BEYOND WORDS: THE RISKS OF GENERATIVE INTERPRETATION

JONATHAN SCHER*

ABSTRACT

Judges are beginning to use large language models like ChatGPT to interpret legal texts. This Note examines whether they should do so. Prior studies testing LLMs as legal interpreters use survey responses as benchmarks for performance. I offer the first study comparing LLM interpretations to real-world judicial decisions. Across eight Ninth Circuit cases, I test whether GPT-4 Turbo (a model of ChatGPT) correctly identifies legal text as ambiguous or unambiguous. I find that ChatGPT's assessments diverged from the court's determinations 50% of the time. I then advance a novel argument: judicial reliance on LLMs may constitute improper ex parte communication under current judicial ethics rules.

INTRODUCTION

"[I]t no longer strikes me as ridiculous to think that [a large language model] like ChatGPT might have something useful to say about the common, everyday meaning of the words and phrases used in legal texts."[1] With this statement, a federal judge acknowledged what many in the legal profession have begun to recognize: Large Language Models ("LLMs") are reshaping legal interpretation.

In just over three years since their public release, LLMs have been used to draft briefs, generate contracts, and, increasingly, assist judges in interpreting statutes and contracts. Federal judges have already experimented

* Executive Senior Editor, *Southern California Law Review*, Volume 99; J.D. Candidate, 2026, University of Southern California Gould School of Law; B.S., Chemistry, 2021, University of Michigan. Replication code and data for this Note are available at https://github.com/jscher07/llm-legal-ambiguity-replication. I am grateful to Professor Jonathan H. Choi for providing insightful feedback on the drafts of this Note, and to the editors of the *Southern California Law Review* for their hard work in preparing my Note for publication.

1. Snell v. United Specialty Ins. Co., 104 F.4th 1208, 1234 (11th Cir. 2024) (Newsom, J., concurring).

with LLMs to determine the ordinary meaning of words when traditional tools—such as dictionaries—fail to provide a definitive answer.

This Note makes two contributions to the growing debate over whether judges should use LLMs. First, it presents an original empirical study evaluating GPT-4 Turbo's ability to detect textual ambiguity across eight Ninth Circuit cases. Unlike prior studies that rely on repeated prompts and "temperature" variance, this study extracts the model's internal log-probabilities, providing a more accurate measure of its confidence. The findings reveal a 50% divergence from judicial ambiguity findings. Second, this Note advances a novel argument that judicial reliance on LLMs raises a serious risk of being an improper *ex parte* communication.

This Note proceeds in four parts. Part I provides a technical overview of how LLMs function and examines the tools judges currently use to interpret legal texts, with a focus on the challenge of determining ambiguity. It then discusses three recent cases in which judges have used LLMs to interpret words in legal texts. Part II reports the empirical findings of the GPT-4 Turbo study. Part III examines the normative and ethical implications of these findings, particularly whether the use of LLMs by courts constitutes *ex parte* communication. Part IV addresses a counterargument.

I. BACKGROUND

Since OpenAI's launch of ChatGPT in November of 2022 and the subsequent introduction of other large language models, such as Thompson Reuter's CoCounsel, Anthropic's Claude, and Google's Bard, the legal profession has profoundly changed. Despite well-publicized "hallucinations," in which lawyers cited to fake cases in legal briefs prepared by ChatGPT,[2] the use of artificial intelligence ("AI") by legal professionals in the past year has skyrocketed. According to a recent study, 79% of lawyers use AI in their daily practice, up from 19% in 2023.[3] LLMs can be prompted to summarize cases, produce first drafts of legal documents, and identify and

2. *See* Sara Merken, *New York Lawyers Sanctioned for Using Fake ChatGPT Cases in Legal Brief*, REUTERS (June 26, 2023, 1:28 AM), https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22 [https://perma.cc/MMB6-YSFV]; Owen R. Wolfe, Eddy Salcedo & Jamie Anderson, *Use of ChatGPT in Federal Litigation Holds Lessons for Lawyers and Non-Lawyers Everywhere*, SEYFARTH (June 1, 2023), https://www.seyfarth.com/news-insights/use-of-chatgpt-in-federal-litigation-holds-lessons-for-lawyers-and-non-lawyers-everywhere.html [https://archive.ph/0yaP9]. *See also* Blake Brittain, *Anthropic's Lawyers Take Blame for AI 'Hallucination' in Music Publishers' Lawsuit*, REUTERS (May 15, 2025, 12:25 PM), https://www.reuters.com/legal/legalindustry/anthropics-lawyers-take-blame-ai-hallucination-music-publishers-lawsuit-2025-05-15 [https://perma.cc/YC4L-3G33].

3. Marin McCall, *Clio 2024 Legal Trends Report Identifies AI, Alternative Billing Structures, and Client Engagement as Focus Areas for Today's Lawyers*, 2CIVILITY (Oct. 25, 2024), https://www.2civility.org/clio-2024-legal-trends-report-identifies-ai-alternative-billing-structures-and-client-engagement-as-focus-areas-for-todays-lawyers [https://perma.cc/SLT2-GGF4].

interpret ambiguous terms in contracts.[4]

A. OVERVIEW OF LARGE LANGUAGE MODELS

Simplistically, LLMs are chatbots that predict the next word in a sentence or prompt.[5] To perform this task, they proceed through a three-step process: tokenization, embedding, and word prediction. First, an LLM breaks input text into smaller units called tokens.[6] For example, "unbelievable" becomes "un" + "believ" + "able." Next, each token is converted into a vector (called "embedding")[7] that captures multiple aspects of its meaning,[8] including its grammatical role (for example, whether it is a noun or verb), its contextual relationships with other words, and common usage patterns. Embeddings represent the overall semantic meaning of words and tokens.[9] Next, the embeddings pass through various layers of neurons that modify them based on contextual information in the sentence.[10] For example, the word "bank" will receive a different embedding in the phrase "river bank" compared to the phrase "financial bank." A neuron is akin to a judge: it processes multiple inputs and weighs their importance just as a judge considers different pieces of evidence and arguments and assigns them different weights (or levels of importance) before reaching a ruling. The process of contextualizing embeddings by passing them through layers of neurons is called transformation, and it is achieved by programs called "transformers."[11] The "GPT" in ChatGPT, for example, stands for Generative Pre-Training Transformer.[12] After embeddings are contextualized, the LLM outputs the word(s) that are most likely to come next, based on the contextualized embeddings and its training data.[13]

---

4. *See, e.g.*, Daniel Schwarcz & Jonathan H. Choi, Essay, *AI Tools for Lawyers: A Practical Guide*, 108 MINN. L. REV. HEADNOTES 1, 12–30, 38–39 (2023).

5. *See, e.g.*, Matthew Burtell & Helen Toner, *The Surprising Power of Next Word Prediction: Large Language Models Explained, Part 1*, CSET (Mar. 8, 2024), https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1 [https://perma.cc/27UP-7L7X].

6. *Id.*

7. *Id.*

8. *See* Jonathan H. Choi, *Measuring Clarity in Legal Text*, 91 U. CHI. L. REV. 1, 20 (2024).

9. *Id.*

10. Amit Prakash, *What Is Transformer Architecture and How Does It Power ChatGPT?*, THOUGHTSPOT (Feb. 22, 2023), https://www.thoughtspot.com/data-trends/ai/what-is-transformer-architecture-chatgpt [https://perma.cc/UR7E-JCJW].

11. *See* Choi, *supra* note 8, at 5 n.13.

12. *See* Prakash, *supra* note 10.

