 2025), https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Second, there are claims that LLMs will be more predictable because they are trained on huge corpora of human-written text.[251] This argument fails for the same reason as claims that LLMs will be more accurate because they are trained on huge corpora of human-written text. The training process by itself guarantees nothing. Whether or not an LLM's outputs are predictable is an empirical question, and our empirical demonstrations in Part III show that frequently they are not.

Third, there are claims that the low cost and high accessibility of LLMs will give people access to them in advance of litigation, so that they can predict how courts or arbitrators would rule.[252] This is not necessarily a claim based on accuracy. It is not necessary that the LLM be correct for it to be predictable; all that is required is that parties can obtain the same results *ex ante* by consulting the LLM as judges would obtain *ex post* by consulting it. The fly in the ointment is that for parties to predict outcomes they need to anticipate courts' prompting and configuration choices—many of which are dispute-specific. The more sensitive a method of generative interpretation is to those choices, the more precisely parties must predict those choices. If parties must negotiate or litigate dispute-specific choices to achieve predictable results, the efficiency benefits of generative interpretation shrink. And if those negotiations now require access to a new, expensive kind of additional technical expertise, generative interpretation might simply exacerbate challenges indigent parties face in gaining access to justice.

Finally, it is important to remember that predictability alone is not sufficient for legal legitimacy. An interpretive method that always reads text in the way that most favors the party with the longer name is cheap, accessible, *and* predictable. It would improve access and reduce costs. Yet we suspect most readers would agree that name length is not a legitimate factor in adjudicatory reasoning.[253]

V. REWORKING GENERATIVE INTERPRETATION

We aren't writing to announce unbridled AI pessimism. Our concern is that generative-interpretation proponents' influential proposals treat LLM output as probative of facts that they do not actually signify. Although these particular uses of LLMs are problematic—or rather, the problems they raise have not yet been surmounted—the good news is that there are many legal applications where LLMs can realize these benefits without running into the same problems.

---

[251] *Generative Interpretation*, *supra* note __ at 485 (referring to "the vast corpus on which [GPT-4] sits"). *Cf.* David Hoffman, *Consumers' Unreasonable Textual Expectations*, Harv. Bus. L. Rev. (forthcoming) at 6, *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5127354 ("The gist of our claim [in *Generative Interpretation*] is that large language models are particularly good at offering majoritarian readings of phrases read in the context of contracts, and can probabilistically estimate what meaning the ordinary reader of that document will take.")

[252] *E.g.*, Gandall, Kieffaber, and McLaren, *supra* note __, at 4–5.

[253] *Cf.* Branerd Currie, *Conflict, Crisis, and Confusion in New York*, 1963 DUKE L.J. 1, 11–13 (arguing that if uniformity and predictability were all that mattered, a choice of law rule selecting Alaska law for all disputes, whether or not they had anything to do with Alaska, would be ideal).

A. How AI-Generated Artifacts can Acquire Normative Weight

Broadly speaking, we see ways in which LLMs can be made fit for use in the legal process: for their purely rhetorical value, when they have been experimentally validated, or when they are subject to adversarial testing.

1. Rhetorical Value

LLM proponents often appeal to AI's strength as a rhetorician—and sometimes being rhetorically effective is exactly what one wants, and *all* that one wants. Unikowsky, for example, writes that after he presented Claude with the merits briefs of adjudicated cases from the Supreme Court's most recent term, Claude decided all but ten of them "in the same way the Supreme Court did. In the other 10 . . ., I frequently was more persuaded by Claude's analysis than the Supreme Court's." If Claude is as persuasive a legal writer as Unikowsky suggests, we see no downsides to harnessing LLMs' rhetorical power *per se*. LLMs are legitimate assets to the legal profession at least insofar as they are harnessed for their rhetorical power. Lawyers could ask LLMs to generate persuasive arguments for their client's positions, then personally ensure that these outputs adhere to the other formal requirements of legal briefing. And lawyers and judges alike could ask LLMs to generate compelling counterarguments to their conclusions, then use those counterpoints to strengthen their briefs and opinions. Deployed for their rhetorical power, LLMs might be valuable tools to help judges and advocates strengthen their reasoning. Our argument here lines up with Waldon *et al.*'s recommendation for "dialectical legal AI," which similarly endorses LLM usage to "generate novel perspectives that challenge or expand a judge's theory of a disputed legal text," but not the "delegat[ion] [of] . . . authority to the model."[254]

To return to our fair-coin analogy one last time, consider the practice of tossing a coin to make a major decision. Deciding to follow the coin's "choice" is arbitrary.[255] But tossing the

---

[254] Brandon Waldon et al., *Large Language Models for Legal Interpretation? Don't Take Their Word for It* 53-55 (Feb. 3, 2025), https://papers.ssrn.com/abstract=5123124.

