Harvard Journal of Law & Technology
Volume 39, Number 1 Fall 2025

TRADITIONAL AND COMPUTATIONAL CANONS

*Eric Martínez**

ABSTRACT

As part of the rise of modern textualism, dictionaries and linguistic canons have become a ubiquitous part of legal interpretation. One longstanding question is whether judges citing these tools sincerely attempt to uncover ordinary meaning, or if their invocation is merely window-dressing for a preferred outcome. The practical significance of this question extends across all major doctrinal areas, and with the Supreme Court's overturning of Chevron deference, its importance is only to grow, as courts are now instructed to use every tool at their disposal to resolve ambiguity when interpreting a law. This Article is the first to formally investigate this question by measuring the extent to which courts align with linguistic consensus — as judged by both ordinary and expert readers — when asserting plain meaning by appealing to one interpretive tool over another.

After documenting the rise of plain meaning and interpretive canons in a sample of over two million opinions, the Article operationalizes several longstanding accounts regarding canon usage into formal models with testable predictions. The Article next tests these predictions via a large-scale, pre-registered experiment examining how lawyers (n = 2,373) and non-lawyers (n = 4,533) interpret the words at issue in 180 disputes in which a judge purported to follow the plain meaning of the text by invoking one canon at the expense of another. Three-quarters of participants tended to converge on the interpretation dictated by one canon over that of a counter-canon in a given case, and the court chose this interpretation in four out of five of cases. Incentivized lawyer-participants correctly predicted the majority interpretation at a similar rate. The results support new

---

* Earl B. Dickerson Fellow & Instructor in Law, University of Chicago Law School. JD, Harvard Law School; PhD, MIT Brain & Cognitive Sciences. Thanks to Jonathan Masur, Kevin Tobia, Hajin Kim, Jacob Goldin, Lee Fennell, Richard McAdams, Bridget Fahey, Lior Strahilevitz, Adam Chilton, Jamie Macleod, Dan Schwarcz, Tara Grove, Jon Choi, Holger Spamann, Sonja Starr, Brian Leiter, Joe Schottenfeld, Jared Mayer, Austin Peters, Fernanda de la Torre, and audiences at the American Association for the Advancement of Sciences Annual Meeting; Georgetown Law Symposium on Legal Interpretation and Data; American Law & Economics Association Annual Meeting; Connecticut Insurance Law & AI Conference; UChicago Law Junior Scholars Colloquium; and Chicago/Michigan PALS workshop. I am grateful to Jacob Levine and the entire editorial staff of the *Harvard Journal of Law & Technology* for their careful work on this piece. Thanks is also owed to the lawyers and non-lawyers who generously participated in the study. Anonymized data and code are viewable here: https://osf.io/jqk8p/?view_only=8f2733160ecc496c9bfc316dfed609c7.

interpretations of competing formalist and realist views regarding canon usage and suggest that judges tend to sincerely (albeit imperfectly) attempt to follow the meaning of the text when claiming to do so.

With the advent of Artificial Intelligence ("AI") models purportedly equipped with legal and linguistic competence, a second question concerns whether novel computational tools might offer a useful aid to uncovering the best reading of a legal text. Prompting flagship AI models on the same materials, this Article is the first to show that their predictions of linguistic consensus reliably match, though do not exceed, those of human judges invoking canons and dictionaries in real-world cases. This is true, even when controlling for factors such as knowledge of the case, suggesting that carefully deployed computational tools can offer an efficient, if not more effective, supplement to traditional interpretive tools.

Finally, leveraging both sets of findings, the Article advances a computational research program to both formalize new canons and illuminate the old ones, with the aim of transforming them from opaque slogans into interpretable predictions about language.

TABLE OF CONTENTS

I. INTRODUCTION

Dictionaries and linguistic canons of construction have become a ubiquitous part of legal interpretation and adjudication. As observed by commentators and formally documented by academics, the last several decades have seen a rise in the interpretive philosophy of textualism in both the Supreme Court and the federal judiciary writ-large, such that courts increasingly rely on analyses of text and plain meaning in arriving at a judicial decision regarding statutory interpretation.[1]

1. *See* Anita S. Krishnakumar, *Cracking the Whole Code Rule*, 96 N.Y.U. L. REV. 76, 97 (2021) (finding that between 2005 and 2017, the Roberts court relied on "text" and "plain meaning" in nearly 50% of majority opinions); Abbe R. Gluck & Richard A. Posner, *Statutory*

As part of the federal judiciary's increased reliance on text and plain meaning to resolve interpretive disputes, the Supreme Court in its majority opinions now tends to reference "at least one interpretive canon in resolving a question of statutory meaning."[2]

Throughout the state court system, reliance on these tools has been equally prevalent, if not more so. As a leading treatise on statutory interpretation notes: "No matter which version of the [plain-meaning] rule a court uses . . . all courts accept that standard, recognized, contemporary dictionaries are a valuable source to understand a word's approved, common meaning."[3] And even when canons "fell out of favor with federal judges [in the mid-1900s], the linguistic canons remained important to state court judges."[4]

With the Supreme Court's overturning of *Chevron* deference in *Loper Bright Enterprises v. Raimondo*,[5] the invocation of canons and dictionaries to resolve linguistic ambiguity is likely only to grow, as courts are now instructed to "use every tool at their disposal to determine the best reading of the statute and resolve the ambiguity"[6] instead of granting deference to an administrative agency's "reasonable interpretation" of a statute.[7]

Similarly, these tools also have been demonstrated to play a pivotal role in the interpretation of other legal documents, aside from statutes,

---

Interpretation on the Bench: A Survey of Forty-Two Judges on the Federal Courts of Appeals, 131 HARV. L. REV. 1298, 1300 (2018) ("The younger judges, who attended law school and practiced during the ascendance of textualism, are generally more formalist and accepting of the canons of construction."); Paul Killebrew, *Where Are All the Left-Wing Textualists?*, 82 N.Y.U. L. REV. 1895, 1895 (2007) ("Statutory textualism has adherents on the Supreme Court, throughout the federal judiciary, and, increasingly, in academia as well."); Kevin Tobia, *Dueling Dictionaries and Clashing Corpora,* 71 DUKE L.J. 146, 146 (2022) ("Textualism has broad support — at the Supreme Court, within the lower federal courts' new cohort of young 'Trump judges,' within many state courts, and even within the legal academy.").

2 . Nina A. Mendelson, *Change, Creation, and Unpredictability in Statutory Interpretation: Interpretive Canon Use in the Roberts Court's First Decade*, 117 MICH. L. REV. 71, 73 (2018).

3 . NORMAN SINGER & SHAMBIE SINGER, SUTHERLAND STATUTES AND STATUTORY CONSTRUCTION § 47:28 (7th ed. 2025).

4. LINDA JELLUM, STATUTORY INTERPRETATION IN THE FEDERAL AND STATE COURTS § 3.05 (LexisNexis ed., 2024).

5. 603 U.S. 369 (2024).

6. *Id.* at 373.

7. *Id.* at 472 (Kagan, J., dissenting).

such as contracts,[8] wills,[9] trusts,[10] deeds,[11] constitutions,[12] treaties,[13] jury instructions,[14] and court orders.[15]

At the same time, despite their ubiquity as interpretive tools by judges, both dictionaries and linguistic canons have long been derided by scholars as merely a smokescreen for judges' policy preferences. For example, Karl Llewellyn's seminal work, *Remarks on the Theory of Appellate Decision and the Rules or Canons about How Statutes Are to Be Construed*, famously argued that, in virtually any appellate case in which a judge turns to canons, there exist two equally applicable canons available to a judge that, if invoked, would lead to opposite results.[16] This observation has since been used to suggest canons "to be so malleable in their application as to operate mostly as pretext."[17]

Although modern textualists employing linguistic canons have sought to insulate themselves from such critique, in part by characterizing canons as context-sensitive generalizations of language usage as opposed to exceptionless rules,[18] and despite recent evidence supporting the existence of at least some such generalizations in everyday language,[19] commentators have argued that judges continue to use linguistic canons as mere "window-dressing," to rationalize decision-making on other means.[20] Similar critiques have recently been

---

8. *See, e.g.*, Epps v. Fowler, 351 S.W.3d 862, 866 (Tex. 2011) ("[W]e consult dictionaries to discern the natural meaning of a common-usage term not defined by contract, statute, or regulation."); *see generally* Ethan J. Leib, *The Textual Canons in Contract Cases: A Preliminary Study*, 2022 WIS. L. REV. 1109 (2022).

9. *See, e.g.*, *In re* Horner's Will, 82 N.Y.S.2d 491, 493 (1948) (invoking *ejusdem generis* canon to interpret the phrase "personal effects" in a will).

10. *See, e.g.*, *In re* Thomas P. & Janet A. Hendy Revocable Tr., No. 257228, 2006 WL 859445, at *2 (Mich. Ct. App. Apr. 4, 2006) (using Black's Law Dictionary to interpret a word in a trust).

11. *See, e.g.*, Vogel v. Cobb, 141 P.2d 276, 280 (Okla. 1943) (invoking ejusdem generis canon to interpret the phrase "other minerals" as used in a deed).

12. *See, e.g.*, District of Columbia v. Heller, 554 U.S. 570, 582 (2008) (relying on contemporaneous Johnson and Webster dictionary definitions to interpret the phrase "keep Arms" in the Second Amendment).

13. *See* Martinez v. United States, 828 F.3d 451, 459 (6th Cir. 2016) (relying on bilingual dictionaries to interpret the phrase "prescripción" in a treaty).

14. *See, e.g.*, Fennell v. State, 128 S.E.2d 43, 45 (Ga. 1962) (applying the conjunctive/disjunctive canon to interpret jury instructions).

15. *Cf.* Cal. Lumbermen's Council v. FTC, 115 F.2d 178, 184–85 (9th Cir. 1940) (applying the conjunctive/disjunctive canon to the interpretation of a court order).

16. *See* Karl N. Llewellyn, *Remarks on the Theory of Appellate Decision and the Rules or Canons About How Statutes Are to Be Construed*, 3 VAND. L. REV. 395, 401–06 (1950).

17. Ryan Doerfler, *Late Stage Textualism*, 2021 SUP. CT. REV. 267, 267 (2021).

18. *See, e.g.*, ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 51 (2012).

19. *See* Kevin Tobia, Brian G. Slocum & Victoria Nourse, *Statutory Interpretation from the Outside*, 122 COLUM. L. REV. 213, 250–52 (2022) (finding, for example, that people implicitly invoke the gender and number canons when reading a rule or law).

20. *See* Abbe R. Gluck & Richard A. Posner, *Statutory Interpretation on the Bench: A Survey of Forty-Two Judges on the Federal Courts of Appeals*, 131 HARV. L. REV. 1298, 1330

levied at judges' use of dictionaries, with scholars accusing the Supreme Court of using dictionaries to engage in what "purports to be careful, detailed linguistic analysis but what is, upon closer inspection, mildly elaborate obfuscation."[21]

Yet amidst this mismatch between the increased internal adoption of these tools by judges and widespread external criticism of their usage, it remains an open question to what extent judges successfully arrive at a legal text's plain meaning when claiming to follow a legal text's plain meaning.

Similarly, to the extent that judges citing traditional tools are unable or unwilling to arrive at the plain meaning of a legal document, it remains an open question whether novel computational tools such as large language models ("LLMs") — themselves designed to predict language usage in a given context — might offer a useful supplement or alternative.

Recent literature has documented the rise of capabilities of LLMs on both domain-general [22] and law-specific [23] language tasks. In addition, some leading cognitive scientists and language experts have gone as far as arguing that LLMs are not only good at using language, but can also challenge and inform theories of human linguistic

---

(2018) ("Linguistic canons especially, as opposed to policy canons, seem to be of [the] 'window-dressing' variety."); *see also* REED DICKERSON, THE INTERPRETATION AND APPLICATION OF STATUTES 234 (1975) (criticizing Latin canons that "masquerade as rules . . . while . . . describing results reached by other means"); Einer Elhauge, *Preference-Eliciting Default Rules*, 102 COLUM. L. REV. 2162, 2206 (2002) (describing canons sometimes used as "makeweight" support for conclusions about statutory meaning or legislative intent reached on other grounds); Evan C. Zoldan, *Canon Spotting*, 59 HOU. L. REV. 621, 638 (2022) ("[T]here are those who argue persuasively that the canons are mere window dressing."); KENT GREENAWALT, STATUTORY INTERPRETATION: 20 QUESTIONS (2008) ("[J]udges who are actually deciding in a particular way because they favor the policy it promotes may choose to put forward a canon that supports their conclusion, portraying it as much more decisive than it really is for them."); WILLIAM N. ESKRIDGE, JR., DYNAMIC STATUTORY INTERPRETATION 275 (1994) (noting that the legal realists considered the canons "window dressing"); Mendelson, *supra* note 2, at 76 ("Judicial reliance on canons has, however, lately drawn its own share of blistering criticism as capricious and potentially manipulative."); Anita S. Krishnakumar, *Dueling Canons*, 65 DUKE L.J. 909, 959 (2016) ("[T]he Justices could be using the canons as just window-dressing for results reached for other reasons.") (citation omitted).

21. *See, e.g.*, Doerfler, *supra* note 17, at 305.

22. *See generally* OpenAI, *GPT-4 Technical Report* (Mar. 4, 2023) (unpublished manuscript), https://arxiv.org/pdf/2303.08774 [https://perma.cc/2EJX-ACP8].

23. *See, e.g.*, Daniel Martin Katz, Michael James Bommarito, Shang Gao & Pablo Arredondo, *GPT-4 Passes the Bar Exam*, 382 Phil. Transactions Royal Soc'y A 1 (2024) (finding that the large language model GPT-4 from OpenAI passed the Uniform Bar Exam); Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana et al., *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models* (Sept. 26, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4583531 [https://perma.cc/Y4MU-DXGG] (reporting on studies of LLM success at various lawyering tasks).

cognition.[24] In light of these developments, and in spite of documented shortcomings,[25] several legal academics[26] and even some judges[27] have begun advocating for using LLMs as a tool to uncover the ordinary meaning of a phrase in a legal document.

Others have vehemently opposed this practice, recently claiming that LLMs are "unreliable judges"[28] that "are not up to the task"[29] of uncovering ordinary meaning, and, by extension, that "[j]udges should not rely on direct queries to [LLMs] about the meaning of legal texts."[30] However, it remains to be seen to what extent LLMs, as compared to human judges invoking linguistic canons, are able to uncover the consensus meaning of a term at issue in real-world cases.

The purpose of this Article is to investigate these questions, via a combination of large-scale text analysis, original behavioral experiments, and state-of-the-art LLM-prompting, assessing the degree of linguistic consensus, as well as judge and LLM alignment with said consensus, in real-world cases involving questions of interpretation.

---

24. *See, e.g.*, EDWARD A.F. GIBSON, SYNTAX: A COGNITIVE APPROACH 15–18 (2025) (stating that the best models of human language are arguably LLMs); Steve Piantadosi, *Modern Language Models Refute Chomsky's Approach to Language*, *in* FROM FIELDWORK TO LINGUISTIC THEORY 384 (Edward Gibson & Moshe Poliak eds., 2024) (LLMs "make language the most exciting arena in all of cognitive science and AI," and "are also a tool that will help linguistics to refine theories and compare leading ideas to strong alternatives").

25. *See, e.g.*, Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,* 22 J. EMPIRICAL LEGAL STUD. 216, 224–25, 232 (2024) (finding that cutting-edge LLMs continue to hallucinate at a high rate); Eric Martínez, *Re-Evaluating GPT-4's Bar Exam Performance*, 32 A.I. & L. 581, 584–86, 589–90 (2024) (finding that reports of LLM performance on the Uniform Bar Exam were overinflated); Brandon Waldon, Nathan Schneider, Ethan Wilcox, Amir Zeldes & Kevin Tobia, *Large Language Models for Legal Interpretation? Don't Take Their Word for It*, 114 GEO. L.J. 115, 115 (2025) ("[L]egal practitioners run the risk of inappropriately relying on LLMs to resolve legal interpretative questions."); Lisa Larrimore Ouellette, Amy R. Motomura, Jason Reinecke & Jonathan S. Masur, *Can AI Hold Office Hours?*, J. LEGAL EDUC., 1 (forthcoming) (finding that a "substantial number of responses" of LLM "were unacceptable in the sense of being harmful for learning").

26. *See, e.g.*, Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451, 455 (2024) (using LLMs to resolve contested meaning of contract terms); Christoph Engel & Richard H. McAdams, *Asking GPT for the Ordinary Meaning of Statutory Terms*, 2024 U. ILL. J.L. TECH. & POL'Y 235, 237 (2024); *cf.* Daniel Schwarcz & Jonathan H. Choi, *AI Tools for Lawyers: A Practical Guide*, 108 MINN. L. REV. HEADNOTES 1 (2023) (describing how lawyers can use LLMs to produce high-quality legal writing); Jonathan H. Choi, *Measuring Clarity in Legal Text*, 91 U. CHI. L. REV. 1 (2024) (using vector-based language model to measure clarity in legal text).

27. *See* Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1226–34 (11th Cir. 2024) (Newsom, J., concurring).

28. *See, e.g.*, Jonathan H. Choi, Off-the-Shelf Large Language Models Are Unreliable Judges (Nov. 26, 2025) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5188865 [https://perma.cc/XN6L-72Y7].

29. Thomas R. Lee & Jesse Egbert, *Artificial Meaning*, 77 FLA. L. REV. (forthcoming 2025).

30. Waldon, Schneider, Wilcox, Zeldes & Tobia, *supra* note 25, at 4.

Part II begins by documenting the prevailing role of plain meaning in judicial decision-making, as well as the role of various tools in uncovering the plain meaning of legal text. Analyzing every published opinion of the U.S. Supreme, circuit, and district courts between 1881 and 2020, as well as a comparably sized sample of state-level decisions, this Article finds, in line with the prior literature, that decisions referencing issues of "plain meaning" and relevant synonyms have formed an increasingly substantial proportion of all judicial decisions. The same holds true for linguistic tools purported to aid in the resolution of these cases, such as *expressio unius*, *noscitur a sociis*, analysis of grammar and punctuation, the last-antecedent rule, and both legal and non-legal dictionaries. Part II also provides background on the primary empirical questions addressed in this Article.

Part III begins by formalizing these debates into models with testable predictions. Part III then tests these predictions via a well-powered, pre-registered experiment investigating the degree of linguistic consensus — as judged by lawyers (n = 2,373) and laypeople (n = 4,533) — and judicial alignment with said consensus across 180 real-world cases in which two apparently equally plausible canons pulled in opposite directions. The behavioral experiments revealed a relatively high degree of linguistic consensus among both lawyers and ordinary people regarding the meaning of the words at issue. On average, 75–80% of participants converged on the interpretation dictated by a certain canon over that of a counter-canon.

When comparing participant responses to the court interpretation, the experiments further revealed that the interpretation reached by a judge aligned with that of the majority of participants in approximately 80% of cases. This alignment rate descriptively matched the rate at which incentivized lawyer participants accurately predicted the majority interpretation among other lawyers. Similar levels of alignment were observed when breaking down results by demographic variables such as gender, race, politics, and age, as well as (in the case of lawyers) potential familiarity with the case. In addition, analyses further revealed surprisingly similar levels of consensus, as well as alignment with consensus, across court level, jurisdiction, and genre of legal text, though there was some notable variation between certain canons.

Part IV presents the results of an LLM-prompting experiment investigating the extent to which cutting-edge AI models, such as OpenAI's GPT-4.1 and o3, align with ordinary and expert consensus. Analyses revealed the predictions of these models aligned with consensus at rates similar to each other and to those of human judges invoking traditional tools. These results were similarly robust to a host of control analyses, including potential data contamination in the model's training set.

Part V addresses possible objections to the study's design and lays out seven implications on the basis of the empirical results.

First, the data show that when courts invoke the plain-meaning doctrine, a relatively large consensus of ordinary and legally trained readers converge on the same interpretation, though a substantial minority still disagrees. In H. L. A. Hart's terms, these findings sharpen our sense of where such disputes fall on the spectrum between the "central core" of settled meaning and the surrounding "penumbra" of indeterminacy.[31] For critics in the Llewellyn-style tradition — who argue that every canon is neutralized by an equal-and-opposite counter-canon — this pattern might spell bad news: *linguistic* canons, it turns out, are not mere coin-flips. Conversely, insofar as one assumed the canons to always yield a single, indisputable answer, the persistent 25% dissensus should commensurately update one away from this position. Regardless of where one locates the boundary between determinacy and indeterminacy, this study provides estimates and formal modeling to tie judicial assertion of "plain meaning" interpretation to observable thresholds of reader convergence and model-predicted error.

Second, these data supply crucial evidence for weighing the longstanding charge that linguistic canons and dictionaries are merely political window-dressing. Insofar as one began convinced that judges cherry-pick linguistic tools solely to advance policy preferences, these numbers should move one toward a more charitable view; to the extent that one assumed courts unfailingly reach the one correct answer, the remaining 20% misalignment should likewise temper that confidence. On balance, the pattern suggests that judges usually, though not infallibly, choose the interpretive aid that best tracks linguistic consensus. And where courts diverge from consensus, the difference looks less like strategic, policy-driven manipulation and more like the predictable by-product of cognitive constraints, such as limited attention, salience effects, noisy inference, and overreliance on familiar heuristics.