13. *See id.*

During training, GPT is given text and asked to predict the next word in the sentence.[14] If it gets the prediction wrong, it adjusts its parameters via a process called backpropagation.[15] Adjustments could be made to the embeddings themselves or to the weights assigned to each neuron.[16] While we know what gets adjusted during training, experts do not understand how LLMs make decisions.[17] The sheer number of parameters (in GPT, billions) makes it extremely difficult to determine precisely the effect each variable has on the overall output.[18] For this reason, LLMs are considered "black boxes."[19]

An LLM is only as good as the data on which it is trained.[20] However, since most LLMs are proprietary, their training data is not disclosed to the public.[21] The proprietary nature of LLMs, in addition to their black-box nature, poses problems when they make bad decisions. For example, LLMs tend to hallucinate.[22] A hallucination occurs when an LLM generates a response that is factually incorrect or nonsensical despite sounding coherent and grammatically correct.[23] Another concern is that LLMs may engage in algorithmic discrimination because they are trained on biased data.[24]

B. STATUTORY INTERPRETATION AND THE AMBIGUITY PROBLEM

Textualism is a method of statutory interpretation that says legal text should be interpreted by its ordinary, everyday meaning at the time it was written, and extratextual sources such as legislative history should only be considered when the text is ambiguous.[25] A word or phrase is ambiguous if it can reasonably have more than one meaning.[26] For example, the word "bat," out of context, is ambiguous because it could mean a baseball bat or an animal. Textualists claim that their approach is less biased and results in

14. *See* Burtell & Toner, *supra* note 5.
15. *Id.*
16. *Id.*
17. Matthew Kosinski, *What Is Black Box AI?*, IBM (Oct. 29, 2024), https://www.ibm.com/think/topics/black-box-ai [https://perma.cc/3SRK-8PAZ].
18. *See* Prakash, *supra* note 10.
19. *See* Kosinski, *supra* note 17.
20. Christopher Engel & Richard H. McAdams, *Asking GPT for the Ordinary Meaning of Statutory Terms*, MAX PLANCK INST. FOR RSCH. ON COLLECTIVE GOODS, 2024/5, at 1, 10.
21. *Id.* at 12.
22. *Id.* at 10.
23. *Id.*
24. *See generally* Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky & Sharese King, *AI Generates Covertly Racist Decisions About People Based on Their Dialect*, 633 NATURE 147 (2024) (finding that racial prejudices embedded in web-scraped training data caused LLMs to make discriminatory decisions, including assigning lower-prestige jobs and harsher criminal sentences to speakers of African American English).
25. *See* ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 74 (2012).
26. Choi, *supra* note 8, at 11.

fewer policy-based outcomes.[27] Textualism is also said to put parties on notice as to the text's semantic scope.[28] This concept is related to the principle of fair notice in due process jurisprudence—the idea that people should have adequate notice of what the law requires so that they can conform their behavior accordingly.

A main critique of textualism is that there is no bright-line rule or standard for determining ambiguity.[29] This is problematic, Justice Kagan notes, because "[s]ome people think things are clear in circumstances in which other people think there's still a lot of question marks."[30] In practice, according to Justice Kavanaugh, judges have different thresholds for how clear a statute must be to be unambiguous.[31] Because of the lack of defined ambiguity thresholds, Professor Ward Farnsworth said that "judgments about ambiguity . . . are dangerous, because they are easily biased by strong policy preferences that the makers of the judgments hold."[32]

Modern purposivists are generally more open than textualists to consider Congress's aims in passing a law;[33] however, like textualists, when the statute is unambiguous, they tend to consider only the text.[34] Purposivists claim that textualism is burdensome as it requires Congressmembers to explicitly state all the ways that a law should be applied.[35] Secondly, Congress cannot foresee some issues that arise many years in the future.[36]

---

27. Sam Capparelli, Comment, *In Search of Ordinary Meaning: What Can Be Learned from the Textualist Opinions of Bostock v. Clayton County?*, 88 U. CHI. L. REV. 1419, 1424 (2021).

28. *See, e.g.*, Choi, *supra* note 8, at 56.

29. Brett M. Kavanaugh, *Fixing Statutory Interpretation*, 129 HARV. L. REV. 2118, 2138 (2016) (reviewing ROBERT A. KATZMANN, JUDGING STATUTES (2014)).

30. Harvard Law School, *The 2015 Scalia Lecture: A Dialogue with Justice Elena Kagan on the Reading of Statutes*, at 58:20 (YouTube, Nov. 25, 2015), https://www.youtube.com/watch?v=dpEtszFT0Tg [https://perma.cc/5Y3X-RFB9].

31. Kavanaugh, *supra* note 29, at 2137–38; *see* Brian G. Slocum, *The Importance of Being Ambiguous: Substantive Canons, Stare Decisis, and the Central Role of Ambiguity Determinations in the Administrative State*, 69 MD. L. REV. 791, 806–07 (2010).

32. Ward Farnsworth, Dustin F. Guzior & Anup Malani, *Ambiguity About Ambiguity: An Empirical Inquiry into Legal Interpretations*, 2 J. LEGAL ANALYSIS 257, 290 (2010).

33. STEPHEN BREYER, MAKING OUR DEMOCRACY WORK: A JUDGE'S VIEW, 94, 96 (2010); *see* Kavanaugh, *supra* note 29, at 2129 ("[T]extualists tend to find language to be clear rather than ambiguous more readily than purposivists do.").

34. Kavanaugh, *supra* note 29, at 2129 (quoting Chief Judge Katzmann) ("Purposivists tend not to go beyond the words of an unambiguous statute.").

35. *See* Paul Killebrew, Note, *Where Are All the Left-Wing Textualists?* 82 N.Y.U. L. REV. 1895, 1914 (2007) ("Sending statutes back to Congress for more definite answers whenever circumstances arise in which it is not clear whether the statutes apply . . . would create such a logjam in the legislative process that Congress may not be able to provide all of the specific answers requested.").

36. *See* Caleb Nelson, *What Is Textualism?* 91 VA. L. REV. 347, 415 (2005).

Finally, textualism may not necessarily serve a notice function for statutes passed many years ago.

C. CURRENT INTERPRETATIVE TOOLS

1. Dictionaries

Despite being the traditional tool for determining ordinary meaning, dictionaries have significant limitations.[37] First, judges sometimes "dictionary shop," selecting definitions that align with their preferred interpretations.[38] Despite the large number of dictionaries from which to choose and their frequently dispositive impact on a case, judges offer little explanation or methodology for their choices.[39] This concern is particularly weighty because it undermines textualism's underlying goals of predictability and judicial restraint.[40] Second, dictionaries define words in isolation and are not appropriate guides for how words are actually used in phrases and sentences.[41] Third, dictionaries written at the time of a statute may not accurately reflect contemporary usage, as they often borrowed definitions from earlier editions.[42] These limitations, along with the rising recognition that "meaning is an empirical fact,"[43] have led scholars and judges to explore empirical methodologies like surveys, corpus linguistics,[44] cosine similarities,[45] and now LLMs, to determine ordinary meaning.

2. Corpus Linguistics

Corpus linguists assess the meaning of words by examining the frequency of words (collocates) that are close to them.[46] For example, if "dog" and "puppy" frequently co-occur with the words "leash" and "bone," they are probably similar in meaning. One problem with corpus linguistics is that word frequencies are an inadequate measure of semantic meaning,

37. Note, *Looking It Up: Dictionaries and Statutory Interpretation*, 107 HARV. L. REV. 1437, 1454 (1994).

38. *See* William N. Eskridge, Jr., *The New Textualism and Normative Canons*, 113 COLUM. L. REV. 531, 534 (2013).

39. Jason Weinstein, *Against Dictionaries: Using Analogical Reasoning to Achieve a More Restrained Textualism*, 38 U. MICH. J.L. REFORM 649, 657 (2005).

40. *See* Capparelli, *supra* note 27, at 1425.

41. Weinstein, *supra* note 39, at 656.

42. *Id.* at 661.

43. *See* Gary S. Lawson, *Reflections of an Empirical Reader (Or: Could Fleming Be Right This Time?)* 96 BOS. U. L. REV. 1457, 1475 (2016); Larry Alexander, *Connecting the Rule of Recognition and Intentionalist Interpretation: An Essay in Honor of Richard Kay*, 52 CONN. L. REV. 1513, 1525 (2021) ("Interpretation of legal texts is an empirical, not a normative, endeavor.").

44. *See, e.g.*, Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 795 (2018).

45. *See generally* Choi, *supra* note 8.