[255] A surprising number of judges have been sanctioned for decision–by–coin–flip. *E.g., In re Daniels*, 340 So. 2d 301 (La. 1976) (sanctions for pretending to decide guilt or innocence of criminal defendants by coin-flip); *Matter of Friess v. N.Y. State Comm'n*, 91 A.D.2d 554 (N.Y. App. Div. 1982) (sanctions for determining sentence-length by coin flip); *In re Brown*, No. 123687 (Mi. 2003) (sanctions for deciding which parent has child custody on Christmas by coin-flip); *Judicial Inquiry v. Shull*, 274 Va. 657, 651 S.E.2d 648 (Va. 2007) (sanctions for determining child visitation rights by coin-flip). Some legal theorists have considered the (narrow) circumstances under which randomness is preferable to human judgment. S*ee*, *e.g.,* Bernard E. Harcourt, *Post-Modern Meditations on Punishment*, *in* CRIMINAL LAW CONVERSATIONS 163, 167-68 (Paul H. Robinson, Stephen Garvey, & Kimberly Kessler Ferzan eds., Oxford U. Press, Nov. 2011) ("there always came a moment when the empirical facts ran out…*and yet the reasoning continued*. There was always this moment when the moderns—those paragons of reason—took a leap of faith… Where our social scientific theories run out, where our principles run dry, we should leave the decision-making to chance. We should no longer take that leap of faith, but turn instead to the coin toss."); Jeph Loeb & Tim Sale, *Batman: The Long Halloween*, DC COMICS (1996) (depicting comic book villain Harvey Dent arguing that coin flips are normatively preferable to courts or legislatures because random chance is impartial); Adam M Samaha, *Randomization and Adjudication*, 51 W&M L.

coin and introspecting on whether the outcome feels right or wrong is a popular trick for getting insight into your internal compass. In that case, the coin toss's outcome is arbitrary, but the insights it generates are not.[256] And, just as importantly, the coin's value as a tool to probe one's own thoughts does not depend on whether the coin actually is unbiased.

Of course, rhetorical use of AI isn't risk-free—but the risks are largely that rhetorical use might invite unfounded assumptions about LLMs' authority. For example, a lawyer or a judge may ask an LLM to produce competing arguments about an issue in a case. Any of the arguments that the LLM produces can help its interlocutor think through the issues and refine a brief or an opinion. The LLM user may, however, find that the LLM is able to argue for one result, A, more persuasively than for a competing result, B. If the LLM user takes this behavior as evidence that the LLM "believes" in the result it argued for more persuasively, and thus that A is correct and B is incorrect, then the user has stopped using the LLM as a rhetorical tool and has instead begun to defer to it, unfoundedly, as an epistemic authority. An LLM might craft a more persuasive argument for A than B for any number of arbitrary reasons, and the arguments that an LLM produces are not perforce the strongest ones. That an LLM argued for A more persuasively than B is not evidence that the strongest argument for A beats the strongest argument for B, and it would be an error to interpret as such.

But evaluating arguments is precisely what lawyers and judges know how to do best. So long as they approach LLMs as rhetoricians--not as oracles whose minds need to be read–lawyers and judges can use LLMs' powers of persuasion to test their own reasoning and make their writing more persuasive.

2. Experimental Validation

Limiting LLM use to purely rhetorical applications would obviate one of the most exciting potential benefits of AI: using automation to reduce judicial and administrative backlogs. By limiting our qualitative assumptions about LLM-produced text to those that are susceptible to objective benchmarks, we can expand the set of situations in which LLM use is justifiable. These applications of AI do not involve the splashy, novel issues of interpretation that make for exciting technical demonstrations. Instead of answering unfamiliar interpretive questions, LLMs can assist with the exact opposite: run-of-the-mill cases, where fact patterns rarely deviate from a handful of common situations, and voluminous precedent provides ready-made benchmarks for testing predictive accuracy.

