

Large Language Models for Legal Interpretation? Don't Take Their Word for It

BRANDON WALDON*, NATHAN SCHNEIDER**, ETHAN WILCOX***, AMIR ZELDES****, AND KEVIN TOBIA*****†

Recent breakthroughs in statistical language modeling have impacted countless domains, including the law. Chatbot applications such as ChatGPT, Claude, and DeepSeek—which incorporate “large” neural network-based language models (LLMs) trained on vast swathes of internet text—process and generate natural language with remarkable fluency. Recently, scholars have proposed adding AI chatbot applications to the legal interpretive toolkit. These suggestions are no longer theoretical: in 2024, a U.S. judge queried LLM chatbots to interpret a disputed insurance contract and the U.S. Sentencing Guidelines.

We assess this emerging practice from a technical, linguistic, and legal perspective. This Article explains the design features and product development cycles of LLM-based chatbot applications, with a focus on properties that may promote their unintended misuse—or intentional abuse—by legal interpreters. Next, we argue that legal practitioners run the risk of inappropriately relying on LLMs to resolve legal interpretive questions. We conclude with guidance on how such systems—and the language models which underpin them—can be responsibly employed alongside other tools to investigate legal meaning.

TABLE OF CONTENTS

INTRODUCTION	117
I. TEXTUALISM, EMPIRICISM, AND THE TURN TO LLMs	120
A. TEXTUALISM AND EMPIRICAL FACTS ABOUT MEANING	120

* Assistant Professor of Linguistics and Philosophy, University of South Carolina. © 2025, Brandon Waldon, Nathan Schneider, Ethan Wilcox, Amir Zeldes & Kevin Tobia.

** Associate Professor of Linguistics and Computer Science, Georgetown University.

*** Assistant Professor of Computational Linguistics, Georgetown University.

**** Associate Professor of Computational Linguistics, Georgetown University.

***** Professor of Law, Georgetown Law.

† For excellent editorial assistance, thanks to *The Georgetown Law Journal*, especially Trent Dowell, Daniel Gonzalez, Galen Green, Allyson Rosenblum, and Jordan Williams; for helpful comments, we also thank Cleo Condoravdi, Beth Levin, Emily Manetta, Madelaine O'Reilly-Brown, William D. Richard, Kevin Newsom, Jonathan Choi, and Jonathan Wroblewski, as well as audiences at George Mason University, the University of South Carolina, the 2025 North American Summer School on Logic, Language and Information (NASSLLI 2025), the 2025 Symposium on Legal Interpretation and Data at Georgetown Law (SOLID 2025), and the 2025 International Language and Law Association General Conference (ILLA 2025).

B.	TEXTUALISM’S METHODOLOGICAL REVISIONS	122
C.	THE RECENT TURN TO LLMS	125
II.	HOW LLMs WORK: LEGAL IMPLICATIONS OF MODEL DESIGN AND DEVELOPMENT	131
A.	WHAT IS AN LLM?	131
1.	Language Models: Probabilistic Text Generators	131
2.	Language Model Platforms: Usable Solutions for Interacting with Language Models.	134
3.	Distinguishing “Closed” vs. “Open” Source.	135
B.	FOUR MYTHS ABOUT LLMS	137
1.	Myth 1: LLM Training Data Reflects <i>Naturalistic</i> Language Use	138
2.	Myth 2: LLM Chatbots Are Optimized to Produce Correct Statements About Ordinary Language	144
3.	Myth 3: LLM Chatbots Are Capable of Producing Accurate <i>Abstractions</i> of Their Training Data	146
4.	Myth 4: LLMs Are Less Biased than Traditional Interpretive Tools	149
C.	STUDY: RED-TEAMING LLMS FOR BIASED ORDINARY MEANING JUDGMENTS.	153
III.	RECOMMENDATIONS AND BEST PRACTICES FOR USING LLMs TO AID LEGAL INTERPRETATION	160
A.	A GENERAL CHECKLIST FOR RESPONSIBLE LLM USE	161
1.	Use Open Models and Interfaces	161
2.	Document Interaction Settings	162
3.	Use Multiple Models and Prompts.	163
4.	Independently Verify Specific Claims in LLM Output.	164
5.	Recruit Domain Experts.	165
B.	TOWARDS DIALECTICAL LEGAL AI	165
1.	A Demonstration of Dialectical AI with “Physically Restrained”	167

2025]	LARGE LANGUAGE MODELS FOR LEGAL INTERPRETATION	117
	2. A Linguistic Perspective on “Physically Restrained”	172
	C. LESSONS FOR TEXTUALISM	180
CONCLUSION		183

INTRODUCTION

There is much hype surrounding “AI” and the law. Elon Musk announced that he is adding “all court cases” to the training set of “Grok,” his AI chatbot.¹ Musk proposed that Grok will render “extremely compelling legal verdicts.”² Skeptics might be surprised that similar proposals have come from beyond the tech industry’s enthusiasts. In the past year, legal scholars and judges have proposed using large language models (LLMs)³ to inform interpretation of contracts,⁴ statutes,⁵ and the Constitution.⁶ And a U.S. circuit court judge has twice used a chatbot to aid textualist interpretation of the “ordinary meaning” of an insurance contract⁷ and the U.S. sentencing guidelines.⁸ Many U.S. judges are textualists, and the use of chatbots as evidence of meaning is likely to grow. Our Article is the first to assess this emerging practice in equal measure from three disciplines—computer science, linguistics, and the law.⁹

1. Elon Musk (@elonmusk), X (Nov. 30, 2024, at 01:01 ET), <https://x.com/elonmusk/status/1862753703897219527> [<https://perma.cc/Q7ZP-L2HY>]; see Joe Patrice, *Elon Musk Feeds AI ‘All Court Cases,’ Promises It Will Replace Judges Because He’s an Idiot*, ABOVE THE L. (Dec 2, 2024, at 15:40 ET), <https://abovethelaw.com/2024/12/elon-musk-feeds-ai-all-court-cases-promises-it-will-replace-judges-because-hes-an-idiot> [<https://perma.cc/JK5X-FYNH>].

2. See Musk, *supra* note 1.

3. We provide a definition and in-depth explanation of LLMs and related concepts below. See *infra* Section I.A.

4. See generally Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451 (2024) (demonstrating how AI models can “ascertain ordinary meaning in context, quantify ambiguity, and fill gaps in parties’ agreements”).

5. See generally Christoph Engel & Richard H. McAdams, *Asking GPT for the Ordinary Meaning of Statutory Terms*, 2024 J.L. TECH. & POL’Y 235 (2024).

6. See generally Suzanne Monyak, *AI to Make Originalist Historical Analysis Easier, US Judge Says*, BLOOMBERG (Apr. 1, 2024, at 17:55 ET), <https://news.bloomberglaw.com/us-law-week/ai-to-make-originalist-historical-analysis-easier-us-judge-says> [<https://perma.cc/59VV-NK38>] (reporting that Judge John Bush suggested that AI can help with originalist analysis).

7. *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1221 (11th Cir. 2024) (Newsom, J., concurring).

8. *United States v. Deleon*, 116 F.4th 1260, 1270 (11th Cir. 2024) (Newsom, J., concurring).

9. One prior paper also assesses this emerging practice from the perspective of corpus linguistics, see Thomas R. Lee & Jesse Egbert, *Artificial Meaning?*, 77 FLA. L. REV. (forthcoming), although our evaluations diverge. Lee and Egbert focus primarily on a proposed distinction between “artificial meaning,” and empirical evidence of meaning. *Id.*; see also Lawrence B. Solum, *Artificial Meaning*, 89 WASH. L. REV. 69, 69–70 (2014) (investigating the “concept of artificial meaning”). A recently published student note also analyzes Judge Newsom’s two decisions. See Parker Miller, Note, *Extraordinary Meaning: Judge Newsom’s A.I. Experiments in Textualist Interpretation*, 9 GEO. L. TECH. REV. 538, 564 (2025). Our analysis in Part II and recommendations in Part III are distinct from these prior discussions.

This Article provides a detailed analysis of LLM chatbots in U.S. legal interpretation, but this issue and our analysis have much broader significance.¹⁰ Judges have also used LLM chatbots to interpret law in Colombia,¹¹ India,¹² and Pakistan,¹³ and for other judicial purposes in Argentina,¹⁴ Peru,¹⁵ Mexico,¹⁶ the Netherlands¹⁷ and the U.K.¹⁸ The practice is on the rise globally, and it will likely proliferate as legal-specific chatbots enter the market.¹⁹ Judges will continue to ask ChatGPT, Claude, Harvey AI and other chatbots about the “ordinary meaning” of legal texts, taking the chatbots’ responses as relevant evidence of such meaning.

Drawing on our team’s expertise in computer science, linguistics, and law, we explain this emerging practice from a technical, linguistic, and legal perspective. This perspective leads us to two overarching theses about LLMs and legal interpretation. First, a critical claim: *Judges should not rely on direct queries to ChatGPT (or similar chatbots) about the meaning of legal texts.* By “direct queries,” we mean questions such as, “What is the ordinary meaning of [term in a statute]?” Second, a constructive claim: *Judges could query LLMs as dialectical partners, in an effort to uncover overlooked relevant arguments about meaning.*

10. In this Article, we will use the term “LLM” to refer to both a class of statistical language models and the systems capable of generating text from such models. However, we will distinguish between LLMs and productized “chatbot” applications based on LLMs. Drawing this distinction will help to isolate risks associated with LLM-based legal interpretation in general and chatbot-based legal interpretation in particular.

11. See Juzgados Laborales de Circuito [Juzg. Circ.] [Labor Circuit Court], enero 30, 2023, Juan Manuel Padilla García, Sentencia No. 032 (Colom.), <http://forogpp.com/wp-content/uploads/2023/01/sentencia-tutela-segunda-instancia-rad.-13001410500420220045901.pdf> [<https://perma.cc/WZ6W-BF23>]. For a critique of these uses, see Juan David Gutiérrez, *ChatGPT in Colombian Courts*, VERFASSUNGSBLOG (Feb. 23, 2023), <https://verfassungsblog.de/colombian-chatgpt> [<https://perma.cc/HS3T-V9NW>].

12. See generally Pallavi Pundir, *This Court Used ChatGPT to Decide Bail in a Murder Case*, VICE (Mar. 29, 2023, 07:12 ET), <https://www.vice.com/en/article/india-court-chatgpt-bail-murder-case> [<https://perma.cc/XA6T-K9X9>].

13. See *State v. AM* (a juvenile), (2023) 5 (Pak.).

14. See Victoria Mendizabal, *Courts in Buenos Aires Are Using ChatGPT to Draft Rulings*, REST OF WORLD (Nov. 22, 2024), <https://restofworld.org/2024/buenos-aires-courts-adopt-chatgpt-draft-rulings> [<https://perma.cc/ASK3-2C75>].

15. See Juan David Gutiérrez, *Judges and Magistrates in Peru and Mexico Have ChatGPT Fever*, TECH POL’Y PRESS (Apr. 19, 2023), <https://www.techpolicy.press/judges-and-magistrates-in-peru-and-mexico-have-chatgpt-fever> [<https://perma.cc/BN6M-CRHK>].

16. See *id.*

17. See Linda A. Thompson, *Dutch Judge’s Use of ChatGPT Prompts Outrage, Disbelief in the Netherlands*, LAW.COM (Aug. 7, 2024, at 18:48 ET), <https://www.law.com/international-edition/2024/08/07/dutch-judges-use-of-chatgpt-prompts-outrage-disbelief-in-the-netherlands>.

18. See Gareth Corfield, *British Judge Uses ‘Jolly Useful’ ChatGPT to Write Ruling*, THE TELEGRAPH (Sep. 14, 2023, 22:14 BST), <https://www.telegraph.co.uk/business/2023/09/14/british-judge-uses-jolly-useful-chatgpt-to-write-ruling>. The practice has been approved by the Courts and Tribunals Judiciary. See Brian Melley, *Judges in England and Wales Are Given Cautious Approval to Use AI in Writing Legal Opinions*, ASSOCIATED PRESS (Jan. 8, 2024, 00:04 ET), <https://apnews.com/article/artificial-intelligence-ai-guidance-england-wales-judges-c2ab374237a563d3e4b56876955f7> [<https://perma.cc/C67S-92WR>].

19. See, e.g., HARVEY AI, <https://www.harvey.ai> [<https://perma.cc/DV4A-UX8C>] (last visited Sep. 15, 2025).

With appropriate caution about the reliability of synthetic text, LLMs can be utilized as tools—alongside traditional tools such as dictionaries, corpora, and surveys—in a way that retains judicial sovereignty over interpretive reasoning. We first offer a brief roadmap that highlights the Article’s novel contributions. Section I.A. introduces modern textualism. This theory is the basis for the Supreme Court’s resolution of dozens of blockbuster statutory cases each year.²⁰ The Justices are all committed to “textualism,”²¹ although they subscribe to subtly different versions of the theory.²² Nevertheless, they share a general commitment to interpret statutory language in line with what its text would communicate to an “ordinary reader.”²³ Next, Section I.B documents textualism’s recent methodological revisions. Today’s Supreme Court has questioned some of the theory’s traditional tools, like substantive and linguistic canons.²⁴ The Court has also been open to new methods to discern ordinary meaning, like corpus linguistics and surveys of ordinary Americans.²⁵ In light of these developments, scholars have suggested that courts also look to language models and chatbots like ChatGPT that are built on language models.²⁶ Section I.C introduces these scholarly proposals and the two U.S. examples in which a judge queried ChatGPT: *Snell*²⁷ and *Deleon*.²⁸

Part II develops and defends our primary critical thesis: Judges should not “directly query” LLM chatbots for judgments about meaning. We begin with a thorough explanation of what an LLM is in Section II.A. Next, in Section II.B, we articulate and debunk four myths about LLMs in interpretation. Finally, in Section II.C, we present an original empirical demonstration of “red-teaming” LLM chatbots to produce biased judgments of ordinary meaning. We demonstrate how extremely subtle, possibly unintentional, variations in a judge’s prompting of a chatbot can lead to opposite results. Chatbots have been framed as a way for textualists to move beyond dueling canons and cherry-picked dictionaries and to achieve greater predictability and objectivity. However, we argue, this possibility is undermined by the possibility of “prejudiced prompting.”

20. See generally Victoria Nourse, *The Paradoxes of a Unified Judicial Philosophy: An Empirical Study of the New Supreme Court: 2020-2022*, 38 CONST. COMMENT. 1 (2023).

21. See HARVARD LAW SCHOOL, *The 2015 Scalia Lecture | A Dialogue with Justice Elena Kagan on the Reading of Statutes*, at 08:21–08:30 (YouTube, Nov. 25, 2015), <https://youtu.be/dpEtszFT0Tg> (Justice Kagan says, “we’re all textualists now.”). But see *West Virginia v. EPA*, 597 U.S. 697, 779 (2022) (Kagan, J., dissenting) (“It seems I was wrong [that we are all textualists now].”). See generally Kevin Tobia, *We’re Not All Textualists Now*, 78 N.Y.U. ANN. SURV. AM. L. 243 (2023).

22. See generally William N. Eskridge, Jr., Brian G. Slocum & Kevin Tobia, *Textualism’s Defining Moment*, 123 COLUM. L. REV. 1611 (2023) (describing the “major categories of doctrinal and theoretical choices” defining modern textualists).

23. Anita S. Krishnakumar, *Textualism in Practice*, 74 DUKE L.J. 573, 579 (2024).

24. See *infra* Section I.B.

25. See *id.*

26. See Arbel & Hoffman, *supra* note 4, at 454–56; Engel & McAdams, *supra* note 5, at 237.

27. See *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1221 (11th Cir. 2024) (Newsom, J., concurring).

28. See *United States v. Deleon*, 116 F.4th 1260, 1270 (11th Cir. 2024) (Newsom, J., concurring).

Part III develops and defends a more constructive perspective. First, in Section III.A, we elaborate a checklist of responsible LLM use, itemizing practices that should be followed if a compelling use of LLMs in interpretation were developed. Next, in Section III.B, we develop a picture of how legal interpreters might employ AI tools to assist in interpretation. Finally, in Section III.C, we reflect more broadly on the implications that the Article’s analysis has for textualism.

I. TEXTUALISM, EMPIRICISM, AND THE TURN TO LLMs

A. TEXTUALISM AND EMPIRICAL FACTS ABOUT MEANING

In interpreting statutes, the Supreme Court Justices are all textualists.²⁹ Justice Scalia’s textualist theory,³⁰ once a minority view, is now the theory of interpretation applied in virtually *every* statutory case.³¹ Each year, there are dozens of such cases that reach the Supreme Court,³² impacting laws concerning the environment, health, education and student loans, crime, tax, and gun regulation, just to name a few. To understand the Supreme Court, you have to understand textualism.

Textualism might seem simple. Justice Scalia proposed the theory nearly thirty years ago, instructing judges to interpret a law in line with its *text*—and nothing else.³³ However, textualism is more complicated than it may first appear; today’s textualism is not identical to Justice Scalia’s, and the theory continues to quickly change.³⁴ To give just a few examples: today’s textualist Supreme Court has overruled the principle of *Chevron* deference,³⁵ paving the way for the courts “to determine the best reading” of ambiguous statutes—rather than deferring to the federal agency tasked with administering the statute;³⁶ today’s textualists are increasingly skeptical of substantive canons—like the rule of lenity or veteran benefit canon;³⁷ and they increasingly appeal to broader notions of

29. See HARVARD LAW SCHOOL, *supra* note 21.

30. See 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, 23–25 (Amy Gutmann ed., 1997); ANTONIN SCALIA & BRYAN A. GARNER, *READING LAW: THE INTERPRETATION OF LEGAL TEXTS* 41–42 (2012).

31. See Eskridge, Jr. et al., *supra* note 22, at 1615 (“In constitutional cases, there are intense debates between . . . textualist Justices and . . . pragmatic Justices . . . but in statutory cases, it is textualism all the way down.”).

32. See Anita S. Krishnakumar, *Cracking the Whole Code Rule*, 96 N.Y.U. L. REV. 76, 76 (2021) (analyzing 532 statutory interpretation cases decided during the Roberts’ Court’s first twelve-and-a-half terms).

33. See Scalia, *supra* note 30, at 23–25, 31–32.

34. See Richard M. Re, *Legal Realignment*, 92 U. CHI. L. REV. (forthcoming 2025) (manuscript at 19–20) (on file with authors). See generally Eskridge, Jr. et al., *supra* note 22.

35. See *Loper Bright Enters. v. Raimondo*, 603 U.S. 369, 412 (2024).

36. See *id.* at 373.

37. See *Wooden v. United States*, 595 U.S. 360, 379 (2022) (Kavanaugh, J. concurring); *Rudisill v. McDonough*, 601 U.S. 294, 314–18 (2024) (Kavanaugh, J., concurring). *But see Wooden*, 595 U.S. at 397 (Gorsuch, J. concurring).

context-sensitive interpretation.³⁸ There are other complexities: the textualist Justices often cite other non-textualist modalities, like consequences³⁹ and purpose.⁴⁰ Textualism in practice does not always mirror textualism in theory.⁴¹ And even the theory is more complicated: The textualist Justices employ different versions of textualism.⁴²

Nevertheless, there is some common ground among textualists. Broadly speaking, textualists seek to interpret statutes in line with what the statutory text would communicate to an ordinary or reasonable reader.⁴³ And although there is some scholarly disagreement about whether empirical facts about language are relevant to textualists' inquiries,⁴⁴ the vast majority of textualist scholars and judges understand the study of ordinary meaning and a statute's communicative content as a partly empirical question. For example, Professor Lawrence B. Solum describes a statute's communicative content as an empirical fact.⁴⁵ Professor Randy E. Barnett, writing about "the New Originalism"—a fusion of textualism and originalism, identifies "Originalism's Descriptive Claim" as the claim that "identifying original public meaning is an empirical inquiry."⁴⁶

In line with this view, current Supreme Court Justices have claimed that textualist practice should be empirically grounded. In writing about the linguistic canons, Justice Barrett notes that "[w]hether the canons actually capture patterns of ordinary usage is an empirical question. If they do not track common usage, then the textualist rationale for using them is undermined."⁴⁷ This line of argument also emerged in the *Facebook v. Duguid* oral argument from Justice Alito.⁴⁸ He noted in his concurring opinion that "interpretive canons attempt to identify the

38. See *Biden v. Nebraska*, 600 U.S. 477, 507–21 (2023) (Barrett, J., concurring). By contrast, "traditional textualists . . . flexibly contract interpretive contexts in ways that ignore relevant context." Kevin Tobia, Brian G. Slocum & Victoria Nourse, *Progressive Textualism*, 110 GEO. L.J. 1437, 1442 (2022).

39. See Nourse, *supra* note 20, at 2.

40. See generally Anita S. Krishnakumar, *Backdoor Purposivism*, 69 DUKE L.J. 1275 (2020).

41. Krishnakumar, *supra* note 23, at 591.

42. See Tara Leigh Grove, *Which Textualism?*, 134 HARV. L. REV. 265, 279 (2020); Eskridge, Jr. et al., *supra* note 22, at 1611 ("[T]he new textualist Court is riven with internal divisions and sends less-than-clear messages to the lower courts.").

43. See Frank H. Easterbrook, *The Role of Original Intent in Statutory Construction*, 11 HARV. J.L. & PUB. POL'Y 59, 65 (1988) ("The meaning of statutes is to be found . . . in the understanding of the objectively reasonable person."); SCALIA & GARNER, *supra* note 30, at 33 ("[A] reasonable reader, fully competent in the language, would have understood the text at the time it was issued."); John F. Manning, *The Absurdity Doctrine*, 116 HARV. L. REV. 2387, 2458 (2005) ("[T]extualists interpret statutory language by asking how 'a skilled, objectively reasonable user of words' would have understood the statutory text . . .").

44. See Tara Leigh Grove, *Testing Textualism's "Ordinary Meaning"*, 90 GEO. WASH. L. REV. 1053, 1073–74 (2022).

45. See Lawrence B. Solum, *Disaggregating Chevron*, 82 OHIO ST. L.J. 249, 265 (2021).

46. Randy E. Barnett, *The Gravitational Force of Originalism*, 82 FORDHAM L. REV. 411, 415 (2013) (alterations omitted).

47. Amy Coney Barrett, *Congressional Insiders and Outsiders*, 84 U. CHI. L. REV. 2193, 2204 (2017) (citation omitted).

48. See Transcript of Oral Argument at 11–12, *Facebook, Inc. v. Duguid*, 592 U.S. 395 (2021) (No. 19–511).

way in which ‘a reasonable reader, fully competent in the language, would have understood the text.’”⁴⁹ Since Barrett’s suggestion in 2017, scholars have attempted to empirically evaluate the canons in precisely this way.⁵⁰

These aspects of textualism—a focus on the “ordinary” reader and empirical facts about language—have been present from Scalia, Easterbrook, and Manning through the writings of modern textualist Justices like Barrett, Kavanaugh, and Gorsuch. One important change, however, has concerned which methodologies textualists adopt in their interpretive toolkit.

B. TEXTUALISM’S METHODOLOGICAL REVISIONS

Textualists—both traditional and modern—have sought to support the framework’s empirical enterprise with tools that promote values central to evidence-based inquiry: consistency, accuracy, and replicability. Consistency demands that interpretive methods yield similar results when applied by different judges to the same text. Accuracy requires that these tools faithfully capture ordinary language understanding. Replicability ensures that interpretive conclusions can be independently verified and tested against available linguistic evidence rather than resting on a judge’s idiosyncratic intuitions. These methodological commitments reflect the premise that statutory meaning depends in part on empirical facts to be discovered through systematic investigation, not merely on normative considerations to be resolved exclusively through judicial discretion.

Though textualists agree in broad strokes on the above point, one of the most striking recent developments in textualism is its shifting methodological toolkit. Traditional textualists, like Scalia, relied heavily on dictionaries and linguistic canons of interpretation. The newest textualists on the Court today have grown more skeptical about these traditional tools.