46. *See id.* at 15–16.

that is, the meaning between words.[47] Words that are semantically similar may nevertheless be viewed as dissimilar via the collocation method because they appear among different groups of words.[48] For example, the words "pilot" and "driver" are likely to appear among different words in a sentence, even though semantically they are very similar in meaning.[49]

3. Surveys

Some scholars have conducted wide-ranging surveys of everyday citizens, seeking to demonstrate that dictionaries do not always capture ordinary understandings of legal texts.[50] Surveys, however, have several limitations. Practically speaking, it is untenable to conduct a survey every time a word is disputed in litigation. Surveys are time-consuming and expensive—some survey providers charge up to thousands of dollars.[51] Second, surveys are prone to many biases. While surveys may be distributed to a representative slice of the population, those who respond are not representative.[52] Additionally, surveys are prone to primacy bias: individuals may spend more time thinking about their response to the first option than later ones.[53] Finally, those surveyed may be susceptible to acquiescence bias in which they answer questions in a way that they think will make them look socially desirable.[54]

D. RECENT JUDICIAL USE OF LLMS

1. *Snell v. United Specialty Ins. Co.*

Recently, scholars and judges have turned to LLMs like ChatGPT as possible tools for determining the ordinary meaning of legal texts. For example, in *Snell v. United Specialty Ins. Co.*, Eleventh Circuit Judge Newsom discussed his use of ChatGPT to determine the ordinary meaning of "landscaping" in an insurance policy.[55] In *Snell*, an insurance company refused to defend and indemnify a landscaper who was sued by a homeowner

47. *Id.*
48. *Id.*
49. *Id.*
50. Kevin P. Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726 (2020).
51. *See What Is Your Pricing?*, PROLIFIC (last visited Nov. 30, 2025), https://researcher-help.prolific.com/en/articles/445239-what-is-your-pricing [https://perma.cc/T9CB-BGET].
52. *See* Choi, *supra* note 8, at 52.
53. Sean P. Mackinnon & Mengyao Wang, *Response-Order Effects for Self-Report Questionnaires: Exploring the Role of Overclaiming Accuracy and Bias*, 16 J. ARTICLES SUPPORT NULL HYPOTHESIS, 114, 114 (2020).
54. Yphtach Lelkes & Rebecca Weiss, *Much Ado About Acquiescence: The Relative Validity and Reliability of Construct-Specific and Agree-Disagree Questions*, RSCH. & POLS., July–Sep. 2015, at 1, 1–2.
55. Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1224–25 (11th Cir. 2024) (Newsom, J., concurring).

for negligent installation of a ground-level trampoline in the homeowner's backyard.[56] The landscaper's insurance policy limited coverage to "landscaping" projects, but the policy on its face did not define the word.[57] Ultimately, the case was not resolved on a statutory interpretation issue because the landscaper admitted in his insurance application that his work did not include recreational or playground equipment, and, under Alabama law, insurance applications are part of the insurance contract.[58]

In a concurring opinion, Judge Newsom agreed with the majority and sought to determine the ordinary meaning of "landscaping" from dictionary definitions.[59] *Webster's* said that landscaping had to be *natural* while *Oxford* said landscaping required an *aesthetic purpose*.[60] These definitions were unhelpful because they did not give a single controlling criterion, nor did they square with Judge Newsom's intuitive notions of landscaping.[61] If landscaping had to be natural, then it would presumably exclude walkways and accent lights, both of which intuitively seemed to qualify as landscaping.[62] If landscaping had to have an aesthetic purpose, then it would rule out a project to regrade a yard, something he also considered to be landscaping.[63]

"As a lark" Judge Newsom asked ChatGPT for the ordinary meaning of "landscaping."[64] The ChatGPT version said that landscaping involved changing an area of land or outdoor space for aesthetic or practical purposes.[65] This definition squared with Judge Newsom's intuition that landscaping included more than just natural improvements and covered both aesthetic and functional purposes.[66]

2. *United States v. Deleon*

In a subsequent case, *United States v. Deleon*, a concurring opinion explored the potential use of LLMs to determine the ordinary meaning of *phrases* in criminal statutes. In *Deleon*, an individual pointed a gun at a cashier during an armed robbery of a convenience store.[67] The defendant was

56. *Id.* at 1213.
57. *Id.* at 1213–14.
58. *Id.*
59. *Id.* at 1223 (Newsom, J., concurring).
60. *Id.*
61. *Id.*
62. *Id.*
63. *Id.*
64. *Id.* at 1224–25.
65. *Id.* at 1225 (ChatGPT's definition: " 'Landscaping' refers to . . . altering the visible features of an area of land, typically a yard, garden or outdoor space, for aesthetic or practical purposes . . . includ[ing] . . . planting trees, shrubs, flowers, or grass, as well as installing paths, fences, water features.").
66. *Id.*
67. United States v. Deleon, 116 F.4th 1260, 1261–62 (11th Cir. 2024).

indicted by a grand jury for armed robbery and brandishing a firearm during and in relation to a crime of violence.[68] The armed robbery sentence was enhanced (increased by two months), as mandated by the Sentencing Guidelines, because the defendant "physically restrained" the cashier by holding him at gunpoint.[69] The Guidelines defined "physically restrained" as the "forcible restraint of the victim such as by being tied, bound, or locked up."[70] Even though the defendant did not literally physically restrain the cashier, the court held that the defendant's conduct created circumstances that practically restrained him, leaving him with no choice but to stay in place and comply with the defendant's demands.[71]

Agreeing with the majority in full, Judge Newsom wrote a concurring opinion in which he described an experiment using three LLMs (ChatGPT, Claude, and Gemini) to analyze the ordinary meaning of "physically restrained." Each model was prompted ten times.[72] While the definitions varied slightly across iterations,[73] they consistently required tangible force—either bodily contact or the use of a physical device—to qualify as a physical restraint.[74]

Judge Newsom's work in *Snell* and *Deleon* are part of a growing body of research testing the capabilities and limitations of LLMs in legal interpretation. For example, Arbel & Hoffman investigated whether LLMs can determine the existence of ambiguity in contracts. In their study, LLMs were used to determine if early prepayment was allowed based on the face of a loan agreement, despite it not being a term of the contract.[75] In this study, LLMs were fed with the entire loan agreement and asked to give a number from zero to one hundred representing their confidence that the agreement allowed for early prepayment.[76] After running the prompt many times, the models found the agreement clearly excluded early prepayment as an option.[77] However, the court held the agreement was ambiguous as to whether prepayment was allowed—its language was "reasonably susceptible" to several interpretations offered by the parties.[78]

This divergence highlights a broader concern: LLMs may provide

68. *Id.* at 1262.
69. *Id.* at 1263–64.
70. *Id.* at 1263.
71. *Id.*
72. *Id.* at 1273 (Newsom, J., concurring).
73. *Id.* at 1273–74.
74. *Id.* at 1275.
75. Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451, 486–87 (2024).
76. *Id.* at 488.
77. *Id.*
78. *Id.* at 486–87.

confident and consistent-sounding answers that do not align with judicial reasoning. A comprehensive study by Jonathan Choi offers systematic evidence of these reliability problems across interpretative tasks. Testing eight different LLMs on identical "ordinary meaning" questions, Choi found near-zero agreement between the models ($p = 0.051$) in their responses.[79] When one model concluded that a word had a particular meaning, another model was just as likely to reach the opposite result. By contrast, a survey of 1,007 human respondents answering the same questions showed significantly higher agreement ($p = 0.285$), suggesting that people share common intuitions about linguistic meaning that current LLMs do not replicate.[80] LLMs are highly sensitive to how questions are phrased. Across more than 2,000 slight variations of the same legal question, GPT-4o sometimes expressed complete confidence in one answer and, with equal confidence, endorsed the opposite answer.[81]

Choi also identified a deeper flaw in how LLM reliability has been tested. Most studies have analyzed the text the LLM produces, rather than the probabilities the model assigns internally to possible answers. That output text is the product of an arbitrary "temperature" setting that governs how often the model chooses lower-probability responses. At temperature 0, the model will always choose its top-ranked answer; increase the temperature, and it will output more low-probability choices, creating the appearance of variability without revealing true uncertainty.[82] The variation researchers observe in a model's actual responses is not genuine uncertainty; it is simply an artifact of the model's "temperature" setting, which controls how often the model selects lower-probability responses over higher-probability ones.[83]

Judge Newsom and Lee & Egbert both examined the LLM's text outputs rather than its underlying probability distributions. Judge Newsom, finding only "minor variations in structure and phrasing," concluded that "the responses did coalesce, substantively, around a common core."[84] Lee and Egbert, on the other hand, found substantial variation upon repeated queries, concluding that ChatGPT's "lack of internal consistency raises serious questions about replicability."[85] Both approaches assumed that the words the model displays reveal its confidence or reliability. That

79. Jonathan H. Choi, *Off-the-Shelf Large Language Models Are Unreliable Judges* (Oct. 21, 2025) (unpublished manuscript at 46) (on file with the author).