REV. at 34 ("In a subset of these cases, however small, randomization will be the theoretically superior option for reasons of practical indivisibility, equality norms, nagging uncertainty, incentive effects, and/or experimental value. That we have difficulty identifying this class of cases with precision is no reason to think it is an empty set"); Ronen Perry & Tal Zarsky, *'May the Odds Be Ever in Your Favor': Lotteries in Law*, 66 ALA. L. REV. 1035 (2015).

[256] *See,* Mariela E. Jaffé, Leonie Reutner, and Rainer Greifeneder, "Catalyzing Decisions: How a Coin Flip Strengthens Affective Reactions," ed. Baogui Xin, *PLOS ONE* 14, no. 8 (August 14, 2019); Mariela E. Jaffé and Rainer Greifeneder, Mariela E. Jaffé and Rainer Greifeneder, "Deciding Advantageously after Flipping a Coin," 223 *Acta Psychologica* 103511 (March 1, 2022) (finding that the coin-flip technique may improve decision-making under certain controlled conditions).

In these cases, inductively validating the factual accuracy of LLM outputs is well within the technical capabilities of most social scientists. For example, an AI that can detect choice-of-venue clauses in a contract, or specific, common but invalid clauses in rental agreements from certain jurisdictions, could be tested against a curated list of similar agreements. A tool that performs well against that benchmark might be able to organize dockets and direct judicial attention in ways that increase efficiency without supplanting any part of the adjudicatory process in any individual case. Similarly, AIs could be tested on their ability to detect or correct procedural deficiencies in court and agency filings, allowing petitioners to amend those filings without an extra visit to the courthouse.

The common thread across these examples is *verifiability*. By using LLMs to perform rote tasks with verifiable results, we can establish *ex ante* the likelihood that those results are accurate and can monitor *ex post* whether the system remains accurate over time. With well-designed and continually re-tested benchmarks, LLMs could automate portions of the legal process while conforming to the same principles and standards used for other kinds of scientifically verified mechanization and measurement.[257]

Inductively validating LLMs is at least a tractable problem; this is a kind of science that computational legal scholars know how to do. But a comparison with immensely simpler and yet still highly controversial empirical interpretive projects—trademark surveys and corpus linguistics—shows that here the mountains are high and we are barely in the foothills.[258] Effective legal benchmarks would need to provide a set of tests representative of the variety and complexity of interpretation tasks that face courts, and would need some well-defined "ground truth" to measure against.[259]

Today's high-profile legal-reasoning benchmarks are far from representative.[260] An LLM's ability to pass the bar exam, for instance, is impressive but uninformative.[261] The bar exam is designed to test humans, and it assumes that test-takers are capable of lateral applications of knowledge. No LLM has been shown to be capable of that kind of lateral "thinking." Unless courts need an LLM to do *exactly* what a benchmark tests (e.g., answer bar exam questions from the same corpus, using the same format, writing style, level of complexity, and subject matter), that benchmark is not necessarily predictive of LLM accuracy. Though initial work on building

---

[257] We don't mean to imply that all the questions surrounding the use of scientific and empirical evidence in courts are settled. Just that these situations—unlike blind faith in an LLMs output—don't require courts to assume any *new* descriptive or epistemic assumptions.

[258] *See generally, e.g.*, Barton Beene and Jeanne C. Fromer, *Are We Running Out of Trademarks? An Empirical Study of Trademark Depletion and Congestion*, 131 HARV. L. REV. 945 (2018) (thoughtful empirical study of trademark depletion); Jens Frankenreiter and Michael A. Livermore, *Computational Methods in Legal Analysis*, 16 ANN. REV. L. & SOC. SCI. 39 (2020) (overview of computational textual methods).

[259] *Cf.* Michael J. Hasday, *Accuracy and the Robot Judge*, 25 J. Appellate Prac. & Process 1 (2025) (arguing that AI judges could be benchmarked by comparing their decisions to decisions made by human judges).

[260] Florencia Marotta-Wurgler & David Stein, *Can Machines Read Like Lawyers? Building Benchmarks for Legal AIs* [under submission].