Third, the results help pinpoint which interpretive canons most faithfully mirror ordinary language use in the kinds of cases that reach court. Canons such as *expressio unius*, *noscitur a sociis*, and the "may versus shall" canon tended to strongly track reader consensus in the instances in which judges relied on them, whereas the last-antecedent rule and *ejusdem generis* aligned with audience judgments only inconsistently. For judges who ground their methodology in textual fidelity, these findings supply concrete guidance on which canons

31. H. L. A. Hart, *Positivism and the Separation of Law and Morals*, 71 HARV. L. REV. 593, 607 (1957) ("There must be a core of settled meaning, but there will be, as well, a penumbra of debatable cases in which words are neither obviously applicable nor obviously ruled out.").

deserve greater weight — and which call for caution — when the professed aim is to capture plain meaning.

Fourth, by operationalizing plain meaning in terms of both ordinary readers (laypeople) and well-informed readers (lawyers) and comparing the difference, this dual benchmark clarifies the practical gap between competing strands of textualism and pinpoints which conception better predicts how judges actually decide cases.

Fifth, by comparing the ability of state-of-the-art LLMs to predict linguistic consensus with that of lawyers, laypersons, and judges, this study gauges whether computational tools can complement — or someday supplant — traditional aids such as linguistic canons and dictionaries. Because the models generally matched but did not exceed the rate at which human respondents agreed with the courts, they appear poised to serve as helpful supplements rather than full replacements for conventional interpretive methods.

Sixth, analyses across prompts, hyper-parameters, and model families point to concrete "dos and don'ts" for judicial use of LLMs. Overall accuracy remained stable across temperature and model family but climbed noticeably under few-shot prompts that supplied the model with several example items annotated with the full distribution of human responses — worked examples of how people in fact answered — before it made its own predictions. The systems were consistently better, relative to lawyer participants, at forecasting lay consensus than lawyer consensus, suggesting that they may be most useful as a guide to ordinary rather than expert usage. Importantly, however, this was heavily contingent on prompt. The LLMs estimated the full distribution of reader judgments far more accurately when asked to report probabilities directly instead of being queried repeatedly for binary answers, suggesting an advantage of direct probability elicitation over repeat yes/no prompting.

Seventh, because linguistic canons, properly understood, are an open set of heuristics that proxy language use, the same computational methods that apply existing maxims can also uncover and formalize new ones. LLMs, syntactic parsers, and other machine-learning techniques can mine statutory and judicial corpora for lexical or structural patterns, yielding "emergent canons" to enrich the interpretive toolkit. Crucially, these tools can also illuminate the old canons. Traditional maxims, like AI models, [32] have long been criticized as black-boxes:[33] Llewellyn's famous "thrust-and-parry" list suggests a canon (and counter-canon) for every proposition, yet the

32. *See generally* Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889 (2018).

33. Thomas R. Lee & Stephen C. Mouritsen, *The Corpus and the Critics*, 88 U. CHI. L. REV. 275, 297 (2021) (arguing that ordinary-meaning analysis "resides within the black box of judicial intuition").

generative logic behind a judge's selection often remains opaque.[34] Drawing on insights from computational cognitive science, this Article proposes avenues to reverse-engineer the cues that drive both LLM outputs and judicial use of canons, revealing *why* a particular maxim fits a given context. By pairing discovery pipelines for new heuristics with explanation-oriented interfaces for old ones, this research program promises not only to expand judges' interpretive resources but also to render their deployment more transparent and auditable in the courtroom.

II. SETTING THE STAKES

This Part provides background on the empirical questions being investigated in the Article. Section II.A (1) documents the importance and prevalence of the "plain-meaning rule" in a sample of 2.5 million published opinions across the federal and state judiciary; and (2) details the open empirical questions related to the rule.

Section II.B (1) provides an overview of canons, dictionaries, and other tools used by judges when purporting to uncover plain meaning; (2) reviews longstanding, empirically unexplored criticisms of judicial use of these tools; and (3) distinguishes these criticisms from a third possibility — left for future work — that canons do track ordinary readers' judgments but operate chiefly as post hoc labels for judges' preexisting linguistic intuitions.

Section II.C (1) documents the rise in the capabilities of LLMs on domain-general and law-specific tasks; and (2) explores the open empirical questions regarding their usage.

A. Plain Meaning

1. Doctrinal Importance of Plain Meaning

Today, courts regularly interpret a legal document according to the "plain" or clear meaning of the text as judged by an ordinary or reasonable reader. The importance of this rule extends to the interpretation of virtually every interpretive instrument, such as contracts, wills, trusts, deeds, constitutions, treaties, jury instructions, and court orders.[35] The importance also extends across jurisdictions,[36]

34. Llewellyn, *supra* note 16, at 401–06.

35. *See supra* notes 6–13 and accompanying text.

36. For example, in Texas, courts begin with a word's "plain and common meaning" and go no further if the text is unambiguous. *See* City of Houston v. Bates, 406 S.W.3d 539, 543–45 (Tex. 2013) (applying the plain meaning of the statute's language). Similarly strong commitments are found within the courts of other states such as: Florida, *see* State Farm Mut.

both within federal and state courts. Moreover, the rule is widely endorsed across interpretive theories, consistent not only with textualism[37] but also with intentionalism[38] and pragmatism,[39] for example.

Against this doctrinal backdrop, a natural empirical question is how often courts actually invoke the rule. Previous work has investigated this question as applied to cases involving statutory interpretation in particular courts at a more limited scale, such as the United States Supreme Court and state supreme courts over the last few decades.[40]

For robustness purposes, this Article replicates and extends this work by documenting references to plain meaning across various levels of both the federal and state judiciaries in a sample of over two million

---

Auto. Ins. Co. v. Spangler, 64 F.4th 1173, 1179 (11th Cir. 2023) ("Faced with an undefined term in the Policy, we defer to Florida's rules of contract construction, which instruct us to give the term its . . . plain and ordinary meaning."); Wisconsin, *see* Sharpe v. Hasey, 114 N.W. 1118, 1119 (Wis. 1908) ("We are not to depart from the plain ordinary sense unless it is manifest that such was the legislative purpose."); California, *see* Universal Pictures Corp. v. Superior Ct. of L.A. Cnty., 50 P.2d 500, 502 (Cal. 1935) ("[I]t is a universal rule that . . . words in a statute should be given their ordinary meaning."); Wyoming, *see* Garton v. State, 910 P.2d 1348, 1353 (Wyo. 1996) ("If the language of a statute is clear and unambiguous, we apply the plain and ordinary meaning of the words."); and Michigan, *see* People v. Labbe, 168 N.W. 451, 453 (Mich. 1918) ("In statutory construction it is the rule that unless some expression in the statute clearly indicates otherwise, words in general use should be given a common construction according to their generally accepted meaning.").

37. *See, e.g.*, Amy Coney Barrett, *Substantive Canons and Faithful Agency*, 90 B.U.L. REV. 109, 164 (2010) ("The bedrock principle of textualism . . . is its insistence that federal courts cannot contradict the plain language of a statute."); *see also* John F. Manning, *Second-Generation Textualism*, 98 CALIF. L. REV. 1287, 1309–10 (2010) (describing the "defining feature of 'second-generation textualism'" to be the "proposition that courts must respect the terms of an enacted text when its semantic meaning is clear, even if it seems contrary to the statute's apparent overall purpose").

38. Intentionalists and purposivists often see the text as a primary indicator of intent or purpose. *See, e.g.*, 12W RPO, Ltd. Liab. Co. v. Affiliated FM Ins. Co., 353 F. Supp. 3d 1039, 1052 (D. Or. 2018) ("Without a policy definition, I determine the plain meaning of the term as an 'aid[] of interpretation to discern the parties' intended meaning.") (quoting Groshong v. Mut. Of Enumclaw Ins. Co., 985 P.2d 1284, 1287 (Or. 1999)); Estrada v. McDowell, No. 16-CV-02827, 2017 WL 5068641, at *22 (N.D. Cal. Nov. 3, 2017) ("The word 'act' has a plain meaning. If the Legislature had intended to restrict the word 'act' in section 1101, subdivision (b) to 'bad' acts, it would have done so."); BedRoc Ltd., LLC v. United States, 541 U.S. 176, 183 (2004) ("The preeminent canon of statutory interpretation requires us to 'presume that [the] legislature says in a statute what it means and means in a statute what it says there.'") (quoting Conn. Nat'l Bank v. Germain, 503 U.S. 249, 253–54 (1992)).

39. *See, e.g.*, Alan Schwartz & Robert E. Scott, *Contract Interpretation Redux*, 119 YALE L.J. 926, 932 (2009) ("A plain meaning default that presumes the parties have written in the standard language reduces [the risk of strategic behavior] by requiring parties to specify the terms that take technical meanings."); David A. Strauss, *Why Plain Meaning?*, 72 NOTRE DAME L. REV. 1565, 1565 (1997) ("Plain meaning is the place to begin in interpreting statutes, not because the meaning best reflects the legislature's intentions, or for any of the other reasons usually given, but just because the ordinary meaning is an obvious point of agreement in circumstances in which disagreement is too costly.").

40. *See, e.g.*, Krishnakumar, *supra* note 1, at 77–78; Austin Peters, *Are They All Textualists Now?*, 118 NW. U.L. REV. 1201, 1209–11 (2023) (analyzing the prevalence of textualism in state courts).

published opinions. These opinions were scraped from the Harvard Caselaw Access Project and included (a) every available federal opinion; and (b) a comparably sized sample of state law opinions. Descriptively, the prevalence has gone up dramatically as a proportion of all written opinions at all three levels of the federal judiciary over the last several decades, and to varying degrees in the state judiciary as well, as shown in Figure 1.

Full details of the method underlying this analysis are documented in Appendix 1.E.

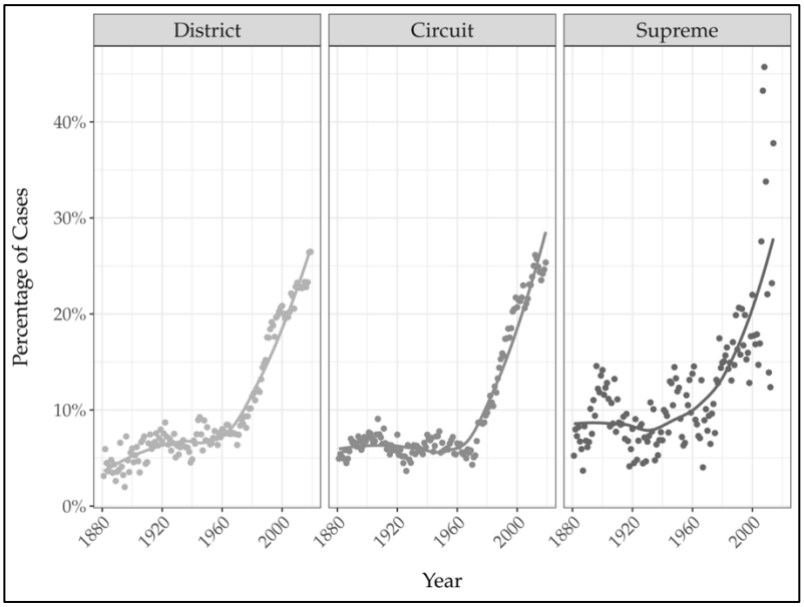

Figure 1: Proportion of published opinions within the federal judiciary referencing "plain meaning" or relevant synonym over time. Dots represent proportion for a given year. Trend lines represent smoothed LOESS regression lines.

2. Polemics of Plain Meaning

Despite the ubiquity and importance of plain meaning, there are several open debates regarding its usage, including (a) an internal debate regarding who is the relevant readership in judging the meaning; and (b) an external debate regarding how clear the legal text actually is in real-world cases in which the plain meaning rule is invoked. This Subsection details each of these debates in turn, as well as the relevance of empirical scholarship in investigating them.

a. Whose Meaning?

There exist three main sets of views of the relevant readership in judging plain meaning as exercised by courts and argued by jurists: the ordinary view, the expert view, and the hybrid view.

According to the ordinary view, the relevant readership is that of an ordinary person. As explained by Oliver Wendell Holmes in *The Theory of Legal Interpretation*, when interpreting law: "[W]e ask . . . what those words would mean in the mouth of a normal speaker of English . . . ."[41] This view is particularly prominent among modern textualists such as Justice Amy Coney Barrett, who "view themselves as agents of the people rather than of Congress" and thus "approach language from the perspective of an ordinary English speaker" in order to construe statutory language consistent with its "ordinary meaning."[42]

Similarly, when interpreting the constitution, the Supreme Court has written that "we are guided by the principle that '[t]he Constitution was written to be understood by the voters; its words and phrases were used in their normal and ordinary as distinguished from technical meaning.'"[43] And in the interpretation of contracts, courts have held that a term must be given "its plain meaning as understood by the '[person]-on-the-street,' that is, the term's plain and ordinary meaning."[44]

According to the expert view, the relevant readership is that of a reasonable or well-informed reader. Under this view, a judge tasked with interpreting words in a legal text attempts to "hear the words . . . as they would sound in the mind of a skilled, objectively reasonable user of words."[45] This view has been articulated by several prominent textualists, such as Antonin Scalia, who in his seminal work with Bryan Garner sought a return to the "oldest and most commonsensical interpretive principle" as "words mean what they conveyed to reasonable people at the time they were written,"[46] and that "[t]hrough accurate knowledge of language and proper education in legal method,

---

41. Oliver Wendell Holmes, *The Theory of Legal Interpretation*, 12 HARV. L. REV. 417, 417–18 (1899).

42. Amy Coney Barrett, *Congressional Insiders and Outsiders*, 84 U. CHI. L. REV. 2193, 2194–95, 2209 (2017).

43. District of Columbia v. Heller, 554 U.S. 570, 576 (2008) (quoting United States v. Sprague, 282 U.S. 716, 731 (1931)).

44. State Farm Mut. Auto. Ins. Co. v. Spangler, 64 F.4th 1173, 1179 (11th Cir. 2023) (quoting State Farm Fire & Cas. Co. v. Castillo, 829 So. 2d 242, 244 (Fla. Dist. Ct. App. 2002)).

45. John F. Manning, *What Divides Textualists from Purposivists?*, 106 COLUM. L. REV. 70, 79 n.32 (2006); *see also id.* at 100–01 ("[T]extualists quite reasonably believe that a federal court fulfills its obligation as Congress's faithful agent by trying to 'hear the words [of the statute] as they would sound in the mind of a skilled, objectively reasonable user of words.'") (second alteration in original) (quoting Frank H. Easterbrook, *The Role of Original Intent in Statutory Construction*, 11 HARV. J.L. & PUB. POL'Y 59, 65 (1988)).

46. Scalia & Garner, *supra* note 18, at 36.

lawyers ought to have a shared sense of what meanings words can bear and what linguistic arguments can credibly be made about them."[47]

Even judges who do not explicitly endorse the expert view may implicitly adopt it when, for example, determining plain or ordinary meaning based on definitions of words in Black's Law Dictionary[48] or legal precedent,[49] or when interpreting words based on prescriptive criteria associated with high-status speakers, such as "the rules of grammar"[50] or "proper usage."[51]

Finally, under the hybrid view, the relevant readership depends on the context or genre of the provision. In many instances, this hybrid view manifests itself as a presumption that words will be taken in their ordinary sense, unless it is clear that a word is a term of art.[52] For example, in the case of contract drafting, courts often adopt the ordinary view for most terms in an agreement but make an exception for words such as "person," which to an ordinary reader would include only natural persons but nonetheless are interpreted by courts (even in the absence of an explicit definition) as including (as they would to a lawyer) corporations.[53] Similarly, judges who explicitly endorse the ordinary view may necessarily incorporate some elements of the expert view when interpreting the language of complex legal provisions given the difficulty that non-lawyers face when interpreting such language.[54]

---

47. *Id.* at 47; *see also* Jennifer Nou, *Regulatory Textualism*, 65 DUKE L.J. 81, 85 (2015) (proposing a theory of regulatory textualism that asks "how the reasonable reader of [a rule or regulation] would have understood its meaning as negotiated by the President, Congress, and other authoritative regulatory actors").

48. *See, e.g.*, United States v. Melvin, 948 F.3d 848, 852 (7th Cir. 2020) (using Black's Law Dictionary to find the plain meaning of "disclose"); *see also* United States v. Riscajche-Siquina, 30 F. Supp. 3d 580, 583 (S.D. Tex. 2014) (using "legal and other well-accepted dictionaries" to find the "plain meaning" of the words at issue).

49. *See, e.g.*, Young v. Nevada Gaming Control Bd., 473 P.3d 1034, 1035, 1037 (Nev. 2020) (determining the ordinary meaning of the term "patron" based on its definition in Black's Law Dictionary and its interpretation in prior cases); State v. Marsh, 196 N.W. 930, 931 (Minn. 1924) ("[L]exicographers and judges agree that the word 'prostitution,' as ordinarily used in its application to lewd women, does not refer to sexual intercourse with but one man.").

50. Anhydrides & Chemicals, Inc. v. United States, 130 F.3d 1481, 1483 (Fed. Cir. 1997) ("The rules of grammar apply in statutory construction . . . .").

51. Scalia & Garner, *supra* note 18, at 109.

52. *See, e.g.*, Shell Oil Co. v. Winterthur Swiss Ins. Co., 12 Cal. App. 4th 715, 746 (1993) ("Judicial interpretation is controlled by the clear and explicit meanings of words in their ordinary and popular senses, unless the parties adopted a special or technical usage.").

53. Scalia & Garner, *supra* note 18, at 77.

54. *See, e.g.*, Eric Martínez, Francis Mollica & Edward Gibson, *Poor Writing, Not Specialized Concepts, Drives Processing Difficulty in Legal Language*, 224 COGNITION 1, 1–2 (2022) (finding that contracts are laden with features such as center-embedded clauses, passive voice structures, non-standard capitalization and low-frequency words at a higher rate than other genres of English); Eric Martínez, Francis Mollica & Edward Gibson, *So Much for Plain Language: An Analysis of the Accessibility of United States Federal Laws Over Time*, 224 J. EXP. PSYCHOL. GEN. 1153, 1154 (2024) (finding similar rates of these structures in statutes passed by Congress); Eric Martínez, Francis Mollica & Edward Gibson, *Even*

It is worth noting that despite the plurality of views regarding the relevant readership of a legal text in cases of interpretation, courts are often unclear about which of the two views they are adopting in a given case.

Despite this lack of clarity, most empirical scholarship to date has focused primarily on the ordinary view. The vast majority of experimental jurisprudence studies, investigating questions of interpretation have examined how ordinary people (non-lawyers) judge the meaning of words in a fictional legal text.[55] It thus remains an open question what the practical distinctions are between the different versions of plain meaning, as well as which view offers a better account of judicial behavior.

b. How Plain?

Regardless of which view is adopted, an additional question concerns to what extent the meaning of the text is in fact "plain" in the cases in which courts assert it to be so.

For example, many have speculated that textualists often exaggerate the determinacy of statutes and other legal texts when invoking the plain meaning doctrine in order to achieve ideological ends.[56] Even Justice Scalia acknowledged that "[w]illful judges might use textualism to achieve the ends they desire, and when the various indications of textual meaning point in different directions, even dutiful judges may unconsciously give undue weight to the factors that lead to what they consider the best result."[57]

Despite this speculation, however, the level of clarity in plain-meaning cases, under either the ordinary or expert view, remains

---

Lawyers Do Not Like Legalese, 2023 PROC. NAT'L ACAD. SCI. U.S.A. 1, 1–2 (2023) (finding that lawyers, like laypeople, struggled to understand legal content written in a complex register relative to a simplified register, and that lawyers preferred contracts written in a simplified register).

55. *See* Kevin Tobia, *Experimental Jurisprudence*, 89 U. CHI. L. REV. 737, 766 (2022) (discussing and defending the practice that "there are many experimental-jurisprudence studies of only laypeople, but it is much less common to find studies of only experts"). *See generally* James Macleod, *Finding Original Public Meaning*, 56 GA. L. REV. 1 (2022) (as an additional example of important work intentionally studying laypeople). For examples of this practice in other important experimental jurisprudence work in other areas, see James A. Macleod, *Ordinary Causation: A Study in Experimental Statutory Interpretation*, 94 IND. L.J. 957 (2019). For exceptions to this practice, see, for example, Kevin Tobia, *Legal Concepts and Legal Expertise*, 203 SYNTHESE 1, 3–4 (2024) (comparing how lawyers and non-lawyers interpret the meaning of "intentional"); Ivar R. Hannikainen, Kevin P. Tobia, Guilherme da F. C. F. de Almeida, Noel Struchiner, Markus Kneer, Piotr Bystranowski et al., *Coordination and Expertise Foster Legal Textualism*, 2022 PROC. NAT'L ACAD. SCI. U.S.A. 1, 6 (2022) (comparing legal interpretations of those with and without legal training cross-culturally).

56. *See, e.g.*, Doerfler, *supra* note 17, at 305–12 (referring to two Supreme Court Cases in which textualists read narrow meaning into indeterminate linguistic phrases) (citing Niz-Chavez v. Garland, 593 U.S. 155 (2021); Van Buren v. United States, 593 U.S. 374 (2021)).

57. Scalia & Garner, *supra* note 18, at 37.

unclear. Recall that plain meaning is defined as clarity according to some subset of human readers (ordinary or well-informed). Although recent work has proposed measuring clarity using computational methods,[58] the dearth of available human benchmarking data from real-world, plain-meaning cases prevents one from evaluating the external validity of these methods (and, by extension, from evaluating the level of clarity in plain-meaning cases).