80. *Id.* at 41, 46.

81. *Id.* at 18.

82. *Id.* at 11

83. *See id.*

84. United States v. Deleon, 116 F.4th 1260, 1274–75 (11th Cir. 2024).

85. Thomas R. Lee & Jesse Egbert, *Artificial Meaning?*, 77 FLA. L. REV. (forthcoming) (manuscript at 42) (on file with the author).

assumption is misplaced: those words reflect a post-processing choice, not the model's actual certainty.[86]

Building on Choi's insight, this study follows a probability-based approach: instead of measuring variation in the displayed text, it queries the model's internal log probabilities for each answer. This method eliminates distortions from temperature settings and gives a clearer picture of how strongly the model leans toward a particular interpretation. I apply this approach to real appellate cases, evaluating not only whether GPT-4 Turbo's outputs are stable but also whether they align, or fail to align, with judicial findings on ambiguity.

II. EMPIRICAL ANALYSIS OF GPT-4 TURBO'S ALIGNMENT WITH JUDICIAL AMBIGUITY DETERMINATIONS

With this improved method, the study examines GPT-4 Turbo's performance in assessing textual ambiguity across eight Ninth Circuit cases. The original aim was to test whether an LLM could serve as a tool for judges to "double-check" ambiguity determinations. The results, however, reveal concerning patterns in how LLMs assess textual ambiguity, underscoring the risks of relying on these tools in legal decision-making.

A. METHODS

This study examines GPT-4 Turbo's performance in assessing textual ambiguity in eight Ninth Circuit cases decided between January 2023 and January 2025. The cases were selected through a two-stage process. First, all Ninth Circuit opinions in that period containing the words "ambiguous," "ambiguity," or "interpretation" were identified. Second, cases from this list were narrowed based on the following criteria: (1) the interpretation issue was dispositive; (2) the court made an explicit ambiguity determination; and (3) the disputed text was short enough to present in full to GPT-4 without exceeding its context limit. This process yielded eight cases: three statutory interpretation cases, two regulatory interpretation cases, two contract interpretation cases, and one treaty interpretation case. GPT-4 Turbo was chosen as the test model because, unlike GPT-4o, it lacks web browsing capability and therefore cannot draw on real-time case information that could bias the analysis.

The study uses a modified version of Choi's "confidence estimation" method for evaluating the reliability of LLMs in legal interpretation. In Choi's approach, the model is asked to state its confidence, on a scale from 0 ("no") to 100 ("yes"), for a binary interpretative question.[87] The confidence

86. *See* Choi, *supra* note 79, at 11.
87. Choi, *supra* note 79, at 13.

score is then calculated as a weighted average of the numeric values the model is most likely to output, with each value weighted by the log probability the model assigns to it.[88] My adaptation differs in two ways: (1) I use GPT-4 Turbo instead of GPT-4o, and (2) I present the model with one hundred paraphrased versions of each question, rather than Choi's 2,000.[89]

To ensure systematic data collection and analysis, I developed a Python script interfacing with the OpenAI API. For each case, GPT-4 Turbo was prompted 100 times with a paraphrased version of the relevant legal interpretation question and asked to give its confidence on the 0–100 scale described above. The prompts were generated in Claude 3.7 Sonnet. For each prompt, GPT-4 Turbo's top five most probable numeric confidence values were recorded, along with their associated probabilities. The prompt's "average confidence level" was then calculated as the sum of each confidence value multiplied by its probability.

B. RESULTS

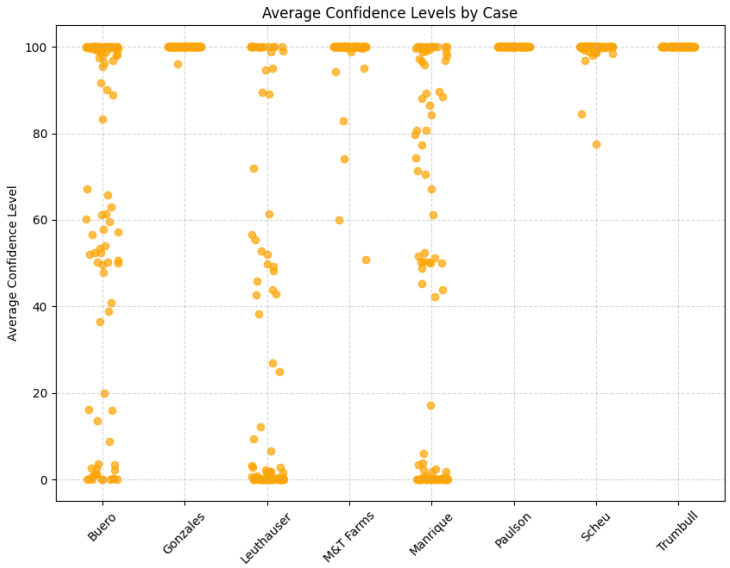
The Ninth Circuit unanimously found ambiguity in two cases and no ambiguity in five cases. In one case, the court was divided: the majority found no ambiguity, while the concurrence and dissent concluded the text was ambiguous (*see* Table 1 and Figure 1). GPT-4 Turbo's ambiguity assessments matched the Ninth Circuit's majority's determination in five of the eight cases. However, notable discrepancies emerged. In two cases in which the court found no ambiguity (*Leuthauser* and *Toledo*), GPT-4 Turbo found ambiguity. In one case in which the court found ambiguity (*M&T Farms*), GPT-4 Turbo reached the opposite conclusion, treating the text as ambiguous.

88. *Id.* at 12–13.
89. *Id.*

TABLE 1. Comparison of Ambiguity Assessments by Ninth Circuit and GPT-4 Turbo

Case Name	*Court Finding*	*GPT-4 Finding*
Buero v. Amazon.com Servs., Inc., 61 F.4th 1031 (9th Cir. 2023)	Ambiguous	Ambiguous
Gonzales & Gonzales Bonds & Ins. Agency, Inc. v. U.S. Dep't of Homeland Sec., 107 F.4th 1064 (9th Cir. 2024)	Majority: Not Ambiguous Concurrence/Dissent: Ambiguous	Not Ambiguous
Leuthauser v. United States, 71 F.4th 1189 (9th Cir. 2023)	Not Ambiguous	Ambiguous
M&T Farms v. Fed. Crop Ins. Corp., 103 F.4th 724 (9th Cir. 2024)	Ambiguous	Not Ambiguous
Toledo v. Kolc, 65 F.4th 1037 (9th Cir. 2023)	Unambiguous	Ambiguous
United States v. Paulson, 68 F.4th 528 (9th Cir. 2023)	Majority: Not Ambiguous Dissent: Ambiguous	Not Ambiguous
United States v. Scheu, 83 F.4th 1124 (9th Cir. 2023)	Not Ambiguous	Not Ambiguous
United States v. Trumbull, 114 F.4th 1114 (9th Cir. 2024)	Majority: Ambiguous Concurrence: Not Ambiguous	Not Ambiguous

IMAGE 1. Average Confidence Levels From 0 to 100 Across Eight Ninth Circuit Cases



Notes: Each orange dot represents a different prompt. For each case GPT-4 Turbo was prompted 100 times with a textual interpretation question. See Methods above for complete methodology.