At the same time, judges have expressed interest in *new* tools, like corpus linguistics, surveys, and even ChatGPT. Corpus linguistics, “a linguistic methodology that analyzes language function and use by means of an electronic database called a corpus,”⁵¹ is the most established of these recent tools.⁵² About fifteen years ago, Stephen Mouritsen wrote a student note proposing the use of corpus linguistics in legal interpretation,⁵³ and Utah Supreme Court Justice Thomas Lee

49. *Facebook, Inc. v. Duguid*, 592 U.S. 395, 412–413 (2021) (Alito, J., concurring) (quoting SCALIA & GARNER, *supra* note 30, at 33).

50. See, e.g., Kevin Tobia, Brian G. Slocum & Victoria Nourse, *Statutory Interpretation from the Outside*, 122 COLUM. L. REV. 213, 224–25 (2022); Janet Randall & Lawrence Solan, *Legal Ambiguities: What Can Psycholinguistics Tell Us?*, in THE CAMBRIDGE HANDBOOK OF EXPERIMENTAL JURISPRUDENCE 1, 2 (Kevin Tobia ed., 2025). See generally Eric Martínez, *Traditional and Computational Canons*, 39 HARV. J.L. & TECH. (forthcoming 2026); Jonathan H. Choi & Nina A. Mendelson, *A Comprehensive Survey of Canon Usage at the Supreme Court, 1791-2025* (unpublished manuscript) (on file with authors).

51. Stephen C. Mouritsen, Note, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915, 1954 (2010).

52. For an introductory textbook on corpus linguistics, see DOUGLAS BIBER, SUSAN CONRAD & RANDI REPPEN, *CORPUS LINGUISTICS: INVESTIGATING LANGUAGE STRUCTURE AND USE* (1998).

53. See generally Mouritsen, *supra* note 51.

used corpus linguistics in a concurring opinion to inform the court’s analysis of the word “custody” in a federal statute.⁵⁴ Lee and Mouritsen published a manifesto on the use of corpus linguistics in law,⁵⁵ which inspired much debate and critique,⁵⁶ but also many uses in other courts.⁵⁷

Over the past fifteen years, dozens of judges have used or looked to corpus linguistics as evidence in legal interpretation.⁵⁸ The “law and corpus linguistics movement” largely proceeded from scholarly journals and lower courts to the Supreme Court, and there are now several examples of Supreme Court references to the tool.⁵⁹

Within the past year, the Supreme Court has further expanded its interpretive toolkit by considering another source of data: *survey evidence*. This suggestion had surfaced earlier at the Supreme Court. In the oral argument of *Facebook v. Duguid*, Chief Justice Roberts asked:

[O]ur objective is to settle upon the most natural meaning of the statutory language to an ordinary speaker of English, right? . . . So the most probably useful way of settling all these questions would be to take a poll of 100 ordinary—ordinary speakers of English and ask them what [the statute] means, right?⁶⁰

Last term, some of the Justices looked to survey evidence of ordinary meaning for the first time.⁶¹ Legal scholars have increasingly used survey methods to

54. See *J.M.W., III v. T.I.Z.* (In re Adoption of Baby E.Z.), 266 P.3d 702, 724 n.21 (Utah 2011) (Lee, J., concurring).

55. See generally Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 *YALE L. J.* 788 (2018).

56. See Anya Bernstein, *Democratizing Interpretation*, 60 *WM. & MARY L. REV.* 435, 452 (2018); Anya Bernstein, *What Counts as Data?*, 86 *BROOK. L. REV.* 435, 439 (2021); Anya Bernstein, *Legal Corpus Linguistics and the Half-Empirical Attitude*, 106 *CORNELL L. REV.* 1397, 1397 (2021); John S. Ehrett, *Against Corpus Linguistics*, 108 *GEO. L.J. ONLINE* 50, 64 (2019); Ethan J. Herenstein, *The Faulty Frequency Hypothesis: Difficulties in Operationalizing Ordinary Meaning Through Corpus Linguistics*, 70 *STAN. L. REV. ONLINE* 112, 113 (2017); Donald L. Drakeman, *Is Corpus Linguistics Better than Flipping a Coin?*, 109 *GEO. L.J. ONLINE* 81, 82 (2020); Stanley Fish, *The Interpretive Poverty of Data*, *BALKINIZATION* (Mar. 2, 2018, at 12:11 ET), <https://balkin.blogspot.com/2018/03/the-interpretive-poverty-of-data.html> [<https://perma.cc/3SRN-BTPD>]; Carissa Byrne Hessick, *Corpus Linguistics and the Criminal Law*, 2017 *BYU L. REV.* 1503, 1504 (2017); Brian G. Slocum & Stefan Th. Gries, *Judging Corpus Linguistics*, 94 *S. CAL. L. REV. POSTSCRIPT* 13, 13 (2020); Kevin P. Tobia, *Testing Ordinary Meaning*, 134 *HARV. L. REV.* 727, 746–747 (2020); Evan C. Zoldan, *Corpus Linguistics and the Dream of Objectivity*, 50 *SETON HALL L. REV.* 401, 403 (2019).

57. See generally Kevin Tobia, *The Corpus and the Courts*, 2021 *U. CHI. L. REV. ONLINE* 1 (2021) (describing best practices for use of corpus linguistics in court).

58. See *id.*, at 23–28.

59. See, e.g., *Carpenter v. United States*, 585 U.S. 296, 347 n.4 (2018) (Thomas, J., dissenting); *Moore v. United States*, 602 U.S. 572, 607 (2024) (Barrett, J., concurring).

60. Transcript of Oral Argument at 51–52, *Facebook, Inc. v. Duguid*, 592 U.S. 395 (2021) (No. 19–511).

61. See *Pulsifer v. United States*, 601 U.S. 124, 124–25 (2024). *Pulsifer* concerned the interpretation of 18 U.S.C.A. §3553(f)(1), which entitles certain criminal defendants to sentencing relief if “the defendant *does not have* (A) more than 4 criminal history points, excluding any criminal history points resulting from a 1-point offense, . . . (B) a prior 3-point offense, . . . and (C) a prior 2-point violent offense. . . .” 18 U.S.C.A. §3553(f)(1) (emphasis added). Writing as *amici curiae*, a team consisting of

inform statutory cases,⁶² and *Pulsifer* may portend a new inclusion of survey methods in certain debates about statutory meaning.

As courts look to new tools, like corpus linguistics and surveys, they also grow more skeptical of some old tools. Scholars have long critiqued textualists' heavy reliance on dictionaries.⁶³ Today, textualists and non-textualists alike recognize that the "dictionary is not a fortress."⁶⁴ Even Justice Thomas rejected a dictionary definition in *Diaz v. United States*, since "a word's meaning is informed by its surrounding context."⁶⁵ However, Thomas relied on other dictionary definitions later in the opinion.⁶⁶ To be sure, textualists still cite dictionaries frequently.⁶⁷ But the Court's use of dictionaries is evolving. Today, the Court is less likely to rely on one dictionary; instead, the Court frequently cites multiple dictionaries in defense of a linguistic conclusion.⁶⁸

A second traditional tool is the "linguistic canons of interpretation." These are rules of thumb that the Court uses as measures of statutory meaning.⁶⁹ The Court certainly cites these rules frequently, but some of the Justices have grown skeptical that the traditional canons are necessarily linguistically valid. For example, in *Facebook, Inc. v. Duguid*, Justice Alito's concurring opinion noted that "[linguistic] interpretive canons attempt to identify the way in which 'a reasonable reader,

professors of law and applied linguistics reported a survey experiment in which ordinary Americans were asked to interpret structurally analogous phrases of the form such as "does not have, A, B, and C." Brief of *Amici Curiae* Professors Thomas R. Lee et al. in Support of Neither Party at 15, *Pulsifer v. United States*, 601 U.S. 124 (2023) (No. 22-340) (quoting Kevin Tobia, Jesse Egbert & Thomas R. Lee, *Triangulating Ordinary Meaning*, 112 GEO. L.J. ONLINE 23, 25–26 (2023)). The results of this study were referenced by Justice Barrett in oral argument, as well as by Justice Gorsuch in his dissenting opinion. See Transcript of Oral Argument at 9, *Pulsifer v. United States*, 601 U.S. 124 (2024) (No. 22-340); *Pulsifer*, 601 U.S. at 163 (Gorsuch, J., dissenting).

62. See generally James A. Macleod, *Finding Original Public Meaning*, 56 GA. L. REV. 1 (2021) (reporting experimental survey studies asking participants to apply statutory language); Tammy Gales & Lawrence M. Solan, *Revisiting a Classic Problem in Statutory Interpretation: Is a Minister a Laborer?*, 36 GA. ST. U. L. REV. 491 (2020) (presenting a corpus linguistic study on understand of legal phrase "labor or service").

63. See generally Ellen P. Aprill, *The Law of the Word: Dictionary Shopping in the Supreme Court*, 30 ARIZ. ST. L.J. 275 (1998) (critiquing the Supreme Court's use of dictionaries); John Calhoun, *Measuring the Fortress: Explaining Trends in Supreme Court and Circuit Court Dictionary Use*, 124 YALE L.J. 484 (2014) (describing concerns with increased dictionary usage in federal appellate courts).

64. Mouritsen, *supra* note 51, at 1925.

65. *Diaz v. United States*, 602 U.S. 526, 536 (2024).

66. See *id.* at 537, 548.

67. Krishnakumar, *supra* note 23, at 651–79.

68. See Kevin Tobia, Brian G. Slocum & Victoria Nourse, *Ordinary Meaning and Ordinary People*, 171 U. PA. L. REV. 365, 443 (2023) (finding, for example, that "[i]n 45% of [Supreme Court] opinions citing a legal dictionary, an ordinary dictionary is also cited for the meaning of the same term defined by the legal dictionary.").

69. Barrett, *supra* note 47, at 2203–05. Examples of linguistic canons include the *ejusdem generis* canon. SCALIA & GARNER, *supra* note 30, at xiv ("Where general words follow an enumeration of two or more things, they apply only to persons or things of the same general kind or class specifically mentioned (*ejusdem generis*)."). Another example is the conjunctive/disjunctive canon. See *id.* at xii ("And joins a conjunctive list, or a disjunctive list—but with negatives, plurals, and various specific wordings there are nuances.").

fully competent in the language, would have understood the text.”⁷⁰ He proposed further that which of these canons is linguistically valid could be tested, in the future, through corpus linguistics.⁷¹ Scholars have taken up this project, suggesting that some canons—for example, *ejusdem generis*—reflect ordinary understanding well.⁷² For other canons, there is more debate. For example, writing as amicus curiae in *Campos-Chaves v. Garland*, a team of linguists identified that the conjunctive/disjunctive canon’s “negative proof” corollary—which states in part that “negative disjunctions” of the form “not A or B” are interpreted as not A and not B—obscures the fact that such phrases are inherently ambiguous and sometimes permit a not A or not B reading as a matter of ordinary interpretation.⁷³

As a final example, consider “substantive canons.” Justice Scalia endorsed the use of many of these,⁷⁴ but current textualists have grown much more skeptical.⁷⁵ Last year, Justice Kavanaugh wrote a concurring opinion, joined by Justice Barrett, critiquing the (substantive) veterans canon⁷⁶ as anti-textualist.⁷⁷

In sum, textualists are growing more skeptical of some “old tools,” like dictionaries, linguistic canons, and substantive canons. And they are turning to some new tools, like corpus linguistics and surveys. In light of these developments, it is sensible to ask whether other tools, such as LLM-based chatbots, might also feature in tomorrow’s textualism.

C. THE RECENT TURN TO LLMS

Against the backdrop of textualism’s methodological evolution, scholars and judges have been looking to artificial intelligence and machine learning tools. Recently, scholars have studied legal and ordinary language with word embeddings, numerical representations of words based on their usage in a large corpus.⁷⁸ Some even propose informing questions of ordinary meaning based on

70. 592 U.S. 395, 412–413 (2021) (Alito, J., concurring) (quoting SCALIA & GARNER, *supra* note 30, at 33).

71. *See id.* at 412.

72. Tobia, Slocum & Nourse, *supra* note 50, at 262.

73. *See* Brief for *Amici Curiae* Lucas Champollion et al. in Support of Petitioners at 33, *Campos-Chaves v. Garland*, 602 U.S. 447 (2024) (Nos. 22-674 & 22-884). For more discussion of the linguistic canons and their empirical limitations, see, for example, Tobia, Slocum & Nourse, *supra* note 50; Randall & Solan, *supra* note 50.

74. *See generally* SCALIA & GARNER, *supra* note 30.

75. *See* Amy Coney Barrett, *Substantive Canons and Faithful Agency*, 90 B.U. L. REV. 109, 110 (2010).

76. The veterans canon states that when a court interprets a law providing benefits to veterans, the court should resolve ambiguities in the law in favor of veterans.

77. *See* Rudisill v. McDonough, 601 U.S. 294, 318 (2024) (Kavanaugh, J., concurring) (joined by Barrett, claiming that “any canon that construes benefits statutes in favor of a particular group—rather than just construing the statutes as written—appears to be inconsistent both with actual congressional practice on spending laws and with the Judiciary’s proper constitutional role in the federal spending process”).

78. *See* Julian Nyarko & Sarath Sanga, *A Statistical Test for Legal Interpretation: Theory and Applications*, 38 J.L. ECON. & ORG. 539, 545–46 (2021).

these representations.⁷⁹ These proposals have been met with some calls for caution,⁸⁰ and at least some of the proponents of these tools also express similar caution about whether courts should use these tools today.⁸¹

Even more recently, there has been a movement to inform interpretation with “AI.” Two scholarly Articles propose this possibility.⁸² And recently, Judge Kevin Newsom, of the Eleventh Circuit, has twice used LLM-based chatbots (for example, ChatGPT) to aid the interpretation of the ordinary meaning of legal texts: in *Snell v. United Specialty Insurance Co.*⁸³ and in *United States v. Deleon*.⁸⁴

Snell concerned an insurance contract which was interpreted differently by the insurance company, “United,” and James Snell, the policyholder.⁸⁵ Snell was hired to perform various tasks for a family, including the installation of an in-ground trampoline, the installation of shrubs and trees, and the installation of a sprinkler system.⁸⁶ To install the trampoline, Snell dug a pit, installed a trampoline, and constructed a concrete block wall to prevent erosion and a wood cap around the walls of the pit.⁸⁷ A friend’s daughter was later injured on the trampoline, when she allegedly struck her face on the wooden panel surrounding the pit.⁸⁸

Snell had applied for and received an insurance policy from United.⁸⁹ The policy’s broad coverage was limited to “Specified Operations, Premises, or Projects;” and the relevant operation to this case was that the “insured performs landscaping.”⁹⁰ United claimed that the injury arose from the trampoline

79. For example, Professor Jon Choi used word embeddings to measure the similarity between terms (technically, “cosine similarity” scores, which measure the geometric distance between multidimensional corpus-derived numerical representations of words). See Jonathan H. Choi, *Measuring Clarity in Legal Text*, 91 U. CHI. L. REV. 1, 21–25 (2024).

80. For example, because word embedding models create de-contextualized representations of words, embedding-based analyses risk obscuring the role of context in ordinary linguistic interpretation. See Kevin Tobia, *Algorithmic Interpretation*, 91 U. CHI. L. REV. ONLINE 1, 8 (2024); see also Peter Henderson, Daniel E. Ho, Andrea Vallebuena & Cassandra Handan-Nader, *Corpus Enigmas and Contradictory Linguistics: Tensions Between Empirical Semantic Meaning and Judicial Interpretation*, 25 MINN. J.L. SCI. & TECH. 127, 131–32 (2024) (arguing that corpus-based approaches, including word embedding analyses, “may rely on foreign law to give meaning to U.S. constitutional or statutory provisions,” “may . . . be offering subjective or strategic forms of legislative history that textualists would disavow . . . for statutory interpretation,” and “may represent elite rhetoric, not ordinary original public meaning”). This Article also demonstrates that small changes in the composition of a corpus can have large downstream consequences for word embedding analyses based on that corpus. See Henderson et al., *supra* note 80, at 148–49.

81. See, e.g., Choi, *supra* note 79, at 23 n.85.

82. See Arbel & Hoffman, *supra* note 4, at 498; Engel & McAdams, *supra* note 5, at 296.

83. See 102 F.4th 1208, 1221 (11th Cir. 2024) (Newsom, J., concurring).

84. See 116 F.4th 1260, 1270 (11th Cir. 2024) (Newsom, J., concurring).

85. 102 F.4th at 1211.

86. *Id.*

87. *Id.* at 1211–12.

88. *Id.* at 1212.

89. *Id.* at 1212.

90. *Id.* at 1213.

installation, which is not *landscaping* and is therefore not covered.⁹¹ Snell argued that the injury stemmed from landscaping work, which was covered under the insurance policy.⁹²

For our purposes, the relevant topic from *Snell* is the debate about the ordinary meaning of *landscaping*.⁹³ Judge Newsom explores this question in his concurring opinion. He notes that he “spent hours and hours (and hours) laboring over the question whether Snell’s trampoline-installation project qualified as ‘landscaping’ as that term is ordinarily understood.”⁹⁴ Judge Newsom describes himself as a “plain-language guy,” an interpreter who prioritizes the ordinary meaning rule as “the most fundamental semantic rule of interpretation.”⁹⁵

After consulting dictionaries and other standard textualist sources, Judge Newsom considered results from inquiries to ChatGPT.⁹⁶ The first prompt was: “What is the ordinary meaning of ‘landscaping,’” to which ChatGPT replied:

“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 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.⁹⁷

Judge Newsom reports, in his concurrence, that this was “more sensible” than expected and also a response that “squared” with his impression of language.⁹⁸ He notes, “I definitely didn’t want to fall into the trap of embracing ChatGPT’s definition just because it aligned with my priors.”⁹⁹ He then reports a second question, “Is installing an in-ground trampoline ‘landscaping,’” to which ChatGPT replied:

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

91. *Id.*

92. *Id.*

93. An additional wrinkle in the case, which is not relevant to our primary linguistic concern, is Snell’s insurance application. The application asked “[d]o you do any recreational or playground equipment construction or erection?” Snell checked the “no” box. *Id.* The majority decision in *Snell* relies on this fact, and Alabama’s state law, which provides that “[e]very insurance contract shall be construed according to the entirety of its terms and conditions as set forth in the policy and as . . . modified by any rider, endorsement or application which is a part of the policy.” ALA. CODE § 27-14-17 (a) (1975); see *Snell*, 102 F.4th at 1217 (Newsom, J., concurring).

94. *Snell*, 102 F.4th at 1222 (Newsom, J., concurring).

95. *Id.* (quoting SCALIA & GARNER, *supra* note 30, at 69). See generally BRIAN G. SLOCUM, ORDINARY MEANING: A THEORY OF THE MOST FUNDAMENTAL PRINCIPLE OF LEGAL INTERPRETATION (2015).

96. See *Snell*, 102 F.4th at 1223–24 (Newsom, J., concurring).

97. *Id.* at 1225.

98. *Id.*

99. *Id.*

the outdoor environment, often aimed at enhancing the overall landscape and usability of the area.¹⁰⁰

Judge Newsom also notes that he posed the same questions to Google’s Bard chatbot,¹⁰¹ which reported “pretty similar” answers to his first question and a “more equivocal” answer to the second.¹⁰² In sum, “both models’ answers indicated that the trampoline-related work Snell had performed . . . just might *be* landscaping.”¹⁰³ From this exercise, Judge Newsom concludes that “[h]aving initially thought the idea [of using chatbots like this] positively ludicrous, I think I’m now a pretty firm ‘maybe.’”¹⁰⁴ The remainder of the concurring opinion discusses pros of LLM-based chatbots—they “train on ordinary-language inputs,” “can ‘understand’ context,” “are accessible,” allow “relatively transparent” research, and “hold advantages over other empirical interpretive methods” like surveys or corpus linguistics.¹⁰⁵ He also discusses potential drawbacks: LLM-based chatbots “can ‘hallucinate,’” “don’t capture offline speech,” may be subject to manipulation attempts, and may even “lead us into dystopia.”¹⁰⁶

To many, the *Snell* concurrence is shocking: A judge suggests informing a legal-interpretive question by asking ChatGPT! At the same time, the concurrence sounds notes of caution. Judge Newsom’s bottom line is that these tools’ “highest and best use is (like a dictionary) helping to discern how normal people use and understand language, not in applying a particular meaning to a particular set of facts to suggest an answer to a particular question.”¹⁰⁷

The second example came in *United States v. Deleon*.¹⁰⁸ Judge Newsom wrote a concurrence, a self-described “sequel of sorts” to his concurring opinion in *Snell*.¹⁰⁹ *Deleon* concerned the interpretation of the U.S. Sentencing Guidelines, which impose a two-level enhancement for armed robbery, “if any person was physically restrained to facilitate commission of the offense or to facilitate escape.”¹¹⁰ Joseph Deleon entered a store, asked the cashier for cigarettes and handed him \$10, then pointed his gun at the cashier as he demanded money from the register, received the money, and left. All of these events took place in about a minute.¹¹¹ The interpretive question is whether the cashier was “physically restrained” by Deleon to facilitate his commission of the armed robbery.¹¹² The majority opinion answered yes, on the basis of controlling Eleventh Circuit

100. *Id.*

101. Google has since rebranded Bard as Google Gemini.

102. *Snell*, 102 F.4th at 1225 n.4. (Newsom, J., concurring).

103. *Id.* at 1225.

104. *Id.*

105. *Id.* at 1226–30.

106. *Id.* at 1230–32.

107. *Id.* at 1232–33.

108. 116 F.4th 1260 (11th Cir. 2024).

109. *Id.* at 1270 (Newsom, J., concurring).

110. U.S. SENT’G GUIDELINES MANUAL § 2B3.1(b)(4)(B) (U.S. SENT’G COMM’N 2024).

111. *Deleon*, 116 F.4th at 1261–62.

112. *See id.* at 1263.

precedent,¹¹³ but two concurring judges noted that this line of cases and “[t]he issue in this case is ripe for en banc review.”¹¹⁴

Most relevant to our purposes is Judge Newsom’s concurring opinion on the ordinary meaning of *physically restrained*. Here again, Judge Newsom first asked ChatGPT, “What is the ordinary meaning of ‘physically restrained’?” to which it answered:

The ordinary meaning of “physically restrained” refers to the act of limiting or preventing someone’s movement by using physical force or some kind of device. This can involve holding someone down, using handcuffs, straps, or other mechanisms to restrict their ability to move freely. The term implies an active and physical intervention to control or confine a person’s movements.¹¹⁵

Here again, Judge Newsom notes that this response matches his priors.¹¹⁶ He also asked Claude the same question, twice, to which it gave two slightly different answers.¹¹⁷ Judge Newsom reports being shaken by this variability, so he then asked his question of “the three leading models—GPT, Claude, and Gemini—ten times apiece.”¹¹⁸ Judge Newsom reports that “reassuringly, the 30 results I received—10 apiece from each of the three leading LLMs—largely echoed the initial response that I got from ChatGPT.”¹¹⁹ The “gist,” writes Judge Newsom, is that:

When defining “physically restrained,” the models all tended to emphasize “physical force,” “physical means,” or “physical barriers.” ChatGPT and Claude specifically used one (or more) of those phrases in every one of their responses. For whatever reason, Gemini was a little different. It didn’t invariably employ one of those terms explicitly, but even when it didn’t, the concept of what I’ll call *corporeality* (via either human touch or a tangible object) pervaded and tied together its example-laden answers.¹²⁰

Judge Newsom concludes that:

I continue to believe—perhaps more so with each interaction—that LLMs have something to contribute to the ordinary-meaning endeavor. They’re not perfect, and challenges remain, but it would be myopic to ignore them. . . . [But] [w]e should give careful thought to how we assess and account for LLMs’ sometimes varying answers to user queries. . . . [and] I’m not advocating that we give up on traditional interpretive tools—dictionaries, semantic canons, etc. But I do think—and increasingly so—that LLMs may well serve a valuable auxiliary role as we aim to triangulate ordinary meaning.¹²¹

113. See *id.* at 1265; *United States v. Jones*, 32 F.3d 1512, 1518–19 (11th Cir. 1994); *United States v. Victor*, 719 F.3d 1288, 1289–90 (11th Cir. 2013).

114. *Deleon*, 116 F.4th at 1265 (Rosenbaum, J., concurring).

115. *Id.* at 1272 (Newsom, J., concurring).

116. *Id.*

117. See *id.*

118. *Id.* at 1273.

119. *Id.* at 1274.