B. Traditional Tools

What tools do courts invoke when purporting to uncover plain meaning? Courts have traditionally appealed to three broad categories of interpretive tools: linguistic canons, dictionaries, and judicial authority. An analysis of over two million published opinions confirms that just as references to plain meaning have risen over the decades, so too have references to these tools.[59]

This Section first provides an overview of these tools. The Section next provides an overview of two longstanding questions regarding their usage that are empirically tested in this Article: (1) Whether judges invoke canons as mere rhetorical camouflage for outcome-based policy preferences; and (2) whether legal language is indeterminate, such that the interpretation dictated by one canon is no more or less plausible than that dictated by a counter-canon. Finally, the Section distinguishes these two questions from a third question that is left open by the study: Whether judges turn to canons and dictionaries ex ante — as an aid to forming their own sense of ordinary meaning — or ex post, as a way to articulate and defend that sense.

Categories of Tools

What tools do courts invoke when purporting to uncover plain meaning? Courts have traditionally appealed to three broad categories of interpretive tools, including linguistic canons, dictionaries, and judicial authority. This Subsection provides an overview of these tools.

---

58. *See generally* Choi, *supra* note 26.

59. In doing so, this Section replicates and extends previous work examining similar trends at the federal level over a shorter timescale. *See generally* Aaron-Andrew P. Bruhl, *Statutory Interpretation and the Rest of the Iceberg: Divergences Between the Lower Federal Courts and the Supreme Court*, 68 DUKE L.J. 1 (2018); Aaron-Andrew P. Bruhl, *Communicating the Canons: How Lower Courts React When the Supreme Court Changes the Rules of Statutory Interpretation*, 100 MINN. L. REV. 481 (2015). *See infra* Appendix 1.E for details.

a. Linguistic Canons

Linguistic canons, sometimes referred to as "semantic" canons, are generalizations about how particular linguistic constructions are purportedly used and understood by readers of English in a given context.[60] Linguistic canons are often contrasted with substantive canons, which are non-linguistic considerations that weigh in favor of particular legal results.[61]

This Section provides an overview of the different categories of linguistic canons as used in the materials for the two studies.

1. *Expressio Unius* and Implied Exemptions

One type of canon concerns whether the omission of some things in a legal text implies their exclusion. Perhaps the most famous of these canons is *expressio unius*, according to which expression of one thing excludes another.

Consider the Virginia Supreme Court case of *Tate v. Ogg*,[62] in which a law expressly forbade letting any "'horse, mule, cattle, hog, sheep, or goat' to run at large upon lots or lands inclosed . . . ."[63] The court ruled that the plain language of the statute excluded the applicability of this rule to dogs or fowls, since they were not among the animals that were expressly prohibited.[64]

On the other hand, courts have also invoked a counter-canon to this rule which states that some things in a text may be listed only by way of example, and that in such cases an item that is not expressly included may nonetheless be implied. This is especially the case when a provision includes the word "include," which courts often interpret as implying a non-exhaustive list. Consider the example of *United States v. Hawley*,[65] which involved the interpretation of a sentencing guidelines provision that stated that prior sentences are to be computed in the criminal history score, "including uncounseled misdemeanor sentences where imprisonment was not imposed."[66] The court rejected the defendant's argument, that under *expressio unius*, the guidelines

60. *See, e.g.*, Benjamin Eidelson & Matthew C. Stephenson, *The Incompatibility of Substantive Canons and Textualism*, 137 HARV. L. REV. 515, 516 (2023). Note also that some divide linguistic canons further into further categories such as "semantic" and "syntactic" canons. For a discussion of this distinction, see Scalia & Garner, *supra* note 18, at 69.

61. Eidelson & Stephenson, *supra* note 60, at 516. *But see* Brian G. Slocum & Kevin Tobia, *The Linguistic and Substantive Canons*, 137 HARV. L. REV. F. 70, 70 (2023) ("We question the traditional dichotomy between linguistic and substantive canons.").

62. 195 S.E. 496, 499 (Va. 1938).

63. *Id.* at 499.

64. *Id.*

65. 919 F.3d 252 (4th Cir. 2019).

66. *Id.* at 255 (emphasis omitted).

excluded sentences beyond those listed, concluding that the language of the guidelines "[does] not support such an inference in this case."[67]

2. *Noscitur a Sociis* and the Interpretation of General Terms

In addition to the implication of words that are not expressly listed in the law, other canons relate to the interpretation of words that *are* expressly listed in the law. One such class of canons concerns whether the interpretation of general terms such as "object" and "case" may be limited by context. On the one hand, one commonly invoked canon states that general terms must be given a general interpretation. In the case of *Smith v. United States*,[68] the Supreme Court was tasked with interpreting a statute that imposed additional penalties on the "use" of a firearm during and in relation to a drug trafficking crime.[69] Rejecting the defendant's argument that the term "use" should have been interpreted narrowly to mean "discharge" given the surrounding context, the Supreme Court determined that "use" was to be interpreted in its general, "everyday meaning," and thus included the defendant's handling of the firearm as barter for drugs.[70]

On the other hand, courts often invoke a separate canon known as *noscitur a sociis* ("a word is known by its associates") in cases where they deem the language to support a narrower interpretation. Consider, for example, the case of *Nehme v. Smithkline Beecham Clinical Labs*,[71] in which the Florida Supreme Court was tasked with interpreting a statute that extended the statute of repose for medical malpractice claims if "fraud, concealment or intentional misrepresentation of fact prevented discovery of the injury."[72] The question was whether "concealment" included negligent misdiagnosis of the injury.[73] Applying the doctrine of *noscitur a sociis*, the court concluded that the plain meaning of the term "concealment," given the surrounding words, implied intentional acts, and therefore did not encompass negligent diagnosis.[74]

3. *Ejusdem Generis* and Enumerated Lists

A third and related class of canons specifically concerns the interpretation of general terms at the beginning or end of an enumerated list. The most famous of these canons is *ejusdem generis*, which states that "where general words follow the enumeration of particular classes

67. *Id.* at 256.
68. 508 U.S. 223 (1993).
69. *Id.* at 225.
70. *Id.* at 228.
71. 863 So. 2d 201 (Fla. 2003).
72. *Id.* at 203.
73. *Id.* at 203–04.
74. *See id.* at 209.

of persons or things, the general words will be construed as applicable only to persons or things of the same general nature or class as those enumerated."[75] Consider the Pennsylvania Commonwealth Court case of *S.A. by H.O. v. Pittsburgh Public School District*,[76] where the court was tasked with interpreting whether a sharpened pencil counted as a weapon under a school policy.[77]

The policy defined a weapon as any "knife, cutting instrument, cutting tool, explosive, mace, nunchaku, firearm, shotgun, rifle" or any other "tool, instrument or implement capable of inflicting serious bodily injury."[78] The court determined that the meaning of the general terms "tool," "instrument," and "implement" were plainly limited by the specific terms, "all of which contain a metal blade, discharge projectiles, or are otherwise traditional weapons that serve no innocuous purpose when brought onto school grounds."[79] Consequently, the court concluded that the plain meaning of "weapon" under the statute did not include a pencil.[80]

On the other hand, courts have also developed a counter-canon to *ejusdem generis*, which states, for example, that "where the particular words exhaust the class, the general words must be construed as embracing something outside of that class."[81] For example, consider the case of *Corby Baking Co. v. Commonwealth*,[82] in which the Supreme Court of Appeals of Virginia was tasked with interpreting a law that exempted from peddler licenses those who sold "ice, wood, meats, milk, butter, eggs, poultry, fish, oysters, game, vegetables, fruits, or other family supplies of a perishable nature . . . ."[83] The court concluded that the statute expressed, "in language too plain to be misunderstood," that the exception applied to all family supplies of a perishable nature, including, for example, bread, as opposed to only certain supplies that were enumerated in a list.[84]

4. Rules of Grammar and the Interpretation of "and" and "or"

Whereas the above two classes of canons relate to the meaning of content such as nouns and verbs, a fourth class of canons concerns the interpretation of grammatical constructions involving "and" and "or."

The default canon in this class, described by Scalia & Garner in *Reading Law*, is the conjunctive/disjunctive canon, which states that

75. State v. Russell, 187 So. 540, 543 (Miss. 1939).
76. 160 A.3d 940 (Pa. Commw. Ct. 2017).
77. *Id.* at 944.
78. *Id.* at 942.
79. *Id.* at 947.
80. *See id.*
81. United States v. Mescall, 215 U.S. 26, 31 (1909).
82. 96 S.E. 133 (Va. 1918).
83. *Id.* at 133.
84. *Id.*

the plain meaning of "and" is generally conjunctive, whereas "or" is generally understood disjunctively.[85] For example, in the case of *Petrohawk Props v. Heigle*,[86] the Arkansas Court of Appeals was tasked with interpreting a lease agreement whose duration was for five years "and as long thereafter" as production continued "and as long thereafter" as operations continued.[87] Invoking the conjunctive canon, the court concluded that the language of the agreement "plainly means that both the production and operation conditions" had to be satisfied as opposed to just one in order for the lease to be extended.[88]

In contrast, courts have also invoked an exception to this canon, which states that sometimes "and" is to be interpreted disjunctively, and that "or" is sometimes to be read conjunctively. For example, in *Ex parte Jordan*,[89] the Alabama Supreme Court concluded that the meaning of "or" in a law referring to "a mother or father" was the conjunctive sense and to hold otherwise would "defy common sense."[90] Additionally, in *California Lumbermen's Council v. Federal Trade Commission*,[91] the Ninth Circuit concluded that there was "no question" that the meaning of the word "and" in a court order was disjunctive.[92]

5. Last Antecedent and Series Qualifier

Another class of syntactic canons relates to the referents of modifying words that occur at the end of a series of terms. Under the "last-antecedent" rule, when some limiting phrase follows an item in a list, it usually modifies only that immediately preceding item as opposed to all of those in the list. The last-antecedent rule is based on a cognitive hypothesis (proposed by lawyers) about how people interpret language in relevant contexts — namely, that "it is easier to apply that modifier only to the item directly before it," and is said to apply especially in cases where "it takes more than a little mental energy to process the individual entries in the list, making it a heavy lift to carry the modifier across them all."[93]

In other cases, courts invoke the series-qualifier canon, which states that "[w]hen there is a straightforward, parallel construction that involves all nouns or verbs in a series, a prepositive or postpositive modifier normally applies to the entire series."[94] For example, in the

85. *See* Scalia & Garner, *supra* note 18, at 116–25.
86. 386 S.W.3d 657 (Ark. Ct. App. 2011).
87. *Id.* at 659.
88. *See id.* at 660.
89. 592 So. 2d 579 (Ala. 1992).
90. *Id.* at 581.
91. 115 F.2d 178 (9th Cir. 1940).
92. *Id.* at 185.
93. Lockhart v. United States, 577 U.S. 347, 351 (2016).
94. Scalia & Garner, *supra* note 18, at 130.

case of *Facebook, Inc. v. Duguid*,[95] the Supreme Court concluded that the "most natural reading" of a provision referring to equipment with the capacity to "store or produce" telephone numbers "using a random or sequential number generator" was that the modifying phrase "using a random or sequential number generator" referred to both "store" and "produce," not just "produce."[96]

6. Words of Permission and Obligation

One specific class of linguistic canon pertains specifically to the words "may" and "shall." The default canon is that the meaning of the word "may" indicates permission, whereas the word "shall" indicates an obligation. For example, in the case of *Lopez v. Davis*,[97] the Supreme Court invoked this canon to conclude that a provision stating that the sentence of a non-violent offender after completing a substance abuse program "may" be reduced by the Bureau of Prisons plainly meant that the Bureau was permitted, but not required, to reduce the offender's sentence.[98]

In contrast to the default rule, courts occasionally depart from this presumption by interpreting either "shall" as permissive or "may" as obligatory. Although sometimes the justification for this is policy-based, courts also draw this conclusion based on plain-meaning. For example, in the case of *Julian Depot Miami, LLC v. Home Depot U.S.A., Inc.*[99] the Eleventh Circuit concluded that "'shall' plainly meant 'may'" in a contract provision stating that a retail tenant "shall, at its election and at its expense," rebuild the store in the case of casualty loss.[100]

7. Punctuation

The final class of canons relates not to individual words or grammar but to punctuation of the text. Historically, the dominant canon in this class was that "[p]unctuation marks are no part of an act," and courts would either disregard punctuation or even "repunctuate, if that be necessary, in order to arrive at the natural meaning of the words employed."[101]

As shown in Figure A.3, however, over the past several decades references to punctuation within judicial decisions have gone up, with

95. 592 U.S. 395 (2021).
96. *Id.* at 402–03.
97. 531 U.S. 230 (2001).
98. *Id.* at 241.
99. 824 F. App'x 609 (11th Cir. 2020).
100. *Id.* at 612–13.
101. *See, e.g.*, United States v. Shreveport Grain & Elevator Co., 287 U.S. 77, 82–83 (1932).

courts frequently taking the position that "the meaning of a statute will typically heed the commands of its punctuation."[102]

Some judges have more aggressively defended the use of punctuation, stating that ignoring punctuation "would make English teachers cringe," and that "stuffing punctuation to the bottom of the interpretive toolbox would run the risk of distorting the meaning of statutory language."[103] In particular, courts pay special attention to the use and placement of commas. For example, in the case of In re *Moschell*,[104] the court interpreted a provision of the bankruptcy code excepting from discharge any debt that is "for fraud or defalcation while acting in a fiduciary capacity, embezzlement or larceny." Based on the placement of the comma, the court concluded that the "plain language of the statute" provided that "general claims for embezzlement or larceny are excepted from discharge in bankruptcy without regard to whether or not the debtor was (or is) a fiduciary."[105]

b. Dictionaries

The second category of tools judges invoke when purporting to uncover plain meaning is dictionaries. As explained by a leading treatise on statutory interpretation: "No matter which version of the [plain-meaning] rule a court uses . . . all courts accept that standard, recognized, contemporary dictionaries are a valuable source to understand a word's approved, common meaning."[106]

In recent years, judges have appealed to dictionaries at increasing rates, invoking them in approximately 10% of federal district court cases; 12% of federal circuit court cases; and 20% of United States Supreme Court cases — up from under 5% at each of these levels in the 1960s. Similar (though varying) increases have been observed in state courts. [107] Judges reference legal dictionaries (such as Black's, Ballentine's, and Bouvier's) and ordinary non-legal dictionaries (such as Webster's, Oxford, and American Heritage) at remarkably similar rates.[108]

---

102. *See, e.g.*, U.S. Nat'l Bank of Or. v. Indep. Ins. Agents of Am., Inc., 508 U.S. 439, 454 (1993).

103. *See, e.g.*, NACS v. Bd. of Governors of the Fed. Rsrv. Sys., 746 F.3d 474, 486 (D.C. Cir. 2014), *cert. denied*, 574 U.S. 1121 (2015).

104. No. 19-21819, 2020 WL 5998166 (Bankr. W.D. Pa. Oct. 9, 2020).

105. *Id.* at *9.

106. Singer & Singer, *supra* note 3, at § 47:28.

107. *See infra* Appendix 1.E.

108. *Id.*

c. Judicial Authority

The third category of tools includes judicial authority. The primary form of judicial authority consists of common-law precedent — in particular, a different court's interpretation of the meaning of a word in a purportedly analogous context. As explained by a leading treatise on statutory interpretation: "In addition to dictionary definitions, courts also agree that a word's common law definition may be synonymous with its common meaning."[109]

For example, as detailed by Tara Grove: "Supreme Court precedent can help an interpreter determine the meaning of statutory terms and phrases." [110] In the Supreme Court case of *Bostock v. Clayton County*,[111] for instance, the Court relied on precedent to conclude that, "[i]n the language of law," "because of" signals but-for causation.[112] Similarly, in the case of *Young v. Nevada Gaming Control*,[113] the Nevada Supreme Court relied on both dictionary definitions and prior cases to explain the plain meaning of "patron" in the context of a regulation governing the use of chips and tokens.[114]

The second form of judicial authority consists of the court's own judgment, without purporting to rely on other tools. For example, in the case of *Lutz Appellate Services, Inc. v. Curry*,[115] the district court determined the "ordinary meaning" of the term "unsolicited advertising" based on its own reading of the text, without appealing to a canon, dictionary, or prior case interpreting the same.[116] Likewise, in the case of *Solon v. Gray*,[117] the Eighth District Court of Appeals of Ohio used a similar approach to define the "common, ordinary meaning" of the word "structure."[118]

2. Critiques

To what extent do judges invoking these tools successfully arrive at a text's plain meaning? Despite longstanding speculation on this topic, the question essentially remains empirically untested.

---

109. Singer & Singer, *supra* note 3, at § 47:28.
110. Tara L. Grove, *Is Textualism at War with Statutory Precedent?*, 102 TEX. L. REV. 639, 642 (2024).
111. 590 U.S. 644 (2020).
112. *Id.* at 656.
113. 473 P.3d 1034 (Nev. 2020).
114. *Id.* at 1037.
115. 859 F. Supp. 180 (E.D. Pa. 1994).
116. *Id*. at 181–82.
117. 660 N.E.2d 509 (Ohio Ct. App. 1995).
118. *Id.* at 510.

Prior work has asked whether congressional staffers generally think of certain interpretive tools when drafting statutes[119] and whether non-lawyers generally interpret rule-like language in line with linguistic canons.[120] However, there exists no systematic evaluation of whether, in the real-world cases that go to litigation, the court's interpretation, as assisted by interpretive tools, tracks the meaning ordinary or legally informed readers actually assign.

The debate centers on two distinct issues. The first issue concerns whether real-world disputes confronting judges are linguistically determinate. If texts at litigation's edge admit no single best reading, then canons cannot be invoked to support such a reading. If one or more canons converge on the sense that ordinary or well-informed readers would pick, the canons retain explanatory power. The second issue concerns judicial motive and behavior: Even when one linguistic canon does reflect ordinary meaning, do judges merely invoke it opportunistically to mask policy preferences?

Both debates are associated with Karl Llewellyn's seminal work *Remarks on the Theory of Appellate Decision and the Rules or Canons About How Statutes Are to Be Construed*, which has been described as "one of the most celebrated law review articles of all time."[121] Llewellyn argued that in virtually any appellate case in which a judge turns to canons, there exist two equally applicable canons ("thrust" and "parry") available to a judge that, if invoked, would lead to opposite results.[122]

Indeterminacy proponents see the chart and other sources of evidence as support for the view that language in litigated disputes often "runs out,"[123] given that "communicative or assertive content . . . is often sparse, minimal, or indeterminate."[124] Proponents of this view "roundly bemoan the determinacy of communicative content,"[125]

---

119. *See* Abbe R. Gluck & Lisa S. Bressman, *Statutory Interpretation from the Inside — An Empirical Study of Congressional Drafting, Delegation, and the Canons: Part I*, 65 STAN. L. REV. 901, 905–06 (2013).

120. *See* Tobia, Slocum & Nourse, *supra* note 19, at 246, 296–97; Janet Randall & Lawrence Solan, *Legal Ambiguities: What Can Psycholinguistics Tell Us?*, *in* THE CAMBRIDGE HANDBOOK OF EXPERIMENTAL JURISPRUDENCE (Kevin Tobia ed., 2025).

121. William N. Eskridge, Jr. & Philip P. Frickey, *Quasi-Constitutional Law: Clear Statement Rules as Constitutional Lawmaking*, 45 VAND. L. REV. 593, 595 (1992).

122. Llewellyn, *supra* note 16, at 401.

123. *See* Doerfler, *supra* note 17, at 293; *see also* William Baude & Stephen E. Sachs, *The Law of Interpretation*, 130 HARV. L. REV. 1079, 1129 (2017) (acknowledging that the communicative content of statutory language often runs out, even if background legal principles determine systematically how judges ought to rule in such cases).

124. Richard H. Fallon, Jr., *The Meaning of Legal "Meaning" and Its Implications for Theories of Legal Interpretation*, 82 U. CHI. L. REV. 1235, 1269 (2015).

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

which they view as a mere "fiction,"[126] leaving judges who deploy canons with "no option but to exaggerate the determinacy of linguistic meaning."[127]

Smokescreen critics focus instead (or in addition) on judicial *use*, taking Llewellyn's thrust and parry as evidence that canons are "so malleable…as to operate mostly as pretext";[128] that they are little more than "window dressing" [129] or disingenuous "figleaves" [130] that "masquerade as rules" [131] to conceal policy preferences. Parallel complaints cast dictionaries as tools that might "enable political values or bias to influence interpretation"[132] or supply only "mildly elaborate obfuscation,"[133] and view textualism itself as "a rhetorical smokescreen for extremely conservative results."[134]

Canon defenders counter on both fronts. On the determinacy front, for example, Scalia & Garner lament the "prevailing confusion" that leads many lawyers to treat language as endlessly elastic, [135] while Thomas Lee and Stephen Mouritsen similarly argue against the view of scholars who question judges' ability to ascertain linguistic meaning.[136] Courts go a step further, arguing, as Chief Justice Roberts did in *Loper Bright v. Raimondo*, that "statutes, no matter how impenetrable, do — in fact, must — have a single, best meaning," and that courts must "use every tool at their disposal to determine the best reading of the statute and resolve ambiguity."[137]

Similarly, on the smokescreen front, canon defenders reject the "hyper-realist" caricature — that courts invariably reach their preferred result and then shop for a canon — arguing instead that "at the margin,

126. Erwin Chemerinsky, *The Myth of Plain Meaning*, ABA J. (Oct. 31, 2017, at 08:00 ET), https://www.abajournal.com/news/article/chemerinsky_plain_meaning_is_a_myth [https://perma.cc/EVN9-BENG].

127. Doerfler, *supra* note 17, at 305.

128. *Id.* at 267.