1. *Buero v. Amazon.com Servs., Inc.*

Buero was a class action brought by Amazon employees who alleged that the company violated Oregon wage and hour laws by failing to compensate them for time spent waiting and passing through mandatory security screenings before and after work shifts.[90] Oregon law requires employers to pay employees minimum wage for "work time," defined to include both "time worked" and "time of authorized attendance."[91] The dispute turned on whether time spent in security screening lines constituted "time of authorized attendance."[92]

The court found this phrase was ambiguous because the statute did not define the phrase and dictionary definitions supported multiple reasonable interpretations.[93] "Authorized attendance" could mean (1) any time an employee is permitted on the employer's premises, or (2) only time spent

90. Buero v. Amazon.com Servs., Inc., 61 F.4th 1031, 1032 (9th Cir. 2023).
91. *Id.* at 1038.
92. *Id.* at 1
93. *Id.* at 1046.

waiting for a work assignment.[94]

Presented with one hundred paraphrased versions of the legal question, GPT-4 Turbo produced confidence scores clustering around 0, 50, and 100, with an average of 66.07. This means the model sometimes gave high-confidence "no" answers, sometimes high-confidence "yes" answers, and sometimes midrange responses.[95] While the average suggests no strong lean toward either side, consistent with the court's classification of the term as ambiguous, the bimodal distribution indicates high prompt sensitivity. The model's "agreement" with the court is therefore unstable: small changes in question phrasing push it toward opposite confident answers. GPT-4 Turbo therefore reached the same ambiguity classification as the court, but for unreliable reasons. The underlying pattern reflects prompt-driven swings, not consistent recognition of genuine ambiguity.

2. *Gonzales & Gonzales Bonds & Ins. Agency, Inc. v. U.S. Dep't of Homeland Sec.*

In *Gonzales*, the Ninth Circuit considered whether Secretary of Homeland Security Mayorkas could ratify a rule initially promulgated by Chad Wolf, who had been improperly serving as Acting Secretary.[96] Under the Federal Vacancies Reform Act ("FVRA"), ratification is prohibited for actions "taken by any person who is not acting under" the FVRA "in the performance of any function or duty of a vacant office."[97] The FVRA defines "function or duty" as: (1) a function or duty "established by statute[,]" and (2) a function or duty "required by statute to be performed by the applicable officer (*and only that officer*)[.]"[98] The dispute was whether "function or duty" applied only to expressly nondelegable duties.[99]

The majority held that "function or duty" applies only to duties expressly made nondelegable by statute, so the ratification was lawful.[100] The concurrence and dissent found the phrase ambiguous and favored a broader reading, noting that most federal duties are delegable and that Congress did not include the term "delegate" in the statutory definition despite using it elsewhere in the FVRA.[101]

---

94. *See id.* at 1045–46.

95. For the raw data, see Jonathan Scher, gpt4_logprobs_100prompts.xlsx, https://github.com/jscher07/llm-legal-ambiguity-replication (Jonathan Scher, LLM Legal Ambiguity Replication, GitHub) [hereinafter Data].

96. Gonzales & Gonzales Bonds & Ins. Agency, Inc. v. U.S. Dep't of Homeland Sec., 107 F.4th 1064, 1078 (9th Cir. 2024).

97. 5 U.S.C. § 3348(d)(1).

98. *Id.* §§ 3348(a)(2)(A)(i–ii) (emphasis added).

99. *Gonzales*, 107 F.4th at 1090 (Christen, J., dissenting).

100. *Id.*

101. *Id.* at 1081 (Johnstone, J., concurring); *id.* at 1090–91 (Christen, J., dissenting).

Across 100 paraphrased prompts that included the statutory definition, GPT-4 Turbo returned confidence scores almost exclusively at 100, with the lowest at 99.79.[102] This indicates that the model treated the phrase as unambiguously referring to nondelegable duties. The output showed no prompt sensitivity: the model's position was consistent and maximally confident across all variations. However, the model also failed to register the interpretative division reflected in the concurrence and dissent. The absence of any recognition of the alternative interpretations in the case record indicates that the model's agreement may be superficial, omitting important competing arguments that informed judicial disagreement.

3. *Leuthauser v. United States*

In *Leuthauser*, the Ninth Circuit considered whether Transportation Security Officers ("TSOs") are "investigative or law enforcement officers" under the Federal Tort Claims Act (FTCA), which waives sovereign immunity for certain intentional torts committed by such officers.[103] The FTCA defines an investigative or law enforcement officer as "any officer of the United States who is empowered by law to execute searches, seize evidence, or make arrests for violations of Federal law."[104]

The court held that TSO agents fall within this definition because they routinely conduct screenings, which courts have recognized as Fourth Amendment "searches."[105] Since executing searches is one of the enumerated functions, TSO agents qualify as officers under the FTCA, and the court found the statute's language unambiguous.

Across 100 paraphrased prompts containing the statutory definition, GPT-4 Turbo's confidence scores clustered near 0, 50, and 100, with a heavier concentration at 0. The average score was 31.71,[106] suggesting the model leaned toward "no" but with substantial variation across prompts. The bimodal distribution indicates high prompt sensitivity. This diverges from the court's holding, which held that TSA agents unambiguously qualify as officers of the U.S. due to their authority to conduct Fourth Amendment searches.

---

102. Data, *supra* note 95.
103. Leuthauser v. United States, 71 F.4th 1189, 1192 (9th Cir. 2023).
104. *Id.* at 1193.
105. *Id.* at 1197.
106. Data, *supra* note 95.

GPT-4 Turbo appears to have misread the statute by treating the three listed functions, "execute searches," "seize evidence," and "make arrests," as conjunctive rather than disjunctive, requiring all three to be satisfied for officer status. This error led it to hesitate in classifying TSA agents as officers, since they primarily perform searches and not the other enumerated functions.

GPT-4 Turbo's classification diverged from the Ninth Circuit's "not ambiguous" holding and was unstable across prompt variations. While the court relied on the disjunctive reading supported by other circuit decisions, the model's conjunctive misinterpretation and sensitivity to wording produced inconsistent and incorrect outcomes.[107] This raises an important question: Would GPT-4 Turbo interpret the officer definition differently if it were trained on lower court case law?

4. *M&T Farms v. Fed. Crop Ins. Corp.*

In *M&T Farms*, the Ninth Circuit considered whether a storefront partnership's marketing and selling of its partners' farm commodities constituted "farming activity" under a federal crop insurance policy.[108] The court found the term "farming activity" was genuinely ambiguous in this context and deferred to the Federal Crop Insurance Corporation's ("FCIC") interpretation that such activities fall within the definition.[109]

Across 100 paraphrased prompts, GPT-4 Turbo almost uniformly returned a "yes" response with an average confidence score of 98.47. Probabilities for alternative answers were negligible (all under 0.8%).[110] This indicates that the model treated the term as unambiguous and showed no prompt sensitivity.

GPT-4 Turbo's classification diverged from the Ninth Circuit's ambiguity finding. It appears to have relied on a rigid textualist reading of "farming activity" that naturally encompassed the sale of products grown by the seller, without considering administrative law principles, such as recognizing the FCIC's specialized expertise in administering federal crop insurance, or policy reasons, such as the need to prevent insurance fraud and "double dipping[,]" to ensure consistent eligibility criteria.[111] This case illustrates a limitation of LLM-based statutory interpretation: GPT-4 Turbo defaulted to a rigid textualist approach, treating the phrase as if it had a single, plain meaning, while the court engaged in an administrative law

107. *Id.*
108. M&T Farms v. Fed. Crop Ins. Corp., 103 F.4th 724, 730 (9th Cir. 2024).
109. *Id.* at 731.
110. Data, *supra* note 95.
111. *Id.* at 730.

analysis that accounted for agency expertise and policy considerations.

5. *Toledo v. Kolc*

In *Toledo*, the court held that formal charges were not required for a citizen to be extradited under the United States-Peru Extradition Treaty (the "Treaty").[112] In this case, Peru issued a formal accusation (Acusación Fiscal) against its former president, alleging multiple crimes.[113] The Treaty provides for the extradition of individuals who have been "*charged with* . . . the commission of an extraditable offense" by authorities in the requesting state.[114] To determine the meaning of "charged with," the court relied on dictionary definitions of "charge," similar extradition treaties, and other provisions within the Treaty.[115] Ultimately, the court held that "charged with" does not require the issuance of formal charges, meaning that an Acusación Fiscal was sufficient to satisfy the Treaty's requirement.[116]

GPT-4 Turbo was prompted with 100 variations of the following question:

> Does the term "charged with" in an extradition treaty include non-formal charges? Answer with only a number from 0 to 100 representing a confidence level. 0 being "no" and 100 being "yes." Don't let your prior responses influence your future responses. Only answer with a number from 0 to 100.