120. *Id.*

121. *Id.* at 1277. On triangulating ordinary meaning, see generally Tobia et al., *supra* note 61.

These are very recent developments, and there are only a few reactions. Many have been positive, expressing excitement.¹²² There is only one other sustained scholarly analysis of these opinions, and it is more skeptical.¹²³ We discuss these critiques, as well as our own, throughout Parts II–III. First, in Part II, we take a deep dive into LLMs and the applications that make use of them, highlighting common language model “myths” which—once dispelled—problematize the use of these technologies in pursuit of ordinary linguistic meaning.

We pay particular attention to one proposed use case of LLMs, the one reported by Judge Newsom in his recent opinions: directly querying chatbots for linguistic judgments, for example, “Is installing an in-ground trampoline ‘landscaping’?”; “What is the ordinary meaning of ‘physically restrained’?”¹²⁴ On this “direct query” approach, the query is a question about language—a “metalinguistic” question—and the chatbot’s response is a similarly metalinguistic statement that answers the question; for example, “Yes, installing an in-ground trampoline can be considered a part of landscaping.”¹²⁵ The metalinguistic judgments produced by the chatbot are taken to be relevant to judicial decision-making.

This is just one possible use case, and it is not the only one on the market. However, we focus on this method for three reasons. First, it is the method defended by Judge Newsom.¹²⁶ Second, it is a relatively straightforward method which lawyers and judges might pursue as a naïve default strategy for interacting with an LLM. Finally, we hope that our evaluation of this method in Part II will help readers critically assess alternative proposals for applying LLMs to legal interpretation—to that end, we offer our own alternatives in Part III. Relatedly, many of the best practices we identify in Part III apply beyond the “direct query” method and thus can inform best practices for future methodologies.

122. See Joe Patrice, *Federal Judge Uses ChatGPT to Make His Decision and It Makes More Sense than You'd Think*, ABOVE THE LAW (June 5, 2024, at 12:44 ET), [https://abovethelaw.com/2024/06/federal-judge-chatgpt-opinion/\[https://perma.cc/QT2K-RSJM\]](https://abovethelaw.com/2024/06/federal-judge-chatgpt-opinion/[https://perma.cc/QT2K-RSJM]); Justin P. Gunter, *Using ChatGPT to Interpret Insurance Policies? Eleventh Circuit Opens the Door to AI's Role in Policy Interpretation*, BRADLEY (July 17, 2024), <https://www.itpaystobecovered.com/2024/07/using-chatgpt-to-interpret-insurance-policies-eleventh-circuit-opens-the-door-to-ai-s-role-in-policy-interpretation> [https://perma.cc/MFV4-887L]; Eugene Volokh, *Judge Suggests Courts Should Consider Using “AI-Powered Large Language Models” in Interpreting “Ordinary Meaning.”* REASON: THE VOLOKH CONSPIRACY (May 30, 2024, at 08:01 ET), <https://reason.com/volokh/2024/05/30/judge-suggests-courts-should-consider-using-ai-powered-large-language-models-in-interpreting-ordinary-meaning> [https://perma.cc/A55V-YAV2].

123. See Lee & Egbert, *supra* note 9, at 53; Isha Marathe, *Judges React to 11th Circuit's Gen AI Use: 'Creative,' Occasionally 'Misleading,' and 'Brave.'* LAW.COM: LEGALTECH NEWS (June 6, 2024, at 15:42 ET), <https://www.law.com/legaltechnews/2024/06/06/judges-react-to-11th-circuits-gen-ai-use-creative-occasionally-misleading-and-brave>.

124. *Snell*, 102 F.4th at 1225 (Newsom, J., concurring); *Deleon*, 116 F.4th at 1272 (Newsom, J., concurring).

125. *Snell*, 102 F.4th at 1225 (Newsom, J., concurring).

126. “For better or worse, I tend to default to asking, more generally, for a term’s ordinary meaning—probably because that’s what I’m doing when I consult dictionaries. If there’s a better way—a better question—I’m fine with that.” *Deleon*, 116 F.4th at 1274 n.6 (Newsom, J., concurring).

II. HOW LLMs WORK: LEGAL IMPLICATIONS OF MODEL DESIGN AND DEVELOPMENT

In this Part, we examine popular misconceptions about LLMs that, when properly understood, raise questions about their usefulness for textualist analysis. We start in Section II.A with a comprehensive overview of LLMs, including the essential structural properties of these models and the multiple training regimes employed to parameterize them. In Section II.B, we present four common myths about LLMs and discuss how these myths serve to oversell the potential of LLMs to triangulate ordinary meaning. In Section II.C, we present an empirical original study that underscores one risk of judicial reliance on direct queries of LLMs: their susceptibility to manipulation.

In our presentation of these LLM myths, we focus heavily on Judge Newsom's defense of the "direct query" method in *Snell* and *Deleon*, at times quoting liberally from his opinions in those two cases. Our intention is not to "pile on" the judge; we appreciate that he has staked a clear position which can serve as a point of reference for our critical analysis. Judge Newsom's experiments in *Snell* and *Deleon* are bold and novel applications of a paradigm-shifting technology to the field of law. As discussed below, we find many of Judge Newsom's pro-LLM arguments compelling even if the "direct query" method is, as currently practiced, deficient. Subsequent sections of this paper offer guidance for judges and scholars who, like Judge Newsom, are interested in the responsible use of LLMs for legal interpretation. First, however, it is crucial to understand the fundamentals of these models, including their design features and development lifecycles, in order to fully appreciate the risks and challenges presented by these technologies.

A. WHAT IS AN LLM?

LLMs are inextricably linked to chatbot applications such as ChatGPT, Claude, and recent competitors such as DeepSeek. However, chatbots are not equivalent to the LLMs that power them; rather, LLMs are just one part of a larger infrastructure involving multiple technological components. The interaction of these components enables the expressive "chat" behavior observed by the end user. Our critiques of the direct query method will involve a critical examination of each component.

There is substantial variation in how LLMs are discussed. To help clarify the terms of debate and set the stage for the subsequent arguments laid out in this Article, it will be useful to distinguish between two interrelated concepts: (1) (large) language models and (2) language model platforms.

1. Language Models: Probabilistic Text Generators

LLMs are part of a larger class of models known as *language models*.¹²⁷ A language model is any model that assigns a probability to a sequence of text—a

127. See generally Daniel Jurafsky & James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (Jan. 12, 2025) (unpublished manuscript) (on file with author); SOFIA SERRANO, ZANDER BRUMBAUGH & NOAH A. SMITH, *LANGUAGE MODELS: A GUIDE FOR THE PERPLEXED* (2023), <https://arxiv.org/pdf/2311.17301> [<https://perma.cc/PF8L-PQQ8>].

string.¹²⁸ LLMs assign string probabilities via the, ultimately simpler, task of assigning probabilities to *tokens*—individual words and sub-word units—given a preceding “context” sequence of tokens,¹²⁹ in a task known as *autoregressive* language modeling. For example, given the context “*For a long time I went to bed . . .*”, a well-tuned language model should assign a high probability to “*early*” or “*late*,” but should assign a low probability to “*crackers*” or “*Tolstoy*.” Autoregressive language modeling can be implemented in a variety of ways, but most modern applications do so with deep neural network architectures. A *neural network* is a model that relates an input to an output by passing the input through one or more levels of intermediate representation, known as “hidden layers.”¹³⁰ A “deep” neural network is a neural network with several hidden layers; the output from one intermediate layer becomes input for the subsequent layer.¹³¹ In the case of LLMs, the original inputs to our neural networks are vectors of numbers that represent the preceding text, and the final model outputs are vectors of numbers that represent the model’s probability distribution over possible next words.¹³² At each layer, the network changes the inputs according to a set of learned parameters, which we will refer to as *weights*. The network’s weights determine how the input is modified as it passes through the network and ultimately how successful the network is. By passing the inputs through several hidden layers as it makes a prediction, and thus changing the original input through updates by several sets of weights, the deep neural network can learn complex relationships between the inputs and the outputs.

The weights that constitute the model’s “knowledge” of language have to be “learned” via exposure to data, in a process known as “training.” At the beginning, a language model is initialized with random weights, and its predictions are rubbish. In order to make the network better, one shows the network several examples and adjusts its weights so that it can make better predictions for these examples.¹³³ What makes LLMs “large” is that they are deep neural networks with billions of such weights (or “parameters”) that have been shown trillions of such examples.¹³⁴ By adjusting the network’s weights in response to these training examples, the model has learned to become an excellent next-word prediction engine. But not only. It turns out that when trained to predict the next word, large enough networks exposed to enough text display a wide variety of emergent behaviors—for example, learning to follow various grammatical rules, and

128. See Jurafsky & Martin, *supra* note 127, at 32.

129. See *id.* at 4–30. A related type of language modeling involves assigning probabilities to words given their surrounding context; this is known as *masked* language modeling. For explanatory purposes, we will stick to discussing autoregressive language modeling.

130. See *id.* at 168.

131. See *id.* at 140.

132. See *id.*

133. See *id.* at 145.

134. See *id.* at 220.

representing a word differently when it is used with different meanings in different contexts.¹³⁵

In the rest of this discussion, we will refer to this base language model as the *pretrained language model*, to distinguish it from the final LLM, which involves several additional components discussed immediately below.

A raw, pretrained language model must be further refined in order to use it in chatbot applications, typically through two processes. In the first, called *instruction fine-tuning*, the model is further trained on datasets of carefully curated texts that exemplify instruction-response pairs.¹³⁶ This steers the model to produce outputs that satisfy user requests. Second, the model is subjected to a process called *preference alignment*. In preference alignment, the model is trained on datasets which couple instruction-response pairs with ratings of the responses.¹³⁷ These ratings reflect preferences elicited either from human annotators or from separate artificial systems. Training on preference datasets ideally guides the model to produce outputs that would receive high ratings from the annotator population, while suppressing low-rated outputs.

To generate text using a pretrained, fine-tuned, or preference-aligned LLM, one requires two crucial ingredients: an input string (a *prompt*) and a method of sampling from probability distributions over tokens to generate as output. As the linguist and AI researcher Chris Potts recently put it, “Once you choose [a prompt and a sampling strategy], you have a *system*.”¹³⁸ In this paper, we will use the terms “large language model” and “LLM” for both the mathematical objects which map text to probabilities, as well as for the *systems* capable of generating text from an underlying text-probability mapping. An LLM system, when fed an initial prompt, computes a distribution over likely next words.¹³⁹ A word is selected from this distribution and appended to the prompt. The whole thing is then fed back into the model. From here the loop repeats itself, with the model predicting new words which are then consumed as further context, in a process called *inference*.¹⁴⁰

Of course, the strategy one uses for selecting the next word from the language model’s distribution has a huge effect on the grammaticality, quality, and diversity of the generated text. There is no such thing as a single output from a

135. *See id.* at 202–23; *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1227–28 (11th Cir. 2024) (Newsom, J., concurring) (“So, for instance, modern LLMs can easily discern the difference—and distinguish—between the flying-mammal ‘bat’ that uses echolocation and may or may not be living in your attic, on the one hand, and the wooden ‘bat’ that Shohei Otani uses to hit dingers, on the other.”).

136. *See* Jurafsky & Martin, *supra* note 127, at 249–53.

137. *See id.* at 249. A prominent example of preference alignment is reinforcement learning from human feedback (RLHF). In RLHF, the model is given a prompt, or instruction, from which two completions are sampled. These completions are shown to human annotators, who rate which of the two they prefer. The preferences are used as a learning signal, to further update the model’s weights, guiding it to produce text that is highly rated by the population of annotators.

138. Christopher Potts (@ChrisGPotts), X (Sep. 13, 2024, at 13:07 ET), <https://x.com/ChrisGPotts/status/1834640151500538110> [<https://perma.cc/D82Q-Z3WP>] (emphasis added).

139. *See* Jurafsky & Martin, *supra* note 127, at 199–201.

140. *See id.* at 207–10.

language model; rather, outputs exist under a certain sampling strategy. If one keeps the language model the same, but varies the sampling strategy or the prompt, the result could be vastly different generated texts.¹⁴¹ The systems that power chatbot applications employ instruction-tuned, preference-aligned pre-trained LLMs, which are prompted with a user query plus a global *system prompt* (or *initial prompt*) hidden from the user.¹⁴² System prompts instruct the model to be helpful, unbiased, balanced, and non-offensive.¹⁴³

2. Language Model Platforms: Usable Solutions for Interacting with Language Models

LLMs, as we've defined them, specify solutions for generating text given a prompt, a sampling method, and a text-probability mapping. Developers implement systems using programming libraries which enable computers to efficiently perform the many billions of calculations required for training and inference.¹⁴⁴ Most users, however, encounter LLM-generated text from a further degree of separation from the model. We will use the term *platform* to refer to software which includes both an LLM and an interface that mediates between the user and the LLM.

The platforms most relevant to our discussion are chatbot applications that are readily usable without programming experience.¹⁴⁵ The user queries the platform via a chat window, and the platform returns a chat "response" that features text generated by the underlying LLM on the basis of the user query plus the global system prompt. Importantly, the chatbot's response may be subject to further refinement or filtering not observed by the user. For example, chatbot applications frequently feature *blacklists*, lists of words or phrases that the chatbot is not allowed to mention or substantively discuss, so as to prevent harmful, biased, or toxic outputs.¹⁴⁶ Subsequent chatbot responses reflect LLM sampling on the basis of a more elaborate prompt, which consists of the global system prompt, the user's most recent query, and all prior interactions with the model in the chat session.

We schematize the language model platform in [Figure 1](#).

141. *See id.* As demonstrated in Section II.C *infra*, this property of language models poses a particularly significant challenge for the direct query method pursued by Judge Newsom.

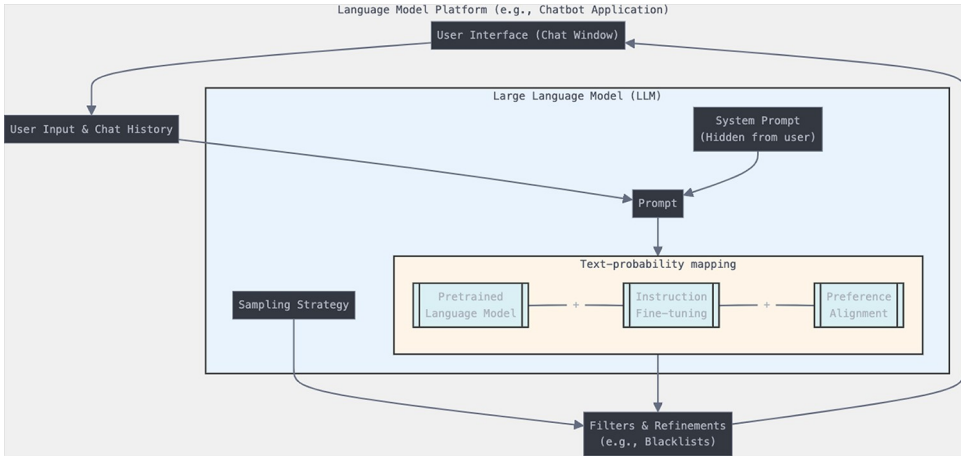
142. *See id.* at 323.

143. *See id.* at 211, 249, 323.

144. *See id.* at 229.

145. A second important variety of platform makes use of an Application Programming Interface (API), which provides programmers with a standardized set of computer-code commands for sending requests to and receiving responses from an LLM. In Section III.A, we advocate for the use of APIs as a general best practice for interacting with LLMs.

146. *See* Jurafsky & Martin, *supra* note 127, at 211, 249.

Figure 1: Schematic representation of a language model platform.

3. Distinguishing “Closed” vs. “Open” Source

Before we move on, we wish to highlight a key difference in how various LLMs (as well as associated platforms) are released to the public. In commercial settings, it is common to encounter chatbots built on *closed-source* data and technology.¹⁴⁷ Closed-source data is private and proprietary, in the sense that users are unable to view, modify, or distribute the sources of text used in training.¹⁴⁸ Portions of a closed-source training dataset may in fact be publicly available, but when a dataset is closed-source, users have no means of verifying the dataset’s composition. Similarly, developers of closed-source technology treat development procedures and implementation details as trade secrets.¹⁴⁹ Users can interact with the final product but cannot examine or build upon the core technology.¹⁵⁰

At the time of writing, chatbots developed by OpenAI, Google, and Anthropic are all closed source: we have no way of inspecting the code the developers wrote

147. See Rishi Bommasani et al., *The 2024 Foundation Model Transparency Index*, TRANSACTIONS ON MACH. LEARNING RSCH., Dec. 2024, at 8–9, <https://arxiv.org/pdf/2407.12929> [<https://perma.cc/G83S-Z7DE>]. The term *closed source* is the antonym of *open source*. In software development, *open source* describes software whose code is accessible to the public, with terms of distribution that allow anyone to examine, modify, and further disseminate the code. The Open Source Initiative (OSI), a California-based public benefit corporation, offers a widely cited definition of open-source software. See *The Open Source Definition*, OPEN SOURCE INITIATIVE (Feb. 16, 2024), <https://opensource.org/osd> [<https://perma.cc/239P-3W73>]. For the OSI’s definition of open-source AI, see *The Open Source AI Definition 1.0*, OPEN SOURCE INITIATIVE, <https://opensource.org/ai/open-source-ai-definition> [<https://perma.cc/ES72-HXHL>] (last visited Sep. 20, 2025).

148. *Closed-Source AI*, TENSORWAVE (July 31, 2024), <https://tensorwave.com/glossary/closed-source-ai> [<https://perma.cc/7XHQ-P35W>].

149. *See id.*

150. *See id.*

to produce the chatbot platform. Moreover, users of these platforms cannot fully inspect the underlying LLM due to access restrictions associated with the associated training data, the model parameters, the prompt, and/or the sampling strategy.

Alternatively, research groups may opt to release their projects as *open-source* contributions to the broader scientific or engineering community. *Open models* are models whose weights and/or training procedures are made public.¹⁵¹ At the time of writing, notable firms developing open models include Mistral AI, a French Artificial Intelligence company, and Meta.¹⁵² Finally, some groups also publicly release models' training data.¹⁵³ LLMs for which code, weights, and data are all publicly available are known as *open data models* and are considered the gold standard for open-access science. At the time of writing, the Allen Institute for AI, a US-based nonprofit, and EleutherAI, a non-profit AI research lab, are prominent research teams engaged in open model development.¹⁵⁴

The distinction between open vs. closed source is not categorical: for example, one could conceivably develop a chatbot platform which makes use of an open data model. If the code used to implement the chatbot platform—including the code that specifies the system prompt and LLM sampling strategy—is kept private, then the platform is closed but the underlying pretrained model is open. Similarly, an open-source platform may allow users to interact with proprietary, closed-source models. In this paper, we will advocate for fully open-source solutions for interacting with LLMs, as these solutions are the only ones that enable users to report LLM interactions in a fully transparent manner.¹⁵⁵

The preceding discussion should make it clear that LLM-based chatbots are not one single piece of technology. Rather, they link together several interrelated types of technologies, including, at their core, a language model trained on multiple separate sources of data.¹⁵⁶ Modern chatbot applications can even draw on other functionality, like querying the web or calling a math engine.¹⁵⁷ While each component part of the application is included to make it a more useful platform

151. See *The Open Source AI Definition*, *supra* note 147.

152. See Bommasani et al., *supra* note 147, at 5.

153. See Steve Lohr, *An Industry Insider Drives an Open Alternative to Big Tech's A.I.*, N.Y. TIMES (Oct. 19, 2023), <https://www.nytimes.com/2023/10/19/technology/allen-institute-open-source-ai.html>; Craig S. Smith, *EleutherAI: When Open AI Isn't Enough*, IEEE SPECTRUM (Mar. 21, 2022), <https://spectrum.ieee.org/eleutherai-openai-not-open-enough> [<https://perma.cc/NQ4W-6FTC>].

154. See Smith, *supra* note 153.

155. For readers seeking deeper technical knowledge about LLMs and related technologies, there are numerous high-quality resources, including ones geared to legal readers. See, e.g., Paul Ohm, *Focusing on Fine-Tuning: Understanding the Four Pathways for Shaping Generative AI*, 25 COLUM. SCI. & TECH. L. REV. 214 (2024). However, to fully appreciate the common misconceptions about LLMs discussed in this paper, it is essential to first understand their fundamental characteristics. This Section synthesizes the key properties of LLMs to provide the reader with this necessary foundation.

156. See *id.* at 231.

157. For discussion of querying capabilities, see Jurafsky & Martin, *supra* note 127, at 324. For demonstration of an LLM system that incorporates a math calculator, see Timo Schick et al., *Toolformer: Language Models Can Teach Themselves to Use Tools*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 36 (A. Oh et al. eds., 2023).

for responding to user inputs, each component is also a locus of objections to the direct query method as pursued by Judge Newsom. We present these objections in the following subsection, via four common myths about what LLMs are and how they function.

B. FOUR MYTHS ABOUT LLMs

Textualists interpret statutes, and some other legal texts, by interpreting the text through its linguistic meaning. Many textualists prioritize “ordinary” meaning, for example by presuming that terms in a statute express their ordinary, non-technical meanings, unless context indicates otherwise.¹⁵⁸ To a textualist judge vexed by thorny interpretive questions, LLMs may appear to be sages of the English language, capable of calculating an answer by weighing far more data than a human with limited reading capacity. Below, we argue that much of the optimism for turning to LLMs as “super-judges” of ordinary meaning rests upon fundamental myths about how these tools work. These myths include: (1) that LLM training data reflects naturalistic language use; (2) that LLM chatbots are optimized to produce correct statements about ordinary language; (3) that LLM chatbots are capable of formulating accurate generalizations of their training data; and (4) that LLMs are less biased than alternative interpretive tools.

In both *Snell* and *Deleon*, Judge Newsom explicitly invokes the Ordinary-Meaning canon: “[w]ords are to be understood in their ordinary, everyday meanings—unless the context indicates that they bear a technical sense.”¹⁵⁹ Quoting Scalia and Garner, Judge Newsom describes himself as “firmly of the view that ‘[t]he ordinary meaning rule is the most fundamental semantic rule of interpretation.’”¹⁶⁰ In *Snell*, Judge Newsom characterizes his use of ChatGPT as supplementing “initial efforts to pinpoint the ordinary meaning of the term ‘landscaping’”;¹⁶¹ similarly, in *Deleon*: “our aim is to discern ‘ordinary meaning.’”¹⁶² In both cases, Judge Newsom’s methods speak to this objective in a direct and straightforward manner. He asks: “*What is the ordinary meaning of ‘landscaping’?*”¹⁶³; “*Is installing an in-ground trampoline ‘landscaping’?*”¹⁶⁴; and “*What is the ordinary meaning of ‘physically restrained’?*”¹⁶⁵.

As we discuss below, the opinions employing this direct-query method showcase the four aforementioned myths. Moreover, we show that these myths unravel upon closer examination of the design features of LLMs, the practices involved in their development, and the platforms which employ them.

158. See generally SLOCUM, *supra* note 95.

159. SCALIA & GARNER, *supra* note 30, at 69.

160. *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1222 (11th Cir. 2024) (Newsom, J., concurring) (quoting ANTONIN SCALIA & BRYAN A. GARNER, *READING LAW: THE INTERPRETATION OF LEGAL TEXTS* 69 (2012)).

161. See *id.*

162. *United States v. Deleon*, 116 F.4th 1260, 1276 (11th Cir. 2024) (Newsom, J., concurring).

163. *Snell*, 102 F.4th at 1232 (Newsom, J., concurring).

164. *Id.* at 1234.

165. *Deleon*, 116 F.4th at 1272 (Newsom, J., concurring).

1. Myth 1: LLM Training Data Reflects *Naturalistic* Language Use

For Judge Newsom, model training data is “the best reason to think that LLMs might provide useful information to those engaged in the interpretive enterprise” of ordinary meaning textualism:¹⁶⁶

LLMs are quite literally “taught” using data that aim to reflect and capture how individuals use language in their everyday lives . . . and at least as I understand LLM design, those data run the gamut from the highest-minded to the lowest, from Hemmingway novels and Ph.D. dissertations to gossip rags and comment threads. Because they cast their nets so widely, LLMs can provide useful statistical predictions about how, in the main, ordinary people ordinarily use words and phrases in ordinary life.¹⁶⁷

On first impression, Judge Newsom’s data argument is a compelling one, particularly when one compares the richness of LLM training data to the relatively impoverished empirical basis of traditional textualist tools. For example, as presented by Scalia and Garner in *Reading Law*, the canons of construction are frequently justified by postulated generalizations of ordinary linguistic usage.¹⁶⁸ However, it is not clear which canons hold up to empirical scrutiny.¹⁶⁹ Because the datasets used to train LLMs are incredibly vast—it would take many thousands of lifetimes to read all the text contained within them—we might worry less about empirical “blind spots” compared to other, more established interpretive techniques.