129. ESKRIDGE, *supra* note 20, at 275.

130. Continental Cas. Co. v. Pittsburgh Corning Corp., 917 F.2d 297, 300 (7th Cir. 1990).

131. DICKERSON, *supra* note 20, at 234.

132. Kevin P. Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726, 750 (2020).

133. Doerfler, *supra* note 17, at 305.

134. Neil H. Buchanan & Michael C. Dorf, *A Tale of Two Formalisms: How Law and Economics Mirrors Originalism and Textualism*, 106 CORNELL L. REV. 591, 640 (2021); *see also* William N. Eskridge, Jr. & Philip P. Frickey, *The Supreme Court, 1993 Term — Foreword: Law as Equilibrium*, 108 HARV. L. REV. 26, 77 (1994) (arguing that the "new, tougher" text-based approaches to statutory interpretation ultimately serve "as a cover for the injection of conservative values into statutes"); Richard A. Posner, *The Incoherence of Antonin Scalia*, THE NEW REPUBLIC (Aug. 24, 2012), https://newrepublic.com/article/106441/scalia-garner-reading-the-law-textual-originalism [https://perma.cc/EW7N-CL3R] ("[T]ext as such may be politically neutral, but textualism is conservative.").

135. Scalia & Garner, *supra* note 18, at 34.

136. Lee & Mouritsen, *supra* note 125, at 793.

137. 603 U.S. 369, 400 (2024).

an interpretive rule could" make the difference in "countless cases."[138] Antonin Scalia and Bryan Garner open *Reading Law* in the same spirit, insisting that linguistic canon usage is "not a device calculated to produce socially or politically conservative outcomes" but a neutral method grounded in the ordinary meaning of enacted text.[139] Michael Sinclair similarly argued that careful examination of Llewellyn's cited examples "undermines the claim that canons may be chosen to suit one's ends, whatever they may be."[140]

Regardless of where the "conventional wisdom"[141] ultimately settles, the sheer coexistence of sharply conflicting theories, ranging from canons-as-fig-leaves to canons-as-partial-constraints, coupled with the aforementioned evidentiary gap, underscores the need for systematic empirical evidence about how judges actually employ linguistic canons in practice.[142]

Part III therefore develops and implements a methodology to address these unresolved controversies simultaneously. To do so, it first translates the competing accounts of each inquiry into explicit mathematical models from which one can derive testable empirical predictions. It then tests those predictions in an experimental paradigm to uncover whether: (1) litigated texts are in fact determinate for ordinary or legally trained audiences, and (2) judges' stated canons track those audience judgments or diverge in ways suggestive of strategic deployment.

3. Causality?

It is important to keep separate the two issues discussed above — whether judges invoke canons to disguise policy-driven choices and

138. Nicholas Quinn Rosenkranz, *Federal Rules of Statutory Interpretation*, 115 HARV. L. REV. 2085, 2154 (2002).

139. Scalia & Garner, *supra* note 18, at 16.

140. Michael Sinclair, *Only A Sith Thinks Like That: Llewellyn's Dueling Canons, One to Seven*, 50 N.Y.L. SCH. L. REV. 919, 921 (2005). It is also worth noting that not all critiques are equivalent in their scope or severity. William Eskridge, for instance, writes that "[f]or any *difficult* case, there will be as many as twelve to fifteen relevant 'valid canons' cutting in different directions, leaving considerable room for judicial cherry-picking." William N. Eskridge, Jr., *The New Textualism and Normative Canons*, 113 Colum. L. Rev. 531 (2013) (emphasis added) (reviewing ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS (2012)). Cass Sunstein similarly describes canons as helpful but highly defeasible heuristics whose weight depends on context; the risk of strategic invocation, he notes, is often present but not omnipresent. *See generally* Cass R. Sunstein, *Interpreting Statutes in the Regulatory State*, 103 HARV. L. REV. 405 (1989).

141. Alexander Volokh, *Choosing Interpretive Methods: A Positive Theory of Judges and Everyone Else*, 83 N.Y.U. L. REV. 769, 771 (2008).

142. Note also that the majority of these discussions have centered around judges' use of dictionaries and linguistic canons in the context of *statutory* interpretation. *See supra* notes 120–142. However, as noted in the previous Subsections, the use of these tools by judges to arrive at plain meaning pervades the interpretation of legal text of all genres. With respect to those genres, these issues remain empirically untested.

whether legal language rarely yields a single, knowable meaning — from a third possibility that may otherwise be conflated with them. Under this third view, the traditional linguistic canons do in fact track how ordinary readers resolve meaning in context, and judges rely on them in good faith. What those canons do not do, however, is cause the judge's conclusion ex ante. A judge first consults her own linguistic intuition — much as any fluent speaker would — to decide whether a particular reading is the text's "plain" or "ordinary" sense. Under this view, only afterward does the judge cite the relevant canon or dictionary to describe, classify, or bolster that pre-existing linguistic judgment.

Nothing about this explanatory, ex post use of a label would affect either the canons' descriptive accuracy or the judge's sincerity. Grammar handbooks, for example, do not create the future-perfect tense; they merely supply a name and a rule to capture native-speaker usage ("I will have finished . . ."). [143] In the same way, insofar as canons reflect linguistic usage in attested triggering conditions, a court's reference to *ejusdem generis* may simply articulate a commonsense tendency shared by ordinary readers to treat "gin, bourbon, vodka, rum, and other beverages" [144] as limited to alcoholic drinks, or a citation to the rule of the last antecedent may capture how English speakers ordinarily attach a modifier to the nearest noun phrase — even if none of those readers has ever heard the associated Latin terms.

This descriptive or explanatory account is most frequently the sense in which linguistic canons are discussed by the invoking judge. Justice Scalia once remarked that "[a]ll of this is so commonsensical that, were the canons not couched in Latin, you would find it hard to believe anyone could criticize them." [145] In a similar vein, Sotomayor's majority opinion in *Lockhart v. United States* [146] adopted the last-antecedent rule, reflecting the "basic intuition that when a modifier appears at the end of a list, it is easier to apply that modifier only to the item directly before it," [147] whereas Justice Kagan's dissent defended the competing series-qualifier canon as "the completely ordinary way

143. *See, e.g.*, BROCK HAUSSAMEN, GRAMMAR ALIVE! A GUIDE FOR TEACHERS 4 (2003) (describing "Standard English" as the variety that grammar books codify rather than create); Eli Hinkel, *Descriptive Versus Prescriptive Grammar*, *in* 1 THE TESOL ENCYCLOPEDIA OF ENGLISH LANGUAGE TEACHING 1–2 (Hossein Nassaji ed., 2018) (defining descriptive grammar as a study that "describes the language, its structure, and the syntactic rules that govern sentence and phrase constructions"); *What Is the Future Perfect?*, WALL ST. ENG., https://www.wallstreetenglish.com/exercises/what-is-the-future-perfect/ [https://perma.cc/582V-V4JG] (illustrating the English future perfect with examples such as "I'll have finished the report by lunchtime").

144. Tobia, Slocum & Nourse, *supra* note 19, at 237.

145. ANTONIN SCALIA, A MATTER OF INTERPRETATION 26 (1997).

146. 577 U.S. 347 (2016).

147. *Id.* at 351.

that people speak and listen, write and read."[148] Likewise, the plurality in *Yates v. United States*[149] relied on *noscitur a sociis* and *ejusdem generis* on the ground that a word's sense is "appropriately read" in light of "the words immediately surrounding" it — underscoring the implied importance that the canons simply "reflect" how English speakers take meaning from contextual companions.[150]

Because this explanatory use of canons is compatible with both accurate communication and sincere judging, the present study focuses on testing (i) the window-dressing model and (ii) the canon indeterminacy model — leaving to future work the question of when, if ever, the canons operate merely as ex post labels for judgments already reached through ordinary linguistic intuition.

C. Computational Tools

To the extent that judges invoking traditional interpretive tools fail to successfully arrive at plain meaning, a related open question concerns whether novel computational tools, such as LLMs, might prove to be a useful supplement or alternative. This Section documents the rise of LLMs and the need to empirically evaluate their potential promise in questions of interpretation.

1. Rise of LLMs

LLMs pertain to a class of AI models sometimes referred to as generative AI. Unlike traditional AI models, generative AI models are not only capable of performing a specific task with a limited set of possible rules or output (e.g., playing chess or recognizing faces) but are able to "generate" new content based on their training data in response to human prompting.[151] LLMs in particular are a form of generative AI that is trained on massive amounts of language in order to recognize and generate language content (as opposed to or in addition to images, for example).[152]

Over the past few years, LLMs, such as OpenAI's GPT models, have achieved human-level performance on a number of domain-

148. *Id.* at 364 (Kagan, J., dissenting).

149. 574 U.S. 528 (2015).

150. *Id.* at 543–44.

151. Kim Martineau, *What Is Generative AI?*, IBM (Apr. 20, 2023), https://research.ibm.com/blog/what-is-generative-AI [https://perma.cc/R2LR-8A9].

152. *See* Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman et al., A Comprehensive Overview of Large Language Models 2 (Oct. 17, 2024) (unpublished manuscript), https://arxiv.org/pdf/2307.06435 [https://perma.cc/G4SL-ZMYM].

general natural language processing tasks[153] and domain-specific language-based benchmarks, including those in law.[154] For example, recent work by Jon Choi and colleagues,[155] as well as other work by Blair-Stanek and colleagues,[156] found various GPT models to have passed law school exams in a blind grading format. Meanwhile, other work found LLMs to pass law licensing exams, such as the Uniform Bar Exam.[157]

More recent work has attempted to investigate AI systems' abilities to assist on more substantive legal work, with multiple studies finding efficiency gains for law students on various tasks,[158] and one survey finding self-reported productivity gains among lawyers using AI.[159]

In addition to task performance, some leading cognitive scientists and language experts have gone as far as arguing that LLMs are not only good at using language but can also inform theories of human linguistic cognition.[160]

Despite the impressive capabilities of LLMs, recent literature has also documented a few of their shortcomings. For example, shortly after the announcement by OpenAI that GPT-4 had achieved 90th percentile on the Uniform Bar Exam, it was revealed that this estimate was significantly overinflated, with its actual performance nearing the 62nd percentile and its score on the essay section falling within the bottom 20th percentile of practicing lawyers.[161] In addition, multiple

---

153. *See*, *e.g.*, Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydlo, Joanna Baran et al., *ChatGPT: Jack of All Trades, Master of None*, 99 INFO. FUSION 1, 9 (2023) (finding ChatGPT and GPT-4 to achieve near, though consistently below, state-of-the-art performance on classical NLP tasks).

154. *See generally* OpenAI, *supra* note 22, at 5.

155. Jonathan H. Choi, Kristin E. Hickman, Amy B. Monahan & Daniel Schwarcz, *ChatGPT Goes to Law School*, 71 J. LEGAL EDUC. 387, 391 (2022) (finding that GPT 3.5 passed law exams with a grade of roughly C+).

156. *See* Andrew Blair-Stanek, AI Gets Its First Law School A+s 1 (May 29, 2025) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5274547 [https://perma.cc/8Z23-2WLP]; *see generally* Andrew Blair-Stanek, Anne-Marie Carstens, Daniel Goldberg, Mark A. Graber, David Gray & Maxwell L. Stearns, GPT-4's Law School Grades: Con Law C, Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B (May 9, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4443471 [https://perma.cc/TM8E-LYJM].

157. *See* Daniel Martin Katz, Michael James Bommarito, Shang Gao & Pablo David Arredondo, *GPT-4 Passes the Bar Exam*, 382 PHIL. TRANSACTIONS ROYAL SOC'Y A 1, 5 (2024).

158. Aileen Nielsen, Stavroula Skylaki, Milda Norkute & Alexander Stremitzer, *Building a Better Lawyer: Experimental Evidence that Artificial Intelligence Can Increase Legal Work Efficiency*, 21 J. EMPIRICAL LEGAL STUD. 979, 980, 1002 (2024); *see* Jonathan H. Choi & Daniel Schwarcz, *AI Assistance in Legal Analysis: An Empirical Study*, 73 J. LEGAL EDUC. 384, 411 (2025).

159. Colleen V. Chien & Miriam Kim, Generative AI and Legal Aid: Results from a Field Study and 100 Use Cases to Bridge the Access to Justice Gap 27 (Mar. 14, 2024) (unpublished manuscript), https://ssrn.com/abstract=4733061 [https://perma.cc/HQ5E-26RP].

160. *See supra* note 25.

161. *See* Eric Martínez, *Re-Evaluating GPT-4's Bar Exam Performance*, 33 A.I. & L. 581, 598 (2024).

studies have found LLMs to hallucinate at an alarmingly high rate, with models such as GPT-3.5 estimated to hallucinate over 50% of the time,[162] and even more recent models touted to be "hallucination-free" have been estimated to hallucinate more than 20% of the time in certain contexts.[163] Finally, in some of the aforementioned work that examined AI's ability to assist in legal tasks, those on the right-side of the distribution (i.e. highest performers) tended to show worse performance when using AI compared to without using AI, suggesting that AI assistance may not be so effective as a cognitive aid for those who are already competent at a legal task.[164]

2. LLMs and Legal Interpretation

In light of the capabilities of LLMs, and in spite of the aforementioned shortcomings, legal academics[165] and even some judges[166] have begun advocating for using LLMs as a tool to uncover the meaning of a phrase in a legal document. For example, in a recent piece in the NYU Law Review, Yonathan Arbel and David Hoffman proposed using LLMs to resolve the meaning of contested contract terms.[167] Similarly, Christoph Engel and Richard McAdams recently examined the potential of ChatGPT to uncover the ordinary meaning of statutory terms such as "vehicle."[168] Meanwhile, in the judiciary, Eleventh Circuit concurrences by Judge Newsom have likewise speculated that using LLMs, such as ChatGPT, could "maybe" aid in uncovering the ordinary meaning of statutory terms.[169]

On the other hand, an equally vocal cohort of scholars has urged extreme caution — or outright rejection — of turning to LLMs for semantic guidance. Jon Choi, in a recent working paper, warns that model outputs can swing widely with slight tweaks in post-processing

---

162. Matthew Dahl, Varun Magesh, Mirac Suzgun & Daniel E. Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, 16 J. LEGAL ANALYSIS 64, 66–68 (2024).

163. Magesh, Surani, Dahl, Suzgun, Manning & Ho, *supra* note 25, at 225–26.

164. *See* Choi & Schwarcz, *supra* note 158, at 387. *But see* Daniel Schwarcz, Sam Manning, Patrick Barry, David R. Cleveland, JJ Prescott & Beverly Rich, *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice* 1, 51 (Minn. L. Stud. Rsch. Paper, Paper No. 25–16, 2025) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162111 [https://perma.cc/ZL35-RZ8R] (finding less of an equalizing effect with law students using more advanced reasoning models, such as o1-preview, relative to previous work with GPT-4).

165. *See, e.g.*, Arbel & Hoffman, *supra* note 26, at 455.

166. *See* Snell, *supra* note 27, at 1226–34.

167. *See* Arbel & Hoffman, *supra* note 26, at 455.

168. *See* Engel & McAdams, *supra* note 26, at 239–43.

169. *See* Snell, *supra* note 27, at 1225; *cf.* United States v. Deleon, 116 F.4th 1260, 1270, 1277 (11th Cir. 2024) (Newsom, J., concurring) (calling the opinion a "sequel" to his *Snell* concurrence and concluding "that LLMs may well serve a valuable auxiliary role as we aim to triangulate ordinary meaning").

choices or training data, and that post-training "techniques render LLMs less biased and more practically useful, but they also cause LLM outputs to deviate from empirical predictions of language use in practice."[170] A multidisciplinary team led by Brandon Waldon argues, in a similar vein, that "[j]udges should not rely on direct queries to ChatGPT (or similar chatbots) about the meaning of legal texts."[171] Echoing this skepticism, Thomas Lee and Richard Egbert conclude that existing AI tools "as currently constituted, cannot" tell judges the ordinary meaning of a word or phrase,[172] while a recent Lawfare commentary by a group of legal-tech scholars bluntly asserts that "[j]udges [s]houldn't [r]ely on AI for the [o]rdinary [m]eaning of [t]ext."[173]

Many of these pieces themselves acknowledge that plain meaning is typically defined as clarity according to some relevant population of human readers. That said, the value of LLMs as tools in uncovering plain meaning is largely, if not directly, related to how well LLMs serve as a proxy for the meaning gleaned by some relevant population of human interpreters. Yet in no study to date has there been a systematic comparison between how well LLMs are able to emulate the interpretations of human readers regarding the meaning of words at issue in real-world cases.

Insofar as LLMs are able to emulate the interpretations of human readers, an additional potential use case of computational tools is not only to help judges determine the plain meaning in a given case alongside their use of canons but more broadly to refine and discover new linguistic canons beyond those currently in existence. After all, as pointed out by Kevin Tobia, Brian Slocum, and Victoria Nourse,[174] linguistic canons, properly understood, are an open set of proxies for language usage and understanding. Therefore, to the extent that one can employ computational tools such as LLMs to uncover reliable proxies of language usage and understanding, it follows that one can use these tools to uncover new canons, thus adding to the set of traditional tools used by judges to find plain meaning.

III. STUDY 1: TRADITIONAL TOOLS

This Part presents an experimental study that attempts to empirically move the needle on the theoretical debates traced in Part II.

---

170. Choi, *supra* note 28, at 9.

171. Waldon, Schneider, Wilcox, Zeldes & Tobia, *supra* note 25, at 4.

172. Lee & Egbert, *supra* note 29, at 56.

173. Justin Curl, Peter Henderson, Kart Kandula & Faiz Surani, *Judges Shouldn't Rely on AI for the Ordinary Meaning of Text*, LAWFARE (May 22, 2025, at 13:00 ET), https://www.lawfaremedia.org/article/judges-shouldn-t-rely-on-ai-for-the-ordinary-meaning-of-text [https://perma.cc/4WGK-6W4F].

174. *See* Tobia, Slocum & Nourse, *supra* note 19, at 225.

Section III.A formalizes several competing accounts regarding the canons into mathematical models that yield testable empirical predictions. Section III.B tests those predictions via an experimental jurisprudence study, and Section III.C details the results.

A. Questions & Hypotheses

Part II traced two long-running debates about how legal interpreters use so-called ordinary meaning. Each debate yields a pair of rival hypotheses that Box 1 and Box 2 (found below) restate in mathematical form and translate into observable predictions. The prose below summarizes those questions and hypotheses before turning to data and methods.

Model assumptions and notation are further expanded upon in Appendix 4.

1. Are the Canons Indeterminate?

As explained in Part II, many scholars have debated the determinacy of language and interpretive tools in cases where judges purport to determine the ordinary meaning of a legal text. In particular, whereas some have speculated that such cases are inherently indeterminate, such that for every canon there exists an equal canon of opposite force (*canon indeterminacy model*), others have asserted that, in practice, a coherent subset of canons apply and yield a clear reading (*canon determinacy model*).[175] Given that ordinary meaning is defined by judges and scholars as the meaning a relevant population of readers would attribute to the text,[176] this controversy is ultimately empirical.

If the canon indeterminacy hypothesis were true, then one would not expect there to be consensus in plain meaning cases — that is, readers will *divide* between interpretations linked to competing canons; no single reading will command consensus support. Conversely, if the linguistic determinacy hypothesis were true, then under the same conditions, one would expect that ordinary and well-informed readers would converge on the interpretation supported by a single canon (or mutually compatible set of canons).

Box 1 formalizes these hypotheses and predictions in mathematical notation (with further expansion in Appendix 4).

175. *See supra* notes 124–141 and accompanying text.
176. *See supra* Section II.A.2.

2. Are the Canons a Smokescreen?

A related line of scholarship asks how judges use canons. Whereas some have taken the position that these tools are "so malleable as to operate mostly as pretext," and that judges merely use them as "window-dressing"[177] tools to camouflage their policy preferences,[178] others have taken the position that canon usage is a neutral method grounded in the ordinary meaning of enacted text, albeit subject to the skill of the interpreter.

These two accounts of the mechanisms of judicial behavior likewise make competing predictions. In particular, if the *window-dressing model* were true, then one would not expect that judges would consistently align with ordinary and well-informed readers when invoking these tools. After all, the individual policy preferences of a judge are unlikely to consistently track the collective linguistic consensus of ordinary and well-informed readers, and a judge can simply invoke a counter-canon if they don't like the policy outcome dictated by the canon that does track linguistic consensus.[179]

Conversely, if judges invoked canons to support a sincere attempt to get at plain meaning, one would expect a greater-than-chance alignment with linguistic consensus (see Box 2 and Appendix for more details).[180]

177. *See supra* note 20.

178. John F. Manning, *Legal Realism & the Canons' Revival*, 5 GREEN BAG 2d 283, 283 (2002).

179. *See* Posner, *supra* note 134 (acknowledging that "text as such may be politically neutral," even if "textualism is not"). The assumption is further validated in the experiment. *See infra* Appendix 2.I.

180. Note that this assumes that interpreters can infer linguistic consensus at a rate higher than chance. The experiment validates this assumption. *See infra* Section III.C.

Box 1: Canon (In)Determinacy Hypotheses

Definitions

Let $C = \{1, \ldots, N\}$ be the universe of "plain-meaning" cases and $S \subseteq C$ a survey sample ($|S| = s$). For any case k there are two competing interpretations $\mathcal{I}_k = \{i_{k1}, i_{k2}\}$, each supported by at least one canon ($\mathcal{T}(i_{kj}) \neq \varnothing$). Let π_{kj} denote the share of the relevant readership Ω choosing i_{kj}.