[261] See, Will Heaven, *AI Hype is Build on High Test Scores. Those Tests are Flawed*, MIT TECH. REV. (Aug. 2023).

better benchmarks is underway, those efforts currently focus on questions with well-defined right answers—like checking outputs for rhetorical and factual consistency.[262] Subjective tasks—like the interpretation of ambiguous or contested terms—have no well-defined right answers to test against.[263]

3. Adversarial Testing

When a judge uses an LLM, they are recruiting an expert without the *Daubert* rule, without impeachment, without a jury, without the adversarial checks on which our legal system relies. By contrast, if a litigant introduces LLM outputs as evidence, they are forced to justify that use in a way that can survive evidentiary checks and adversarial challenge. In contrast to generative interpretation, where parties have no opportunity to challenge factual assumptions that judges make about the LLM-generated text, we take no similar issue with the use of LLM outputs by litigants. In those cases, the trial process *is* the verification process.[264]

B. Two Examples

Two other approaches to interpretation—trademark surveys and corpus linguistics—show what it takes to make empirical measurements of linguistic meaning normatively legitimate in adjudication. In each case, a field of expert specialists have grappled seriously with the problem of providing both descriptive and epistemic justifications for their methods. They have provided descriptive justifications by developing a consistent and relatively standardized set of methods; they have provided epistemic justifications by being precise about what their methods measure and modest in their ambitions. They show a possible road forward for generative interpretation, but that road is a long and hard one.

1. Trademark Surveys

Survey evidence is broadly accepted and widely used in trademark and false advertising cases.[265] Many core issues in trademark litigation involve questions of how consumers

---

[262] *See, e.g.,* Guha, Nyarko, Ho, et. al., *LegalBench: A collaborative Built Benchmark for Measuring Legal Reasoning in Large Language Models*, 26 NEURIPS (2023).

[263] Again, there is some nascent progress here; some recent benchmarks make the initial steps towards measuring an LLM's ability to navigate the grey areas in legal interpretation. *See*, *e.g., See* Florencia Marotta-Wurgler & David Stein, *Building a Long Text Privacy Policy Corpus with Multi-Class Labels*, 63 PROC. ACL __ (2025), https://aclanthology.org/2025.acl-long.908 [editors: link will become active in late July, 2025]

[264] *Cf.* Arbel and Hoffman, *supra* note 2, at 506-9 ("[Judges will] want to be careful about parties' manipulative [use of LLMs]").

[265] *See* J. Thomas McCarthy, 5 McCarthy on Trademarks and Unfair Competition § 32:158 (2025) ("Survey Evidence is Routinely Received and Weighed by the Courts") Shari Seidman Diamond and David J. Franklyn, *Trademark Surveys: An Undulating Path*, 92 Tex. L. Rev. 2029, 2040 (2014) (over 1000 trademark surveys in reported cases between 1991 and 2012)

understand words or other marks, including whether a mark describes a product feature or the source of the goods, whether the defendant's use of a mark refers to their own goods or to the plaintiff's, and what an challenged statement conveys to consumers. These questions are all fundamentally empirical. While a judge's intuitions about language may play a role, the legal question is not whether "Fish-Fri" in the abstract uniquely refers to the plaintiff's coating mix or to any coating mix for frying fish, but whether *consumers understand* "Fish-Fri" narrowly to refer to the plaintiff's coating mix or broadly to refer any coating mix. Indeed, even evidence that specific individual consumers have been confused about the source of the goods they are buying is only one factor in the tests used to ask whether consumers in general are likely to be confused. The interpretive community of consumers is sovereign in trademark law.[266]

To establish descriptive validity, judges, scholars, and experts have built up an extensive body of standardized best practices for trademark surveys.[267] A typical survey design must consider how to identify consumers in the relevant market, how to present the marks to them, how and in what order to ask questions about their perceptions of those marks; what control questions to include to establish a relevant baseline; and what statistical analyses to apply to the raw data of participants' answers. All of these constrain the process so that it is more reliable and replicable; for example, there is a strong preference for closed-ended questions that do not require manual coding by the survey's administrators, because manual coding introduced a source of variation and possibly of bias. Even so, litigated trademark cases frequently feature the parties' trying to impeach each others' surveys, requiring judges to probe the fine details of how questions are worded.[268]