However, recent scholarship has raised a number of worries regarding the data used to train LLMs. As Lee & Egbert explain, “[w]e currently know next to nothing about the target language domain or sampling decisions behind the LLMs that widely used chatbots are based on.”¹⁷⁰ They further note that LLM judgments do not transparently account “for important linguistic variables such as register and speech community.”¹⁷¹ In *Snell*, Newsom acknowledges related concerns, including that training datasets “don’t capture offline speech” and that those without internet access “may be less likely to contribute to the sources from which LLMs draw in crafting their responses to queries.”¹⁷²

One could register these criticisms but still accept the main thrust of Judge Newsom’s data argument: however unbalanced or incomplete their training data may be, LLMs are useful because they have been trained on the language of “ordinary people” in “ordinary life.”¹⁷³ However, we raise objections that have not yet been identified in legal scholarship but which cut at the heart of Judge Newsom’s data argument: the LLMs that power chatbots are, by design, trained

166. *Snell*, 102 F.4th at 1226 (Newsom, J., concurring).

167. *Id.*

168. See generally SCALIA & GARNER, *supra* note 30.

169. See generally Tobia, Slocum & Nourse, *supra* note 50; Randall & Solan, *supra* note 50.

170. Lee & Egbert *supra* note 9, at 34.

171. *Id.*

172. *Snell*, 102 F.4th at 1231 (Newsom, J., concurring).

173. *Id.* at 1226. Moreover, when it comes to the training data of proprietary models, “[w]e know nothing about the target language domain,” how that domain was operationalized, or how the creators sampled from that operational domain. Lee & Egbert *supra* note 9, at 34.

on specialized data that cannot plausibly be taken to reflect “everyday” linguistic usage. In fact, this data is often not even produced by human beings, let alone “ordinary” people, however we might want to define that term. Perhaps most worryingly, *LLMs are optimized to predict this artificial data* in the fine-tuning stages that follow pretraining.

This non-naturalistic language data takes three primary forms: (a) instruction data used to train LLMs to “follow” instructions specified in prompts; (b) human preference data used to optimize LLMs for any number of developer objectives (e.g., safety); and (c) language model outputs that find their way into the training data of subsequent models.

a. Non-naturalistic Instruction Data

LLM-based chatbots are useful in part because we perceive their outputs as responses to our queries. For example, we can prompt GPT-4o to “Write a haiku about the Georgetown Hoyas football team” and observe the following completion:

“Hoyas on the field,
Echoes of past triumphs rise,
Autumn whispers hope.”

The same prompt, passed to OpenAI’s legacy davinci-002 model, yields a far different completion:

“Winning contributions will receive a free pass to attend the next game, The Chains”

The second completion is clearly deficient, not just because it is a bad haiku (in fact, it is not a haiku at all), but because it does not even strike us as an attempt to fulfill our poetry request. If anything, the second completion is an almost plausible—though not entirely coherent—elaboration of a context in which we might encounter our original poetry request within a larger document—it continues the text by motivating readers to write the haiku in order to win a prize.

This difference in behavior between these two models can be attributed in part to differences in the training data used to parameterize each one: davinci-002 is a “base” model parameterized exclusively via pretraining.¹⁷⁴ By contrast, we can assume that GPT-4o—like its predecessors—is additionally instruction-tuned¹⁷⁵ on a vast number of prompt-completion pairs in which the prompt expresses a request fulfilled by the completion. For illustration, the prompt-completion pairs in [Table 1](#) come from an open-source dataset developed by the US-based AI firm HuggingFace.¹⁷⁶

174. *davinci-002*, OPENAI, <https://platform.openai.com/docs/models/davinci-002> (last visited Sep. 20, 2025).

175. See Jason Wei et al., *Finetuned Language Models are Zero-Shot Learners*, ICLR, 2022, at 9, <https://arxiv.org/pdf/2109.01652> [<https://perma.cc/8MUT-ATPQ>]; *Aligning Language Models to Follow Instructions*, OPENAI (Jan. 27, 2022), <https://openai.com/index/instruction-following>.

176. *Instruction-Dataset*, HUGGINGFACE, <https://huggingface.co/datasets/HuggingFaceH4/instruction-dataset> [<https://perma.cc/6VRR-577T>] (last visited Sep. 20, 2025).

TABLE 1. PROMPT-COMPLETION PAIRS FROM HUGGING FACE’S INSTRUCTION DATASET

Prompt	Completion
Give me the answer to this trivia question: what team was LeBron James playing for when he broke the NBA all time scoring record?	Los Angeles Lakers
Does the United States use Celsius or Fahrenheit?	The United States still uses Fahrenheit, while most of the world uses Celsius.
Write a poem about the sun and moon.	<p>The sun and the moon, the guards from the sky One works the day, other watches the night As the rays of sunshine hide their face The moon gracefully hurries to take their place One watches us play, the other watches us sleep Both are reminders of the beauty we keep</p>

The core idea behind instruction tuning is that models are trained on many thousands of such pairs (exemplified in Table 1), such that their completions begin to take the form of helpful answers to user requests. However, this instruction data is not “naturalistic” in the sense of having come from “normal people us[ing] language in their everyday lives.”¹⁷⁷ Rather, this data is *synthetic*, in the sense of having been produced by researchers and language model developers to overcome a specific deficiency of the language models, sometimes by automatic means, such as using templates. As the examples in Table 1 illustrate, most instruction tuning examples contain single-turn interactions; the language is not embedded in larger contexts. Early pioneers in the field of instruction tuning trained their instruction-tuned language model on sixty-two different datasets adapted to the prompt-completion format seen in Table 1.¹⁷⁸ Many of these datasets were originally designed to train and evaluate systems for highly particular real-world tasks, e.g., text summarization, sentiment analysis, or translation.¹⁷⁹ They were never intended to faithfully reflect patterns of “everyday” language use, and in fact, NLP researchers increasingly acknowledge that contemporary instruction datasets suffer from a sort of linguistic “monoculture,” reflecting a limited range of style and content in a sparse array of contexts.¹⁸⁰

177. *Snell*, 102 F.4th at 1227 (Newsom, J., concurring).

178. See generally Wei et al., *supra* note 175.

179. See SHENGYU ZHANG ET AL., INSTRUCTION TUNING FOR LARGE LANGUAGE MODELS: A SURVEY (2024), <https://arxiv.org/pdf/2308.10792> [<https://perma.cc/CAG9-LKAB>].

180. See, e.g., *id.*

Though there exist open-source instruction-tuned models—and open-source instruction datasets, such as the HuggingFace dataset shown in Table 1—the vast majority of commercial chatbot applications have been trained on proprietary instruction data unavailable to the wider research community. With closed-source models such as OpenAI’s GPT-4o, we cannot know what kinds of synthetic language data the model was trained on, who produced the data, whether and how the data was quality controlled, and what kinds of usage patterns the data reflect. These development properties make it impossible to evaluate whether and to what extent synthetic training data obscures a model’s insights into “how people talk” in everyday contexts.¹⁸¹

b. Non-naturalistic Preference Data

Chatbot developers want their platforms to not just answer users’ queries but to also provide *good* answers to queries. To that end, developers train models on preference data designed to steer models towards “good” behaviors.

For a user interested in employing chatbots to investigate ordinary linguistic meaning, the term “good” should stand out as worryingly vague.¹⁸² In practice, “good” does not stand in for “faithful to ordinary linguistic usage”; rather, it denotes an ever-evolving set of desiderata, which different organizations may define in different ways.¹⁸³ For much of the AI community, “good” includes “safe”:¹⁸⁴ for example, models should not produce outputs that could abet unethical activities or are offensive, insulting, defamatory, or marginalizing. Of course, “safe” is also an inherently vague term. For a particular development team, a “good”/“safe” model might also be one that does not spread mis- or disinformation, however the developers operationalize those two notions.

For illustration, Table 2 presents a sample row from the publicly available HelpSteer preference dataset developed by Nvidia.¹⁸⁵ To develop this dataset, completions sampled from “an early version of an inhouse LLM” were rated by annotators on a scale of 0 to 4 along five dimensions: helpfulness, correctness, coherence, complexity, and verbosity.¹⁸⁶ Nvidia then fine-tuned an LLM on this data, such that the fine-tuned LLM’s completions could be optimized for, for example, high correctness and low verbosity.¹⁸⁷

181. *United States v. Caniff*, 916 F.3d 929, 941 (11th Cir. 2019) (Newsom, J., concurring in part and dissenting in part).

182. See HEATHER M. ROFF, EXPECTED UTILITARIANISM 1 (2020), <https://arxiv.org/pdf/2008.07321> [<https://perma.cc/C8V2-WJC3>] (“We want AI to be ‘good’ for humanity. We want it to help, not hinder, humans. Yet what exactly this entails in theory and in practice is not immediately apparent.”).

183. See *id.*

184. See, e.g., JOSEF DAI ET AL., SAFE RLHF: SAFE REINFORCEMENT LEARNING FROM HUMAN FEEDBACK 1 (2023), <https://arxiv.org/pdf/2310.12773> [<https://perma.cc/Q3BL-CQZB>] (“With the development of large language models (LLMs), striking a balance between the performance and safety of AI systems has never been more critical.”).

185. *HelpSteer Dataset*, HUGGINGFACE, <https://huggingface.co/datasets/nvidia/HelpSteer/viewer/default/train?p=3&row=393> [<https://perma.cc/V2RN-ZASH>] (last visited Sep. 20, 2025).

186. *HelpSteer*, HUGGINGFACE, <https://huggingface.co/datasets/nvidia/HelpSteer> [<https://perma.cc/ZC2X-53RZ>] (last visited Sep. 20, 2025).

187. See *id.*

TABLE 2. EXAMPLE FROM THE HELPSTEER PREFERENCE DATASET DEVELOPED BY NVIDIA.

Prompt			Completion		
What is the most useful resource in the world today?			The most useful resource in the world today is still water. It is essential for sustaining life, powering energy sources, and sustaining the environment.		
Helpfulness	Correctness	Coherence	Complexity	Verbosity	
3	3	4	1	1	
<p>Completions are sampled from a model and then rated by naïve participants along several dimensions. Preference data is not “naturalistic,” in that it involves an exchange between a human and an LLM—or an interaction of two LLMs. Moreover, participants often do not represent ordinary language users—e.g., they may have little or no real-world experience with American English.</p>					

We will return to preference alignment in the following subsections, as it presents additional conceptual challenges for users interested in using LLMs as a window into “how . . . ordinary people ordinarily use words and phrases in ordinary life.”¹⁸⁸ For the time being, we note that the synthetic nature of preference data presents a similar conceptual challenge to the one presented by instruction data. Preference data reflects judgments about either human-computer interaction, in which a human produces a prompt and an LLM generates the completion, or computer-computer interaction, in which the prompt is also generated by an LLM.¹⁸⁹ Though these interactions are increasingly common, they are not even exchanges between people, let alone “normal people.”¹⁹⁰

c. Non-naturalistic Language “Spillover”

Finally, LLMs—even pretrained models that have not undergone instruction tuning or additional parameterization with human preference data—are likely trained on the outputs of previous models. This is a necessary outcome of continuing to train models on vast swathes of web-based documents, as the internet

188. *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1226 (11th Cir. 2024) (Newsom, J., concurring).

189. See YUNTAO BAI ET AL., CONSTITUTIONAL AI: HARMLESSNESS FROM AI FEEDBACK 5 (2022), <https://arxiv.org/pdf/2212.08073> [<https://perma.cc/ZFR4-4C94>].

190. Moreover, preference alignment exacerbates the “monoculture” problem presented by instruction tuning: like instruction data, preference data often reflects a limited range of styles and contexts. Additionally, recent findings suggest that the human annotations in preference data may further steer LLM outputs towards an idiosyncratic style in which certain words and phrases are over-represented (e.g., “delve,” “intricate,” “underscore”). See Tom S. Juzek & Zina B. Ward, *Why Does ChatGPT “Delve” So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models*, PROCS. 31ST INT’L CONF. ON COMPUTATIONAL LINGUISTICS 6397, 6404 (2025).

becomes increasingly populated by machine-generated text.¹⁹¹ In the extreme, models trained on model-generated text may come to “forget the true underlying data distribution,” suggesting that “the use of LLMs at scale to publish content on the Internet will pollute the collection of data to train them.”¹⁹²

At the same time, there is growing concern in the AI research community that there is simply not enough human-generated text data to develop models of ever-increasing size and ever-improving performance.¹⁹³ One proposed solution to this impending data deficit is to augment human-generated data sources with synthetic text.¹⁹⁴ In other words, there is no guarantee that an LLM is free of “pollution” from non-human sources; in fact, contemporary and future models may be intentionally trained on machine-generated text to supplement the finite supply of text generated by people.¹⁹⁵

Similarly, because LLMs are trained on large scrapes of internet data, they are exposed to large amounts of non-linguistic data, most notably computer code.¹⁹⁶ This exposure is one reason that LLMs are valuable coding assistants for professional software developers and computer science students alike. However, training on large amounts of computer code has been found to affect LLMs’ *linguistic* abilities, in sometimes unpredictable ways.¹⁹⁷

In *Snell*, Judge Newsom writes that “ordinary-meaning interpretation aims to capture how normal people use language in their everyday lives—and the bulk of the LLMs’ training data seem to reflect exactly that.”¹⁹⁸ However, as we have shown, much of this training data is highly abnormal.¹⁹⁹ This is frequently a

191. See ILIA SHUMAILOV ET AL., THE CURSE OF RECURSION: TRAINING ON GENERATED DATA MAKES MODELS FORGET 1 (2024), <https://arxiv.org/pdf/2305.17493> [<https://perma.cc/7ENB-PKNW>] (“If most future models’ training data is also scraped from the web, then they will inevitably come to train on data produced by their predecessors.”).

192. *Id.* at 2.

193. See PABLO VILLALOBOS ET AL., WILL WE RUN OUT OF DATA? LIMITS OF LLM SCALING BASED ON HUMAN-GENERATED DATA 1 (2024), <https://arxiv.org/pdf/2211.04325> [<https://perma.cc/6BVU-PH7Q>].

194. *See id.*

195. Henderson et al. point out that one could also—at least in theory—*intentionally* alter the composition of a training dataset for nefarious purposes. The authors review NLP research which suggests that “simply inserting a handful of modified documents can change how the model would interpret a given text,” and they observe that “[t]here is nothing stopping a model creator from doing the same thing to manipulate empirical analyses of meaning. One could imagine encoding arbitrary statutory preferences through this approach.” *See* Henderson et al., *supra* note 80, at 150.

196. JACKSON PETTY, SJOERD VAN STEENKISTE & TAL LINZEN, HOW DOES CODE PRETRAINING AFFECT LANGUAGE MODEL TASK PERFORMANCE? 1 (2024), <https://arxiv.org/pdf/2409.04556v1> [<https://perma.cc/4CR3-9UMQ>].

197. *See id.* at 10.

198. *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1227 (11th Cir. 2024) (Newsom, J., concurring).

199. Judge Newsom does not purport to know “*exactly* what data LLMs use for training.” He identifies this as “an area ripe for a transparency boost, especially as LLMs become increasingly relevant to legal work.” *Id.* at 1226 n.6. In this section, we’ve taken up his call to shed light on the kinds of training data common in LLM development. Of course, we can only hope for so much “transparency” when it comes to platforms developed by private companies using proprietary datasets.

design feature, not a bug: synthetic training data prioritizes system performance and the ethical objectives of model developers over every day conversational norms. This should raise significant questions about the ability of these models to faithfully characterize “everyday” language use.²⁰⁰

2. Myth 2: LLM Chatbots Are Optimized to Produce Correct Statements About Ordinary Language

Though LLMs’ training datasets are full of highly unordinary language, Judge Newsom reports having successfully used LLM-based chatbots in two separate interpretation cases. This raises questions about the force of our criticisms above: should judges care about LLMs’ non-naturalistic training data, if LLM chatbots were nevertheless capable of delivering interpretive results in *Snell* and *Deleon* that reflect the linguistic understanding of a representative population of ordinary readers? In those cases, the chatbots queried by Judge Newsom seemed to reflect intuitions that would plausibly generalize to a broad sample of ordinary English speakers. Synthetic instruction data and preference data, one might argue, merely enable an LLM to package its robust knowledge of ordinary language in the form of helpful, well-formed query responses. One might also argue that instruction tuning and preference alignment, which reward the production of factually grounded content, serve to steer models towards producing accurate statements about ordinary language.

However, we have reasons to doubt that these optimization procedures consistently help models produce sound metalinguistic judgments. Notably, both instruction data and preference data often encode judgments *about language itself*. Thus, when we query chatbots for judgments about ordinary language, we are in fact relying on models which have likely been steered to reflect the linguistic judgments of the researchers and annotators who created the data.²⁰¹ These judgments are sometimes far from robust. To illustrate, the exchange in [Table 3](#) comes from the same HuggingFace instruction dataset introduced above:²⁰²

200. Though we have raised a number of observations problematizing the myth that LLM training data reflects naturalistic language use, we do not attempt to empirically measure the effect of fine-tuning on LLM-generated ordinary meaning judgments. In a recent systematic comparison of pretrained vs. fine-tuned LLMs, Professor Jon Choi reports findings which suggest that “the impact of post-training [fine-tuning] on ordinary meaning assessment is not uniform but depends on both the specific question being asked and the model family being evaluated.” JONATHAN H. CHOI, OFF-THE-SHELF LARGE LANGUAGE MODELS ARE UNRELIABLE JUDGES 41 (2025).

201. *See id.*

202. HUGGINGFACE, *supra* note 176.

TABLE 3. A PROMPT-COMPLETION PAIR FROM HUGGINGFACE’S INSTRUCTION DATASET.

Prompt	Completion
Think of alternatives and paraphrases for the underlined word. Input: what we have <u>expected</u>	+ what we have projected + what we have anticipated + what we have waited + what we have hoped for + what we have accustomed
Such datasets can reflect substandard linguistic judgments, as this pair demonstrates.	

This exchange asks for alternatives and paraphrases of *expected* in the phrase *what we have expected*. Though each paraphrase in the completion roughly approximates the meaning of *expected* in context, two paraphrases—*what we have waited*, *what we have accustomed*—strike us as incomplete or ungrammatical. This example suggests that instruction-tuned LLMs are, at least in some instances, optimized to reflect substandard linguistic judgments.

This concern extends to preference alignment. For illustration, we examine an entry from Nvidia’s HelpSteer preference dataset discussed in the previous section. In Table 4 below, we see that when asked to produce words that rhyme with “flung,”²⁰³ Nvidia’s un-tuned LLM produced four words, only one of which actually rhymes with “flung.” Nvidia’s annotators gave this exchange a rating of 3 for helpfulness, the second-highest possible rating, and a 2 for correctness, the middle of the scale.²⁰⁴ This datapoint illustrates how annotator judgments in preference datasets may feature substandard linguistic judgments:

203. *HelpSteer Dataset*, HUGGINGFACE, <https://huggingface.co/datasets/nvidia/HelpSteer/viewer/default/train?p=3&views%5B%5D=train&row=319> [<https://perma.cc/9C7S-92S3>] (last visited Oct. 30, 2025).

204. *See id.* Notably, Nvidia’s HelpSteer preference dataset was developed with numerous quality controls. To develop this dataset, Nvidia employed 200 US-based annotators contracted through Scale AI, a second US-based artificial intelligence company. The publishers of this dataset report that annotators were first screened for proficiency in English and that each annotation was checked by at least two human reviewers (as well as by an “automated” review process). *See id.* Despite these processes, we observe questionable linguistic judgments in the final dataset, underscoring the difficulty of procuring high-quality data for fine-tuning LLMs. As Lee and Egbert caution, we often know far less about the raters employed to create preference datasets, and non-U.S. raters might dominate OpenAI’s rater pool. *See Lee & Egbert, supra* note 9, at 35; *see also Henderson et al., supra* note 80, at 139–140 (“[Human] annotations may have significant influence on meaning encoded in the [large language] model, . . . yet may reflect the arbitrary preferences of annotators located across the world.”).

TABLE 4. AN EXAMPLE FROM NVIDIA'S HELPSTEER DATASET.

Prompt			Completion	
What are some words that rhyme with "flung"?			Ring, Wing, King, and Stung	
Helpfulness	Correctness	Coherence	Complexity	Verbosity
3	2	3	1	1
Top: Prompt (Words that Rhyme with "Flung") and Completion from Nvidia's un-tuned LLM. Bottom: Annotator judgments for this Prompt-Completion pair.				

Instruction tuning and each preference alignment occur after pretraining, the development phase where we expect to see the most naturalistic language in the LLM's training data.²⁰⁵ This means that while LLMs may be initially pretrained to model naturalistic text, they are optimized to predict synthetic data. For this reason, it is critical to examine that data to assess whether it reflects ordinary patterns of linguistic usage. Our examination, which reveals evidence of dubious judgments in instruction tuning and preference alignment datasets, problematizes the assumption that these optimization steps help LLMs produce correct statements about ordinary language.²⁰⁶

Due to the closed nature of proprietary training datasets, we often cannot be sure about the exact extent to which commercially available chatbot platforms are based on LLMs fine-tuned on deficient judgment data (including questionable linguistic judgments). However, the select examples presented in this subsection suggest that in general, we should approach claims about LLMs' linguistic capabilities with heightened skepticism. Even basic judgments about, for example, rhyming or paraphrasing may be shaped by problematic training, which raises broader concerns about the reliability of more complex linguistic behaviors exhibited by these models.

3. Myth 3: LLM Chatbots Are Capable of Producing Accurate *Abstractions* of Their Training Data

The last subsection identifies an under-appreciated risk associated with the direct-query method: the training data used to fine-tune LLMs may contain highly questionable judgments about ordinary language. However, a defender of the method could raise two objections to the above discussion. The first objection questions the magnitude of the risk: even if there are occasional imperfections in

205. See Ohm, *supra* note 155, at 218. As discussed above, even pretraining data may—intentionally or unintentionally—include synthetic language data.

206. Our investigation is admittedly limited in its scope, identifying a limited number of problematic data points from two publicly available datasets. Moreover, we note that instruction data and preference data take many forms, and the model training regimes involving such data are constantly evolving. We leave a more comprehensive investigation of fine-tuning datasets—and the linguistic generalizations they reflect—to future work.

an LLM’s training data, a judge might still want to know what the model has “learned” about language from the data.²⁰⁷ So long as we are interested in an LLM’s linguistic generalizations, why not just ask the model directly: *What is the ordinary meaning of “landscaping”?* or *What is the ordinary meaning of “physically restrained”?*

The second objection questions whether the risk is a short-term obstacle for the direct query method or a long-term conceptual challenge to it. Though current datasets may exhibit flaws, perhaps a future “legal textualist” LLM could be fine-tuned on highly quality-controlled datasets that feature verifiably robust linguistic judgments. If we had access to such a model, wouldn’t the direct query method be the most straightforward way of interfacing with it?

Though simple and intuitive, the direct query method is flawed even in a perfect world in which we have painstakingly curated the model’s fine-tuning data. The reason is that this approach mistakenly assumes that LLMs can faithfully articulate plain language characterizations of their underlying generalizations. As Judge Newsom articulates this assumption in his opinion in *Snell*: “[T]he model learns what words are most likely to appear where, and which ones are most likely to precede or follow others—and by doing so, it can make probabilistic, predictive judgments about ordinary meaning and usage.”²⁰⁸

To understand why this assumption is flawed, it is useful to consider how LLM-based chatbots are able to provide coherent, though not necessarily accurate, answers to questions about language. As discussed above, this behavior is likely due in part to the presence of similar exchanges in the datasets used to train the underlying LLM in instruction tuning and preference alignment. Such exchanges are *metalinguistic*, in that they involve discourses about language itself. However, the ability of LLMs to talk about language is also due in no small part to the abundance of varied forms of metalanguage—that is, statements about language—in the models’ pretraining data.