GPT-4 Turbo responded with a similar distribution of responses as in *Buero* and *Leuthauser*, with clusters around confidence levels of 0, 50, and 100.[117] Interestingly, this time there was another cluster of confidence scores between 50 and 100, a pattern that was absent in *Buero* and *Leuthauser*. The average confidence score was 43.12, suggesting GPT-4 Turbo found the phrase "charged with" to be ambiguous over whether it included informal "accusations" from prosecutors.

GPT-4 Turbo's response differed from the court, which found the phrase "charged with" to unambiguously include informal charges issued in an Acusación Fiscal.[118] The court came to this conclusion because it did not rely solely on the dictionary meaning of "charge" but instead examined the treaty's structure, similar agreements, and how different legal systems define a "charge."[119] GPT-4 Turbo's tendency to find ambiguity in cases when it is

112. Toledo v. Kolc, 64 F.4th 1106, 1110 (9th Cir. 2023) (emphasis added).
113. *Id.*
114. *Id.* at 1109.
115. *Id.* at 1109–10
116. *Id.* at 1111.
117. Data, *supra* note 95.
118. *Id.* at 1109–11.
119. *Id.*

presented with the term in isolation illustrates a potential limitation of LLMs in legal reasoning. Unlike courts, which engage in a dynamic process of statutory interpretation, precedent analysis, and policy consideration, LLMs often default to rigid textualist readings unless explicitly instructed to weigh contextual factors and the structure of the statute or legal text in which it is situated.

6. *United States v. Paulson*

In *Paulson*, an estate owed federal estate taxes after the original owner died; it paid some taxes and arranged to pay the remainder in installments.[120] Later, the estate missed payments, and the trust assets were eventually distributed to beneficiaries with taxes still unpaid.[121] 26 U.S.C. § 6324(a)(2) makes beneficiaries personally liable for estate taxes on property that is "receive[d], or ha[d] on the date of the decedent's death."[122] The case centered on whether the limiting phrase "on the date of the decedent's death" modifies only the immediately preceding verb "has," or also modifies the more remote verb "receives."[123] The majority held that the limiting phrase modifies only "has," meaning the statute imposes personal liability on those who either (1) have estate property on the date of the decedent's death, or (2) receive estate property at any time (whether on or after the decedent's death).[124]

GPT-4 Turbo was prompted with 100 variations of the following question, keeping the statutory definition constant:

> Under estate law, certain individuals—such as a spouse, trustee, surviving tenant, or beneficiary—may be personally liable for unpaid estate tax if they "receive[], or ha[ve] on the date of the decedent's death" property included in the estate. Does the phrase "on the date of the decedent's death" apply to both "receive" and "have"? Answer with a number from 0 to 100 representing a confidence level. 0 being "no" and 100 being "yes." Don't let your prior responses influence your future responses.

When prompted about whether the timing phrase "on the date of the decedent's death" in 26 U.S.C. § 6324(a)(2) applies to both "receives" and "has," GPT-4 Turbo overwhelmingly responded with "yes," indicating strong agreement that the timing phrase applies to both verbs.[125]

This assessment stands in stark contrast to the Ninth Circuit's ruling in

120. United States v. Paulson, 68 F.4th 528, 532–33 (9th Cir. 2023).
121. *Id.* at 533.
122. *Id.* at 536.
123. *Id.*
124. *Id.*
125. Data, *supra* note 95.

Paulson, which held that the timing phrase "on the date of the decedent's death" modifies only the immediately preceding verb "has" and not the more remote verb "receives."[126] Applying the rule of the last antecedent, which says that when a phrase modifies a list of items, it typically only applies to the last item in the list, the court concluded that the provision imposes personal liability for unpaid estate taxes on individuals who either (1) have estate property on the date of the decedent's death, or (2) receive estate property at any time after the decedent's death.[127]

The *Paulson* ruling highlights an aspect GPT-4 Turbo likely did not emphasize as much as the court: the statute's punctuation. The court emphasized that the comma separating "receives" from the phrase "on the date of the decedent's death" supports a reading where the modifier applies only to "has."[128] Further, the court's decision was guided by policy considerations which GPT-4 Turbo likely did not consider (or at least emphasize as much). For example, the court held that its interpretation was guided by the necessity of ensuring effective tax collection and avoiding loopholes that would allow estate property recipients to evade liability.[129] This case demonstrates a recurring limitation of LLM-based statutory interpretation: a tendency to rigidly apply textualist logic while overlooking structural, grammatical, and policy-based reasoning that courts use in practice.

7. *United States v. Scheu*

In *Scheu*, a defendant dragged a woman thirty-five to forty feet into a cornfield, where she could not be seen by cars passing by, and sexually assaulted her.[130] The question in this case centered around whether a sentencing enhancement should apply for abduction.[131] The court concluded that from the facts of the case, there was an abduction based on the plain meaning of the word "abduct."[132]

126. *Id.* at 537.
127. *Id.*
128. *Id.*
129. *Id.* at 545.
130. United States v. Scheu, 75 F.4th 1126, 1128 (9th Cir. 2023).
131. *Id.* at 1127.
132. *Id.* at 1130.

GPT-4 Turbo was prompted with 100 variations of the following question:

> A defendant dragged an individual 35 to 40 feet into a cornfield before sexually assaulting her. Based on the plain meaning of the word "abduct," did the defendant abduct her? Answer with a number from 0 to 100 representing a confidence level. 0 being "no" and 100 being "yes."

Nearly every prompt gave a confidence level score of 100. The average confidence level was 99.46. There were only two outliers, with confidence levels of 84.57 and 77.61.[133] This aligns with the court's holding that the woman was abducted based on the plain meaning of the word "abduct."

8. *United States v. Trumbull*

In *Trumbull*, an individual was indicted under 18 U.S.C. § 922(g)(1) for being a felon in possession of a firearm.[134] U.S.S.G. § 2K2.1(a)(4)(B) sets the base offense level for felon-in-possession violations at twenty if the offense involved a "semiautomatic firearm that is capable of accepting a large capacity magazine."[135] The U.S. Sentencing Commission ("USSG") defined "large capacity magazine" as a magazine that can accept more than fifteen rounds of ammunition.[136] The majority held that "large capacity magazine" was ambiguous because "large" is a comparative term that admits of degree and may vary depending on context.[137] Since the phrase was ambiguous, it deferred to the USSG's interpretation of "large capacity magazine" under *Kisor* deference.[138] The concurrence found the phrase to be unambiguous because the context makes clear that "magazine" refers to ammunition devices, not publications, leaving only one meaning and making the term vague rather than ambiguous.[139]

GPT-4 Turbo was prompted with 100 variations of the following question:

> The defendant was arrested while in possession of a Glock 17 loaded with 17 rounds. Is this a "semiautomatic weapon"that is "capable of accepting a large capacity magazine?" Answer with a number from 0 to 100 representing a confidence level, 0 being "no" and 100 being "yes." Don't let your prior responses influence your future responses.

Interestingly, GPT-4 Turbo responded with a confidence score of 100

133. Data, *supra* note 95.
134. *Id.*
135. *Id.* at 1117.
136. *Id.*
137. *Id.* at 1118.
138. *Id.*
139. *Id.* at 1121 (Bea, J., concurring). A term is vague when its "unquestionable meaning has uncertain application to various factual situations." Scalia & Garner, *supra* note 25, at 32.

to every prompt,[140] suggesting it could consistently apply the term to specific facts of the case. This supports the concurrence's view that the term, while perhaps vague in application, has a sufficiently clear meaning that courts (or AI systems) can apply it without deferring to agency interpretation, rather than the majority's conclusion that the term's ambiguity requires *Kisor* deference.

C. DISCUSSION

GPT-4 Turbo's ambiguity assessment differed from the court's majority opinion in four out of eight cases (*Leuthauser*, *M&T Farms*, *Toledo*, and *Trumbull*). Even when GPT-4 Turbo agreed with the court on the lack of ambiguity (*Gonzales* and *Paulson*), it sometimes reached different outcomes or failed to recognize interpretative divisions that the court acknowledged, suggesting the model came to its conclusions through different reasoning. In *Gonzales*, the model reached the majority's outcome but failed to register the interpretative division reflected in the concurrence and dissent. In *Paulson*, despite finding no ambiguity, GPT-4 reached a different holding because it did not consider canons of construction, including the rule of the last antecedent.