As for epistemic validity, it is striking how modest judicial reliance on trademark surveys is, notwithstanding the immense effort and expense that go into them. Every circuit has a detailed multi-factor test for consumer confusion, including factors that are not susceptible to surveying, such as the defendant's good or bad faith and the quality of their products.[269] Judges admit but discount survey evidence when they have concerns about its methodology, and they weigh survey evidence against other sources of evidence on consumer understandings. In short, the legal system treats all of the issues that trademark surveys speak to as discrete factual questions about the beliefs of actual consumers. Surveys are a powerful and potentially probative source of evidence about those discrete factual questions, but they are not regarded as conclusive even on those questions, let alone on an entire case. And even so, scholars still

---

[266] *See* Stanley Fish, *Interpreting the "Variorium*,", 2 CRITICAL INQUIRY 465 (1976).

[267] *See generally, e.g.*, JAMES T. BERGER AND R. MARK HALLIGAN, TRADEMARK SURVEYS: A LITIGATOR'S GUIDE (2023); TRADEMARK AND DECEPTIVE ADVERTISING SURVEYS: LAW, SCIENCE, AND DESIGN (Shari S. Diamond and Jerre Bailey Swann eds., 2d ed. 2022); JACOB JACOBY, TRADEMARK SURVEYS, VOLUME 1: DESIGNING, IMPLEMENTING, AND EVALUATING SURVEYS (2015).

[268] *See, e.g.*, Novartis Consumer Health v. Johnson & Johnson-Merck Consumer Pharmaceuticals Co., 290 F.3d 578, 590–95 (2002) (considering whether survey questions were improperly leading). Scholars continue to propose refinements to survey methodology. *See, e.g.,* Irina D. Manta, *In Search of Validity: A New Model for the Content and Procedural Treatment of Trademark Infringement Surveys*, 24 CARDOZO ARTS & ENT. L.J. 1027 (2007).

[269] *See generally* Barton Beebe, *An Empirical Study of the Multifactor Tests for Trademark Infringement*, 95 CALIF. L. REV. 1581 (2006).

offer powerful critiques that question whether trademark surveys actually measure what they purport to.[270]

2. Corpus Linguistics

If trademark surveys are a mature empirical interpretive tool, corpus linguistics is in the process of attempting to become one. Corpus methods are well-established in linguistics, where they provide new sources of data on large-scale patterns of linguistic use. Like generative interpretation, corpus linguistics and law developed from the combined efforts of academics[271] and judges.[272] The basic idea is to take a large corpus of existing linguistic data—a collection of texts written by actual humans, such as the Corpus of Contemporary American Usage (1.1 billion words from 1990 to 2010)—and run an algorithm over it to obtain an answer to an interpretive question of interest.

Again, consider this empirical tool through the lenses of descriptive and epistemic validity. A linguistic corpus by itself says nothing; it is simply a collection of things people have written or said. To use a corpus to answer specific interpretive questions—e.g., does a person "discharge" a weapon once for each shot, or once a group of shots fired in close succession?—one must find a way to interpret the corpus. In this example, from *State v. Rasabout*, Justice Lee used a collocation analysis. He identified 86 cases in which "discharge" was used within five words of a term relating to a firearm; of them, 27 said or implied that a "discharge" was a single shot, and only 1 referred to a group of shots.

This is a specific empirical protocol, and like any empirical protocol it can be challenged on grounds of unreliability. Scholars like Kevin Tobia have done significant empirical work questioning whether corpus-linguistic methods for legal interpretation are replicable in the scientific sense; Tobia found that they are "surprisingly unreliable."[273] Prominent proponents have responded in defense of their methods, and the debate continues in the law-review pages.[274] This back-and-forth is a normal part of the scholarly process; the point is simply that

---

[270] *See, e.g.*, Barton Beebe, Roy Germano, Christopher Jon Sprigman, and Joel H. Steckel, *Consumer Uncertainty in Trademark Law: An Experimental Investigation*, 72 EMORY L.J. 489 (2023); Sepehr Shahshahani and Maggie Wittlin, *The Missing Element in Trademark Infringement*, 110 IOWA L. REV. 1247 (2024).