Canon-Consensus Monotonicity (CCM). Under ordinary-meaning doctrine the strength of the canon set backing i_{kj} grows monotonically with reader support:

$$\rho_{kj} = f(\pi_{kj}), \qquad f'(x) > 0 \quad (\therefore\ \pi_{kj} > \pi_{k\ell} \Rightarrow \rho_{kj} > \rho_{k\ell}).$$

Consensus Threshold. Let $\text{Con}(k) = \max\{\pi_{k1}, \pi_{k2}\}$ be the interpretation commanding the highest degree of linguistic consensus. Let $\tau \in [.5, 1]$ denote some pre-specified threshold of consensus needed for determinacy.

Determinacy Indicator. A case is *determinate* when $\text{Con}(k) > \tau$ and *indeterminate* otherwise:

$$A_k = 1\{\text{Con}(k) > \tau\}, \qquad I_k = 1 - A_k.$$

Let $q = \Pr_{k \in C}(I_k = 1)$ be the population share of indeterminate disputes.

Hypotheses

Canon Indeterminacy (H_{CI})
$H_{\text{CI}}(\lambda):\ q \geq \lambda \quad (0 \leq \lambda \leq 1)$
Claim: At least λ cases are indeterminate, such that (from CCM) each canon is opposed by an equally powerful counter-canon (i.e. $\rho_{k1} \approx \rho_{k2}$).

Canon Determinacy (H_{CD})
$H_{\text{CD}}(\lambda):\ q < \lambda$
Claim: At least λ cases are determinate, such that (from CCM) one canon dominates (e.g. $\rho_{k1} > \rho_{k2}$).

Predictions

Let $\bar{I}_s = \dfrac{1}{s} \sum_{k \in S} I_k$ be the sample indeterminacy rate. Then

$$\bar{I}_s \geq \lambda \implies H_{\text{CI}}(\lambda) \text{ is supported}$$

$$\bar{I}_s < \lambda \implies H_{\text{CD}}(\lambda) \text{ is supported}$$

Box 2: Window-Dressing vs Cognitive Constraints Hypotheses

Definitions

For any case k there are two candidate interpretations $\mathcal{I}_k = \{i_{k1}, i_{k2}\}$, each supported by at least one canon ($\mathcal{T}(i_{kj}) \neq \varnothing$). Let i_k^* be the interpretation commanding a higher degree of linguistic consensus. The court adopts interpretation $J_k \in \mathcal{I}_k$ and cites a canon in support T_k.

Policy Pay-Off Vector. Let $\vec{d}_k = \langle d_{k1}, d_{k2} \rangle$ where d_{kj} is the desirability of interpretation i_{kj} to the deciding judge.

Independence Axiom (IA). Judges' policy preferences do not track linguistic consensus:

$$P(\vec{d}_k,\ i_k^*) \ = \ P(\vec{d}_k)\, P(i_k^*), \qquad \text{equivalently } \vec{d}_k \perp\!\!\!\perp i_k^*.$$

Misperception Probability. Let $\hat{i}_k^*$ be the judge's perceived modal meaning in case k and let $\varepsilon_k = \Pr(\hat{i}_k^* \neq i_k^*)$ denote the probability that this is incorrect.

Alignment Indicator. Let $B_k = \mathbf{1}\{J_k = i_k^*\}$ indicate a match between the court's interpretation and the modal meaning, and let $\theta_k = \frac{1}{2}$ be the chance threshold.

Hypotheses

Window-Dressing Realist (H_{WD})
P1: Policy–maximising choice $J_k = \arg\max_j d_{kj}$.
P2: Post-hoc tool citation $T_k \in \mathcal{T}(J_k)$.
Claim: Judge chooses interpretation J_k to maximize policy pay-off d_{kj}, without regard to linguistic consensus, then cites canon T_k as smokescreen.

Cognitively Constrained Formalist (H_{CCF})
Q1: Consensus-maximising choice $J_k = \hat{i}_k^*$ with $\Pr(\hat{i}_k^* \neq i_k^*) = \varepsilon_k$.
Q2: Ex-post tool citation $T_k \in \mathcal{T}(\hat{i}_k^*)$.
Claim: Judge chooses interpretation J_k that she believes reflects consensus $\hat{i}_k^*$, with misperception probability ε_k, then cites canon T_k in support.

Predictions

Let $\bar{B}_s = s^{-1} \sum_{k \in S} B_k$ be the sample alignment rate and let $\bar{\theta}_s = 0.5$ be the chance threshold. From IA, and assuming $\varepsilon_k < 0.5$, then

$$\bar{B}_s \ \leq \ \bar{\theta}_s \implies H_{\text{SS}} \text{ supported}$$

$$\bar{B}_s \ > \ \bar{\theta}_s \implies H_{\text{CF}} \text{ supported}$$

B. Methods

1. Materials

To answer these questions, 180 sets of materials ("items") were constructed from real-world cases, in which the deciding judge

explicitly invoked a linguistic canon or other interpretive tool to privilege one construction over a rival counter-canon while professing fidelity to ordinary meaning.

The number of items is, to the author's knowledge, several times larger than that of any other experimental jurisprudence study to date,[181] and one to two orders of magnitude higher than most other behavioral experiments.[182] This large sample of materials was chosen to maximize the generalizability of the research findings, given the breadth of the research question and the known concern of replicability in experimental research.[183]

The items were derived from real-world cases as opposed to hypothetical ones to maximize ecological validity. In particular, each item presented participants with the legal text considered by the court and asked participants to interpret that provision as applied to the relevant facts as raised by the parties and considered by the court.

a. Categories of Cases

The set of items was constructed to appropriately span the diverse types of plain meaning cases, with a particular emphasis on the contexts in which the hypotheses being tested were most attested to apply. In particular, items and corresponding cases from which the items were adapted consisted of sixteen types of cases divided across three main categories.

The principal category comprised canon cases — instances where the deciding judge explicitly invoked a linguistic canon to privilege one construction over a rival counter-canon while professing fidelity to ordinary meaning. Drawing on Llewellyn's classic "thrust-and-parry" taxonomy, we sampled all seven thrust-parry pairings in rows 20–26 of his 1950 chart and treated the two sides of each pairing as separate

181 . To the author's knowledge, the largest set of stimuli used in experimental jurisprudence research prior to this study was less than three dozen items. *See* Roseanna Sommers, *Commonsense Consent*, 129 YALE L.J. 2232, 2306 (2020) (testing "over two dozen scenarios" across various legal contexts); Kevin Tobia, *How People Judge What Is Reasonable*, 70 ALA. L. REV. 293, 359 (2020) (testing people's reasonableness judgments across nearly three dozen domains).

182. For examples of important and influential studies with fewer than ten stimuli, see, e.g., Amos Tversky & Daniel Kahneman, *The Framing of Decisions and the Psychology of Choice*, 211 SCI. 453, 457 (1981) (using one stimuli to demonstrate framing effects of gains versus losses); Christopher Jaeger, *The Hand Formula's Unequal Inputs*, 135 YALE L.J. 461 (2025) (using five vignettes to test the influence of different aspects of the Hand Formula on people's judgments of reasonableness).

183. *See, e.g.*, Felix Holzmeister, Magnus Johannesson, Robert Böhm, Anna Dreber, Jürgen Huber & Michael Kirchler, *Heterogeneity in Effect Size Estimates*, 121 PROC. NAT'L ACAD. SCI. U.S.A. 1, 1 (2024); Open Science Collaboration, *Estimating the Reproducibility of Psychological Science*, 349 SCI. 943, 4716–1 (2015) (finding that more than 50% of significant results in psychology studies failed to replicate).

sub-types, yielding fourteen canon categories.[184] For each condition we created ten independent items, producing 140 canon items in total.

To maximize generalizability, the design also incorporated two non-canon categories that capture the principal alternative mechanisms judges deploy when resolving lexical disputes. The first covered dictionary cases, in which a court preferred one dictionary definition over a competing tool to justify its interpretation. The second was judicial-authority cases, where a court leaned either on precedent or on its own textual analysis. Twenty items were selected within each of these categories, bringing the corpus to 180 items (140 canon, 40 non-canon). An illustration of the categories is visualized in Table 1.

Table 1: Breakdown of Experimental Conditions

Canon Conditions	
Thrust	**Parry**
Expression of one thing excludes another (*expressio unius*).	Some things may be mentioned only by way of example.
General terms are to receive a general construction.	General terms may be limited by context (*noscitur a sociis*).
Where general words follow an enumeration they apply only to things of the class mentioned (*ejusdem generis*).	General words must operate on something. Further, *ejusdem generis* is only an aid in getting the meaning.
Qualifying or limiting words are to be referred to the next preceding antecedent (Last Antecedent Rule).	Qualifying or limiting words apply to the entire series of nouns or verbs in a construction (Series Qualifier Rule).
The plain meaning of a provision will heed the commands of its punctuation.	Punctuation marks will not control the plain meaning of language.
"And" is to be read conjunctively; "Or" is to be read disjunctively.	"And" or "or" may be read interchangeably.

184. *See* Llewellyn, *supra* note 16, at 401–06.

There is a distinction between words of permission and mandatory words.	Words imparting permission may be read as mandatory and vice versa.
Non-Canon Conditions	
Dictionary	
Dictionaries are a reliable guide to a word's plain meaning.	
Judicial Authority	
Judicial authority is a reliable guide to a word's plain meaning.	

b. Material Selection

Within each condition, cases were chosen from legal treatises and databases and selected to be included in an initial draft of approximately 360 items if they satisfied three main criteria. First, the case had to involve the interpretation of a word or phrase in a legal text of some sort (e.g. statute, will, contract).

Second, in order to faithfully investigate the attested malleability and indeterminacy of interpretive tools to justify ordinary meaning, each case must have purported to interpret the meaning of the words at issue in the text by appeal to one of the categories of traditional interpretive tools outlined above at the expense of another. In each of the canon cases, the judge had to invoke a canon at the expense of the counter-canon (that is, the judge could have invoked a corresponding counter-canon to justify the opposite interpretation/result).[185] In the non-canon cases, the judge had to invoke a tool to rule out one definition in favor of another (e.g., broad vs narrow), even if the tool itself did not always have a consistent parry (e.g., in dictionary cases, the definition rejected by the judge was not always dictated by another dictionary definition but instead sometimes dictated by judicial authority, a technical handbook, or linguistic canon).

Third, the context relevant to the court's determination of plain meaning could be feasibly adapted into a survey format without a loss or distortion of meaning. Although most cases satisfied this criteria, this ruled out a small number of cases involving (a) exceedingly long provisions, or several disparate provisions of a code; (b) inordinate

185. This ruled out cases such as those where, for example, the court invokes *ejusdem generis* but decides that the category referred to by the exemplars encompass the target scenario (such as when a judge decides that the exemplar "airplane" is *ejusdem generis* with the category "vehicles"), since this result is the same result that the court would have arrived at had they decided to reject *ejusdem generis* and interpret the catch-all term broadly.

amounts of background knowledge (e.g. complicated bankruptcy or tax cases);[186] and (c) exceedingly old provisions, such that contemporary understanding would not serve as an accurate proxy of plain meaning at the time of judgment.

From the initial set of 360 items, the final of set of 180 were chosen from this larger set in order to achieve a diversity of materials across jurisdictions (state versus federal), court level (supreme versus circuit versus district), legal text (e.g., statutes, contracts, wills) and subject matter (e.g., criminal and civil disputes).

In addition to achieving a balance across these categories, the final set also included a large number of cases that had been discussed in the prior literature as paradigmatic examples of potential canon misuse, such as *Facebook v. Duguid*,[187] *Smith v. United States*,[188] *Yates v. United States*,[189] *Bostock v. Clayton County*,[190] *Lockhart v. United States*,[191] and *Circuit City Stores v. Adams*.[192] The full list of cases is available in the Appendix 2.C.

The items were informally piloted by a set of fifteen law professors and legal practitioners to ensure that the format and content would be understandable to a general legal audience, and then formally piloted by a set of 120 lay subjects to ensure the same among a general lay audience.

186. However, some cases involving the interpretation of two provisions in tandem were included in the final set of materials.

187. 592 U.S. 395 (2021).

188. 508 U.S. 223 (1993).

189. 574 U.S. 528 (2015).

190. 590 U.S. 644 (2020).

191. 577 U.S. 347 (2016).

192. 532 U.S. 105 (2001).

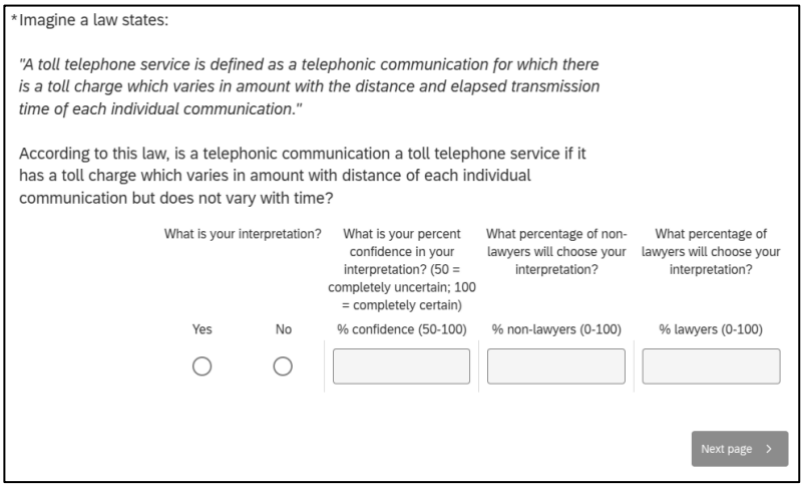

Figure 2: Sample item as viewed via the survey interface.

c. Dependent Variables

Across the sixteen conditions, all items had the same basic structure, with each item presenting participants with the law and facts raised by the parties and reported by the judge as relevant to the plain meaning judgment.

The item first presented participants with a provision or pair of provisions from the relevant legal text from the court case, followed by a question regarding the application of the provision to a set of facts (i.e., those from the court case).

Below is a sample item from the *expressio unius* thrust condition, derived from the Virginia Supreme Court case of *Tate v. Ogg*:[193]

> Imagine a law states:
>
> *"It is unlawful to permit any horse, mule, cattle, hog, sheep, or goat to run at large upon enclosed lands."*
>
> According to this law, is it unlawful to permit a turkey to run at large upon enclosed lands?

Below is a sample item from the *expressio unius* parry condition, derived from the 4th Circuit case *United States v. Hawley*:[194]

> Imagine a law states:

193. 195 S.E. 496, 499 (Va. 1938).
194. 919 F.3d 252, 255 (4th Cir. 2019).

> "*Prior sentences are to be counted in the criminal history score, including misdemeanor sentences where imprisonment was not imposed.*"
>
> According to this law, are misdemeanor sentences where imprisonment was imposed to be counted in the criminal history score?

In order to best operationalize plain meaning judgments, each item contained two separate dependent variables that elicited participant judgments regarding the meaning of the provision as applied to the fact pattern.

The first dependent variable was a binary yes/no question asking about the participants' own interpretation of the meaning of the provision as applied to the fact pattern.

The second dependent variable was a numeric percentage question asking, on a scale of 50–100%, the participant's level of confidence in their preferred interpretation (50% representing a 50-50 coin flip or complete uncertainty between the two interpretations, and 100% representing complete certainty in their interpretation over the other interpretation). The purpose of this dependent variable was to obtain a more fine-grained, within-participant measure of plain meaning in addition to the binary between-participant measure obtained by the first dependent variable.

In addition to the binary and scalar plain meaning dependent variables, each item also contained two additional dependent variables that elicited participants' predictions regarding other participants' responses to the plain-meaning dependent variables.

The first such prediction dependent variable asked, on a scale of 0–100, what percentage of non-lawyer participants would choose their preferred interpretation.

The second prediction dependent variable asked, on a scale of 0–100, what percentage of lawyer participants would choose their preferred interpretation.

A sample item, including all dependent variables and as viewed via the survey interface, is visualized in Figure 2.

2. Participants and Procedure

a. Participant Recruitment

In order to simulate the two populations relevant to a plain-meaning analysis (ordinary and well-informed readers), participants were recruited from two different target populations: lawyers and laypeople. The unprecedentedly large number of materials required a

commensurately large sample size. The target sample of lawyers was 2,250, and the target sample of laypeople was 4,500. Both target sample sizes were specified in advance as part of this study's pre-registration.[195] The lawyer sample was determined based on a power analysis, which computed the relevant sample necessary to obtain a within-condition effect size at least half as large as that observed in a pilot study. From there, the layperson sample was doubled from that number to feasibly allow for the computation of exploratory analyses.

The sample size also ensured that, on average, at least 100 participants in each population would see each item (*see* Procedure, *infra* Section III.B.2.b), in line with Chief Justice Roberts' comment during oral argument of *Facebook, Inc.* v. *Duguid* that "the most probably useful way of settling all these questions [of ordinary meaning] would be to take a poll of 100 ordinary . . . speakers of English and ask them what [the law] means."[196]

Laypeople were recruited via the online recruiting platform Prolific.[197] In particular, participants were recruited via Prolific's demographically representative sample criteria, which ensures that a sample is broadly representative of the United States census on the dimensions of age, gender, ethnicity, and politics.[198]

Lawyers were recruited via direct email invitations, sent to a list of names and email addresses (n~200,000) scraped from various publicly available online sources, including websites of large law firms and state bar directories.

Lay participants were compensated at a prorated rate of $12 per hour for their time. Lawyer' participation was voluntary, though with an additional incentivization structure: lawyers were eligible for up to a $1000 reward, based on how accurate their predictions of other participants would be (in particular, $500 to the participant with the most accurate predictions of other lawyer responses; and $500 to the participant with the most accurate predictions of non-lawyer responses).

---

195. *See Interpretation of Legal Texts, UChicago Law, 2024 (#202821)*, ASPREDICTED (Dec. 4, 2024, at 13:04 ET), https://aspredicted.org/wbck-t8m7.pdf [https://perma.cc/UB5F-M3QP].

196. Transcript of Oral Argument at 51–52, Facebook, Inc. v. Duguid, 592 U.S. 395 (2021) (No. 19-511) ("[If] our objective is to settle upon the most natural meaning of the statutory language to an ordinary speaker of English . . . the most probably useful way of settling all these questions would be to take a poll of 100 ordinary . . . speakers of English and ask them what . . . [the statute] means, right?").

197. *See* PROLIFIC, https://www.prolific.com/ [https://perma.cc/3JZQ-ZF54].

198. *What Are Representative Samples on Prolific*, PROLIFIC (Sep. 16, 2025), https://researcher-help.prolific.com/en/article/95c345 [https://perma.cc/GFQ5-BDS3].

b. Procedure

With respect to procedure, each participant saw eight items in random order — two each of four conditions. The assignment of participants to the four conditions was random, except that participants were never assigned to both a thrust and parry of the same canon pair (e.g., if a participant was assigned to the last-antecedent condition, they would not be assigned to the series-qualifier condition). The format of the items was the same across the two participant groups, except that for the prediction questions, only lawyers were asked to predict other lawyer responses in addition to layperson responses (lay participants were instead asked only to predict the responses of other Prolific participants).

In addition to these eight trials, to ensure data quality, participants also completed two attention checks — one hard, one easy — which looked identical to the main materials, except that the meaning of the provisions as applied to the scenario was uncontroversially unambiguous.

The easy attention check presented participants with the provision "[n]o vehicles are allowed in the park" and asked participants whether, according to the provision, large vehicles were allowed in the park. The purpose of this attention check was to ensure that participants stayed minimally alert throughout the study and understood the basic instructions of the task. Participants who did not respond "no" to this attention check and with at least 90% certainty in their response were excluded from the final analysis.

The hard attention check question asked participants to interpret a paired-down version of a Massachusetts driving under the influence law.[199] The purpose of the hard attention check question was to ensure that participants had a high enough comprehension level of legal syntax to understand the language in the materials. To that end, the law contained multiple center-embedded structures — a type of nested syntactic structure that has been identified in previous studies to be disproportionately common in legal texts relative to standard English and to pose comprehension difficulties for both laypeople and

199. The law, as presented to study participants, read as follows:

> Whoever, upon any way or in any place to which the public has a right of access, or upon any way or in any place to which members of the public have access as invitees or licensees, operates a motor vehicle with a percentage, by weight, of alcohol in their blood of eight one-hundredths or greater, or while under the influence of marijuana must be punished by a fine of five hundred dollars.

See MASS. GEN. LAWS ch. 90, § 24 (2021) (original statute that served as the basis).

lawyers.[200] Participants who did not respond "yes" to this attention check (with whatever level of certainty) were excluded from the final analysis.

3. Analysis Plan

Following best practices of experimental research, the primary analyses were specified in the pre-registration associated with this study.[201]

The analysis plan is reported in full in Appendix 2.B.

C. Results

1. Demographics

This Subsection summarizes the self-reported demographic data from the study's participants. More detailed results are reported in Appendix 2.A.

Following the pre-registration, participants who failed either of the attention checks were excluded from the analyses. Sample sizes of both participant groups post-exclusion met the study's target sample of 4,500 laypeople and 2,250 lawyers.

From the layperson sample, 13.8% of participants failed one or both of the comprehension check questions. After removing 182 participants who reported having been licensed as an attorney (and moving them to the lawyer sample),[202] this resulted in a final sample of 4,533 lay participants.

From the direct-email lawyer sample, 11.4% of participants failed one or both comprehension check questions. After excluding an additional twenty-four participants from this sample who did not report having been licensed as an attorney,[203] as well as after including the lawyers from the Prolific sample, this resulted in a final sample of 2,373 lawyer participants.

---

200. *See* Martínez, Mollica & Gibson, *Even Lawyers Do Not Like Legalese*, *supra* note 54, at 1–2; *see also* Eric Martínez, The Cognitive Underpinnings of Legal Complexity 1, 102 (Sep. 2024) (Ph.D. dissertation, Massachusetts Institute of Technology) (on file at MIT Doctoral Thesis Archive).