These two cases illustrate a glaring issue of relying on an LLM for statutory interpretation. Presented with one hundred paraphrased versions of the legal question, GPT-4 Turbo often produced confidence scores clustering around 0, 50, and 100, with an average of 66.07 in *Buero*; clustering near 0, 50, and 100, with heavier concentration at 0 and an average of 31.71 in *Leuthauser*; and clustering around 0, 50, and 100 (with an additional cluster between 50 and 100) and an average of 43.12 in *Toledo*. This pattern of widely scattered responses across extreme confidence values suggests high prompt sensitivity: small changes in question phrasing between each paraphrased prompt pushed the model toward opposite confident answers. While the average scores in these cases might suggest alignment with the courts' ambiguity findings, the bimodal distributions indicate that GPT-4 Turbo's "agreement" with the court is unreliable and unstable. The model's "agreement" with the court on ambiguity was achieved, but for unreliable reasons: the underlying pattern reflects prompt-driven swings, not consistent recognition of genuine ambiguity.

More glaringly, the distribution of responses that GPT-4 Turbo gave does not reflect real-world responses from surveys of humans. The distributions in *Buero*, *Leuthauser*, and *Toledo* show responses clustered at extreme values (0, 50, and 100) rather than the gradual distribution that one would expect from human respondents. In the real world, humans assessing

140. Data, *supra* note 95.

ambiguous statutory language would typically produce a bell curve. They would give responses between 0 and 50 if they were slightly confident that the answer was "no" or between 50 and 100 if they were slightly confident the answer was "yes." However, in the LLM responses, this nuance is absent. Almost all the data points are at or around 0, 50, and 100. This finding counsels against the assumption that LLMs capture "how normal people use language in their everyday lives."[141]

III. THEORETICAL CONCERNS WITH JUDICIAL LLM USE

The experimental results reveal a troubling pattern: in 50% of cases, GPT-4's assessment of textual ambiguity diverged from judicial determinations. Even more concerning, in two of the five cases in which the model's ambiguity assessment aligned with the court, its interpretation of the text differed.

A. *EX PARTE* COMMUNICATION PROBLEMS

Beyond the empirical concerns, a more fundamental, threshold legal question must be addressed: whether judicial use of LLMs constitutes prohibited *ex parte* communication. While the empirical evidence suggests LLMs may provide unreliable guidance, the legal system's rules about judicial communication could bar their use entirely.

Under the Code of Conduct for U.S. Judges Canon 3(A)(4), judges are prohibited from initiating, permitting, or considering any *ex parte* communications.[142] The rationale behind the prohibition is that *ex parte* communications often cause one party to feel that their opponent gained an unfair advantage, and this undermines the integrity of the judicial process.[143] For this reason, courts have held that *ex parte* contacts that create an appearance of impropriety are prohibited even if they do not result in any actual prejudice to the parties.[144]

The Code of Conduct does not define the term *ex parte* communication. On one hand, as it is traditionally understood, an *ex parte* communication is one that is made by or for one party alone outside the presence of the parties or their lawyers in a pending matter.[145] For example, a judge may not reach out to the District Attorney's office for advice on a pending criminal prosecution. On the other hand, the drafters of the Model Code of Judicial

141. Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1229 (11th Cir. 2024) (Newsom, J., concurring).

142. CODE OF CONDUCT FOR U.S. JUDGES Canon 3(A)(4) (Jud. Conf. U.S. 2019).

143. Cynthia Gray, *Resisting Ex Parte Temptation*, 57 CT. REV. 220, 221 (2021).

144. *Id.*

145. Keith R. Fisher, *New ABA Ethics Opinion Explores the Prohibition on Independent Fact Research by Judges*, NAT'L CTR. FOR STATE CTS.

Conduct (the original source of the Code of Conduct) may have intended for a broader definition of *ex parte* communication by making the prohibition applicable to contacts with law professors and any other person that is not a participant to a proceeding.[146]

Ex parte contacts are not allowed, regardless of how well-intentioned and conscientious the judge may be.[147] For example, in *In re Kaufman*, a judge improperly placed an *ex parte* call to a medical center in an interpleader action involving the disbursement of an insurance payment to an infant who was injured in a motor vehicle accident.[148] The judge argued that he phoned the president to speed along the proceeding.[149] However, the West Virginia Supreme Court of Appeals rejected the judge's argument, holding that "[a]lthough judicial economy is a worthwhile goal, the cost is too great when the integrity of the judicial process is called into question . . . the end does not justify the means."[150] This result suggests that if generative AI is considered an improper *ex parte* communication, judges may not defend their use of it by arguing that they were well-intentioned or cautious in doing so.

If *ex parte* contacts are understood under the broader interpretation, as any communication by a judge to a third party in a pending case, then consulting GenAI is likely an *ex parte* contact for several reasons. First, unlike traditional tools of legal research like the Internet (which are not *ex parte* contacts), LLM outputs are personalized. It is highly unlikely that two people will get the same output from an LLM, even when feeding it the same prompts. Most Internet articles, however, are static, and even when they are updated, everyone accessing them sees the same updates. Additionally, consulting an LLM is like engaging in a conversation with someone: LLMs get prompted with questions, and they respond to them in a conversational manner. Finally, communicating with an LLM likely does not fall under any of the exceptions to the prohibition against *ex parte* contacts: they are not disinterested "experts on the law" or court personnel such as law clerks that a judge may permissibly consult *ex parte* regarding a pending case.[151]

146. *See* MODEL CODE OF JUD. CONDUCT r. 2.9 cmt. at 3 (A.B.A. 2007).
147. *In re* Kaufman, 416 S.E.2d 480, 485 (W. Va. 1992).
148. *Id.* at 481.
149. *Id.* at 485.
150. *Id.*
151. MODEL CODE OF JUD. CONDUCT r. 2.9(A)(2) and 2.9(A)(3)

Under Canon 3(A)(4)(c), a judge may obtain the written advice of a disinterested *expert on the law*, outside the presence of the parties, as long as the judge affords the parties a reasonable opportunity to object and respond to the solicited advice.[152] Neither the Code of Conduct nor the Model Code define the phrase, "expert on the law." However, case law is instructive. In a Florida Supreme Court case, *In re Baker*, computer consultants were not considered legal experts even though they were solicited by the judge to answer an arguably legal question—the calculation of damages.[153] Similarly, in the Tennessee Supreme Court case, *Holsclaw v. Ivy Hall Nursing Home, Inc.*, a university department director whom the judge solicited for information about the job responsibilities of rehabilitation counselors was not an expert on the law.[154] Law professors, on the other hand, have been considered experts on the law.[155] From these cases, it is likely that non-legal AI tools such as ChatGPT and Claude are not "experts on the law" because they are not trained on a sufficient corpus of case law. Unlike law professors, these programs are not likely to accurately recite the law of a given jurisdiction. Non-legal AI is not trained on case law in subscription-based databases like Westlaw and Lexis and thus is likely to misstate the law of local jurisdictions.[156] Moreover, non-legal AI tools are not continuously kept up to date. They are only periodically re-trained on new data and thus may miss recent court cases that are publicly available. Even though LLMs have been used by courts to resolve legal issues, like statutory construction, the programs themselves are not legal experts.

One may argue that LLMs are not *ex parte* contacts because they are not sought out by a judge to benefit one party. For example, in *Snell*, Judge Newsom argued that he used LLMs as a tool to resolve a statutory construction problem that could not be solved by dictionaries alone. He did not consult an LLM to skew the outcome of the case.[157] However, while a judge may have good intentions in consulting a generative AI tool, good intentions are not a defense against a claim of improper *ex parte* contact.[158] What is relevant is whether the *ex parte* contact creates an appearance of impropriety such that it upsets the integrity of the judicial process.[159]

---

152. CODE OF CONDUCT FOR U.S. JUDGES, Canon 3(A)(4)(c) (emphasis added).
153. *In re* Baker, 813 So.2d 36, 37, 37 n.2 (Fla. 2002).
154. Holsclaw v. Ivy Hall Nursing Home, Inc., 530 S.W.3d 65, 71–72 (Tenn. 2017).
155. Time Warner Entertainment Co. v. Baker, 647 So.2d 1070, 1071 (Fla. Dist. Ct. App. 1994).
156. *See Legal AI vs. ChatGPT: What's The Difference?*, LEXISNEXIS (Dec. 4, 2024), https://www.lexisnexis.com/blogs/sg-lnlp/b/ai/posts/legal_2d00_ai_2d00_vs_2d00_chatgpt [https://perma.cc/SU5Q-QJZL].
157. *See* Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1223–25 (11th Cir. 2024) (Newsom, J., concurring).
158. *In re* Kaufman, 416 S.E.2d 480, 485 (W. Va. 1992).
159. *See* Gray, *supra* note 143, at 221.