This metalinguistic text—including online dictionary definitions, textbooks, and discussions in online language learning forums—helps LLMs learn the *form* of a helpful metalinguistic judgment (for example, *Yes, installing an in-ground trampoline can be considered a part of landscaping*). However, LLM predictions also reflect the *substance* of those judgments, not all of which hold up to scrutiny. Some indication of the quality of web-scraped metalinguistic training data can be ascertained by examining the ELQA corpus, which contains thousands of question/answer pairs sourced from English language learning discussion boards.²⁰⁹

207. As discussed below, this is a central question of a nascent and rapidly growing field of scientific study known as AI interpretability. For relevant resources, see *infra* note 228.

208. *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1227 n.7 (11th Cir. 2024) (Newsom, J., concurring).

209. See Shabnam Behzad, Keisuke Sakaguchi, Nathan Schneider & Amir Zeldes, *ELQA: A Corpus of Metalinguistic Questions and Answers About English*, 1 PROC. 61ST ANN. MEETING ASS’N FOR COMPUTATIONAL LINGUISTIC 2031, 2034–35 (2023).

The user-authored answers vary in quality.²¹⁰ The following excerpt is one user’s response to a question titled, *Should nouns after “any” be singular or plural?*:

“To me, the word ‘any’ should always be followed by a singular noun.
Example: Do you have any idea as to when this work will get done?”

Having said that, the below is also correct. Example: Do you know any of these boys, who fled the scene on [sic] the other day?”²¹¹

This answer is clearly incorrect,²¹² but it is likely part of trillions of words of raw text data used to pretrain many LLMs. How exactly metalinguistic training data affects LLMs’ metalinguistic outputs is an open question—but for any given metalinguistic query, there is a real risk of garbage in, garbage out.

That LLMs may, like “stochastic parrots,”²¹³ superficially mimic the ways people on the internet talk about English raises a more general issue for the direct query method: LLMs likely lack the abstract reasoning capabilities needed to arrive at and articulate original, explicit observations about linguistic patterns based on aggregating usage examples in the training data.²¹⁴ In other words, LLMs may produce plausible-sounding metalinguistic claims, but they do so without actually having analyzed the data needed to substantiate such claims.

LLM chatbots answering metalinguistic questions are properly understood not as (artificial) linguists, but as pseudo-linguists. Evidence comes from studies showing a striking disconnect between generated answers and model internals: judgments about language are poorly calibrated to what LLMs have actually “learned” about the statistics of language. For example, Jennifer Hu and Roger Levy find that while LLMs generally assign higher probability to grammatical sentences, *Every child has studied*, than to their ungrammatical counterparts, *Every child have studied*, prompting the model with “Is the following sentence a good sentence of English?” may yield a completion that is inconsistent with the underlying probability.²¹⁵ Hypothetically, then, an LLM might assign very low

210. *See id.*

211. Rajeshprabu Radhakrishnan, ENG. LANGUAGE LEARNERS, *Should Nouns After “Any” Be Singular or Plural?* (July 15, 2019, 04:54 ET), <https://ell.stackexchange.com/questions/1155/should-nouns-after-any-be-singular-or-plural> [<https://perma.cc/R3WY-4HJ4>].

212. A simple counterexample: *I have not eaten any blueberries.*

213. *See* Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Schmittchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FACT ‘21) 610, 616 (2021).

214. To be clear, pretrained models aggregate language evidence across the training data as latent or implicit knowledge—encoded in model parameters—that drives fluent generation. However, encoding general statistical patterns implicitly is not the same as being able to identify individual patterns and explain them explicitly. *See* Andreas Madsen, Sarath Chandar & Siva Reddy, *Are Self-Explanations from Large Language Models Faithful?*, FINDINGS ASS’N FOR COMPUTATIONAL LINGUISTICS: ACL 2024, at 295, 303 (2024).

215. Jennifer Hu & Roger Levy, *Prompting Is Not a Substitute for Probability Measurements in Large Language Models*, PROCS. 2023 CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING 5040, 5045 (2023).

probability²¹⁶ to the string “landscaping work such as installing recreational equipment”—suggesting that the term *landscaping* is a poor fit based on the evidence in the training corpus—while simultaneously answering in the affirmative when asked *directly* if the term fits.

Hu and Levy’s work is an example of research in the nascent scientific field of AI interpretability, which examines neural networks’ behavior in order to make their predictive processes more transparent and comprehensible to humans.²¹⁷ In the context of LLMs, AI interpretability research seeks in part to characterize what exactly LLMs have learned about language and how they manage to learn it.²¹⁸ Though this is an active area of research, it is still in its infancy: AI scientists are not yet capable of saying for sure—at an abstract level—what a given model has “learned” about English from its training data.

At a minimum, judges should be aware that LLM chatbots are not reliably capable of producing accurate abstractions of their training data.²¹⁹ The “direct query” method is therefore a coarse and unreliable means of probing the linguistic knowledge encoded by LLMs. Moreover, judges who defer to the linguistic judgments of LLMs risk putting the cart before the horse: because we don’t know what exactly LLMs learn during training, we don’t yet know how to precisely measure whether and to what extent an LLM’s linguistic competence diverges from “ordinary” competence—however we may choose to define “ordinary.” Until more progress in AI interpretability research has been made, there is no fully responsible way to use these models as proxies for ordinary meaning.

4. Myth 4: LLMs Are Less Biased than Traditional Interpretive Tools

After having reviewed the concerns above, one might still wonder: do the “pros” of Judge Newsom’s direct query method outweigh the “cons,” and how does the method size up against the traditional tools of textualism? Dictionaries,²²⁰ canons, and linguistic corpora introduce many degrees of

216. Low relative to alternative activities of generally similar frequency.

217. For reviews, see Isabel O. Gallegos et al., *Bias and Fairness in Large Language Models: A Survey*, 50 COMPUTATIONAL LINGUISTICS 1097, 1149 (2024); TYLER A. CHANG & BENJAMIN K. BERGEN, LANGUAGE MODEL BEHAVIOR: A COMPREHENSIVE SURVEY 2 (2023), <https://arxiv.org/pdf/2303.11504v2> [<https://perma.cc/2LQ9-AK55>]; DAKING RAI ET AL., A PRACTICAL REVIEW OF MECHANISTIC INTERPRETABILITY FOR TRANSFORMER-BASED LANGUAGE MODELS 35 (2024), <https://arxiv.org/pdf/2407.02646> [<https://perma.cc/WAE7-VR5V>].

218. See Amanda McGrath & Alexandra Jonker, *What is AI Interpretability*, IBM (Oct. 8, 2024), <https://www.ibm.com/think/topics/interpretability> [<https://perma.cc/C3S3-6PLC>].

219. This note of caution also applies to judges who might be interested in employing LLMs for inquiries into *historical* linguistic meaning. For example, even if an LLM is trained in part on a large quantity of text written in 1791, there is no guarantee that responses to direct queries of the form “What was the ordinary meaning of *X* in 1791?” would faithfully reflect that portion of the model’s training data. Moreover, LLMs can be prompted to produce generalizations about historical meaning even if they were not trained on any data from the relevant historical period(s). Finally, there is no guarantee that a metalinguistic generalization produced by an LLM accurately reflects how linguistic usage may have changed over time.

220. See generally Aprill, *supra* note 63 (critiquing the Supreme Court’s use of dictionaries as a textualist tool).

researcher freedom—for example, the choice of dictionary, canon, or corpus—and are thus susceptible to misuse as a result of unconscious confirmation biases, made more severe by false consensus biases in legal interpretation,²²¹ or intentional manipulation. Moreover, these tools are not always available for the lay public to scrutinize. However, Judge Newsom argues that LLMs may represent a significant improvement over presently available alternatives, like dictionaries and canons.

The main claimed advantages over traditional tools are transparency and consistency.²²² First, while LLM-based methods may seem more transparent than the alternatives, these methods often cannot be disclosed with sufficient detail to make them truly reproducible by other courts or scholars. Moreover, as we'll demonstrate, there are significant reasons to doubt whether LLM-based interpretation delivers the kind of consistent, manipulation-resistant results that its advocates suggest. The susceptibility of these models to what we call “prejudiced prompting”—as well as the difficulty of standardizing query methods across different cases—should raise serious questions about whether LLMs actually represent an improvement over traditional interpretive tools when it comes to managing bias.

Many of these challenges stem from properties of the applications which make use of LLMs to process and generate text. In the simplest iterations of chatbot applications, such as OpenAI's ChatGPT or Anthropic's Claude, the model prompt consists in part of a hidden system prompt designed to steer models towards useful outputs.²²³ For example, ChatGPT's system prompt includes the declaration “You are a helpful assistant.”²²⁴ The user's initial query is appended to this hidden system prompt, and the chatbot's initial response is a function of the concatenation of both strings. Subsequent prompts are formed from the system prompt and the user's previous exchanges on the platform.²²⁵

221. In psychology, the false consensus bias is the propensity to overestimate the extent to which one's idiosyncratic traits, preferences, or judgments generalize to a larger population. For experimental findings suggesting that this bias manifests in legal interpretation, see Lawrence Solan, Terri Rosenblatt & Daniel Osherson, *False Consensus Bias in Contract Interpretation*, 108 COLUM. L. REV. 1268, 1299 (2008); and Brandon Waldon, Madigan Brodsky, Megan Ma & Judith Degen, *Predicting Consensus in Legal Document Interpretation*, 45 PROC. ANN. MEETING COGNITIVE SCI. SOC'Y 1101, 1106 (2023).

222. Judge Newsom also argues that LLMs represent a significant accessibility improvement over other tools, noting that “even the [costs of accessing] more advanced LLMs pale in comparison to subscriptions for Westlaw and Lexis, which power most modern legal research, including some involving dictionaries.” *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1228 (11th Cir. 2024) (Newsom, J., concurring). This accessibility argument is compelling on its face—as AI technology becomes more widespread and user-friendly, judges and lawyers, as well as the wider public, can increasingly access these tools with minimal technical expertise. However, the claims about transparency and consistency warrant closer scrutiny.

223. See MINGQIAN ZHENG ET AL., WHEN “A HELPFUL ASSISTANT” IS NOT REALLY HELPFUL: PERSONAS IN SYSTEM PROMPTS DO NOT IMPROVE PERFORMANCES OF LARGE LANGUAGE MODELS 1 (2024), <https://arxiv.org/pdf/2311.10054> [<https://perma.cc/7F6P-Q62Y>].

224. *Id.*

225. *See id.*

This mode of interaction raises serious questions about the transparency and consistency—and, therefore, the reproducibility—of querying exercises such as those undertaken by Judge Newsom in *Snell* and *Deleon*. Chatbot developers may change system prompts over time and in ways not observed by the end user. Chatbot responses will also be affected by previous exchanges in a given chat session.²²⁶ Moreover, at the time of writing, OpenAI’s ChatGPT shares history *across* chat sessions by default: users must activate a “temporary chat” setting to ensure that “ChatGPT won’t be aware of previous conversations or access memories.”²²⁷

Further reproducibility concerns are presented by the procedure by which the underlying model draws samples from its computed probability distributions. Even with identical prompts and system settings, the sampling process introduces inherent randomness controlled by two key factors. First, a random number (called the “seed”) determines the starting point for the generation used in sampling; second, a temperature parameter (a value between 0 and 1) controls how “conservative” the model’s outputs are.²²⁸ A higher temperature (for example, 0.8) leads to more diverse and potentially surprising responses, as the model samples from a broader range of token probabilities, while a lower temperature (for example, 0.2) makes the model more likely to select the highest probability tokens, resulting in more predictable and less diverse outputs.²²⁹ Additionally, the temperature parameter is known to directly impact model decision-making capabilities.²³⁰ Without controlling for both the random seed and temperature settings—information that may not be accessible to researchers—reproducing specific chatbot responses becomes practically impossible, even under otherwise identical conditions.

Judge Newsom—to his credit—makes a concerted effort to report the exact conditions under which his chatbot queries were produced. To that end, in both *Snell* and *Deleon*, he reports the exact prompts he entered into the platforms he used. In *Deleon*, he reports additional details, including information about the underlying LLMs employed by the chatbots,²³¹ the number of queries undertaken on each platform,²³² and the steps taken to reduce interference from previous

226. *United States v. Deleon*, 116 F.4th 1260, 1273 (11th Cir. 2024) (Newsom, J., concurring) (“Each time, I opened a new chat and asked the same question . . . Not perfectly scientific, I know, but better than a rifle shot.”).

227. *Temporary Chat FAQ*, OPENAI, <https://help.openai.com/en/articles/8914046-temporary-chat-faq> (last visited Sep. 20, 2025).

228. See Matthew Renze & Erhan Guven, *The Effect of Sampling Temperature on Problem Solving in Large Language Models*, FINDINGS ASS’N FOR COMPUTATIONAL LINGUISTICS: EMNLP 2024, at 7346, 7347 (2024).

229. See *id.* at 7352.

230. See *id.* at 7347.

231. *United States v. Deleon*, 116 F.4th 1260, 1272 n.2 (11th Cir. 2024) (Newsom, J., concurring) (“I queried the free online version [of ChatGPT], which uses GPT4-o.”).

232. *Id.* at 1273 (“I queried the three leading models—GPT, Claude, and Gemini— ten times apiece.”) (citation omitted).

queries.²³³ Though these disclosures serve to make Judge Newsom’s methods more transparent, they are not enough to fully reproduce his procedures for reasons outlined above.²³⁴

Moreover, even when a particular LLM elicitation is fully reproducible, there are serious questions about the consistency—and therefore the robustness—of the direct query method. Recall that in *Snell* and *Deleon*, Judge Newsom takes care to query multiple chatbots. The responses generated by the various chatbots largely cohere with one another. To Judge Newsom, this result demonstrates a striking advantage of LLMs over traditional tools: whereas judges can “shop” for canons, dictionaries, or corpus entries that support a particular interpretation, Judge Newsom perceives less risk that a judge could engage in “chatbot shopping” or other manipulative behaviors, “at least when coupled with . . . full disclosure of one’s research process.”²³⁵

By contrast, Lee and Egbert report two studies that undermine the idea that LLMs are somehow less susceptible to manipulation than are the traditional tools of textualism.²³⁶ First, they demonstrate that ChatGPT—when queried 100 times using Judge Newsom’s prompt in *Snell*—offers inconsistent judgments.²³⁷ Next, they demonstrate that ChatGPT is highly sensitive to the way in which users pose questions about language:

When asked whether *landscaping* includes the installation of an in-ground trampoline, ChatGPT confidently replies that it can be. When we told it that it was incorrect, however, it modified its answer. Then, when we posed the original question to it again, it confidently states that [landscaping] “typically does not include the installation of an in-ground trampoline.”²³⁸

Both studies demonstrate that the direct query method is—to an extent—prone to irresponsible use. The first study demonstrates that it is not possible to draw robust inferences about an LLM’s linguistic competence on the basis of a single sample from the model. If that single sample accords with a judge’s prior beliefs about ordinary meaning, that judge, however well-intentioned, is susceptible to confirmation bias. A more nefarious actor could—following Lee and Egbert’s proof-of-concept approach—continue to pose the same question to a chatbot until the desired response is produced. The risks exemplified by this study are mitigated by setting random seeds when sampling.

233. *Id.* (“Each time, I opened a new chat and asked the same question: ‘What is the ordinary meaning of “physically restrained”?’ Not perfectly scientific, I know, but better than a rifle shot.”).

234. In Section III.A below, we offer some concrete guidance on how to transparently report LLM-based methods such that they are fully reproducible.

235. *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1232 (11th Cir. 2024) (Newsom, J., concurring).

236. Lee & Egbert, *supra* note 9, at 41–43.

237. The authors “sorted each answer into one of three categories that emerged: (1) Yes, (2) Could be, but not the traditional sense of ‘landscaping’, and (3) Could be, if one or more conditions are met. We observed (1) 68 times, (2) 18 times, and (3) 14 times.” *Id.* at 41 (citation omitted).

238. *Id.* at 43.

The authors' study on prompt sensitivity demonstrates that LLMs—given the right prompting strategy—can be “gamified” to produce text desired by the user. However, when one considers how LLMs are developed, Lee and Egbert's finding is not terribly surprising: instruction-tuned, preference-aligned models are trained to be “cybersycophants,” outputting helpful and pleasing, but not necessarily correct, answers to user queries—regardless of what is asked.²³⁹ Moreover, Lee and Egbert's finding may not be terribly compelling to judges who believe that LLMs can be queried in a careful and responsible manner. Obviously, no judge would ever report in an opinion the kind of exchange described by Lee and Egbert, in which the judge tells the chatbot that its answer is incorrect in an effort to move the chatbot to the opposite verdict. On its face, such an exchange is clearly biased towards a particular desired outcome.

In sum, Lee and Egbert's findings demonstrate real pitfalls of the direct query method. However, a judge can easily address these pitfalls through relatively simple interventions.²⁴⁰ A more compelling challenge to the direct query method would involve demonstrating that it is susceptible to misuse, even when one takes care to pursue it in the most robust, transparent, and replicable way possible—and without the use of obviously biased prompts. We provide such a demonstration below, showing that even under ideal conditions, LLMs can be subtly manipulated to deliver biased interpretive outcomes.

C. STUDY: RED-TEAMING LLMs FOR BIASED ORDINARY MEANING JUDGMENTS

In this section, we show that LLMs' metalinguistic judgments are highly sensitive to subtle prompting variations. The goal of this exercise is to demonstrate 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.²⁴¹ This exercise is an example of *red-teaming*, a common research practice whereby AI systems are stress-tested for vulnerabilities that could yield unsafe or otherwise undesirable system behavior.²⁴² Here, the danger is that a judge—consciously or unwittingly—will use an LLM in a way that merely appears to substantively inform the judge's final decision in a case.²⁴³ We call

239. Proprietary model developers, including OpenAI, recognize that current training practices can lead to sycophancy in LLMs. See, e.g., *Sycophancy in GPT-4o: What Happened and What We're Doing About It*, OPENAI (Apr. 29, 2025), <https://openai.com/index/sycophancy-in-gpt-4o>.

240. We detail a number of best practices for interacting with LLMs in Part III, *infra*.

241. Here and elsewhere in the paper, we frame the issue in terms of the risks posed by lawyers and judges misusing chatbots. However, these risks extend to potential misuse by other relevant actors. For example, an expert report or amicus curiae brief could conceivably pursue the direct-query method to advance a legal interpretation argument.

242. See *What Is Red Teaming for Generative AI?*, IBM: RSCH. (Apr. 11, 2024), <https://research.ibm.com/blog/what-is-red-teaming-gen-ai> [<https://perma.cc/RV7T-GLNC>].

243. Arbel and Hoffman discuss a possible “problem of ‘leading prompts,’ queries that lead the model towards a desired answer.” Arbel & Hoffman, *supra* note 4, at 490. The authors prompt a series of LLM chatbot applications to provide interpretive judgments for a disputed contract. *Id.* at 483–85. In more recent work, Professors James Grimmelman, Benjamin Sobel, and David Stein conduct a critical investigation into the LLM-based interpretive methodologies introduced by Arbel and Hoffman. In line with our own findings reported in this section, Grimmelman, Sobel, and Stein find that Arbel and

this danger “prejudiced prompting.” Textualists often appeal to “rule of law” values, and the seeming predictability and consistency of LLM chatbots is an appealing reason for textualists to turn to them. However, for such textualists, the possibility of conscious or unconscious prejudiced prompting is a substantial threat to the values that justify the use of LLM chatbots and the values that undergird textualism itself.

In keeping with the best practices discussed further in Section III.A, we took several steps to ensure that our sampling methodology is reproducible,²⁴⁴ starting with our choice of model. We posed our queries to a model created by Mistral AI, a developer of open-source language models. Specifically, we employed their instruction-tuned *Ministral-8B-Instruct-2410* model (“Ministral”), which at the time of writing generally outperforms open-source models of similar size developed by Meta and Google on a variety of evaluation metrics.²⁴⁵ When sampling from this model, we specified a relatively low temperature setting of 0.1 and also set a random seed to ensure that our samples could be exactly replicated in the future.²⁴⁶ Our input prompts were queries designed to elicit metalinguistic judgments about the meaning of *landscaping* and *physically restrained*.²⁴⁷ As a control, we first tried the exact queries posed by Judge Newsom to various closed-source chatbots in *Snell* and *Deleon*.

Hoffman’s results depend heavily on minor choices in implementation. See James Grimmelmann, Benjamin Sobel & David Stein, *Generative Misinterpretation*, 63 HARV. J. ON LEGIS. (forthcoming) (manuscript at 21) (on file with authors).

244. Our sampling code can be freely accessed through an online repository at <https://github.com/bwaldon/llms-legal-interp> [<https://perma.cc/JQB2-HMXR>].

245. See *Ministral-8B-Instruct-2410*, HUGGINGFACE, <https://huggingface.co/mistralai/Ministral-8B-Instruct-2410> [<https://perma.cc/59FN-K3XT>] (last visited Sep. 20, 2025). Note that *Ministral* is smaller (in terms of number of parameters) than the models which power the chatbots employed by Judge Newsom. Employing a relatively small model enabled us to execute our queries using modest computational resources, and it will allow others interested in running and/or modifying our code to do the same. One might object that the risk of manipulation may be reduced for larger models, which generally out-perform their smaller counterparts on a variety of evaluation metrics (including those which measure a model’s ability to generate logically coherent text and to generate outputs that satisfy user queries). Of course, we cannot say for certain that *all* models are susceptible to prejudicial prompting, and it is possible (though by no means certain) that larger models are less prone to such practices. These exercises are best understood as a proof-of-concept demonstration of a possibly pervasive risk.

246. As discussed in Section II.B.4 *supra*, a random seed specifies a computer’s starting point for random sampling, ensuring that such samples can be reproduced in the future. Crucially, the LLM-generated text reported in the current subsection reflects individual generations from a single random seed. As such, we cannot make claims in this subsection regarding the representativeness of the samples. We stress, following Section II.B.4 *supra*, that these exercises are best viewed as a demonstration of a phenomenon worthy of future systematic investigation.

247. We intend for this study to be an illustrative demonstration of LLM prompt sensitivity, focusing on just one model and two phrases of interest. For a far broader and more systematic evaluation of LLMs’ prompt sensitivity in what we call “direct query” contexts, with results qualitatively similar to our own, see generally CHOI, *supra* note 194.

TABLE 5. COMPLETIONS SAMPLED FROM MINISTRAL, GIVEN PROMPTS THAT MATCH JUDGE NEWSOM’S ORIGINAL QUERIES IN *SNELL* (TOP AND MIDDLE CELLS) AND *DELEON* (BOTTOM CELL).

<p>Prompt: What is the ordinary meaning of “landscaping”?</p> <p>Completion: The ordinary meaning of “landscaping” refers to the art and practice of designing and creating outdoor spaces, such as gardens, parks, and other green areas. The term encompasses a wide range of activities, including:</p> <ol style="list-style-type: none"> 1. Planting and Arranging Plants: This involves selecting and placing trees, shrubs, flowers, and other vegetation to create a visually appealing and functional landscape. 2. Designing Layouts: Landscapers often create detailed plans that include the placement of plants, hardscaping elements (like paths, patios, and fountains), and other features to enhance the overall aesthetic and usability of the space. 3. Maintaining and Pruning: Regular care is essential to keep the landscape looking its best, which includes pruning plants, removing weeds, and ensuring that the plants receive adequate water and nutrients. 4. Installing Hardscaping: This includes the addition of structures such as walkways, retaining walls, fences, and outdoor lighting to complement the natural elements. 5. Irrigation Systems: Landscapers may also install and maintain irrigation systems to ensure that plants receive the right amount of water. 6. Soil Preparation: This involves preparing the soil to ensure it is suitable for the plants being installed, which may include adding fertilizers, compost, or other amendments. <p>In essence, landscaping is about transforming outdoor spaces into beautiful, functional, and harmonious environments that enhance the aesthetic appeal and usability of a property.</p>
<p>Prompt: Is installing an in-ground trampoline “landscaping”?</p> <p>Completion: Installing an in-ground trampoline can be considered a form of landscaping, but it’s more accurately described as a specific type of hardscaping. Here’s why:</p> <ol style="list-style-type: none"> 1. Landscaping: This term generally refers to the art and practice of designing and maintaining outdoor spaces, including plants, trees, and other natural elements. It can also include hardscaping elements like patios, walkways, and retaining walls. 2. Hardscaping: This is a subset of landscaping that involves the use of non-living elements like stone, brick, concrete, and metal. An in-ground trampoline fits into this category because it’s a non-living structure that’s installed into the ground.