201. *See* ASPREDICTED, *supra* note 195.

202. Note that these lawyer participants then completed a subsequent demographics questionnaire to mirror the one completed by the direct-email lawyer participants.

203. Note that these participants were excluded as opposed to being transferred to the layperson sample, given that the layperson sample as recruited via Prolific was already sufficiently large and already tailored to be representative of the United States population on relevant dimensions (in contrast, participants who took the survey via direct email invites to lawyers and simply did not report having been licensed as an attorney were much less likely to be representative of the lay population as a whole).

Although the direct email method of recruitment meant that there was somewhat less control over the demographics of the eventual lawyer sample, the final sample ended up being broadly representative of the legal population (as documented by official and unofficial methods) on the dimensions of gender,[204] ethnicity,[205] and political affiliation.[206] With respect to age, it skewed slightly older than official estimates of the lawyer population at-large.[207]

In addition to general demographics, lawyers completed two other questionnaires related to their areas of expertise and interpretive philosophies, estimates of which are scarce to non-existent in the prior literature.[208] Results are reported in Appendix 2.A.

2. Linguistic Consensus

Linguistic consensus results are visualized in Figure 3. This Subsection reports the level of consensus in greater detail.

204. 58.2% of the lawyers in the sample identified as male in the sample, relative to 58.4% of lawyers in the United States identifying as male according to the 2024 ABA National Lawyer Population Survey. *See Demographics,* A.B.A., https://www.americanbar.org/news/profile-legal-profession/demographics/ [https://perma.cc/D3QP-DX4L].

205. For example, 20% of the lawyer sample identified as being of color, relative to 23% in the lawyer population at-large in the United States. More specifically, the percentage of lawyers identifying as Black, Asian, Hispanic, and Native American in the sample were 4.1%, 6.3%, 9.1%, and 0.5%, respectively, relative to 5%, 7%, 6%, and 1% in the general lawyer population, respectively. *Id.*

206. For example, 59.5% of lawyers in the sample identified as somewhat liberal in the sample, relative to approximately 62% of lawyers overall according to previous estimates. Adam Bonica, Adam S. Chilton & Maya Sen, *Political Ideologies of American Lawyers*, 8 J. LEGAL ANALYSIS 277, 292 (2015) ("[S]ome 62 percent of the sample of attorneys are positioned to the left of the midpoint between the party means for members of Congress.").

207. *See* A.B.A., *supra* note 204.

208. For estimates of these demographics in the legal academic population, see Eric Martínez & Kevin Tobia, *What Do Law Professors Believe About Law & Legal Academy*, 112 GEO. L.J. 111, 140–45 (2023).

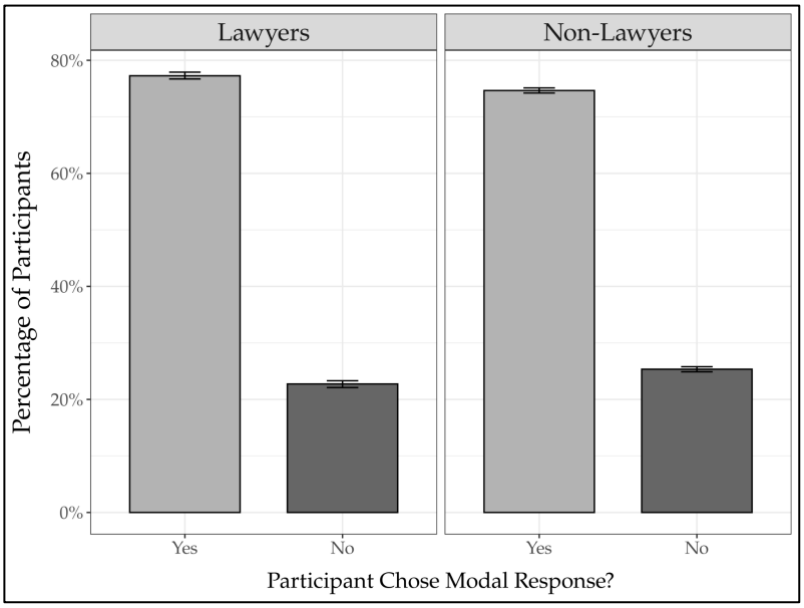

Figure 3: Degree of linguistic consensus among lawyer and non-lawyer participants. Bar heights represent average percentage of participants who converged on the modal interpretation for a given item. Error bars represent 95% bootstrapped confidence intervals.

a. Lay Consensus

Among lay participants, across all trials the percentage of participants who converged on the dominant interpretation was 74.7%.[209] In 127 out 180 (70.5%) of cases, a supermajority (two-thirds or more) of participants converged on one interpretation over another. In the majority of cases, this convergence was at or above 75% of participants.

According to the pre-registered regression model, this tendency was significantly above chance (i.e., lay participants were significantly more likely to converge on one interpretation over another across all trials, after controlling for variation among participants, conditions and items/cases).[210]

When broken down by condition, similar patterns were found. In twelve out of the fourteen canon conditions,[211] and in both of the non-

209. 95% CI: 74.3–75.2.

210. 95% CI of the intercept (after converting from log-odds to probability): 0.757–0.798.

211 . The two canon conditions where convergence was below a supermajority of participants were may versus shall parry condition (65.5%; 95% CI: 63.6–67.9) and the punctuation thrust condition (64.4%; 95% CI: 62.4–66.6).

canon conditions, a supermajority of participants, on average, converged on one interpretation over another.

The canon condition with the highest level of consensus towards one interpretation over another was the "General terms must operate on something" canon (*ejusdem generis* parry),[212] followed by "There is a distinction between words of permission and mandatory words"[213] and *expressio unius*.[214]

Similar results were observed when analyzing participants' level of certainty in responses as opposed to taking the proportion of yes/no responses. Across all cases, the average level of certainty in the consensus/majority interpretation was 70.8%.[215] Even when endorsing the dissensus interpretation, participants tended to be highly confident in their responses, as participants' average level of certainty in their own interpretation (without regard to whether this was the consensus or dissensus interpretation) was 86.5%.[216]

b. Lawyer Consensus

Consensus was similarly high among lawyer participants. Across all trials the proportion of lawyers who converged on the consensus interpretation was 77.3%.[217] According to the pre-registered regression model, this tendency was significantly above chance (i.e., participants were significantly more likely to converge on one interpretation over another across all trials, after controlling for variation among participants, conditions, and cases).[218] In 132 out of 180 (77.3%) cases, a supermajority of lawyer participants converged on one interpretation over another. In the majority of cases, this convergence was above 75% of lawyer participants.

As with laypeople, when broken down by condition, similar levels of consensus were found relative to the overall level. In fact, within each of the sixteen conditions, a supermajority (two thirds or more) of lawyer participants, on average, converged on one interpretation over another. The condition with the highest level of consensus towards one interpretation over another was "There is a distinction between words of permission and mandatory words" (94.0%), [219] followed by

212. 86.7% mean; 95% CI: 85.3–88.2.
213. 83.0% mean; 95% CI: 81.2–84.7.
214. 82.1% mean; 95% CI: 80.3–83.9.
215. 95% CI: 70.4–71.2.
216. 95% CI: 86.3–86.7.
217. 95% CI: 76.6–77.8.
218. 95% CI of the intercept (after converting from log-odds to probability): 0.790–0.832.
219. 95% CI: 92.4–95.3.

expressio unius (84.1%)[220] and "General terms are to receive a general construction" (81.8%).[221]

These results held constant when analyzing participants' level of certainty in responses as opposed to taking the proportion of yes/no responses. For example, across all cases, the average level of certainty in the consensus/majority interpretation was 73.9%.[222] On average, lawyers' level of certainty in their own interpretation (without regard to which interpretation they chose) was 89.4%.[223]

c. Lawyer versus Lay Consensus

Comparing lawyer and lay responses reveals a strong correlation between the interpretations that were converged on by ordinary people and those converged on by those with legal expertise, as well as the strength of convergence on those interpretations between the two groups. Lawyers and laypeople converged on the same consensus interpretation in 146 out of 180 (81.1%) cases, and in the majority of cases/items in all conditions except for one ("'May' and 'shall' may be read interchangeably").[224]

Not only did the two groups converge on similar interpretations of the words at issue across the 180 cases, but the two groups also showed similar levels of consensus around those interpretations. Across all conditions, the correlation between lawyer and lay consensus was 0.831,[225] though lawyers showed slightly higher levels of within-person certainty and between-person convergence on the consensus interpretation.

d. Exploratory Analyses and Robustness Checks

Control analyses revealed that these main results were robust to demographic variables. Participants tended to converge on the same

220. 95% CI: 81.7–86.3.
221. 95% CI: 79.4–84.1.
222. 95% CI: 73.4–74.4.
223. 95% CI: 89.2–89.6.
224. Lawyers and laypeople had different consensus interpretations in six out of ten of these cases.
225. That is, for each item, taking the average number of "Yes" responses among lawyer participants and comparing that to the average number of "Yes" responses among lay participants.

consensus interpretation in the majority of cases, regardless of factors such as age,[226] gender,[227] race[228] or politics.[229]

In addition to consensus across interpretive tools, analyses revealed similarly high levels of consensus across court level — that is, a similar proportion of participants converged on one interpretation over another in lower court cases as in Supreme Court cases.[230] The same was true for jurisdiction (state versus federal cases),[231] and legal text (cases involving the interpretation of public legal documents, such as a statute or regulation, versus private legal documents, such as a contract or will).[232]

3. Judge Alignment with Consensus

Analyses revealed a relatively high degree of alignment between a judge's interpretation and the linguistic consensus of both lawyer and lay participants. This tendency held true regardless of whether analyzed at the case/item level, condition level, or individual participant trial level, and regardless of whether consensus was operationalized as the proportion of participants endorsing one interpretation over another or via the average percentage confidence in the preferred interpretation. This Section details these results, as shown in Figures 4 and 5.

226. For example, participants fifty and older converged with the majority of participants younger than fifty in 93.3% of cases in the lawyer sample, and in 90.6% in the layperson sample.

227. Male-identifying participants converged with the majority of non-male-identifying participants in 93.9% of cases in the lawyer sample, and in 92.8% in the layperson sample.

228. White participants converged with the majority of non-White participants in 94.4% of cases in the lawyer sample, and in 92.2% in the layperson sample.

229. Liberal participants converged with the majority of non-liberal participants in 93.9% of cases in the lawyer sample. In the layperson sample, Democrats converged with the majority of Republican participants in 93.9% of cases.

230. For example, among lawyers, the average level of consensus was 78.7% for Supreme Court cases (95% CI: 77.7–79.6), compared to 76.7% for circuit court cases (95% CI: 75.7–77.7) and 75.6% for district court cases (95% CI: 74.3–77.0).

231. Among lawyers, the average level of consensus was 78.1% for state cases (95% CI: 77.3–78.9), compared to 76.3% for federal cases (95% CI: 75.4–77.2).

232. Among lawyers, the average level of consensus was 77.5% for cases involving the interpretation of public legal documents, such as legislation and regulation (95% CI: 76.8–78.2), compared to 76.5% for cases involving the interpretation of private legal documents, such as contracts and wills (95% CI: 75.2–77.8).

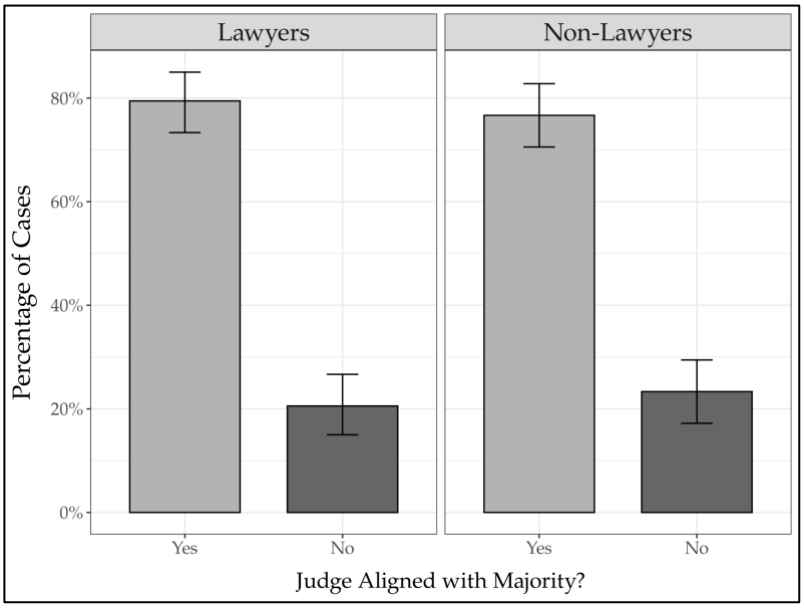

Figure 4: Judge alignment with lawyer and lay consensus. Bar heights represent percentage of cases in which the majority of participants chose ("Yes") or did not choose ("No") the same interpretation as the court. Error bars represent 95% bootstrapped confidence intervals.

a. Lay Alignment

With respect to lay participants, judges aligned with the majority of lay participants in 77.2% of cases (139 of 180). The pre-registered regression model further confirmed this to be significantly above chance (i.e., the rate at which one would expect if judges failed to align with consensus the majority of the time).[233]

Similar levels of alignment were observed when analyzing within each condition. In twelve out of the fourteen canon conditions, and in both non-canon conditions, a majority of participants aligned with the interpretation of the judge. The condition in which the judge had the highest rate of alignment with lay consensus was "General terms must operate on something" condition (*ejusdem generis* parry), with a mean alignment of 86.7% (95% CI: 85.1–88.2), followed by may versus shall ("There is a distinction between words of permission and mandatory words") (82.2%; 95% CI: 80.5–83.8) and *expressio unius* (82%; 95% CI: 80.4–83.7).

233. 95% CI of the intercept (after converting from log-odds to probability): 0.681–0.746.

The only two conditions in which the court failed to align, on average, with the majority of participants were *ejusdem generis* and the rule of the last antecedent.[234]

b. Lawyer Alignment

Judge alignment with lawyer participants was descriptively higher than for lay participants. In particular, the court aligned with the majority of lawyers in 79.4% (143 of 180) of cases, and in 86% of cases when looking only at cases where a clear majority of participants settled on one interpretation over another.

Regressions revealed this to be significantly above chance.[235] Consistent patterns were observed when analyzing within each condition. In thirteen out of the fourteen canon conditions, and in both non-canon conditions, a majority of participants, on average, aligned with the interpretation of the judge. The only condition in which the court failed to align with the majority of participants was *ejusdem generis*.[236]

The condition in which the court had the highest rate of alignment with lawyer consensus was may versus shall thrust ("There is a distinction between words of permission and mandatory words"), with a mean alignment of 94.0%,[237] followed by *expressio unius*, with a mean alignment of 84.0%,[238] and *ejusdem generis* parry condition ("General words must operate on something"), with a mean alignment of 80.4%).[239]

234. The mean alignment with consensus in the *ejusdem generis* thrust condition was 38.7% (95% CI: 36.7–40.8). The mean alignment with consensus in the rule against last antecedent thrust condition was 43.3% (95% CI: 41.2–45.6).

235. 95% CI of the intercept (after converting from log-odds to probability): 0.721–0.787.

236. The mean alignment with consensus in the *ejusdem generis* thrust condition was 44.8% (95% CI: 42.0–48.1).

237. 95% CI: 92.5–95.5.

238. 95% CI: 81.8–86.2.

239. 95% CI: 77.8–82.9.

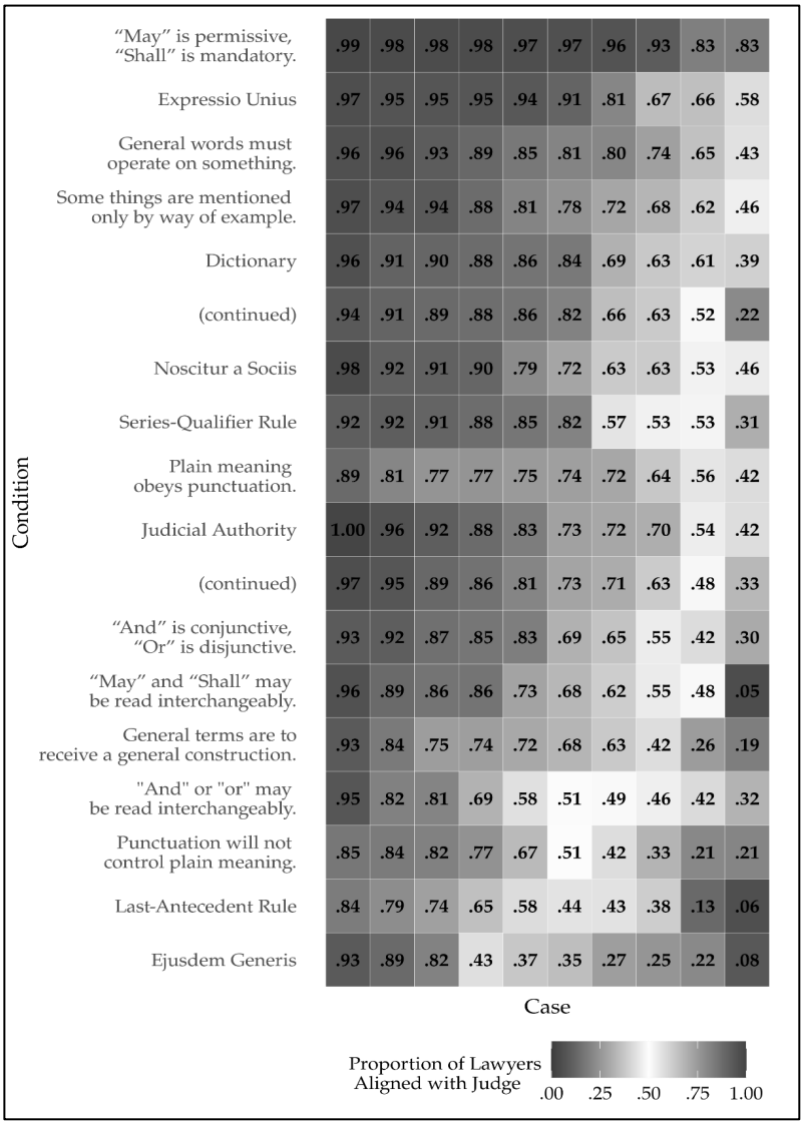

Figure 5: Judge alignment with lawyer consensus, broken down by case and interpretive tool. Each box represents the proportion of lawyers who agreed with the court's interpretation in a given case. Each row represents the cases associated with a given tool.

c. Participant Predictions

Descriptively, courts' alignment with lawyer and lay consensus matched or exceeded participants' own predictions of consensus in the

experiment. For example, the percentage of trials in which lay participants accurately predicted whether the majority of other lay participants chose their interpretation for a given case (and looking only at cases where there was a clear consensus) was 66.7% (95% CI: 66.2–67.2), while the accuracy of lawyers predicting lay consensus was 56.0% (95% CI: 55.3–56.7).[240] Both of these numbers were lower than the aforementioned percentage of cases in which judges aligned with lay consensus using the same exclusion criteria.

In terms of lawyer consensus, lawyers accurately predicted whether the majority of other lawyers would choose their interpretation for a given case (again excluding cases where there was a clear consensus of one interpretation over another) in 77.6% of trials (95% CI: 76.9–78.3). This, too, was lower than the percentage of such cases where judges aligned with lawyer consensus (86%).

These results were robust to several control analyses, such as when looking at all cases as opposed to only those with clear consensus,[241] as well as when including all participants (e.g., those who failed attention checks) as opposed to only those who were retained in the main analysis. The latter results are reported in Appendix 2.E.

d. Robustness Checks and Exploratory Analyses

Control analyses reveal that the main alignment results were robust to demographic factors such as age, politics, gender, and race. As noted in the previous Subsection, participants tended to converge on the same interpretation regardless of which demographic subgroup they pertained to. Similarly, judges tended to align with the consensus interpretation in the majority of cases, regardless of demographic subgroup (and in the case of lawyers, even when controlling for potential familiarity with cases).[242] In addition, analyses revealed similar levels of alignment with consensus across court level, jurisdiction, and legal text. These results are visualized in Figures 6–8.

To further address concerns about sample composition and the possibility that participants were responding based on their policy preferences, a replication study (a) added attention checks requiring

240. This was computed as follows: If a participant's interpretation was the same as the consensus interpretation, a participant was determined to have accurately predicted consensus if their estimated percentage of others who chose their interpretation was above 50%. Conversely, if a participant's interpretation was different from the consensus interpretation, they were deemed to have accurately predicted consensus if their estimated percentage of others who chose their interpretation was less than 50%.

241. For example, the unfiltered accuracy of lay participants predicting lay consensus was 63.4% (95% CI: 63.0–63.9). The accuracy of lawyer participants predicting lay consensus was 56.0% (95% CI: 55.2–56.7). The accuracy of lawyers predicting lawyer consensus was 73.0% (95% CI: 72.3–73.6).

242. These results are reported in full in Appendix 2.F.

participants to interpret provisions that plausibly ran counter to their preferences and (b) stratified responses by education. Results of this study replicated the main analysis and are reported in full in Appendix 2.G.

Furthermore, an additional LLM-prompting experiment was conducted to test concerns that judges might be "stacking the deck" by selectively invoking plain meaning only when it aligns with their policy preferences, or that the experimental materials disproportionately comprised easy, non-politically charged cases in which judges are less likely to hold strong policy views. The analyses found (a) no significant difference in the political valence of outcomes between judges of different ideological leanings; (b) no significant positive correlation between the political valence of outcomes and the political ideological commitments of the judges; and (c) no significant differences in the level of political chargedness or outcome valence between the experimental materials and a random control sample of 2,000 cases. These analyses are reported in full in Appendix 2.H.