The involvement of Gen AI and conflicted third parties in a pending case may create such an appearance of prejudice. For one, both Gen AI and conflicted third parties are biased, the former because it is not trained on representative data. Moreover, judges, like all humans, are prone to confirmation bias and may consult LLMs improperly such that it subverts the appearance of fairness. For example, they may consider Gen AI's responses when it aligns with their views on the merits of a case but disregard a Gen AI's output when it goes against their beliefs. Finally, a judge's use of an LLM may create the appearance of impropriety if people do not trust AI to make judicial decisions. Human adjudications, for all their flaws, are viewed to be fairer than any machine-generated decision.[160]

B. AUTOMATION BIAS AND OVERRELIANCE RISKS

The divergence between LLM outputs and judicial outcomes becomes particularly concerning when considered alongside what we know about human interaction with automated systems. Research on automation bias suggests that judges, like other professionals, may be prone to over-relying on LLM interpretations even when they conflict with traditional legal analysis. Human overreliance on automated systems has long been observed. For example, studies have shown that pilots excessively trust autopilot mode, even when it malfunctions.[161] This results in monitoring failures. In one such study, while almost all students turned the automation off when it failed during a flight, almost half of the trained pilots did not turn it off even though performance of the malfunctioning autopilot was significantly worse than manual human control.[162] Two types of automation failures can occur: omission errors (the operator fails to notice a problem because the program did not inform them) and commission errors (the operator follows the directive of the program, even though it is an inappropriate one).[163] Applied to the LLM context, the most likely errors will be omission errors.

Human overreliance on automation may be due to heuristic bias—mental shortcuts that we unknowingly take to make quick decisions.[164] One such heuristic bias that humans commonly engage in is the availability heuristic: the tendency to view recent events as being more likely to occur than events in the distant past.[165] For example, if a coin is flipped heads twenty times in a row, we will mistakenly think that the next coin flip will

160. Benjamin Minhao Chen, Alexander Stremitzer & Kevin Tobia, *Having Your Day in Robot Court*, 36 HARV. J.L. & TECH. 128, 131 (2022).

161. Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse* 39 HUM. FACTORS 230, 239 (1997).

162. *Id.*

163. *Id.*

164. *Id.*

165. *See id.*

probably be heads, even though the probability of heads being flipped is fifty percent. Applied to the LLM context, legal decisionmakers may have excessive faith in the accuracy of outputs if there have been no recent hallucinations or inaccurate statements of law. This is problematic, as an overreliance on AI systems will cause users to pay less attention to (or completely disregard) evidence that contradicts an AI's output.

A possible example of where this risk arises is in situations like the *Snell* concurrence, where a judge's intuitive understanding of a word's meaning aligns with the LLM, despite conflicting alternative interpretations. Such scenarios present a subtle risk that decisionmakers may give undue weight to AI outputs that confirm preexisting intuitions without giving alternative interpretations due weight.

An excessive trust in automated systems has many negative downstream effects, one of which is skill deterioration.[166] For example, as LLMs are increasingly used for statutory interpretation, legal decisionmakers' ability and willingness to interpret the meaning of words and phrases absent the technology may suffer. This effect, known as cognitive complacency, may lead to even more trust in the AI system creating a "vicious" positive feedback circle.[167]

C. LLMs May Provide False Clarity When the Text is Ambiguous

Many have raised concerns about the lack of an objective benchmark for determining textual ambiguity.[168] What has not yet been considered is how the use of LLMs may inadvertently reduce findings of ambiguity in legal texts. While no binding precedent yet exists, Judge Newsom's concurrence in *Snell* provides a framework for thinking about this risk. In *Snell*, Judge Newsom confronted inconsistent dictionary definitions of "landscaping" and consulted ChatGPT as part of his interpretative process.[169] This methodological proposal raises prospective concerns: if courts facing conflicting traditional sources turn to LLMs for definitional clarity, they might bypass formal findings of ambiguity that would be made in the absence of an LLM. This raises a critical question: would LLMs genuinely resolve ambiguities in these situations or would they simply provide an appearance of clarity that allows courts to avoid the rigorous analysis traditionally required when legal text is genuinely ambiguous?

---

166. *See* Parasuraman & Riley, *supra* note 161, at 243.

167. *Id.*

168. *See, e.g.*, Kavanaugh, *supra* note 29, at 2137–38.

169. Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1223 (11th Cir. 2024) (Newsom, J., concurring).

IV. COUNTERARGUMENT AND RESPONSE: THE "LESSER OF TWO EVILS"

Proponents of LLM use in statutory interpretation argue that despite their flaws, LLMs offer advantages over traditional tools. They contend that LLMs democratize legal interpretation by providing better notice to everyday citizens.[170] Anyone with internet access and a phone or laptop can quickly look up the meaning of words in a statute through ChatGPT. Looking up definitions in a dictionary is much harder, especially if they are not available online and if the text at issue is a phrase that requires parsing together several definitions.

While LLMs may democratize access to legal interpretation tools, it is questionable whether they actually provide better notice because their outputs vary based on their inputs. Everyday citizens, even those who are representing themselves in a case, may ask ChatGPT different questions than a judge when interpreting a statute. For example, an everyday citizen of average American intelligence who is curious about the meaning of the word "possession" would probably ask ChatGPT, "What does 'possession' mean?" However, a textualist judge interpreting the word "possession" in a statute would ask "What is the *ordinary meaning* of the word 'possession?' " These different prompts could lead to subtly different responses. Dictionary definitions, for all their flaws, at least provide static definitions unlike ChatGPT.

CONCLUSION

The use of LLMs in legal interpretation presents serious risks. First, LLMs do not reliably align with judicial determinations of ambiguity. The findings demonstrate that GPT-4 Turbo disagreed with judicial ambiguity assessments in 50% of cases, often misidentifying ambiguous language as clear and vice versa. This raises concerns that courts relying on LLMs may misinterpret statutory language or fail to recognize genuine ambiguity, ultimately distorting legal reasoning.

Second, judicial use of LLMs raises ethical and procedural concerns. LLMs may be considered an *ex parte* communication. Unlike traditional legal research tools, LLMs generate personalized, non-replicable outputs which cannot be meaningfully reviewed by opposing parties. This challenges basic principles of due process and judicial impartiality, as judges consulting

170. *See* Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1228 (11th Cir. 2024) (Newsom, J., concurring).

LLMs could unknowingly receive biased, inaccurate, or misleading legal interpretations.

Third, the psychological phenomenon of automation bias presents a risk of overreliance. Judges and legal professionals may default to LLM-generated interpretations without critically examining their reasoning or methodological limitations. The tendency of LLMs to produce authoritative yet unreliable responses increases the risk that courts may treat ambiguous legal texts as unambiguous and vice versa, undermining rigorous legal analysis and leading to potentially unjust outcomes.

Despite these risks, LLMs may have a limited role in legal interpretation, particularly as secondary tools to cross-check judicial reasoning or test alternative interpretations. However, their use should be approached with caution. Courts should establish clear ethical guidelines, ensure transparency in AI-assisted decision-making, and remain aware of the risks associated with bias, hallucination, and overconfidence in LLM outputs.

Ultimately, while LLMs may reshape how legal texts are analyzed, they should not dictate how they are interpreted. The ambiguity problem in statutory interpretation is complex, and blind deference to AI-driven analysis risks compromising judicial integrity. As the legal system navigates the challenges of integrating artificial intelligence, the core principles of fairness, impartiality, and reasoned decision-making remain paramount.