[271] *See. e.g.,* Steven C. Mouritsen, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpuss-Based Approach to Plain Meaning*, 2010 B.Y.U. L. REV. 1915; Steven C. Mouritsen, *Contract Interpretation with Corpus Linguistics*, 94 WASH. L. REV. 1337 (2019); James C Phillips, Daniel M. Ortner, and Thomas R. Lee, *Corpus Linguistics and Original Public Meaning: A New Tool to Make Originalism More Empirical*, 126 YALE L.J. FORUM 21 (2016).

[272] *See.*, e.g., State v. Rasabout, 356 P. 3d 1258, 1275–90 (Utah 2015) (Lee, J., concurring); Wilson v. Safelite Group, Inc., 930 F.3d 429, 438 (6th Cir. 2019) (Thapar, J., concurring); Richards v. Cox, 450 P.3d 1074 (Utah 2019); People v. Harris, 885 NW 2d 832 (Mich. 2016); Facebook, Inc. v. Duguid, 141 S. Ct. 1163, 1174–75 (Alito. J., concurring).

[273] *See. e.g.,*, Kevin Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726, 799 (2020).

[274] *See, e.g.,* Thomas R. Lee and Steven C. Mouritsen, *The Corpus and the Critics*, 88 U. CHI. L. REV. 275 (2021); Kevin R. Tobia, *The Corpus and the Courts*, 88 U. CHI. L. REV. ONLINE (2021).

it is also a necessary part of establishing an empirical method as reliable in the sense required for use in adjudication.[275]

Note, by way of comparison, how much more tightly constrained the task specification is in corpus linguistics than in generative interpretation. The algorithms used in contemporary legal corpus linguistics are straightforward: they can generally be specified in a single search query or a few lines of code, and they have broadly accepted common implementations. In contrast, generative methods are remarkably open-ended. One must write prompts (a natural-language task!), give them to a model whose internal functioning and semantic correlations are incredibly poorly understood, and then often interpret the outputs as well (another natural-language task!). Similarly, there are only a handful of corpora in wide use for corpus-linguistics purposes in law, while there are dozens of plausible models one could consult for generative interpretation, and more are being introduced constantly. Indeed, two of legal corpus linguistics' most prominent advocates, Thomas Lee and Jesse Egbert, have criticized generative interpretation because it lacks "*transparency* in the methods used and *replicability* and *generalizability* in the findings that are generated."[276]

Just as with trademark surveys, to pick among competing methods, one must have a theory of epistemic validity based on an underlying normative interpretive theory. Here, the theoretical debates among corpus linguists and their critics show why these theories are deeply contested; there is no simple linguistic truth that can be read off unproblematically from a corpus. Tobia points to the deep tension about what to do when corpus linguistics methods conflict with other sources of meaning; the divergence could mean that corpus linguistics comes closer to the true collective linguistic meaning, that corpus linguistics falls further away from collective linguistic meaning, or that collective linguistic meaning itself is a construct that does not exist as a ground truth in the world that can be objectively ascertained. And scholars like Anya Bernstein and Carissa Byrne Hessick observe that there is necessarily a gap between the empirical facts that corpus linguistics can ascertain and the normative claim that its results should be treated as authoritative.[277]

---

[275] Of course, it is possible that this process could cut in favor of generative interpretation compared with other empirical interpretive methods. Tobia's surveys found that dictionaries and corpus linguistics did a worse job of tracking lay survey participants' assessments of lingistuic meaning; judges who did not use these tools came closer to the participants' assessments. Tobia, *Testing Ordinary Meaning*, *supra* note __, at 753–77. It is entirely plausible that generative interpretation could outperform dictionaries and corpus linguistics, precisely because generative systems are engineered to emulate usage patterns in their training data. But this is a fundamentally empirical question, and note what could be measured. No experiment can determine whether a generative method yields correct results, because there is no accessible source of ground truth for legal meaning. (If there were, then legal interpretation would be unnecessary, and so would generative interpretation.) All that can be measured, even in principle, is the extent to which generative methods *track* other measurement of meaning, such as dictionar, survey participants, and judges' opinions.

[276] Thomas R. Lee and Jesse Egbert, *Artificial Meaning?*, 77 FLA. L. REV. (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4973483.

[277] Anya Bernstein, *Legal Corpus Linguistics and the Half Empirical Attitude*, 106 CORNELL L. REV. 1397 (2021); Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 B.Y.U. L. REV. 1503.