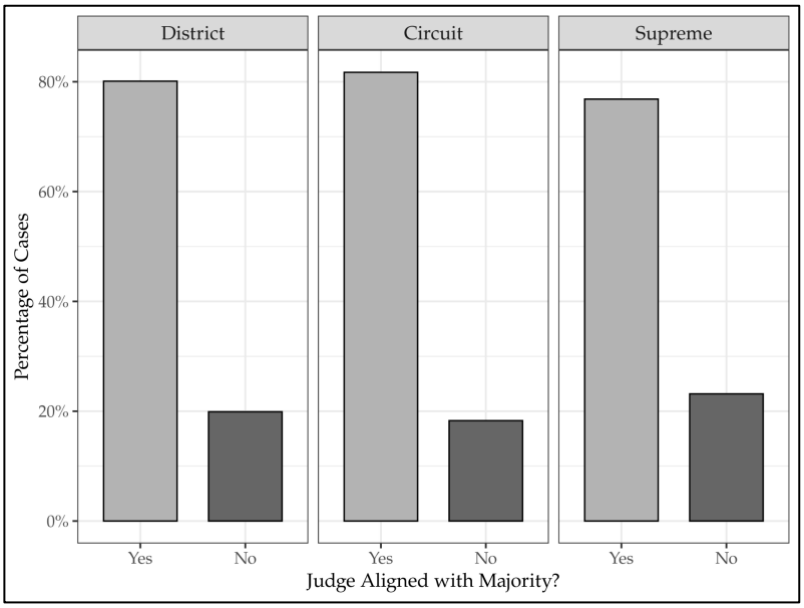

Figure 6: Percentage of lawyer participants aligned with the court, broken down by court level. Bar heights represent the raw percentage of cases.

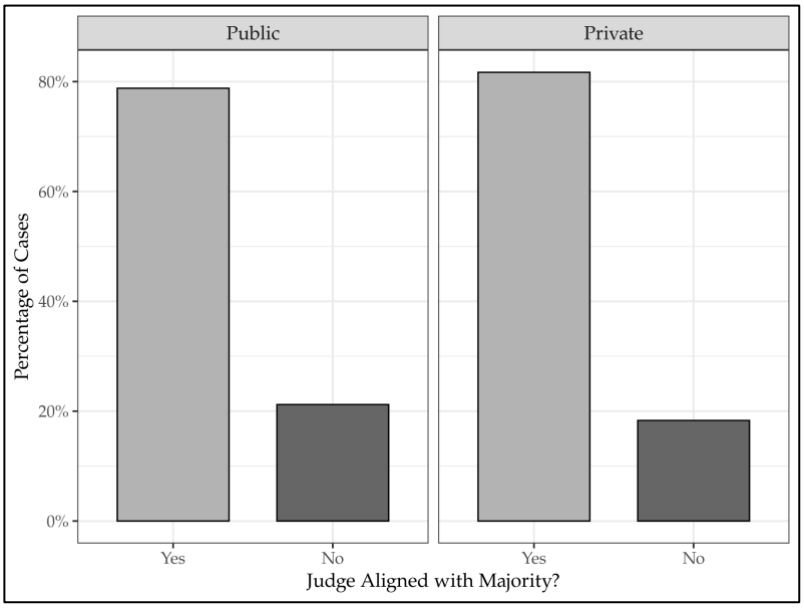

Figure 7: Percentage of cases in which majority of lawyers aligned with the court, broken down by legal text genre.

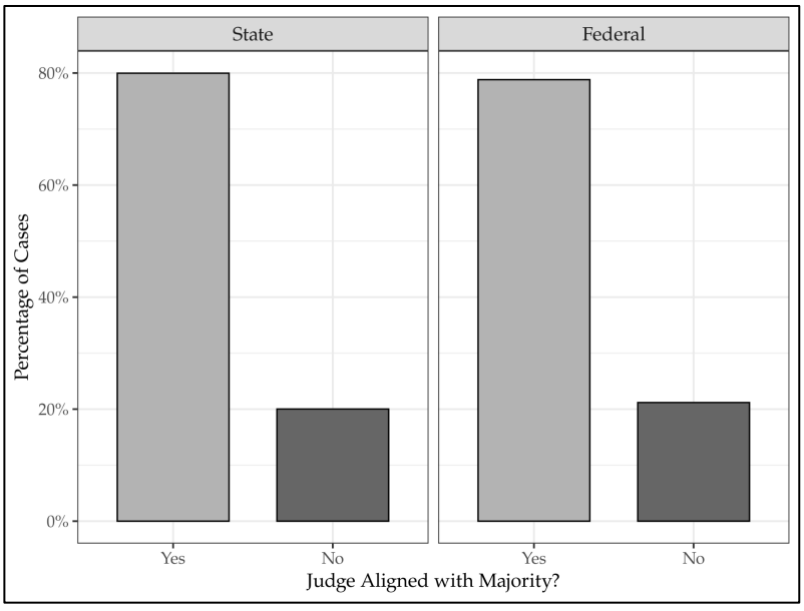

Figure 8: Percentage of cases in which majority of lawyers aligned with the court, broken down by jurisdiction (state/federal).

IV. STUDY 2: COMPUTATIONAL TOOLS

In addition to the question of whether judges successfully arrive at plain meaning when invoking traditional interpretive tools, a related question concerns whether judges would be better served by relying on novel computational tools, such as LLMs, to uncover plain meaning. This Part presents the results of an empirical study aimed at investigating this question.

A. Questions & Hypotheses

As discussed in Part II, academics and judges have recently proposed relying on AI tools as an aid to uncovering the plain meaning of a word at issue in legal disputes.

If novel computational tools were capable of supplementing judicial use of traditional interpretive tools, one would expect that AI tools would align at least as strongly with linguistic consensus as judges using traditional interpretive tools such as canons and dictionaries.

B. Methods

1. Materials

To evaluate this hypothesis, a prompting experiment was designed using the same primary materials as the behavioral experiment, adapted and formatted for the purpose of LLM prompting. Three classes of prompts were used, each designed to imitate one of the questions or dependent variables used in the human experiments. The first prompt class was an interpretation prompt, which asked the LLM to provide their interpretation of the legal text (yes/no) as applied to the facts of a case. The second prompt class was a certainty prompt, which asked the LLM to provide a probability estimate of a given interpretation being the correct one. The third prompt class was a prediction prompt, which asked the LLM to provide a prediction of the percentage of others (either lawyers or non-lawyers) who would endorse a given interpretation.

2. Models and Procedure

As described in Section II.C.1, *supra*, over the past few years, LLMs, such as OpenAI's GPT models, have achieved human-level performance on a number of domain-general natural language

processing tasks[243] and domain-specific language-based benchmarks, including those in law.[244]

Here we tested a number of flagship models from leading AI labs. In particular, initial analyses were conducted using seven flagship models from OpenAI (four LLMs, and three "advanced reasoning models"), including GPT-3.5, GPT-4, GPT-4o, GPT-4.1, o3, o1-mini, and o1.[245] Later exploratory analyses were conducted with flagship models from Anthropic, Google, and Deepseek. All models were prompted via the AI lab's respective API.[246]

When prompting models through the API, there is a temperature setting that controls how random the model's responses are. Lower temperatures (e.g., 0) make outputs more deterministic, while higher temperatures (e.g., 1–2) make them more varied. Given previous research showing a statistically null effect of temperature on performance in legal tasks,[247] temperature was kept to a default value of 1 for all non-reasoning models (reasoning models such as o3, o1, and o1-mini do not have a temperature adjustment option anyway and are always kept at one).

To capture the variability in response of models across trials, for all pre-o1 OpenAI models, each trial included an extraction not only of the AI-model's response but also of the log probabilities assigned to the top twenty most likely tokens.

In other words, with just one trial for each model/item/prompt class combination this allowed for the extraction of the probability that the model would answer with a given response (e.g. 75% of the time "yes") if prompted an indefinite number of times.

For reasoning models such as o3, o1, and o1-mini, log probability extraction is not permitted. Instead, following similar practices in previous work,[248] these models were prompted three times on each prompt.

3. Analysis Plan

Analyses evaluated alignment with linguistic consensus using descriptive statistics and parameter estimates for each model; the

---

243. *See, e.g.*, Kocoń et al., *supra* note 153, at 9 (finding ChatGPT and GPT-4 to achieve near, though consistently below, state-of-the-art performance on classical NLP tasks).

244. *See* OpenAI, *supra* note 22, at 5.

245. *See Models*, OPENAI, https://platform.openai.com/docs/models [https://perma.cc/F3PF-JVNF].

246. *See Introduction*, OPENAI, https://platform.openai.com/docs/api-reference/introduction [https://perma.cc/G5NR-88F9] (summarizing how prompting through an AI lab's API works).

247. *See* Martínez, *supra* note 25, at 594.

248. *See*, *e.g.*, Katz et al., *supra* note 23, at 3 (describing methodology for experiment assessing GPT models' performance on the Multistate Bar Exam, which involves "three runs for each set of prompts and parameters").

outcome variable was agreed_with_LLM (coded one when a participant's answer matched the LLM and zero otherwise), computed probabilistically from the LLM's response distribution. Mixed-effects logistic regressions were then conducted on combined human and LLM data (using the best-performing, pre-o1 model), with agreed_with_LLM_or_judge as the outcome, adjudicator and condition as fixed effects, and participant and item as random intercepts. Parallel analyses covered certainty and prediction, and control analyses assessed robustness to demographics and potential data contamination. Additional mixed-effects logistic regressions compared computational tools with neutral lawyers and laypeople on the 0–100 prediction outcome. In the LLM condition, prediction values were generated to match human sample sizes from log-probabilities, with adjudicator (contrast-coded) and canon (contrast-coded) as fixed effects and participant and item as random intercepts. Confirmatory criteria were no negative main effect for the LLM (or a positive LLM effect), with full estimates reported in the main text and additional details in Appendix 3.A.

C. Results

1. Comparison of AI Models

Performance of different AI models as aligned with lawyer consensus is visualized in Figure 9. Overall, all models aligned with lawyer and lay participants in the majority of cases and at a similar rate with each other.

The best-performing model in terms of aligning with laypeople data was GPT-4, which aligned with lay consensus in 140 out of 180 (77.8%) cases. When including only those cases in which a clear majority of lay participants endorsed one interpretation over another, GPT-4 aligned with lay consensus in 131 of 157 (83.4%) cases. The next best performing model was o1-mini,[249] followed by GPT-4o,[250] GPT-4.1,[251] o3,[252] o1,[253] and GPT-3.5.[254]

---

249. GPT-o1-mini aligned on average with 67.6% of lay participants across all cases (95% CI: 67.2–68.2).

250. GPT-4o aligned on average with 66.6% of lay participants across all cases (95% CI: 66.1–67.1).

251. GPT-4.1 aligned on average with 66.6% of lay participants across all cases (95% CI: 66.1–67.1).

252. o3 aligned on average with 66.0% of lay participants across all cases (95% CI: 65.6–66.5).

253. o1 aligned on average with 65.6% of lay participants across all cases (95% CI: 65.2–66.1).

254. GPT-3.5 aligned on average with 64.6% of lay participants across all cases (95% CI: 64.1–65.0).

In terms of aligning with lawyer consensus, the best-performing model was o3, which aligned with lawyer participants in 140 out of 180 (77.8%) cases. The next best performing model in aligning with lawyer consensus was o1-mini,[255] followed by o1,[256] GPT-4,[257] GPT-4.1,[258] GPT-4o,[259] and GPT-3.5.[260]

Similar levels of alignment were observed when analyzing certainty levels as opposed to yes/no predictions. These results are reported in Appendix 3.B.

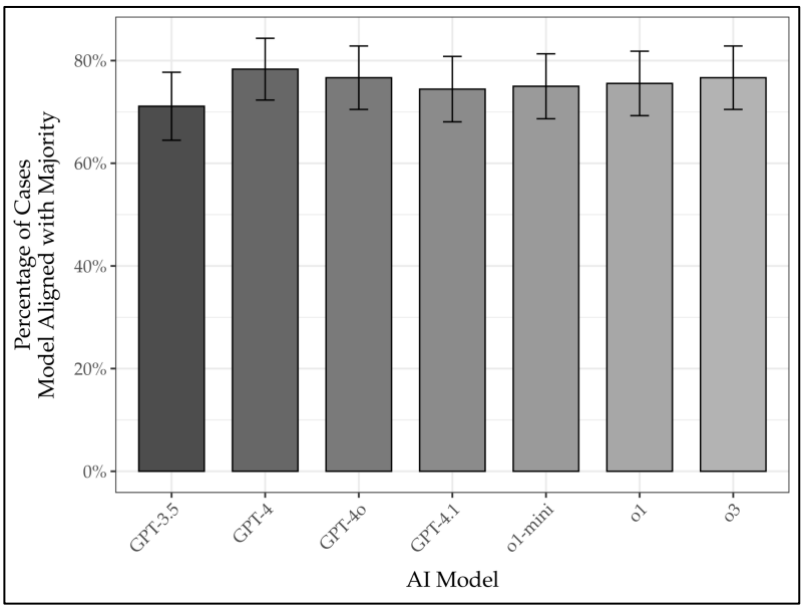

Figure 9: Alignment of AI models with lawyer consensus. Bar heights represent the percentage of cases in which the majority of lawyers

255. o1-mini aligned on average with 71.0% of lawyers across all cases (95% CI: 70.4–71.6).

256. o1 aligned on average with 70.6% of lawyers across all cases (95% CI: 69.9–71.3).

257. GPT-4 aligned on average with 70.4% of lawyers across all cases (95% CI: 69.7–71.1).

258. GPT-4.1 aligned on average with 67.4% of lawyers across all cases (95% CI: 66.6–68.0).

259. GPT-4o aligned on average with 67.2% of lawyers across all cases (95% CI: 66.4–67.9).

260. GPT-3.5 aligned on average with 63.5% of lawyers across all cases (95% CI: 62.8–64.2).

chose the same interpretation as a given AI model. Error bars reflect 95% bootstrapped confidence intervals.

2. AI Versus Judge Alignment

Performance of the best-performing non-reasoning model (GPT-4) in comparison to judges is visualized in Figure 10. In terms of aligning with lay consensus, GPT-4 descriptively aligned with lay participants at a slightly higher rate than did the court across all cases. In terms of alignment with lawyer consensus, GPT-4 descriptively aligned with lawyer participants at a slightly lower rate. According to the regression model, there was no significant difference overall in alignment between AI and the court.[261]

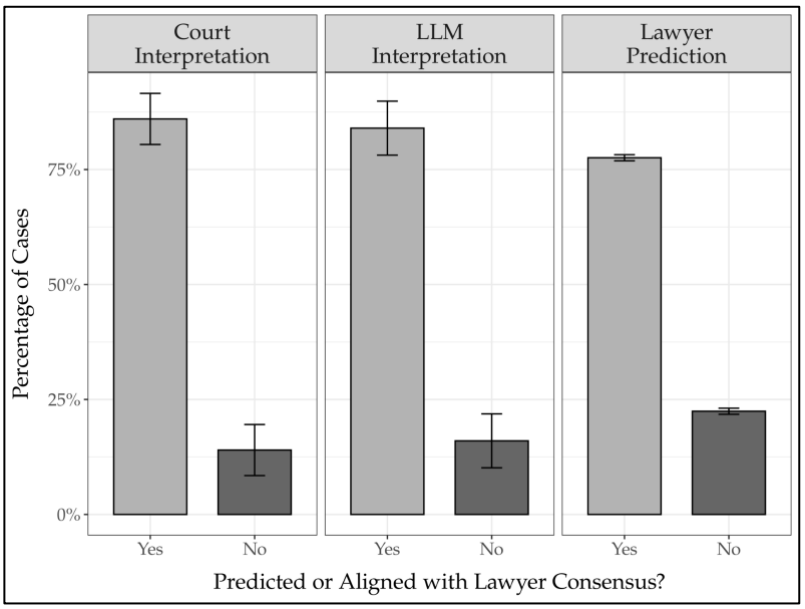

Figure 10: Mean alignment/prediction accuracy of court, LLM and lawyer participant with lawyer consensus. Bar heights, in the case of court and LLM, represent the percentage of cases in which the majority of lawyers chose the same interpretation as the judge or LLM, respectively. "Lawyer Prediction" panel represents the percentage of trials in which the lawyer participant accurately predicted which interpretation would be adopted by the majority of

261. p = 0.112.

other lawyers. Error bars reflect 95% bootstrapped confidence intervals.

3. AI Versus Participant Predictions

In terms of predicting lay consensus, the best-performing AI model (GPT-4) in the prediction prompt descriptively predicted the consensus interpretation among laypeople 73.7% of the time across all trials (95% CI: 73.2–74.1). When filtering out cases without a clear consensus interpretation, this prediction accuracy rose to 79.0% (95% CI: 78.5–79.4). Both were significantly higher than lawyer and lay predictions of lay consensus.[262]

In terms of predicting lawyer consensus, the best-performing AI model (GPT-4o) descriptively predicted lawyer consensus at a rate of 70.4% overall (95% CI: 69.7–71.0) and 75.5% (95% CI: 74.8–76.3) after filtering out cases without a clear consensus. Both of these were significantly lower than the accuracy of lawyer consensus predictions made by lawyer participants.[263]

More detailed results regarding prediction accuracy, such as the absolute error rate of different groups as well as the varying degree of accuracy on different materials, are reported in the Appendix 3.E.

4. AI Versus AI

The main analyses were conducted with OpenAI models. What about output from other AI models, such as those of Anthropic, Google, or DeepSeek? Comparisons between these models and those of the primary analyses reveal a high correlation in their linguistic judgments for a given prompt and similarly high alignment with the human data. These results are reported in Appendix 3.C.

5. Robustness Checks

The main results remained consistent when conducting various control analyses. For example, regression models revealed LLMs to align with both lawyer[264] and lay[265] consensus even when controlling for demographic variables such as age group, gender, race, politics, and (in the case of lawyers) potential familiarity of the case. In addition, as with the judge alignment data, analyses revealed similar levels of

262. Beta = -0.706, SE = 0.017, T = -41.5, $p < .0001$.
263. Beta = 0.203, SE = 0.015, T = 7.08, $p < .0001$.
264. 95% confidence interval of intercept: 0.684–0.760.
265. 95% confidence interval of intercept: 0.702–0.773.

alignment with consensus across court level, jurisdiction, and legal text.[266]

Finally, when assessing whether the results may have been driven by data contamination (i.e., the specific cases in the models' training data), analyses revealed (a) that the highest-performing flagship models (GPT-4 and GPT-4o) overwhelmingly failed to accurately guess the name of the case from which a particular item was derived; and (b) performed similarly well in terms of alignment regardless of whether the model was accurately able to guess the name of the case from which a particular item was derived.[267]

These results are reported in full in Appendix 3.D.

6. Temperature

In the main analyses, temperature values were kept to a default of 1, due to previous work showing a null effect of temperature on model performance.[268] In the case of the basic prompt (i.e. giving the model the same text as that given to humans and asking for its yes/no response), this resulted in models generally giving the same response at an extremely high rate (approximately 97% across all trials).

To test the robustness of the no-temperature effect, as well as to see if higher temperature levels might result in a more human-like distribution of responses, an additional prompting experiment was conducted which prompted a large language model 100 times on each of the 180 study materials at nine different temperature levels (0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, and 2) on the baseline (yes/no) prompt. These results revealed no significant difference in the model's performance to align with consensus in a given case. Moreover, even at the highest temperature setting, the distribution of model responses remained heavily skewed towards the model's modal interpretation, outputting the same response on average on over 95% of trials for a given study material.

Full results are reported in the Appendix 3.G.

7. Few-Shot Prompting

In the main design, all models were prompted using a "zero-shot" method — that is, the model was asked to perform the task without being given any examples to learn from as part of the prompt. Given

266. *See supra* Section III.C.3.d.

267. For example, the mean alignment of GPT-4 with lawyer consensus in cases where it did not accurately guess the case name was 71.0% (95% CI: 70.3–71.7). With respect to layperson consensus in such cases, the mean was 68.9% (95% CI: 68.4–69.4).

268. *See* Martínez, *supra* note 25, at 584 (finding "no significant effect of adjusting temperature settings" on LLMs' performance on bar exam questions).

past work demonstrating the efficacy of "few-shot" prompting in improving accuracy of legal tasks, [269] a follow-up prompting experiment was conducted, in addition to the main analysis, in which an LLM was given, as part of the system prompt, varying amounts of examples of a case, along with the "correct" answer (i.e., the percentage of the relevant readership who responded "yes" to a given question).

The number of examples varied between two and eighteen. All example sets were drawn from the same thrust/parry combination as that of the question that the model was prompted on in a given trial (one-half from the parry, one-half from the thrust).

Results are reported in Appendix 3.I. Descriptively, prediction accuracy in all of the example conditions were higher than in the baseline conditions, both when predicting lay responses and those of lawyers. In both the baseline and example conditions, error rates were lower in the LLM's predictions of laypeople consensus relative to lawyer consensus.

V. GENERAL DISCUSSION

This Part discusses the implications and potential limitations of the empirical findings. Section V.A presents seven sets of implications on the basis of the empirical results: (1) The behavioral data provide novel evidence regarding the determinacy of language in cases in which judges purport to uncover the ordinary meaning of a legal text; (2) The studies provide crucial data to evaluate the long-standing question of whether judges invoke linguistic canons to follow or fabricate linguistic consensus; (3) The results help clarify which linguistic canons best reflect linguistic consensus in real-world cases; (4) The findings clarify the practical differences between different versions of textualism that appeal to different hypothetical readers; (5) The results inform the extent to which traditional interpretive tools may be complemented or supplanted by computational tools, such as LLM prompting; (6) The analyses provide concrete lessons for judicial usage of computational tools for interpretive disputes; (7) The studies indicate that computational techniques can be used to apply and illuminate extant maxims, as well as to discover previously unrecognized ones.