So, while installing an in-ground trampoline is a landscaping project, it's more specifically a hardscaping project. It's important to note that this type of installation can have significant impacts on your yard, so it's often recommended to consult with a professional landscaper or hardscaping specialist before proceeding.

Prompt: What is the ordinary meaning of “physically restrained”?

Completion: “Physically restrained” means to be held back or prevented from moving by physical force or means. This can include being held by someone else, being tied up, or being confined in a way that limits your ability to move freely. It's often used in contexts where someone is being controlled or prevented from doing something, such as in law enforcement, medical settings, or in situations where someone might harm themselves or others.

In each case, Ministeral's completions were largely consistent with the chatbot responses observed by Judge Newsom. First, in Table 5's top cell, Ministeral outputs a plain-language characterization of *landscaping* which, similar to the definition offered to Judge Newsom by ChatGPT, “include[s] more than just botanical and other natural improvements” and “cover[s] both aesthetic and functional objectives.”²⁴⁸ Next, in Table 5's middle cell, when asked whether an in-ground trampoline installation could count as *landscaping*, Ministeral answers in the affirmative, with the caveat that the term *hardscaping*—“a subset of landscaping”—is a more precise label. This, too, is consistent with the response Judge Newsom received from ChatGPT.²⁴⁹ Finally, in Table 5's bottom cell, Ministeral—similar to what Judge Newsom observes with ChatGPT and Claude—defines *physically restrained* in a way that emphasizes the exertion of physical force or the use of tangible barriers.²⁵⁰

We then experimented with subtle variations of those original prompts, in an attempt to elicit qualitatively different metalinguistic judgments. In other words, this exercise involved what might be called “manual” prompt engineering, a process whereby human beings—in this case, we the authors—incrementally edit and adjust prompts to achieve a desired output. We found that with subtle but simple manipulations, it was possible to elicit the opposite judgments from the ones first observed. In Table 6, we present the results of adding a simple, neutral continuation to Judge Newsom's original “trampoline installation” query.²⁵¹

248. Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1225 (11th Cir. 2024) (Newsom, J., concurring).

249. *Id.* at 1225 (“Yes, installing an in-ground trampoline can be considered a part of landscaping.”).

250. United States v. Deleon, 116 F.4th 1260, 1274 (11th Cir. 2024) (Newsom, J., concurring) (“When defining ‘physically restrained,’ the models all tended to emphasize ‘physical force,’ ‘physical means,’ or ‘physical barriers.’”).

251. Note that underlining is added for emphasis and does not appear in the original completion.

TABLE 6. OUR MODIFICATION OF JUDGE NEWSOM’S ORIGINAL *LANDSCAPING* PROMPT, FOLLOWED BY THE COMPLETION SAMPLED FROM MINISTRAL.

Prompt: Is installing an in-ground trampoline “landscaping”? If not, what would you call it?

Completion: Installing an in-ground trampoline is not typically considered “landscaping” in the traditional sense. Landscaping generally refers to the design and maintenance of outdoor spaces, including plants, trees, flowers, hardscaping (like patios, walkways, and retaining walls), and other elements that enhance the aesthetic and functionality of a yard or garden. An in-ground trampoline, while it does involve some groundwork and may require excavation, is more accurately categorized as a recreational or outdoor structure installation. It’s often considered a part of backyard play areas or recreational facilities rather than traditional landscaping. Here are a few terms that might be more appropriate:

1. ****Backyard Installation**:** This term emphasizes that the trampoline is being installed in a backyard setting.
2. ****Recreational Structure Installation**:** This term highlights the recreational nature of the trampoline.
3. ****Outdoor Play Area Development**:** This term suggests that the trampoline is part of a larger effort to create a fun and safe play area.
4. ****Groundwork or Site Preparation**:** These terms focus on the physical work involved in preparing the site for the trampoline, such as excavation and leveling.

While these terms don’t capture the full scope of the project, they provide a more accurate description of the process involved in installing an in-ground trampoline.

Our modification to Judge Newsom’s original prompt explicitly raises the possibility that *landscaping* may not be a suitable label for Snell’s installation activity. However, it is not biased towards that judgment. Nevertheless, Ministrál’s completion in [Table 6](#) contradicts the completion observed in the middle cell of [Table 5](#).

Recall that in *Deleon*, a key question presented to the Eleventh Circuit was whether “physically restraining” a person entails the exertion of physical force upon that person, or the use of a physical barrier to restrict the person—which Judge Newsom calls “corporeality.”²⁵² Given Judge Newsom’s original prompt, Ministrál answered this question affirmatively. However, when provided a slightly altered prompt, Ministrál responds in a way that suggests corporeality is *not* always a necessary condition.

252. *Deleon*, 116 F.4th at 1274 (Newsom, J., concurring).

TABLE 7. OUR MODIFICATION OF JUDGE NEWSOM’S ORIGINAL PROMPT REGARDING THE ORDINARY MEANING OF *PHYSICALLY RESTRAINED*, FOLLOWED BY THE COMPLETION SAMPLED FROM MINISTRAL.

Prompt: In ordinary English, what kinds of meanings are associated with the phrase “physically restrained”?

Completion: The phrase “physically restrained” typically refers to the act of being held back or prevented from moving by physical means. Here are some common meanings associated with this phrase:

1. **Being Held Back:** Someone is physically restrained when they are held in place, often by another person or a device, to prevent them from moving freely.
2. **Controlled Movement:** It can also mean that a person’s movements are controlled or limited, such as when someone is handcuffed, tied up, or placed in a straitjacket.
3. **Preventing Harm:** In some contexts, physical restraint is used to prevent harm to the person being restrained or to others, such as in situations involving mental health crises or dangerous behavior.
4. **Legal or Medical Contexts:** In legal or medical settings, physical restraint might refer to the use of force or devices to control a person’s movements, often as a last resort when other methods have failed.
5. **Sports or Games:** In sports or games, physical restraint might refer to rules or actions that limit a player’s movement or actions, such as tackling in football or tagging in a game of tag.
6. Emotional or Psychological Context: In a broader sense, it can also refer to feelings of being held back or controlled, not just physically, but also emotionally or psychologically.

In all these contexts, the key idea is that the person’s physical freedom is being limited or controlled.

The prompt in [Table 7](#) differs from Judge Newsom’s original prompt in that it presupposes the existence of multiple ordinary meanings. In response to our altered prompt, Ministral provides six “common meanings” of *physically restrained*. The most interesting one for our purposes is (6), which centers the emotional and psychological sensation of “feeling[] . . . held back or controlled.”²⁵³ Similarly, Ministral’s completion in [Table 7](#) states that “the key idea

253. This interpretation is consistent with how the Eleventh Circuit had construed § 2B3.1(b)(4)(B) prior to *Deleon*: in *United States v. Jones* and *United States v. Victor*, the Eleventh Circuit held that § 2B3.1(b)(4)(B)’s sentencing enhancement can apply even in the absence of any physical contact. See *United States v. Jones*, 32 F.3d 1512, 1519 (11th Cir. 1994); *United States v. Victor*, 719 F.3d 1288, 1290–91 (11th Cir. 2013).

is that the person’s physical freedom is being limited or controlled,” with or without the use of physical force upon the physically restrained person. In this sense, Ministeral’s second completion clearly contradicts the first one, and it is in tension with the chatbot responses generally observed by Judge Newsom in *Deleon*.

This section has provided a simple demonstration of the ability to reach opposing conclusions about ordinary meaning through manual prompt engineering, in the exact two cases that have been offered by advocates of LLMs. We have shown that textualists who seek to use these tools in an honest search for ordinary meaning could easily draw opposing conclusions based on subtle, seemingly non-significant, prompt choices. This undermines the tool’s promises of objectivity, consistency, predictability, and transparency.

When it comes to the traditional tools of textualism, judges are rightfully skeptical of their limitations and the potential for manipulation. For example, the well-documented risk of “dictionary shopping”—in which a judge selects from multiple dictionaries to find a definition that supports a desired interpretation²⁵⁴—has long been a concern in textualist jurisprudence. Advocates of LLMs argue that they avoid the “cherry-picking” worries that have been raised against dictionaries.

However, our red-teaming exercise exposes a parallel risk in the context of LLMs: “prompt shopping.” By merely adjusting how we frame our queries, we can elicit substantially different interpretations of the same legal text, as evidenced by Ministeral’s contradictory answers to variations in prompting about the meaning of *landscaping* and *physically restrained*.

The current language modeling paradigm offers no inherent mechanism to mitigate or circumvent this risk. Just as judges must contend with the possibility of selective dictionary usage or selective application of interpretive canons, they will need to grapple with the potential for strategic prompt engineering to support particular interpretive outcomes. Consequently, there is no compelling reason to believe that LLMs—at least as currently employed through the “direct query” method—offer any meaningful advantage over traditional interpretive tools. In fact, they may simply introduce a new vector for the same kind of selective interpretation that already complicates the practice of legal textualism.²⁵⁵

When it comes to unwelcome human biases—conscious or unconscious—there is reason to believe that LLMs may actually pose greater risks than traditional interpretive tools. In general, “[h]umans have a tendency to disregard or not search for contradictory information in light of a computer-generated solution

254. James J. Brudney & Lawrence Baum, *Oasis or Mirage: The Supreme Court’s Thirst for Dictionaries in the Rehnquist and Roberts Eras*, 55 WM. & MARY L. REV. 483, 486–87 (2013).

255. In fact, there is an active domain of language modeling research devoted to *increasing* the variability of LLM outputs. See generally Guy Tevet & Jonathan Berant, *Evaluating the Evaluation of Diversity in Natural Language Generation*, PROCS. 16TH CONF. EUR. CHAPTER ASS’N FOR COMPUTATIONAL LINGUISTICS 326 (2021) (proposing a framework for evaluating diversity in generated text). This enterprise is in tension with LLM-assisted methods—such as the direct querying method—which assume that LLM outputs are generally stable across similar prompting regimes and across models.

that is accepted as correct.²⁵⁶ The propensity for users to anthropomorphize LLM-based platforms may introduce further bias,²⁵⁷ especially when this tendency leads users to falsely assume that LLMs possess human-like—or even “superhuman”²⁵⁸—reasoning skills or linguistic knowledge. Additionally, the general mystique surrounding artificial intelligence may cause legal professionals to be less cognizant of these models’ limitations. While judges have developed a healthy skepticism toward selective dictionary usage over decades of jurisprudence, they may be less equipped to recognize and account for the ways in which LLM responses can be manipulated—for example, through prejudicial prompting.

Finally, recent technological advances amplify the risk of prompt manipulation. Automated prompt optimization tools, which refine prompts with minimal human input, are improving at a rapid rate.²⁵⁹ In the context of legal interpretation, such tools could enable judges, lawyers, or scholars to easily generate neutral-sounding prompts designed to elicit particular interpretive outcomes, all without extensive manual intervention.

III. RECOMMENDATIONS AND BEST PRACTICES FOR USING LLMs TO AID LEGAL INTERPRETATION

In Part II, we have argued that there are strong reasons to doubt the efficacy of the emerging practice of directly querying chatbots for metalinguistic judgments about ordinary meaning. However, LLMs are impressive tools, and they may serve other uses for legal interpreters. Here, we survey some of those possibilities and outline best practices one should adopt when pursuing such uses. We start Section III.A with suggestions of best practices for responsible use of LLM chatbots in general.

Next, in Section III.B, we offer a conceptually sound strategy for LLM-assisted legal interpretation: a “dialectical” approach that avoids pitfalls of the direct query method. Finally, in Section III.C, we reflect on the Article’s broader implications for the interpretive theory of textualism. We argue that in this recent

256. M.L. Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems*, AIAA 1ST INTELLIGENT SYS. TECH. CONF., Sep. 2004, at 1; see also Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse*, 39 HUM. FACTORS 230, 239 (1997).

257. Recent work in the field of medical ethics has expressed similar concerns over LLM usage in medicine. “Adopting LLM-based systems may exacerbate automation bias because LLM outputs are perceived as highly relatable or trustworthy due to their compelling, human-like communication.” Jasmine Ong et al., *Medical Ethics of Large Language Models in Medicine*, 7 NEW ENG. J. MED. 1, 3 (2024).

258. Claims of “superhuman” abilities also pervade the academic AI literature. For a skeptical assessment of these assertions, see generally Simone Tedeschi et al., *What’s the Meaning of Superhuman Performance in Today’s NLU?*, 1 PROCS. 61ST ANN. MEETING ASS’N FOR COMPUTATIONAL LINGUISTICS 12471 (2023).

259. See, e.g., Reid Pryzant et al., *Automatic Prompt Optimization with “Gradient Descent” and Beam Search*, PROCS. 2023 CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING 7957, 7958 (2023).

development, judges' turn to LLM chatbots, and our critical analysis highlights fundamental limitations of current textualist interpretive theory and its conception(s) or "ordinary meaning."

A. A GENERAL CHECKLIST FOR RESPONSIBLE LLM USE

Several of the critiques in Part II reflect challenges with the reliability of LLMs in general—not just for legal interpretation. There are, however, strategies for utilizing LLMs that can encourage—though not guarantee—robustness and reproducibility of the results. We offer the following recommendations:

- (1) Use open models and interfaces;
- (2) Document interaction settings;
- (3) Use multiple models and prompts;
- (4) Independently verify specific claims in LLM output; and
- (5) Recruit domain experts.

These are discussed below in turn.

1. Use Open Models and Interfaces

One commonality in Judge Newsom's *Snell* and *Deleon* opinions is the reliance on closed-source platforms. As discussed in Part II, chatbots developed by private organizations such as OpenAI and Anthropic rely on data unavailable to the public and research community; the platforms rely on models trained under non-transparent conditions; and certain system properties—for example, system prompts—may not be readily accessible by end users.²⁶⁰ In fact, some of the relationships between the platform and the underlying LLM may be proprietary trade secrets, meaning that researchers and courts alike cannot fully scrutinize the inner workings of these systems.²⁶¹ For all these reasons, it is simply not possible to assess the properties of the underlying LLM in terms of its ability to reflect "ordinary" linguistic meaning.

Moreover, closed-source chatbots constantly change, often under opaque circumstances. The default LLM employed by ChatGPT is likely not the same model that platform used a month ago: "upgrades" from, for example, "davinci-003" to "gpt-4o" may reflect substantial changes to the core behavioral properties of ChatGPT.²⁶² Indeed, the model you're using today may not be the same model as yesterday: the underlying LLMs employed by commercial applications are constantly re-assessed and re-trained on user interactions and preference data

260. See *supra* Part II.

261. See *Closed-Source AI*, *supra* note 148.

262. For example, whereas davinci-003 facilitates text generation exclusively, GPT-4o is a "multimodal" model capable of generating text, images, and audio. See *Hello GPT-4o*, OPENAI (May 13, 2024), <https://openai.com/index/hello-gpt-4o>.

collected from OpenAI customers.²⁶³ Documenting the date and time of an interaction does not mitigate this issue: just because we know that a lawyer or judge queried ChatGPT, using model “gpt-4o,” on December 11, 2024, does not necessarily mean that we have any hope of interacting with the same model—with identical parameters—any time in the future.

Another important consideration in preferring open models is that closed models cannot be reliably queried with follow-up questions or asked for clarifications if their performance cannot be reproduced. In other words, a closed model cannot be “called to the stand” for further scrutiny. This is an especially important consideration if a model’s outputs are being used as a proxy judgment on ordinary language usage, as is the case with the direct query method. In such cases, opposing parties must be allowed to interact with the same model under controlled and reproducible conditions. The same applies in the appellate context: decisions based on closed model outputs immediately become inscrutable in a way that is qualitatively different from what open models allow.

2. Document Interaction Settings

Even when using open models, results cannot be reproduced—and are not transparent to scrutiny or appeal—without careful documentation of the interaction settings used. These include:

- (i) The model itself (a model designation, e.g. Mistral’s Ministral-8B-Instruct-2410).
- (ii) A version release date (e.g. December 6, 2024 was the last update to Ministral-8B Instruct-2410 at the time of writing; we employed this version in Part II).
- (iii) For closed models or interfaces, the date on which the model was consulted and the interface used.
- (iv) A complete interaction protocol, i.e. the exact prompt used for the query, as well as full model outputs and any subsequent inputs from the user.

As Judge Newsom correctly identifies, the latter point is particularly important: without full documentation of the entire interaction, it becomes impossible to critically assess the role of context in the particular response obtained.²⁶⁴ It is also imperative to document all variable settings used for the interaction.²⁶⁵ Another important interaction setting that must be documented is the model’s maximum output length. This setting constrains the amount of text generable from the

263. See *How Your Data is Used to Improve Model Performance*, OPENAI, <https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance> (last visited Sep. 20, 2025).

264. Judge Newsom writes, “I would recommend as a best practice: full disclosure of both the queries put to the LLMs (imperfect as mine might have been) and the models’ answers.” *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1229 (11th Cir. 2024) (Newsom, J., concurring).

265. For example, Part II *supra* discussed the “temperature” setting, which governs the “predictability” or “randomness” of model outputs.

model in a single interaction. The constraint has implications for the kinds of answers we can reasonably expect from the model—assuming, for example, that output length and output complexity are at least weakly correlated. Consider a scenario in which a practitioner queries a model limited to an output length of L tokens. One possible answer to the query is readily expressible with less than L tokens; another answer normally is formulated with more than L tokens. The former answer is preferred in this scenario, for the simple reason that outputs greater than L exceed the output length targeted by the system.

Similarly, users must document the amount of prior context that the model takes into account. Especially in cases where an interaction reflects a back and forth between the model and the user, later outputs may be affected by earlier interactions, but only to the extent that the length of the preceding context fits into the maximal input size. Recall as well that a user's chat history across sessions can affect system responses on platforms such as ChatGPT.²⁶⁶ In the interest of reproducibility, we generally recommend querying chatbots from blank session settings, under conditions that do not take into account specific user preferences.²⁶⁷

3. Use Multiple Models and Prompts

A major drawback to the direct query method is that it treats an LLM's output as an authoritative statement on the nature of ordinary language. We discussed the conceptual limitations to this approach in Part II and in Section III.A.1. above. In addition to raising these conceptual challenges to the direct query method, we have identified major practical concerns. In particular, the red-teaming exercise in Part II suggests that the direct query method is highly susceptible to manipulation.

It is tempting to treat the results of directly querying ChatGPT about, for example, the meaning of *landscaping* as indicative of how LLMs generally understand that term. However, as discussed in Part II, a single interaction with a single LLM is not even a robust sample of that single LLM's behavior—let alone a robust sample of LLM behavior in general. To that end, we recommend that users minimally employ multiple models and prompts before drawing general conclusions about LLMs.

In *Deleon*, Judge Newsom makes a concerted effort to query multiple chatbots, to mitigate against overreliance on ChatGPT.²⁶⁸ This is a step in the right direction: it is standard practice in natural language processing research to evaluate

266. See *Temporary Chat FAQ*, *supra* note 227.

267. Users generally have the most control over these settings when using an Application Programming Interface (API), which can provide a programmatic way to send requests to and receive responses from an LLM. If using a programmatic API to interact with an LLM, users should take care to document the details of the API interaction: these include the computer code used to access the API, the user's operating system, the API version, the employed programming language and version (e.g. Python 3.12.2), the relevant programming library used to access the API (along with library versions), and the user's hardware/cloud infrastructure.

268. See *United States v. Deleon*, 116 F.4th 1260, 1273 (11th Cir. 2024) (Newsom, J., concurring).

multiple models on a given task, which may reveal unexpected idiosyncratic variation among the models under comparison.

Beyond making use of multiple models, judges and lawyers interested in pursuing the direct query method should also take care to employ multiple prompts and to report the responses associated with each rephrasing of the query. Doing so mitigates the risk of unintended confirmation bias—or intentional manipulation—by demonstrating that an LLM judgment is robust to multiple plausible ways of eliciting the response.

Even when not pursuing the direct query method, it can be helpful—though not strictly necessary—to make use of multiple models and prompting strategies. For example, the “dialectical” prompting exercises reported below in Section III.B could have perhaps been augmented by asking multiple platforms to provide synthetic examples of *landscaping* used in context, or to provide competing analyses of the phrase *physically restrained*. Our alternative strategies leverage the creative capabilities of LLMs without treating models as authorities to which human interpreters should show deference. For that reason, LLM variability, that is, variability in response both within a single LLM and across different LLMs, is an asset—not a liability—on our favored approaches.

4. Independently Verify Specific Claims in LLM Output

Because LLMs can hallucinate—that is, they can “generat[e] plausible yet nonfactual content”—it is imperative that users independently verify claims and information contained within LLM outputs.²⁶⁹ This suggestion is especially important when users rely on LLM outputs for decision-making in high-stakes domains—such as law. In the context of textualist analysis, hallucinations can manifest as, for example, inaccurate generalizations about linguistic usage or falsified references to other sources such as dictionaries or published research.²⁷⁰ In some cases, there may not be readily-available empirical data that robustly supports or undermines a generalization regarding ordinary linguistic interpretation—for example, a statement of the form *Speakers would ordinarily understand X to mean Y*. If time and resources permit, judges could, with the assistance of experienced researchers, gather relevant data using corpus or experiment-based methods.

Some closed-source LLM chatbots—including, at the time of writing, Google’s Gemini and Microsoft’s Copilot—can search the internet to retrieve websites relevant to a user’s query and generate responses that synthesize those retrieved websites. This kind of internet search functionality does not safeguard

269. Lei Huang et al., *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*, 1 ACM TRANSACTIONS ON INFO. SYS. 1, 1 (2024).

270. In *Snell*, Judge Newsom counters that “hallucinations would seem to be most worrisome when asking a specific question that has a specific answer—less so, it seems to me, when more generally seeking the ‘ordinary meaning’ of some word or phrase.” *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1230 (11th Cir. 2024) (Newsom, J., concurring). However, in Section III.B, we will encounter a hallucinated dictionary definition produced by *Ministral* to support a particular position about the ordinary meaning of *physically restrained*.

LLM chatbots against hallucination; there is no guarantee that the chatbot response will be faithful to the retrieved content. Moreover, the chatbot may retrieve websites of dubious credibility. However, search features can facilitate the user's manual verification of the claims contained within the LLM output. For example, if the chatbot response includes a dictionary definition, coupled with a link to an online version of the dictionary, then the user can quickly verify that the chatbot's response aligns with the original source.

5. Recruit Domain Experts

When used in an unsophisticated manner, and without complementary linguistic research methods, lawyers and judges run the risk of missing crucial linguistic generalizations upon which an interpretive dispute may hinge. We encourage judges to—where possible—look to the field of linguistics for guidance on how to resolve hard cases of textual analysis. When this is not feasible, judges should strive to use LLMs in a way that stimulates robust theory-building rather than simply confirming one's prior beliefs about how language works.²⁷¹ Relatedly, when it comes to the use of AI tools in the courtroom, judges should expect best practices to evolve along with technological advances, which are proceeding at a rapid rate. The best practices offered in this Section reflect norms that are specific to a particular stage of technological development. As these norms evolve, it is imperative that judges seeking to employ AI stay current with advances in AI research.

The average judge, of course, does not have the time to take on the additional roles of linguist and AI scientist. However, there are other ways to expand the professional capabilities of the courtroom. We anticipate that as language technologies come to play a larger role in the textualist toolkit, it will become increasingly important for judges to hire clerks or other staff with domain expertise in AI, linguistics, and language modeling. In a number of interpretation cases heard at the Supreme Court, linguists and AI researchers have offered scientifically grounded insights and recommendations through *amicus curiae* briefs.²⁷² Lastly, there are specialized academic venues—including the longstanding International Conference on Artificial Intelligence and Law (ICAIL)—which present opportunities for professional lawyers and judges to engage with scholarship at the intersection of computer science, machine learning, and legal studies.²⁷³

B. TOWARDS DIALECTICAL LEGAL AI

In Section III.A., we highlighted several shortcomings of the direct query method pursued by Judge Newsom, ranging from fairly low-level characteristics of LLMs—for example, prompt sensitivity—to high-level conceptual dilemmas—for example, that LLMs are optimized to generate highly unordinary natural language. In this subsection, we motivate an alternative to the direct query method,

271. In Section III.B, we will further stress the importance of properly situating LLMs, in a broader context of legal and linguistic inquiry, into the nature of ordinary meaning.