3. Implications

Our point is not that these gaps cannot be overcome, either for corpus linguistics or generative interpretation. Trademark surveys have overcome them, within their domain. They provide a kind of interpretive evidence that is regularly used and is widely accepted as legitimate. Corpus linguistics is doing its level best to do the same. Still, it is being forced to articulate and explain why particular methods are empirically reliable and why the results of those methods should be regarded as legitimate interpretive data. It is slow, laborious, contentious work.

Our greatest fear is that generative interpretation will attempt to circumvent these challenges rather than take them seriously. Given the laboriousness of corpus-linguistic methods, generative interpretation can seem like a remarkable shortcut: rather than specifying a detailed algorithmic query that will provide circumstantial evidence of meaning, why not just ask a model the interpretive question directly and go with what the output says? But that approach has, to quote Bertrand Russell, all the advantages of theft over honest toil. For generative interpretation to fulfill the hopes of its advocates, it will have to confront all of the tedious empirical challenges and all of the contentious normative issues that trademark surveys and corpus linguistics have had to deal with. That will be no mean feat.

Consider how generative interpretation would look if it followed a trademark-survey model. Parties to a case involving an interpretive question would commission their own experts to conduct generative-interpretation studies. The experts would specified detailed experimental methods: the models and systems they use, they prompts they provide to those models, the outputs they examine, how they interpret those outputs, the number and nature of variations on the inputs they try, and the statistical tests they apply to summarize the outputs across numerous prompts. Every choice they made in doing so would be adversarially probed by opposing counsel and experts: Why Claude 4 Opus and not ChatGPT-o3? Why did the prompt use "plaintiff" and "defendant" instead of the parties' names? As we have shown above in Part III, many of these choices can dramatically affect the results of generative interpretation. They will have to be pinned down in a principled way, and that will require hard conversations about what makes one form of generative interpretation more reliable than another. As we have shown in Part IV, there is no way to resolve these debates without a normative theory of what it is that generative interpretation is supposed to be measuring. Trademark surveys have just such a theory, but it is narrower and less ambitious than generative interpretation currently aims for. Corpus linguistics shows what the process of working out such a theory looks like—a drawn-out debate that simultaneously gets deep into the technical weeds and ascends to the highest peaks of jurisprudence.

Or compare Lisa Larrimore Ouelette's proposal to use search results as evidence of a trademark's strength and of the degree of overlap between two marks' likelihood of confusion.[278] Her suggestion combines trademark surveys' domain (consumer understandings of marks), corpus linguistics's quantitative methods (count occurrences with a given word sense), and generative interpretation's use of Internet datasets and reliance on complex black-

[278] Lisa Larrimore Ouelette, *The Google Shortcut to Trademark Law*, 102 CALIF. L. REV. 351 (2014).

box machine-learning systems.[279] But the modesty of Ouelette's proposal leaps off the page: she discusses in extensive detail the ways in which search results might be unstable or unrepresentative of linguistic usage; she individually examines every case in her dataset where a court and Google disagree; she emphasizes the limited nature of her empirical results; and her bottom line is only that search results are probative evidence that courts should not categorically exclude. This is the kind of scholarly foundation on which a persuasive argument for generative interpretation will ultimately need to rest.

CONCLUSION

The legal philosopher Ronald Dworkin famously deployed the figure of Judge Hercules, an interpreter of unlimited competence and intelligence, with perfect access to all of the relevant authorities, and all the time and diligence needed to study and harmonize them. Judge Hercules, Dworkin argued, was the ideal towards which human judging aspires.[280]

Generative AI proponents see AIs as modern miracles, capable of tackling any cognitive task with superhuman skill. So perhaps it should come as no surprise that a version of this boosterism has taken hold in the legal academy. In their various ways, the thinkers we have been calling "LLM proponents" describe ChatGPT, Claude, and the like in terms that recall Dworkin's Judge Hercules. Like him, LLMs have mastered a corpus of text far larger than any human ever could. Like him, they can produce an answer to a complex problem far faster than any human ever could. And like him, they can knit together disparate and conflicting sources into a seamless justification as persuasively as a human could.