272. Brief of *Amici Curiae* Professors Thomas R. Lee et al. in Support of Neither Party, *supra* note 61, at 15.

273. *ICAIL 2025*, NORTHWESTERN: ICAIL 2025, <https://sites.northwestern.edu/icail2025> [<https://perma.cc/P3P4-J92D>] (last visited Sep. 20, 2025).

followed by a demonstration of how our alternative may be deployed in practice and integrated with expert linguistic analysis. Our demonstration focuses on the issue at the heart of *Deleon*—the interpretation of *physically restrained* in the U.S. Sentencing Guidelines.

Before we present our alternative, it is worth reflecting on why the direct query method is so susceptible to the numerous pitfalls we have identified. The fundamental challenge of the direct query method is that it treats LLMs as agentic independent linguistic authorities to which judges can outsource the enterprise of ordinary-meaning textualism. This enterprise starts with a deceptively simple question: *What is the ordinary meaning of “X”?* For the textualist judge, the allure of the direct query method is that it seems to provide a lightning-fast and objective shortcut to a reasonable answer.

This is an old problem for textualism cloaked in the mystique of new technology. Judges have been rightly assailed for leaning too heavily on dictionaries and canons, in ways that mask the complexity of natural language and the essential role of context in interpretation.²⁷⁴ Making sense of this complexity requires rigorous scientific inquiry: formulating well-posed hypotheses about linguistic meaning and carefully testing those hypotheses against real-world linguistic data. The engine of this scientific pursuit is dialectical exchange—a continuous process of evaluating and refining competing theories in the search for understanding of how language “actually” works.

As noted in the previous subsection, judges need not—and *should* not—undertake this search in isolation. Linguists, especially experts in those subfields of linguistics that study natural language meaning—that is, semantics and pragmatics—can provide valuable insights across a broad spectrum of interpretive problems that judges confront.²⁷⁵ These insights are the product of decades—in some cases, centuries—of fruitful work within linguistics and across related disciplines. Our linguistic analysis of *physically restrained*, presented below in this subsection, is illustrative in this regard: we apply analytical insights from contemporary linguistic theory to arrive at a more rigorous account than those accounts attained through LLM querying alone.

However, practical constraints likely make it infeasible to consult linguists for every hard case of textual analysis. Moreover, linguistics—like all sciences—is incomplete and constantly evolving. There remain theoretical gaps that judges may need to bridge in order to render sound decisions. That is, there are cases in which judges may have to exercise their interpretive faculties in the absence of complete scientific guidance.

274. See generally LAWRENCE M. SOLAN, *THE LANGUAGE OF JUDGES* (1993).

275. See, e.g., Brief of *Amici Curiae* Professors Thomas R. Lee et al. in Support of Neither Party, *supra* note 61, at 2–3 (discussing structural ambiguity); Brandon Waldon, Cleo Condoravdi, James Pustejovsky, Nathan Schneider & Kevin Tobia, *Reading Law with Linguistics: The Statutory Interpretation of Artifact Nouns*, 62 HARV. J. LEGIS. (forthcoming 2025) (asserting that solely focusing on lexical ambiguity can lead to overreliance on dictionaries).

In such scenarios, judges can harness LLMs' capacity for dialectical engagement—their ability to generate novel perspectives that challenge or expand a judge's theory of a disputed legal text. In a word: brainstorming. This approach maintains judicial sovereignty over metalinguistic judgments that determine legal outcomes, rather than delegating that authority to the model. Moreover, this approach avoids some other existing critiques of judicial uses of LLM chatbots.²⁷⁶ Because a judge refrains from outsourcing the textualist enterprise, and assuming the judge verifies any factual claims output by the model, we need not worry too much about the peculiarities of how a model was developed, the exact means by which the judge interacted with it, or how exactly the model “works” at a mechanistic level. When employed in this manner, even closed-source LLM-based platforms can constructively inform legal decision-making while preserving the essential role of professional human judgment.²⁷⁷

We find it plausible that law clerks already play a similar role for judges faced with hard cases of legal interpretation: when confronted with ambiguous statutes or conflicting precedents, clerks serve as intellectual sparring partners who help judges stress-test their reasoning and surface alternative readings.²⁷⁸ Crucially, the judge alone must stand behind their final decision, regardless of how they arrived at it: “My clerk said _!” is not a viable justification. Similarly, though LLMs can assist judges in exploring and evaluating different interpretive possibilities, the ultimate responsibility for legal reasoning and judgment must remain with the human judge.

1. A Demonstration of Dialectical AI with “Physically Restrained”

To show how a dialectical approach to legal AI could work in practice, we revisit the interpretive dispute at the heart *Deleon*: the meaning of *physically restrained* in the context of the U.S. Sentencing Guidelines. Using *Deleon* as a case study, we show that judges can employ LLMs to evaluate competing theories of linguistic meaning and to formulate alternatives. This exercise demonstrates the

276. For example, Lee and Egbert argue that LLM chatbots return “artificial meaning,” rather than examples of actual, natural language usage. Lee and Egbert view such examples as irrelevant to textualist judges who seek evidence of (natural) language usage. See Lee & Egbert, *supra* note 9, at 1, 15. This concern is not relevant to our dialectical approach, however. Our approach does not treat LLM output as “data” in the sense of examples of how ordinary speakers have used, or would use, language. Instead, the LLM output is valuable as a resource to aid the judge's own linguistic reasoning.

277. In this sense, our preferred approach coheres with AI best practices which have been proposed in other disciplines. For example, in microbiology research, “AI should be considered a supportive tool rather than a replacement for human judgement and critical thinking.” Andrew J. Page, Niamh M. Tumelty & Samuel K. Sheppard, *Navigating the AI Frontier: Ethical Considerations and Best Practices in Microbial Genomics Research*, MICROBIAL GENOMICS, June 7, 2023, at 2. In medical ethics, “it is clear that AI has potential to complement, but not replace, human expertise.” Michael Balas et al., *Exploring the Potential Utility of AI Large Language Models for Medical Ethics: An Expert Panel Evaluation of GPT-4*, 50 J. MED. ETHICS 1, 6 (2023).

278. The point of this analogy is to emphasize that LLMs may be used in generating ideas for arguments, but should not be considered authoritative. We do not mean to suggest that LLMs can replace human clerks, or that ideas in LLM output are necessarily as valuable to a judge as human-generated ideas.

dialectical capabilities of LLM-based platforms—that is, it tests their ability to engage in reasoned analysis of competing interpretations and to generate novel interpretations. This case study is also of great practical significance. There is a substantial circuit split—among the 4th, 6th, 7th, 9th, 10th, 11th and D.C. Circuits—about the meaning of *physically restrained* in the sentencing guidelines.²⁷⁹

For this exercise, we again create a fully reproducible procedure around Mistral’s Ministral model. In *Deleon*, a key interpretive question was whether physical restraint always entails what Judge Newsom calls “corporeality”—acting upon the physically restrained individual “via either human touch or a tangible object.”²⁸⁰ We first prompted Ministral to provide the best justification it could for this theory of the meaning of *physically restrained*. The resulting completion is shown in [Table 8](#):

TABLE 8. MINISTRAL’S ARGUMENT IN SUPPORT OF THE CLAIM THAT *PHYSICALLY RESTRAINED* REQUIRES CORPOREALITY IN ORDINARY ENGLISH.

Prompt: What is the strongest argument for the following claim about English? “The phrase ‘physically restrained’ requires the application of tangible force upon the restrained entity, either through direct bodily contact or some other device or instrument.”²⁸¹

Focus only on how ordinary English speakers understand and use the phrase “physically restrained”.

Completion: The strongest argument for the claim that the phrase “physically restrained” requires the application of tangible force upon the restrained entity is based on the ordinary usage and understanding of the phrase by English speakers. Here are several points that support this argument:

1. ****Common Usage**:** In everyday language, “physically restrained” is typically used to describe situations where someone is held back or prevented from moving by some form of physical contact or barrier. For example, “The police physically restrained the protester” implies that the police used physical force to prevent the protester from moving.

279. Some Circuits hold that merely pointing a gun is insufficient. *See* *United States v. Herman*, 930 F.3d 872, 876–77 (7th Cir. 2019); *United States v. Parker*, 241 F.3d 1114, 1118–19 (9th Cir. 2001); *United States v. Drew*, 200 F.3d 871, 880 (D.C. Cir. 2000). Others, including the Eleventh Circuit, as in *Deleon*, hold that pointing a gun is sufficient. *See* *United States v. Wallace*, 461 F.3d 15, 34–35 (1st Cir. 2006); *United States v. Dimache*, 665 F.3d 603, 608 (4th Cir. 2011); *United States v. Howell*, 17 F.4th 673, 692 (6th Cir. 2021); *United States v. Miera*, 539 F.3d 1232, 1235–36 (10th Cir. 2008); *United States v. Deleon*, 116 F.4th 1260, 1265 (11th Cir. 2024). Judge Rosenbaum on the Eleventh Circuit has called for the circuit to rehear the issue en banc. *See Deleon*, 116 F.4th at 1265 (Rosenbaum, J., concurring).

280. *Deleon*, 116 F.4th at 1274 (Newsom, J., concurring).

281. This is an adaptation of how Judge Newsom formulates this theory of *physically restrained*. *See id.* at 1275.

2. **Dictionary Definitions**: Major dictionaries define “restrain” in terms of physical force. For instance, the Merriam-Webster dictionary defines “restrain” as “to hold back or prevent from moving or acting by physical force or influence.” This definition aligns with the idea that physical restraint involves tangible force.

3. **Legal Context**: In legal contexts, “physical restraint” is often used to describe situations where someone is held back by physical means, such as being handcuffed, held down, or prevented from moving by a physical barrier. This usage further supports the claim that physical force is required.

4. **Contrast with Other Forms of Restraint**: The phrase “physically restrained” is often contrasted with other forms of restraint, such as “psychological restraint” or “verbal restraint.” These contrasts imply that “physical restraint” specifically involves tangible force.

5. **Contextual Clues**: In most contexts where “physically restrained” is used, it is clear that some form of physical force is involved. For example, in a news report about a riot, the phrase “police physically restrained the crowd” would not make sense if it meant that the police merely talked the crowd into dispersing.

6. **Lack of Alternative Interpretations**: There is no widely accepted alternative interpretation of “physically restrained” that does not involve tangible force. If the phrase were to mean something else, it would be unusual and likely to cause confusion.

These points collectively provide a strong argument that ordinary English speakers understand and use the phrase “physically restrained” to mean that tangible force is applied to restrain the entity.

In [Table 8](#), we observe that Minstral produces a fairly cohesive argument for the position that physical restraint entails corporeality. First, the completion contains ordinary-language examples in support of this claim, including “[t]he police physically restrained the protester” and “police physically restrained the crowd.” Moreover, the discussion of these examples accords with our intuitions: the most natural interpretation of both examples is that the means of restraint involved physical contact. Finally, the completion features fairly sophisticated pragmatic observations. In (4), the point is made that there are salient alternative constructions that could be used to convey a lack of corporeality. In (5), the completion notes that it would be irrational for a speaker to use the phrase *physically restrained* in order to convey a “restraint without corporeality” meaning.

In some respects, however, this analysis clearly falls short. First, the completion includes assertions regarding “common” and “typical” usage, as well as whether there are “widely accepted alternative interpretation[s]” of *physically restrained*. Though these generalizations are supported with intuitional examples, the completion does not reference empirical data that robustly supports these generalizations. Second, we asked for an argument that focuses exclusively on

physically restrained as used and understood in ordinary English. What we received was an argument that, in (3), makes a generalization regarding the use of the phrase “[i]n legal contexts”. This generalization stands in tension with the circuit split regarding the meaning of *physically restrained* in the sentencing guidelines. Finally, the dictionary definition in (2) is a hallucination, pure and simple: the quoted entry does not appear anywhere on the internet—a Google search of the exact string returns no results—much less in *Merriam-Webster*.

Next, in a new chat session, we asked Ministeral the opposite question: what is the strongest argument against this position? We then asked for an alternative analysis of *physically restrained* consistent with the counter-argument. The results of this exercise are displayed in [Table 9](#):

TABLE 9. MINISTERAL’S ARGUMENT AGAINST THE CLAIM THAT *PHYSICALLY RESTRAINED* REQUIRES CORPOREALITY IN ORDINARY ENGLISH.

Prompt 1: What is the strongest argument against the following claim about English?

“The phrase ‘physically restrained’ requires the application of tangible force upon the restrained entity, either through direct bodily contact or some other device or instrument.”

Focus only on how ordinary English speakers understand and use the phrase “physically restrained”.

Completion 1: The strongest argument against the claim that “physically restrained” requires the application of tangible force upon the restrained entity is that ordinary English speakers often use the phrase in a broader sense that does not necessarily imply the use of force or physical contact.

For example, consider the following scenarios:

1. ****Non-physical restraint**:** A person might be “physically restrained” by being confined to a room without the use of physical force. For instance, a child might be “physically restrained” by being told to stay in their room as a punishment.
2. ****Implied restraint**:** In some contexts, “physically restrained” can imply a situation where someone is prevented from moving or acting freely, even if no physical force is applied. For example, a person might be “physically restrained” by being in a situation where they feel they cannot leave or act freely, such as being in a crowded room or being held back by social norms.
3. ****Metaphorical use**:** The phrase can also be used metaphorically. For instance, a person might say that their emotions are “physically restrained” when they feel they cannot express them freely.

These examples show that ordinary English speakers often understand and use the phrase “physically restrained” in a way that does not strictly require the application of tangible force. Therefore, the claim that it always requires such force is not accurate.

Prompt 2:²⁸² Now, provide a better description of the phrase’s meaning, consistent with your examples (1) and (2). Ignore the metaphorical use in (3). Your answer should start: “The phrase “physically restrained” requires. . .”

Completion 2: The phrase “physically restrained” requires that a person is prevented from moving or acting freely, either through direct physical contact, the use of a device or instrument, or by being confined to a specific location or situation. This can occur without the application of tangible force, as long as the person is effectively restricted in their movement or actions. For example, a child being told to stay in their room is “physically restrained” even if no physical force is used, and a person feeling unable to leave a crowded room due to social norms is also “physically restrained” in this broader sense.

As with the supporting argument, we observe intuitional examples in support of the main claims. In this case, however, we find these examples a bit questionable: is a child “told to stay in their room as a punishment” or a person “held back by social norms” necessarily *physically restrained*? We have some doubts that these metalinguistic judgments would truly cohere with ordinary intuition. However, the alternative analysis in the second completion is at least consistent with the claims observed in the second completion: according to the proposed analysis, a *physically restrained* individual is “effectively restricted in their movement or actions . . . even if no physical force is used.”

A textualist judge might consider the supporting argument compelling, at least in part. Or they might happen to prefer the dissenting argument and alternative theory. The judge could also choose to run this dialectic forward, asking an LLM to evaluate or synthesize competing analyses generated previously by LLMs. Ultimately, the value of these LLM outputs will be in the eye of the beholder, and it will be the judge’s responsibility to sift through the argumentation on both sides. We see this as a potentially fruitful exercise, insofar as it prompts the judge to reason through alternative ways of thinking about the language under dispute.

282. Note that Prompt 2 additionally consists of Prompt 1 and Completion 1. Further, of the three “scenarios” mentioned in Ministeral’s first completion, one—“Metaphorical use”—was clearly not relevant to our inquiry, so we asked Ministeral to “ignore” this usage when proposing an alternative analysis.

2. A Linguistic Perspective on “Physically Restrained”

In a case study of how dialectical LLM prompting can be employed in conjunction with rigorous linguistic analysis, we offer our own synthesis of the various positions advanced on the meaning of *physically restrained*, both by the human judges of the Eleventh Circuit and by the LLM Ministral. Our headline claim is that as a matter of ordinary meaning, when there is no additional context, *physically restrained* is linguistically indeterminate. In (a), we examine related expressions of the form *Adj-ly restrained*, in which *Adj* is an adjective (for example, *political*) and *Adj-ly* is the corresponding adverbial form (for example, *politically*). We demonstrate that these expressions give rise to two possible interpretations due to a lexical ambiguity inherent to *restrained*. In (b), we show how our analysis of *Adj-ly restrained* helps to make sense of the various interpretations available for *physically restrained*. In (c), we demonstrate that our analysis of *physically restrained* sheds light on the interpretive dispute at the heart of *Deleon*.

a. “Adj-ly Restrained” Is Often Indeterminate

Sequences of the form *Adj-ly restrained* frequently give rise to a subtle and overlooked linguistic indeterminacy whereby they admit two possible interpretations. The Adverb-as-Means interpretation typically associates *Adj-ly* with the means of restraining the restrained entity. The Adverb-as-Domain interpretation, on the other hand, typically associates *Adj-ly* with the nature of the restraint exhibited by the restrained entity; that is, the adverb specifies the domain of behavior in which the relevant entity exercises restraint. To demonstrate, consider the case of *verbally restrained*, which in the following two real-world examples exhibits two distinct interpretations. In example (1) below, the most plausible interpretation is that Bernie Sanders reigned in Markwayne Mullin’s behavior by verbal means—that is, through verbal chastisement or admonition. In (2), the most plausible interpretation is that the pensive pitcher has a tendency to exercise restraint in what he says, how he says it, or both.

1. [Adverb-as-Means] Sen. Markwayne Mullin (R-Okla.) had to be **verbally restrained** by . . . Committee Chairman Bernie Sanders (I-Vt.) from getting into a fight with the president of the Teamsters during a heated committee hearing.²⁸³
2. [Adverb-as-Domain] Because of an ill-considered comment by Frank Francisco, the Mets’ shaky closer, there was a heightened sense of drama in the ninth inning at Citi Field on Friday . . . But the game was really decided

283. Ian Swanson, *McConnell: Not My ‘Responsibility’ to Police Aggressive Behavior of Other Republicans*, THE HILL (Nov. 14, 2023, at 15:02 ET), <https://thehill.com/homenews/senate/4309461-mcconnell-not-my-responsibility-to-police-aggressive-behavior-of-other-republicans> [https://perma.cc/TBC8-VRUD] (emphasis added).

eight innings earlier when a more pensive and **verbally restrained** pitcher [Andy Pettitte] had the worst inning of his season.²⁸⁴

Similarly, consider the case of *politically restrained*. In (3), the U.S. military was *politically restrained* in the sense that their military operations were limited by the U.S. leaders' exercise of political power. (4), on the other hand, suggests that Rehnquist exhibited restraint in his own exercise of political power once Antonin Scalia and Clarence Thomas joined the Supreme Court:

3. [Adverb-as-Means] [D]uring the Korean War, U.S. political leaders restrained tactical operations in Korea to advance attainment of their grand strategic objectives in Europe. U.S. military commanders in Korea were **politically restrained** from conducting operations north of the Yalu River, which limited their ability to achieve their tactical objectives against North Korean and Chinese forces.²⁸⁵
- 4 [Adverb-as-Domain] When he was appointed to that tribunal two decades ago, it was clear that if Justice Rehnquist did not voice strict conservative views, such views would not be heard. Now, with Antonin Scalia and Clarence Thomas on the bench, there is less need for William Rehnquist to speak for the right. The Rehnquist moderation thesis may be little more than wishful thinking by those of a different political stripe. Still, there is evidence of a more **politically restrained** or astute, if not more moderate, William Rehnquist.²⁸⁶

The generalization suggested by (1) through (4) is as follows. In some cases, and as exemplified by (1) and (3), the most plausible interpretation of *Adj-ly restrained* is of being subject to a *restriction* imposed by someone or something through specific—e.g., verbal or political—means. In these cases, *Adj* does not necessarily describe the nature of the limitation itself. In (1), for example, the surrounding context makes clear that Bernie Sanders sought to prevent a physical, not verbal, fight between Markwayne Mullin and the president of the Teamsters.²⁸⁷ Similarly, in (3), it is military operations, not political operations, that are limited by political leaders' exercise of political power.

In other cases and as exemplified by (2) and (4), the most plausible interpretation of *Adj-ly restrained* is that the limitation is ultimately self-imposed by the “restrained” entity—somebody who *exercises restraint* or exhibits temperate

284. David Waldstein, *Mets Make Opening Statement Hold Up*, N.Y. TIMES (June 22, 2012), <https://www.nytimes.com/2012/06/23/sports/baseball/mets-top-pettitte-and-yankees-in-subway-series-opener.html> (emphasis added).

285. MAJOR RONALD BRYAN BELLAMY, SANCTUARY IN THE KOREAN WAR: A MANIFESTATION OF POLITICAL RESTRAINT 4 (2012) (emphasis added).

286. Vincent R. Johnson, *The Moderate Rehnquist*, 17 VT. L. REV. 267, 273 (1992) (emphasis added) (citations omitted).

287. Swanson, *supra* note 283 (“Mullin sprang up out of his chair at the Health Committee hearing earlier on Tuesday and invited Sean O’Brien, the general president of the International Brotherhood of Teamsters, to turn their social media spat into fisticuffs.”).

behavior. In these interpretations, *Adj* describes the nature of the limitation. In (2), the “verbally restrained” pitcher exercises restraint in his verbal behavior, but we do not necessarily draw the inference that the pitcher exercises restraint with respect to other types of action.

Something else serves to further distinguish the use of *Adj-ly restrained* in (1) and (3) from its use in (2) and (4). In (1), *restrained* is the passive participle form of the verb *restrain*. In (1), Senator Mullin is *restrained by* Senator Sanders; similarly in (3), it is clear from context that the U.S. military commanders were *restrained by* U.S. political leaders. By contrast, in (2) and (4), *restrained* is the dimension of comparison between two entities. In (2), Andy Pettitte is *more restrained* than Frank Francisco; in (4), William Rehnquist is *more restrained* than he was before Antonin Scalia and Clarence Thomas joined the bench. The usage of *restrained* in (2) and (4) is adjectival,²⁸⁸ not verbal. The two canonical interpretations of *Adj-ly restrained* exemplified by (1) through (4) are schematized in Table 10.

TABLE 10. TWO CANONICAL READINGS OF *ADJ-LY RESTRAINED*

Reading	Examples	Interpretation of <i>Adj-ly</i>	Part of speech of <i>restrained</i>
Adverb-as-Means	(1), (3)	Means of restraining	Verb
Adverb-as-Domain	(2), (4)	Nature of exhibited restraint	Adjective

For simplicity, Table 10 identifies the two possible readings with the interpretation of the adverb *Adj-ly*. However, the linguistic indeterminacy arises in large part due to a lexical ambiguity inherent to *restrained*, which, as we saw above, has both an adjectival and a verbal construal. These two construals of *restrained* are systematically associated with distinct senses.²⁸⁹ This can be further demonstrated

288. In particular, *restrained* functions in (2) and (4) as a so-called “gradable” adjective, which can appear in comparative constructions such as *more X* and which permit modification by so-called “degree modifiers”—for example, *very X*, *slightly X*.

289. Corroborating this claim, *Merriam-Webster* associates the two readings of “restrained” with different parts of speech. The adjective entry has the definition “marked by restraint: not excessive or extravagant.” *Restrained*, MERRIAM-WEBSTER, <https://www.merriam-webster.com/dictionary/restrained> [<https://perma.cc/J894-B8CF>] (last visited Sep. 20, 2025). The verb entry offers definitions that involve *restricting*, *preventing*, or *limiting* something/someone, or *depriving* someone of liberty. *See Restrain*, MERRIAM-WEBSTER, <https://www.merriam-webster.com/dictionary/restrain> [<https://perma.cc/PTS7-WFPR>] (last visited Sep. 20, 2025). Importantly, the Adverb-as-Means sense of *Adj-ly restrained* can appear as a modifier to a subsequent noun. For example, the *physically restrained* protester was escorted from the building. We assume that in these usages, *restrained* is a participial verb, not an adjective. This analysis is consistent with the *Cambridge Grammar of the English Language*—the verb *restrained* in the *restrained protester* is analogous to the verb *defeated* in the *defeated army*. John Payne & Rodney Huddleston, *Nouns and Noun Phrases*, in *THE CAMBRIDGE GRAMMAR OF THE ENGLISH LANGUAGE*, 323, 444 (Rodney Huddleston & Geoffrey K. Pullum eds., 2002).

through what is known as the *zeugma* test. A *zeugma* is a sentence that strikes native speakers as somehow odd or humorous because a word or phrase within that sentence is associated with distinct interpretations from the same token of usage.²⁹⁰ *Zeugma* is diagnostic of sense ambiguity. This is illustrated by (5), in which *bank*—a paradigmatic case of lexical ambiguity—may refer to a financial institution, as in (5)(a), or the area next to a river, as in (5)(b), but resists simultaneous interpretation according to both senses, as in (5)(c):

5.

- a. [Context: Noel spent the day working as a clerk at a Wells Fargo branch]
Noel spent the day at a bank.
- b. [Context: Liz spent the day fishing along the Potomac river]
Liz spent the day at a bank.
- c. # Noel spent the day at a bank, and so did Liz.²⁹¹

The sentences in (6) demonstrate that *restrained* can similarly exhibit *zeugma* effects. In (6)(a), *restrained* is identified with the adjectival sense in which the “restrained” individual exhibits restraint in his behavior; in (6)(b), the phrase is identified with the verb sense in which an external agent acts to “restrain” the individual. Similar to (5), (6)(c) demonstrates that these two distinct interpretations of *restrained* are not available for a single usage of the phrase.

6.

- a. [Context: Andy exhibited restraint at the press conference]
Andy Pettitte was restrained at his last press conference.
- b. [Context: Someone restrained Markwayne at the Senate hearing]
Markwayne Mullin was restrained at a recent Senate hearing.
- c. # Andy Pettitte was restrained at his last press conference, and at a recent Senate hearing, Markwayne Mullin was as well.

The above discussion suggests that these two readings of *restrained* tend to differ as to whether an external factor can be thought of as the “cause” of the restraint.²⁹² Most relevant for our purposes is that when *restrained* follows an adverb *Adj-ly*, the interpretation of *Adj-ly* correlates with how we resolve the lexical ambiguity of *restrained*.

290. Emanuel Viebahn, *Ambiguity and Zeugma*, 99 PAC. PHIL. Q. 749, 749 (2018) (“In an application of the *zeugma* test, a supposedly ambiguous expression is placed in a sentence in which several of its supposed meanings are forced together If the resulting sentence sounds *zeugmatic* (i.e. funny, strange or infelicitous), that is taken as evidence for ambiguity . . .”).

291. In linguistics research, the use of the “#” symbol preceding a sentence indicates that the sentence is infelicitous—that is, unidiomatic or nonsensical—but not necessarily ungrammatical.

292. A long tradition of linguistic scholarship is dedicated to understanding how notions of causality relate to linguistic interpretation. *See generally, e.g., CAUSATION IN GRAMMATICAL STRUCTURES* (Bridget Copley & Fabienne Martin, eds. 2014). However, we can largely avoid this matter for the present discussion.

b. “Physically Restrained” Exhibits the Ambiguity of “Adj-ly Restrained”

Interestingly, *Minstral* appears to contemplate these two distinct interpretations of *Adj-ly restrained* in its arguments for and against a “corporeality” requirement of *physically restrained*. In [Table 8](#), *Minstral* provides evidence that *physically restrained* is associated with an *Adverb-as-Means* interpretation—“restrained through physical means.” This is the most natural reading of “[t]he police physically restrained the protester” and of “police physically restrained the crowd.” These sentences clearly exhibit the verbal form of *restrained*, and they also strongly imply that the protester and the crowd are restricted in their physical movements as a result of this restraining action. It is hard to conceive of scenarios in which restraint through physical means would not give rise to a restriction in physical movement. On the other hand, one can readily conceive of scenarios in which one exhibits physical restraint even if one was not restrained through physical means. This latter kind of scenario is readily described by *physically restrained* on its *Adverb-as-Domain* reading.

In [Table 9](#), *Minstral* considers such scenarios when it presents its argument against a “corporeality” requirement for *physically restrained*. Both the “child . . . told to stay in their room” and the “person [who] feel[s] unable to leave a crowded room” are clearly exhibiting restraint in their physical movements as a response to external pressures—for the child, the risk of further punishment; for the person in the crowded room, “social norms.” These individuals are not acted upon through physical means in these contexts, but they are clearly exhibiting restraint in their physical movements—consistent with the *Adverb-as-Domain* reading.

Our primary goal in this subsection is simply to convince the reader that *physically restrained* is ambiguous between these two readings—a conclusion that could be discovered from dialectical prompting of LLM chatbots and one that can be developed through expert linguistic analysis. The above analysis, which is also corroborated by a dictionary source which attests to the ambiguity of *restrained*,²⁹³ suggests that corporeality is a necessary implication of just one reading of the phrase. In light of this ambiguity, one must look to context to determine the operable reading of *physically restrained*.

c. The Context of the U.S. Sentencing Guidelines Serves to Resolve the Indeterminacy

Recall that *Deleon* concerned the interpretation of *physically restrained* as it appears in the U.S. Sentencing Guidelines.²⁹⁴ The above discussion suggests that the interpretation depends on how an ambiguity inherent to *physically restrained* is resolved in context. The relevant portion of the Guidelines —Section 2B3.1(b)(4)—is presented below:

293. See *Restrained*, *supra* note 289; *Restrain*, *supra* note 289.

294. *United States v. Deleon*, 116 F.4th 1260, 1261 (11th Cir. 2024).

(4) (A) If any person was abducted to facilitate commission of the offense or to facilitate escape, increase by 4 levels; or (B) if any person was physically restrained to facilitate commission of the offense or to facilitate escape, increase by 2 levels.²⁹⁵

Our analysis makes clear that as a matter of ordinary meaning, *physically restrained*—with no further context—is ambiguous between two possible readings. Thus, as a matter of ordinary meaning, there are two a priori possible interpretations of *physically restrained* in Section 2B3.1(b)(4). On the *Adverb-as-Means* resolution of the ambiguity, the sentencing enhancement contemplated by Section 2B3.1(b)(4)(B) applies if any person was restrained through physical means “to facilitate commission of the offense or to facilitate escape.”²⁹⁶ On the *Adverb-as-Domain* resolution, the enhancement applies if any person exhibited restraint in their physical activity and if this restraint facilitated the offense or facilitated escape.

The statutory context of Section 2B3.1(b)(4) strongly supports the *Adverb-as-Means* reading. Note that the enhancement in Section 2B3.1(b)(4)(A) clearly contemplates an action involving two participants: the offender and *any person* whom the offender abducts to facilitate the crime or to facilitate escape. Moreover, *abducted* in Section 2B3.1(b)(4)(A) only has a passive participle verb construal. The parallel reading of Section 2B3.1(b)(4)(B) similarly construes *restrained* as a passive participle verb, and it similarly contemplates a two-participant action whereby *any person* is acted upon by the offender.

Looking beyond Section 2B3.1(b)(4), the broader context of the U.S. Sentencing Guidelines further supports an *Adverb-as-Means* reading. That phrase is defined in the Guidelines as follows: “‘Physically restrained’ means the forcible restraint of the victim such as by being tied, bound, or locked up.”²⁹⁷ Crucially, and consistent with the language of Section 2B3.1(b)(4), this statutory definition contemplates the *means* by which a victim is forcibly restrained, rather than the *nature* of the limitation exhibited by the victim—that is, a restriction in one’s physical movements.

Finally, it is notable that the *Adverb-as-Means* reading of Section 2B3.1(b)(4)(B) is the only reading that yields a plausible sentencing enhancement without having to stretch the text beyond its literal meaning. The *Adverb-as-Domain* disambiguation, on the other hand, literally contemplates any scenario in which a person, voluntarily or under duress, limits their movements in a way that facilitates the offense. This reading is clearly too broad to be plausible. In order to make the *Adverb-as-Domain* disambiguation of *physically restrained* a sensible one in context, we have to enrich Section 2B3.1(b)(4)(B) with an extra “duress” requirement, which stretches beyond the literal text of the statute.

295. U.S. SENT’G GUIDELINES MANUAL § 2B3.1(b)(4) (U.S. SENT’G COMM’N 2024).

296. *Id.*

297. U.S. SENT’G GUIDELINES MANUAL § 1B1.1 cmt. n.1(L) (U.S. SENT’G COMM’N 2024).

The *Adverb-as-Means* reading is sensible without a comparable enrichment. As Judge Rosenbaum writes in his concurring opinion, “I understand [Section 2B3.1(b)(4)(B)] to reflect Congress’s view that physical bounding adds another dimension to the intimidation a victim of an armed robbery endures. In other words, it’s bad enough to be robbed at gunpoint, but being tied up or otherwise physically bound is invasive, dangerous, and victimizing in yet another way.”²⁹⁸ Section 2B3.1(b)(4)(B), on the *Adverb-as-Means* reading of *physically restrained*, straightforwardly delivers an enhancement in line with this view. As a matter of literal interpretation, it contemplates a particularly “invasive, dangerous, and victimizing” means of restraining a person in the course of a robbery.²⁹⁹

d. How Our Linguistic Argument Contributes to the Debate

To appreciate how our linguistic analysis informs the current controversy surrounding the interpretation of the Sentencing Guidelines, we first review how the Eleventh Circuit has framed the interpretive dispute. The Eleventh Circuit has historically read *physically restrained* in a manner consistent with what we have called the *Adverb-as-Domain* interpretation. The Eleventh Circuit suggests that this reading is “broad” and non-ordinary,³⁰⁰ grounded in an expansive reading of a statutory definition which contemplates “the forcible restraint of the victim such as by being tied, bound, or locked up.”³⁰¹ According to the Eleventh Circuit, the alternative reading—consistent with what we have called the *Adverb-as-Means* interpretation—reflects the actual “ordinary” meaning of physically restrained.³⁰²

While our linguistic analysis arrives at the Eleventh Circuit’s same basic result with respect to the “ordinary” interpretation of Section 2B3.1(b)(4)(B), our analysis suggests a different way of reconciling Section 2B3.1(b)(4)(B)’s ordinary meaning with the statutory definition. Namely, we show how a broader context of the U.S. Sentencing Guidelines—including the statutory definition—suggests a straightforward resolution of an indeterminacy central to the ordinary meaning of *physically restrained*. The statutory definition, according to analysis, is not in tension with this resulting interpretation; rather, it is an additional piece of contextual evidence which suggests that Congress intended a particular ordinary resolution of the linguistic indeterminacy. In sum, Section 2B3.1(b)(4)(B) requires what Judge Newsom calls “corporeality,” not because that notion is

298. *Deleon*, 116 F.4th at 1267 (Rosenbaum, J., concurring).

299. *Id.*

300. *Id.* at 1265.

301. *Id.* at 1263 (“[I]n *United States v. Jones* . . . [w]e noted that the commentary defined ‘physically restrained’ as ‘the forcible restraint of the victim such as by being tied, bound, or locked up’ Then we reasoned that ‘such as’ in that definition ‘indicates that the illustrations of physical restraint are listed by way of example rather than limitation.’”).

302. *Id.* at 1268 (Rosenbaum, J., concurring); *id.* at 1275 (Newsom, J., concurring).

inherent to the ordinary meaning of *physically restrained*, but because it is a necessary implication of that phrase *in context*.³⁰³

e. A Tale of Two LLM Querying Methods (and Two Interpretive Rationales)

Let us take stock of the issues. Judge Newsom used LLM chatbots to inform the “ordinary meaning” of *physically restrained* in *Deleon*.³⁰⁴ His method is what we call a “direct query” method. He asked the chatbot “what is the ordinary meaning of ‘physically restrained’?” and noted that the answers centered around a “common core”: the LLM frequently emphasized the application of tangible force.³⁰⁵ Our critical conclusion argued in Sections III.B and III.C is that this “direct query” method is unsound. The prompting exercise in the current section corroborates our conclusion. The “ordinary meaning” of *physically restrained* does not require the application of tangible force. There are some contexts in which *physically restrained* describes the nature of the restraint, not the means.

In this section, we proposed an alternative approach, which we call a “dialectical” use of legal AI. Rather than treating the chatbot output as the *answer* to questions about ordinary meaning, we suggest that judges use the chatbot as a tool, to brainstorm possible considerations that are relevant to the analysis of ordinary meaning. We demonstrate such a usage in *Deleon*, which led to the discovery of an ambiguity: *physically restrained* is ambiguous between a “means” and “domain” interpretation. The remainder of this section has developed an original linguistic analysis of this ambiguity. In the context of the debate in *Deleon*, about the U.S. Sentencing Guidelines, there are strong contextual indications to favor the means interpretation of *physically restrained*.

This may seem like pedantic or trivial debate: Judge Newsom and we both agree that certain uses of chatbot LLMs can serve useful functions in some ordinary meaning inquiries, and we all agree about the linguistic outcome in *Deleon*—*physically restrained* means that physical force was used.³⁰⁶ Yet the differences in approach are important.³⁰⁷ Most importantly, Judge Newsom begins with a *direct*

303. In February 2025, we summarized this analysis of Section 2B3.1(b)(4)(B) in a public comment letter to the United States Sentencing Commission, co-authored with Professor Jonathan Wroblewski. In April 2025, the Commission announced a proposal to clarify the interpretation of the enhancement provision consistent with our analysis: the Commission’s proposed revision states in part that the enhancement applies only if “any person’s freedom of movement was restricted through physical contact or confinement, such as being tied, bound, or locked up, to facilitate commission of the offense or to facilitate escape.” U.S. SENT’G COMM’N, *Proposed Amendment, Circuit Conflicts*, in AMENDMENTS TO THE SENTENCING GUIDELINES (PRELIMINARY) 1, 3 (2025); see also Jonathan J. Wroblewski, *Lawyers, Linguists, Computer Scientists, and the Meaning of “Physically Restrained,”* SUBSTACK: SENT’G MATTERS (May 12, 2025), <https://sentencing.substack.com/p/lawyers-linguists-computer-scientists> [<https://perma.cc/XDQ2-BBT7>].

304. See *Deleon*, 116 F.4th at 1272 (Newsom, J., concurring).

305. See *id.* at 1272, 1275 (Newsom, J., concurring).

306. See *id.* at 1275 (Newsom, J., concurring).

307. As an analogy, consider the use of dictionaries. A textualist judge might look up *physically restrained* in one dictionary and reach the same bottom-line conclusion as Judge Newsom. But clearly that approach is not so sophisticated. In a similar way, we stress that even though we reach the same

query of an LLM chatbot and takes the answer as *evidence* of ordinary meaning.³⁰⁸ In contrast, we emphasize *dialectical queries* and taking the responses not as evidence of ordinary meaning but rather as *potential considerations* for a judge's linguistic analysis. More broadly, Judge Newsom's approach seems to be premised on the idea that many terms—at least *physically restrained*—have an *ordinary meaning* that is generally stable across contexts. In contrast, our approach emphasizes that language is context sensitive. The ordinary meaning of *physically restrained* often depends critically on context. We develop this theme further in the next section, as a lesson for textualism.

C. LESSONS FOR TEXTUALISM

The preceding analysis has broader implications for textualist theory, the interpretive view employed in the vast majority of U.S. Supreme Court decisions implicating statutory law.³⁰⁹ Textualists generally follow the “ordinary meaning doctrine” or at least a presumption of ordinary meaning.³¹⁰ This doctrine was described by Justice Scalia as “the most fundamental semantic rule of interpretation.”³¹¹ It holds that language has, or should be presumed to have, the meaning that it ordinarily “has in the context at issue.”³¹²

The ordinary meaning rule is also endorsed by Judge Newsom.³¹³ However, we have some disagreement with the particular operationalization of the rule. Although these might first seem like minor quibbles, they reflect broader lessons about the practice of textualism and the role of context in interpretation.

Judge Newsom describes the “standard approach” to ordinary meaning: “[s]pecifically, we break the phrase into its constituent parts, use conventional interpretive tools to gauge the ordinary meaning of each, and then piece those definitions back together into a coherent whole.”³¹⁴ We call this move the “constituent part splitting.” Moreover, he elaborates, in a move we call “abstract metalinguistic querying”:

[T]he *ideal* gauge of a word's or phrase's ordinary meaning would be a broad-based survey of every living speaker of American English—totally unrealistic, but great if you could pull it off. Imagine how that experiment would go: If you walked out onto the street and asked all umpteen million subjects, “What is the ordinary meaning of ‘physically restrained’?”, I think I can confidently

interpretive outcome as Judge Newsom and favor a use of the same tool, we are not advocating identical approaches.

308. *Deleon*, 116 F.4th at 1275 (Newsom, J., concurring).

309. *See, e.g.*, Krishnakumar *supra* note 23, at 575.

310. *See* SLOCUM, *supra* note 95, at 2.

311. SCALIA & GARNER, *supra* note 30, at 69.

312. SLOCUM, *supra* note 95, at 25; Peter W. Schroth, *Language and Law*, 46 AM. J. COMP. L. 17, 26 n.41 (1998).

313. *See, e.g.*, *Deleon*, 116 F.4th at 1271 (Newsom, J., concurring).

314. *Id.* For more on this approach and competing approaches, see generally Eskridge, Jr. et al. *supra* note 22.

guarantee that you would not get the exact same answer spit back at you verbatim over and over and over.³¹⁵

We question both of these moves. The first, “constituent part splitting,” begins interpreting a legal text by taking each term or phrase of the relevant text in isolation and examining its meaning. The second move, “abstract metalinguistic querying,” recommends settling the meaning of each term—in context—by identifying it with what speakers would describe as “the ordinary meaning” of the term—in the abstract. This approach might get the right answer in certain cases, namely in cases where the term has a generally stable, context-invariant meaning. For example, to interpret the meaning of “neither shall any Person be eligible to that Office [of the Presidency] who shall not have attained to the Age of thirty five Years,” we could take “thirty five” out of context and give it the meaning that most speakers would attribute to “thirty five” in the abstract.³¹⁶

But in many other cases, this approach—decontextualizing terms and attempting to set the *in-context* meaning to an abstract context-independent meaning—will fail.³¹⁷ Take the famous example of the rule “No vehicles may enter the park.” Consider four versions of such a rule:

1. No vehicles may enter the park.
2. No cars, trucks, and other vehicles may enter the park.
3. No cars, bicycles, skateboards, wheelchairs, and other vehicles may enter the park.
4. No vehicles may enter Retiro Park.³¹⁸

One view of language—a view we think is suggested by Judge Newsom—holds that there is simply *one ordinary meaning* of *vehicle*. If that were right, we could simply examine “vehicle” and identify its ordinary meaning—what a large survey would indicate is a vehicle. Perhaps most people would say that cars and trucks are vehicles, but roller skates and canoes are not.³¹⁹ And then we conclude that this is the ordinary meaning of “vehicle,” in all four versions of the rule.

We disagree with this view of language. Many contemporary textualists do as well; increasingly, textualists emphasize that they seek ordinary meaning *in context*. Arguably, in the context of rule (2), “vehicle” includes motorcycles but not roller skates; in the context of rule (3), vehicle would include roller skates; and in the context of rule 4—concerning a park with a large lake—vehicle plausibly includes canoes.

315. *Deleon*, 116 F.4th at 1276 (Newsom, J., concurring).

316. U.S. CONST. art II, § 1, cl. 5.

317. See Victoria Nourse, *Picking and Choosing Text: Lessons for Statutory Interpretation from the Philosophy of Language*, 69 FLA. L. REV. 1409, 1414–15 (2017).

318. A park in Madrid, Spain featuring a large lake.

319. See, e.g., Tobia, *supra* note 56, at 753 (presenting a survey of ordinary Americans, law students, and judges about what is a “vehicle”).

These observations are consistent with modern textualism, which emphasizes the role of context.³²⁰ Canons of interpretation, like *ejusdem generis*, reflect this sort of contextual reasoning.³²¹ And surveys of ordinary Americans show that they are sensitive to context in these ways.³²²

Such examples highlight a tension between textualism's commitment to context-sensitive interpretation and approaches to interpretation that posit a free floating "ordinary meaning" of terms—for example, part splitting of statutes, followed by "abstract metalinguistic querying." To assume that *vehicle* in a particular statute has the ordinary meaning that it has in all other uses is to ignore the critical role that context plays in establishing the ordinary meaning of that term in context.

We think there is an easy resolution to this dilemma. Textualists should give up on the notion of domain-invariant "ordinary meanings." Ordinary understanding of language is more sensitive to context, including textual context. Textualists should therefore favor tools that are adequately sensitive to context and be wary of tools that are premised on domain-invariant ordinary meanings. Context-sensitive arguments include some linguistic canons, like *ejusdem generis*. To be clear, we are not recommending textualists to apply these canons "mechanical[ly]" or in a "wooden" way.³²³ But some of the contextual canons reflect how ordinary speakers construe language. On the other hand, we would discourage judges from relying heavily on a dictionary definition as establishing *the* ordinary meaning—across all, or even many, contexts—of a term.³²⁴

We have already discussed this recommendation in the context of LLMs. Textualists should not directly query a chatbot about what *the* "ordinary meaning" of a term is. Whatever answer the chatbot gives may not correctly account for the context in which the term is used. Our skepticism would extend to a survey of many people about *the* ordinary meaning of a term. Such a survey can sometimes tell us interesting facts about language, but people do not always have introspective access to all the linguistic facts—if we did, legal interpretation would be much easier! More importantly, our view about the abstract, context-

320. See SLOCUM, *supra* note 95, at 25.

321. See Tobia, Slocum & Nourse, *supra* note 50, at 262.

322. See *id.*

323. See Ryan D. Doerfler, *Late-Stage Textualism*, 2021 SUP. CT. REV. 267, 269 (2021).

324. We are not arguing that judges should throw out the dictionary altogether. Dictionaries can be valuable resources for exploring the range of interpretations exhibitable by a word or phrase in frequently encountered contexts of use. For example, in Section III.B, Merriam-Webster corroborates our claim that *restrained* is associated with distinct interpretations. See *supra* Section III.B. However, a dictionary might not exhaustively characterize all possible interpretations; therefore, when considering whether a word or phrase has a possible interpretation, absence of corroborating dictionary evidence should not be taken as evidence of absence. Moreover, where multiple closely related descriptions are given in the same entry, they should not necessarily be treated as competing interpretations. See Stefan Th. Gries, Michael Kranzlein, Nathan Schneider, Brian Slocum & Kevin Tobia, *Unmasking Textualism: Linguistic Misunderstanding in the Transit Mask Order Case and Beyond*, 122 COLUM. L. REV. F. 192, 206 (2022).

invariant ordinary meaning of some term—e.g., *vehicle*—is not necessarily suited to the context of interpretation.

As we discussed previously, our recommendations for LLM usage in interpretation are not all critical and skeptical. On our view, language is frequently context sensitive, and it is often context sensitive in subtle ways. LLMs could be a useful supplement to other interpretive tools, particularly when they are leveraged in the “dialectical” mode we develop here.³²⁵

CONCLUSION

In a legal culture in which we are “all textualists,” judges are turning to chatbots that employ advanced language models to inform “ordinary meaning.” This is the first Article to assess these efforts from a technical, linguistic, and legal perspective. We have argued that such an approach is flawed and that it enables judges to “prejudicially prompt” LLMs to reach a preferred interpretive outcome. This is a critical cautionary lesson, as judges in the U.S. and abroad increasingly look to LLM chatbots in interpretation.

The Article also offers constructive contributions. We develop a list of best practices for legal interpreters who use chatbots, and we develop an original approach to using chatbots in interpretation, which we call “dialectical AI.” We demonstrate this approach in a novel case study that reveals a surprising linguistic ambiguity at the center of an important legal interpretive issue.

In the near future, it is likely that other textualist judges will directly query chatbots about ordinary meaning. We underscore our critical conclusion about this approach: LLM chatbots are impressive, but when it comes to ordinary meaning, don’t take their word for it. Instead, we encourage textualist interpreters to use chatbots in more sophisticated ways to help unearth easily overlooked insights about language.

325. See *supra* Section III.B.2.e.