To all of this we say: *perhaps*. Even granting that LLMs are capable of Herculean feats, a deep and fundamental problem remains. *How can we be confident that Judge Hercules really has decided the case before him competently and in good faith*? Anyone could claim to be Judge Hercules, to have mastered the world's accumulated statutes and precedents. And even if Judge Hercules is who he claims to be, the fact that he purports to present the best harmonization of the law is no proof that he is telling the truth about it. It does no good to point out that he can also write iambic pentameter or functional JavaScript code, when the question that needs

---

[279] For additional hybrids, see Jake Linford and Kyra Nelson, *Trademark Fame and Corpus Linguistics*, 45 Colum. J.L. & Arts 171 (2022) (arguing that corpus-linguistic methods have advantages over surveys for assessing trademark fame); Shivam Adarsh, Elliott Ash, Stefan Bechtold, Barton Beebe, and Jeanne C. Fromer, *Automating* Abercrombie*: Machine-Learning Trademark Distinctiveness,* 21 J. EMPIRICAL LEGAL STUD. 826 (2024).

[280] *See generally* RONALD DWORKIN, LAW'S EMPIRE (1986). *Cf.* Kieffaber, *supra* note 99, at 31–32 (arguing that an ideal AI would be "*Scalia's* Hercules, with a complete understanding of *language alone* rather than the various capricious elements that inform a Dworkinian positivist.") But of course, as we have been arguing at length, there is no reason to assume that any particular LLM's outputs reflect "language alone." For one thing, a typical LLM's training process includes extensive pre-training on heavily curated datasets, fine-tuning for specific use cases, and reinforcement learning from human feedback to optimize its responses against human preferences. *See generally* Lee, Cooper, and Grimmelmann, *supra* note __, pts. I.C.2–.8 (discussing model development and the numerous choice points it offers to shape a system's behavior). To claim that this process learns "language" and nothing else is to hold up a randomly selected kitchen tool and assume that it is a can opener.

answering is whether each interpretation he produces really is faithful to the legal authorities. Perhaps he forgot, or misunderstood, or hallucinated the nuances of the ERISA regulations on which the case depends. Perhaps he is a textualist in insurance cases but consults legislative history in admiralty cases. Perhaps he systematically favors corporate plaintiffs. Perhaps he knows how most people would understand a phrase, but ignores that knowledge when a simpler interpretation is available. Perhaps he finds arguments more persuasive when they SOMETIMES BREAK INTO ALL CAPS. Or perhaps the superhuman judge is also a superhuman liar.

LLMs are like humans in some ways and unlike them in others, and whether an LLM should or should not be treated as equivalent to a human in a particular context depends on the LLM, the human, and the context.[281] A merely human judge's written opinion is a kind of proof of work. It is meant to persuade the reader of the legal conclusion it presents, but it is also meant to persuade the reader that the conclusion was arrived at competently and in good faith. A judge can fall short of the judicial ideal not just by being *too bad* at legal writing—writing an opinion that fails to persuade because it lays bare the threadbare reasoning that resulted in it—but also by being *too good* at it—writing an opinion that is too persuasive because it carries the reader along too briskly to notice the better road not taken. True, the test of persuasiveness has never been sufficient to ensure complete honesty and competence in judging. But LLMs threaten to explode it altogether.[282]

Something more is required to make generative interpretation legitimate, some assurance that *this particular* output from among *all possible* outputs was generated fairly, objectively, on the basis of the evidence, and in accordance with the relevant authorities. To date, LLM proponents have not paid sufficient attention to what that something more might be. We hope that they will.

---

[281] *Cf.* Robert Brauneis, *Copyright and the Training of Human Authors and Generative Machines*, 48 COLUM. J.L. & ARTS 1 (2025) (exploring in detail whether LLMs should be regarded as fundamentally human-like, fundamentally inhuman, or somewhere in between for purposes of copyright infringement).

[282] *Cf.* Eric A. Posner and Shiram Saran, *Judge AI: Assessing Large Language Models in Judicial Decision-Making* (Jan. 28, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5098708 (" If the goal is to produce AI judges that operate like human judges, success would be achieved only if the AI judges decide cases in a realist way while using formalist reasoning—meaning that they do not explain how they actually decide the cases. It is hard to imagine such AI judges being acceptable in a democracy or any well-ordered political system."); Re, *supra* note __, at 1582–85 (" The existence of legal norms and elites has always depended on there being a scarcity of persuasive resources and arguments. … By undermining these constraints, a surfeit of persuasiveness threatens the effectiveness of legal norms.").