A. Are Canons Indeterminate?

One longstanding debate concerns the level of linguistic determinacy in legal texts, particularly in cases where courts assert the meaning of a text to be "plain" or "clear." By measuring the degree of consensus in real-world, plain-meaning cases as judged by ordinary and

269. *See* Martínez, *supra* note 25, at 15–16.

well-informed readers, this study's findings provide novel and crucial evidence to resolve that debate.

In particular, the results of the experimental studies in Part III indicate that when judges invoke the plain meaning doctrine to conclude that the meaning of a text would be clear to an ordinary or well-informed reader, the meaning of that text is in fact generally clear to a substantial majority (75–80%) of ordinary and well-informed readers.

Across 180 sets of materials adapted from a diverse sample of real-world, plain-meaning cases, a supermajority of both lawyer and nonlawyer participants tasked with interpreting the text tended to converge on one interpretation of the text over another interpretation with a high degree of certainty. These results were robust to a host of demographic variables such as age, gender, race, and politics, as well as (in the case of lawyers) potential familiarity with the case. Moreover, this tendency was not only true of federal appellate or United States Supreme Court opinions but also generalized across all levels of the federal and state judiciary and across legal text genres (such as public versus private legal documents). This robust finding of linguistic consensus in plain meaning informs several debates in the previous literature.

For scholars who accept the strong *canon indeterminacy hypothesis* (i.e., that virtually every litigated dispute is a 50/50 coin flip between two equal and opposite dueling canons) the results require substantial revision. In the median dispute in which the judge invokes plain meaning, there is typically one interpretation — and one canon underlying that interpretation — that is significantly more reflective of lay and expert language understanding than that of its corresponding counter-canon.

At the same time, the findings undercut the mirror-image confidence sometimes voiced by canon defenders. For proponents of what might be called the strong *canon determinacy hypothesis* — captured in pronouncements that "statutes . . . must have a single, best meaning"[270] — in which canons almost always point unerringly to a single, objectively clear answer, the evidence counsels a commensurate downward update toward more modest confidence.

Many scholars and judges occupy more nuanced ground, believing, for example, that canons are helpful but defeasible heuristics, or that there exists "determinacy" for many, though not all, legal questions.[271] For this center-mass audience, the study supplies not a reversal but a calibration: In ordinary-meaning disputes that reach the courts, the

270. Loper Bright Enters. v. Raimondo, 603 U.S. 369, 400 (2024).

271. *See, e.g.*, Kent Greenawalt, *How Law Can Be Determinate*, 38 UCLA L. REV. 1, 29 (1990).

modal level of public agreement clusters in the mid-to-high 70% range, with some harder cases where consensus erodes and others where it approaches unanimity.

Several caveats remain. First, the project deliberately sampled opinions in which judges claimed textual clarity. Whereas, cases in which courts declared ambiguity might yield a very different consensus profile. Also, note that the results of this study do not necessarily indicate that there is consensus regarding the correct legal outcome in plain-meaning cases. It could be the case that ordinary or well-informed readers disagree on how the judge should decide, plain-meaning aside.[272]

Similarly, evidence of convergence on plain meaning is not necessarily to be construed as evidence of which interpretation a judge should choose to arrive at the correct outcome. As explained in the prior literature, the extent to which a judge relies on the "plain meaning" of a legal text is, at least in part, necessarily a normative judgment.[273]

As documented here and elsewhere, that normative position is widely held among those within the judiciary,[274] the academy[275] and the legal profession at-large.[276] The fact that readers tend to strongly converge on one interpretation over another across a wide variety of cases might be taken as evidence that such reliance is feasible, if not always desirable, in the contexts in which it is most frequently attempted.

B. Are Canons Window-Dressing?

In addition to the level of consensus in plain-meaning cases, a second longstanding debate concerns the extent to which judges successfully arrive at that consensus when invoking certain interpretive tools (such as canons or dictionaries), or instead whether these tools are a smokescreen to advance their policy preferences.

---

272. *See, e.g.*, Bill Watson, *Explaining Legal Agreement*, 14 JURISPRUDENCE 221, 222–23 (2023) ("If legal agreement . . . characterises the bulk of legal practice, then theories of law that are predominantly motivated by legal disagreement start from a strange place."); Brian Leiter, *Explaining Theoretical Disagreement*, 76 U. CHI. L. REV. 1215, 1247 (2009) ("[T]he most striking feature about legal systems is . . . massive agreement about what the law is."); Dennis Patterson, *Theoretical Disagreement, Legal Positivism, and Interpretation*, 31 RATIO JURIS 260, 272–73 (2018) ("Agreement in judgment is both the central fact of legal practice and the most important social fact that any account of law must explain.").

273. *See, e.g.*, Tara Grove, *Testing Textualism's "Ordinary Meaning"*, 90 GEO. WASH. L. REV. 1053, 1073 (2022) (arguing that some debates over interpretive method, even within the confines of textualism, "depend largely on normative considerations, not an empirical investigation").

274. *See* Peters, *supra* note 40, at 1243–44.

275. *See* Martínez & Tobia, *supra* note 208, at 152.

276. *See infra* Appendix 3.A.

Despite the longstanding and contested nature of this debate,[277] prior to this study there existed no formal empirical evaluation of this question. By uncovering the degree of alignment between judges and both ordinary and well-informed readers in a broad sample of such cases, this study thus provides novel and important evidence relevant to evaluating the window-dressing model.

In particular, the fact that judges consistently aligned with both lawyer and lay participants in a supermajority of cases — a finding that was robust to various control analyses — strongly suggests that judges, as a general matter, do not use these tools merely as a smokescreen. Instead, these results indicate that judges invoking canons, dictionaries, and other interpretive tools generally do so to support a sincere (albeit error-prone) attempt to find the meaning that an ordinary or well-informed reader would infer.

Moreover, these results were remarkably stable, not only across participant groups and demographic subgroups, but also across categories such as court level, jurisdiction (state versus federal) and legal text (private versus public), suggesting that judges' use of canons, dictionaries, and judicial authority in this manner may be largely consistent across the judiciary.

Finally, and perhaps most surprisingly, analyses revealed that courts' alignment with consensus was substantially above chance and matched or exceeded participants' own predictions of consensus. In this study, participants were incentivized to provide high-quality, good-faith predictions and were not responding based on their policy preferences. Consequently, the fact that they had similar, if not higher, error rates as courts in their predictions of consensus suggests that courts' error rates can be more easily attributed to good-faith erroneous judgments as opposed to using these tools as a partial smokescreen.

This inference is further supported by various robustness checks finding no evidence of an effect of political affiliation on political valence of case outcome, nor of courts being less likely to align with ordinary meaning in politically charged cases.

That said, it is worth noting that just because courts tend to align with plain meaning when purporting to align with plain meaning does not necessarily imply they arrive at the correct legal outcome. As noted in the previous Subsection, the extent to which a judge relies on the "plain meaning" of a legal text, as well as how exactly plain meaning is operationalized, is, at least in part, necessarily a normative judgment.[278] Insofar as one subscribes to the position that courts should interpret legal text according to its plain meaning as judged by a majority or supermajority of ordinary or expert readers, one might take

277. *See* Doerfler, *supra* note 17, at 267–68.
278. *See* Grove, *supra* note 273, at 1073.

these results as showing that judges in plain-meaning cases tend to practice not only what they preach, but what is doctrinally correct.

In some jurisdictions, consensus among legal officials around plain meaning may be so strong as to be considered a matter of legal doctrine. For example, the Supreme Court of Texas has not taken factors such as "legislative history," and "the consequences of a particular construction," into account even if it may be permitted by the legislature to do so.[279] Instead, Texas courts begin with a word's "plain and common meaning"[280] and go no further if the text is unambiguous. In such jurisdictions with a more universally adopted adherence to plain meaning among legal officials, one might take these results as stronger evidence that courts in plain-meaning cases are often arriving at the doctrinally correct outcome.[281]

At the same time, of course, these results have also revealed that judges do not always align with linguistic consensus. Thus, by the same token, these results suggest, even from a hard-core textualist perspective, that courts in plain-meaning cases are sometimes failing to arrive at the doctrinally correct outcome.

Indeed, some might interpret the error rate revealed in this study as excessively high and insist that judges stop relying on their own interpretive toolset and instead outsource plain-meaning judgments to a panel of human interpreters. This possibility was in fact raised during oral argument of *Facebook, Inc. v. Duguid*, as Chief Justice Roberts suggested that "[if] our objective is to settle upon the most natural meaning of the statutory language to an ordinary speaker of English . . . the most probably useful way of settling all these questions would be to take a poll of 100 ordinary . . . speakers of English and ask them what [the statute] means . . . ."[282]

Similar suggestions have been made in the context of contract interpretation. Omri Ben-Shahar and Lior Strahilevitz proposed and tested a method of resolving interpretation-related contract disputes through surveys of representative respondents.[283] Here, by evaluating

---

279. *See* TEX. GOV'T CODE ANN. § 311.023 (1985) (listing factors that courts are permitted to consider when construing a statute); Jaster v. Comet II Constr. Inc., 438 S.W.3d 556, 563 (Tex. 2014) (omitting factors listed from their consideration in interpreting a statute).

280. *See, e.g.*, City of Houston v. Bates, 406 S.W.3d 539, 544 (Tex. 2013) (applying the plain meaning of the statute's language).

281. *Cf.* William Baude, *Is Originalism Our Law?*, 115 COLUM. L. REV. 2349, 2365 (2015) (arguing that whether our law is originalist is an "empirical question" about consensus among legal officials and practice).

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

283. *See* Omri Ben-Shahar & Lior Jacob Strahilevitz, *Interpreting Contracts via Surveys and Experiments*, 92 N.Y.U. L. REV. 1753, 1782–84 (2017); *see also* Brandon Waldon, Madigan Brodsky, Megan Ma & Judith Degen, *Predicting Consensus in Legal Document Interpretation*, 45 PROC. ANN. MEETING OF THE COGNITIVE SCI. SOCIETY 1101, 1101–02

the extent to which courts align with expert and representative respondents in their interpretation of legal text, this Article's results help clarify the cost of deviating from this strategy.

It is important to note that whereas the focus of this study was on how judges might use canons and dictionaries to justify their preferred policy outcome, it remains an open question to what extent judges might attempt to influence policy outcomes through other methods, such as case selection. For example, commentators have long observed that the Supreme Court has nearly complete control over its docket.[284] For every case that it hears, a thousand are rejected, with virtually no required explanation or justification for why certain cases are not heard.[285]

In such an environment, it seems plausible that the Justices might disproportionately select cases where the plain language of the statute will lead to a preferred policy outcome. If so, the results of this study might be taken as evidence that scholars concerned with the political biases of the court should focus less on how that is manifested through factors such as the linguistic canon chosen and focus more on how that is manifested through factors that occur beforehand.

Beyond plain meaning, some might understand these results more broadly as informing debates regarding the trustworthiness of the content of judicial decisions in reflecting judicial behavior. Dating as far back as the legal realist movement, jurists have debated whether judges mean what they write in their decisions, or if the content of their decisions is merely a form of window-dressing that betrays the real form of reasoning driving the outcome.[286] The importance of this longstanding question has only grown in recent years, as empirical legal studies increasingly rely on the analysis of written opinions and other computational approaches as a way of understanding not just what judges write, but what judges mean when they write.[287]

(2023); Brandon Waldon, Cleo Condoravdi, James Pustejovsky, Nathan Schneider & Kevin Tobia, *Reading Law with Linguistics: The Statutory Interpretation of Artifact Nouns*, 62 HARV. J. LEGIS. 1, 8–9 (2025).

284. *See, e.g.*, Karen M. Tani, *The Supreme Court 2023 Term — Foreword: Curation, Narration, Erasure: Power and Possibility at the U.S. Supreme Court*, 138 HARV. L. REV. 1, 6–7 (2024); Brian Leiter, *Constitutional Law, Moral Judgment, and the Supreme Court as Super-Legislature*, 66 HASTINGS L. J. 1601, 1608 (2015).

285. *Id.* at 1608; Tani, *supra* note 284, at 6–7.

286. *See, e.g.*, Brian Leiter, *Legal Formalism and Legal Realism: What is the Issue?*, 16 LEGAL THEORY 111, 119–21 (2010) (describing the "distinctive realist thesis" that holds that "appellate judges are applying largely non-legal norms to recurring situation types while reciting general legal doctrines that are mere window dressing and obscure the normative considerations influencing their decisions").

287. *See, e.g.*, Jens Frankenreiter & Michael A. Livermore, *Computational Methods in Legal Analysis*, 16 ANN. REV. L. & SOCIAL SCI. 39, 40 (2020) (describing the rise of text-based approaches to empirical legal studies research); Edward Stiglitz & Rosamond Thalken, *Historical Trends in Macro-Jurisprudence: A Language Model Assessment, 1870–2023*, 84 MD. L. REV. 101, 104, 108–09 (2024).

Here, the alignment between what judges say in their opinions — that the plain meaning of the text supports interpretation X — and how ordinary readers in fact understand the text suggests that, in many instances, judges do mean what they say and are not simply invoking "plain meaning" as a rhetorical flourish. By extension, these results may offer some grounds for optimism that analyzing the content of written opinions as evidence of the mechanisms of judicial decision making is not a misguided enterprise.

C. Which Canons Best Track Plain Meaning?

In addition to the general question of whether judges tend to align with plain meaning when invoking interpretive tools, such as linguistic canons and dictionaries, a separate question concerns whether some tools are more reflective of plain meaning than others. By analyzing the extent to which judges tend to align with lay and lawyer consensus when invoking different tools, this study provides the first evidence of this question as applied to real-world legal disputes.

In particular, the empirical data presented here reveals that the vast majority of interpretive tools reflect linguistic consensus in the real-world cases in which they are invoked. In doing so, these results provide the first formal empirical evidence that linguistic canons themselves are in large part not "dueling" but rather complementary. That is, each of the linguistic canon pairs, long mocked as being mutually contradictory, reflect linguistic consensus in the contexts in which they are invoked, and judges by-and-large correctly discern this context when choosing which canon to invoke.

Similarly, despite the widespread criticism of dictionaries as being error-prone and unreliable indicators of plain meaning,[288] this study has demonstrated that in many real-world cases, dictionaries, specifically in the form in which they are invoked by judges, can and do accord with plain meaning as judged by both ordinary and well-informed readers.

That said, two canons in particular — *ejusdem generis* and rule of last antecedent — did not consistently reflect linguistic consensus in the study's materials. In the case of the rule of last antecedent, these results are consistent with recent work showing that lay participants tend to interpret modifying phrases in ordinary contexts more in line with the series qualifier rule as opposed to the last antecedent rule.[289] These results show that this tendency holds in both everyday contexts with lay participants and the real-world contexts in which the last antecedent rule is invoked. In doing so, the findings suggest that the

288. *See, e.g.*, Mouritsen, *supra* note 125, at 805, 867, 873; Kevin Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726, 751 (2020).

289. *See* Randall & Solan, *supra* note 120, at 19.

psycholinguistic grounds on which the last antecedent rule is said to be based — namely, that it can take "more than a little mental energy to process the individual entries in the list, making it a heavy lift to carry the [qualifier] across them all" — are in many relevant instances largely unfounded.[290]

That said, there were still multiple cases in which the court's invocation of the canon strongly aligned with both lawyer and lay consensus, consistent with Justice Alito's admonition that whether a modifier applies to an entire list or just the nearest antecedent may sometimes have "little to do with syntax[,] and everything to do with our common understanding" of what words likely mean in a given context, and that one "can see [the result change] if we retain the same syntax but replace [one of the list items] . . . with any number of other [words] that describe something" the modifier is unlikely to modify.[291]

In the case of *ejusdem generis*, the results are more surprising, given the results of (a) work by Gluck and Bressman, whose seminal article found that many congressional staffers explicitly or implicitly take into account *ejusdem generis* when drafting legislation;[292] and (b) recent work by Kevin Tobia, Victoria Nourse, and Brian Slocum showing that non-lawyers implicitly invoke *ejusdem generis* in hypothetical triggering conditions (interpreting "gin, bourbon, vodka, tequila, rum, and other beverages" as excluding beverages like orange juice).[293]

In contrast, the results of this study indicate that, although there are some legally relevant contexts in which *ejusdem generis* reflects language understanding, those contexts do not uniformly present themselves in the real-world cases that go to litigation and in which the canon is invoked. Judges may either overestimate the extent to which those contexts present themselves or simply "misapply" the canon (e.g. incorrectly inferring that a target word is *ejusdem generis* with the exemplars). Consequently, insofar as judges are of the view that they should rely on text and plain meaning as a factor in interpreting legal provisions, these results provide suggestive evidence that judges may be better served by reducing the frequency with which they rely on *ejusdem generis* and last antecedent rule.

Finally, note that these results have not yet pinpointed whether judges turn to canons and dictionaries ex ante — as an aid to forming their own sense of ordinary meaning — or ex post — as a way to

290. *See* Lockhart v. United States, 577 U.S. 347, 351 (2016) ("The rule reflects the basic intuition that when a modifier appears at the end of a list, it is easier to apply that modifier only to the item directly before it. That is particularly true where it takes more than a little mental energy to process the individual entries in the list, making it a heavy lift to carry the modifier across them all.").

291. Facebook, Inc. v. Duguid, 592 U.S. 395, 412 (2021) (Alito, J., concurring).

292. Gluck & Bressman, *supra* note 119, at 933, 960.

293. Tobia et al., *supra* note 19, at 259–60.

articulate and defend that sense. Consequently, our results should not be taken as a call for judges to renew their Merriam-Webster subscriptions or to keep Scalia and Garner's *Reading Law* within arm's reach of the bench.[294] Future work can clarify whether canons and dictionaries are shaping interpretation or primarily legitimating it. In the meantime, the fact that judges' invocations of most linguistic canons align closely with the plain-meaning judgments of both ordinary readers and lawyers suggests that the canons generally play a genuine linguistic role rather than serving solely as rhetorical camouflage for outcome-driven policy preferences.

D. Which Plain Meaning?

As discussed above, there exists some disagreement among judges and academics regarding the relevant readership in a plain meaning inquiry — that is, to whom must a text be plain in order for the doctrine to apply? Whereas some have explicitly asserted the relevant readership to be that of an ordinary speaker of English,[295] others have posited that it is or should be a reasonable or well-informed reader.[296] By operationalizing plain meaning in terms of both ordinary readers (laypeople) and well-informed readers (lawyers) and comparing the difference, this Article provides novel and crucial insight into both (a) the practical distinctions between different versions of plain meaning, as well as (b) which of the two versions offers a better explanation of judicial behavior.

With regard to (a), recall that in the vast majority of cases, lawyers and non-lawyers arrived at the same interpretation with a high degree of certainty. The same was true when comparing responses of readers from different demographic subgroups, such as age, race, gender and politics. The same holds true when interpreting cases of different categories, from different jurisdictions and levels of the judicial hierarchy. These findings indicate that, in many cases, the question of relevant readership becomes moot — that is, in plain-meaning cases the meaning is plain to an overwhelming majority of readers, regardless of whether they are "ordinary" or "well-informed." By extension, a judge invoking the ordinary-reader version of plain meaning or the well-informed-reader version of plain meaning will arrive at the same outcome.

The lack of practical distinctions between different versions of the plain-meaning doctrine provides a point of optimism to one common objection to text-based approaches to interpretation. Recent critics of new textualism have pointed out that if different groups of relevant

294. *See generally* Scalia & Garner, *supra* note 18.
295. *See, e.g.*, Barrett, *supra* note 42, at 2194.
296. *See, e.g.*, Scalia & Garner, *supra* note 18, at 47.

readerships obtain different meanings from the text, it becomes difficult to proceed without engaging in some additional normative analysis to determine which readership to prioritize over another.[297] In showing the convergence between the interpretations gleaned by lawyers and non-lawyers, the results of this study illustrate that in many cases this critique, though well-reasoned in principle, may be moot in practice.

Moreover, as the first study to systematically compare the interpretations of lawyers and laypeople across a large number of real-world cases, these results also help inform the extent to which lay participants can serve as a proxy for lawyers in studies involving interpretation, as well as experimental jurisprudence studies writ-large. Historically, the vast majority of experimental jurisprudence work has featured solely on lay participants, whereas legal experts have featured far less prominently as subjects.[298] Leading scholars of experimental jurisprudence have claimed that this prioritization of lay subjects is for principled doctrinal reasons, not only cost and efficiency.[299] In contrast, some scholars have criticized this practice on the grounds that legal experts, not laypeople, are those whose judgments matter,[300] and have therefore questioned the usefulness of deriving doctrinal or jurisprudential implications on the basis of experiments featuring lay subjects (or even law students).[301]

Here, the fact that lawyers and laypeople are mostly aligned in their interpretations indicates that, in many cases, laypeople can serve as accurate proxies for lawyers. Of course, as highlighted earlier in the Article,[302] the cases here were chosen to be reasonably comprehensible to a general audience. In fact, participants in both the lawyer and nonlawyer conditions had to pass an attention check that required some degree of competence in interpreting complex legal constructions (and although the vast majority of participants in both groups passed, this number was higher in the lawyer group than the non-lawyer group). It is plausible that in more technical cases there will be a greater difference between lawyer and lay consensus than observed in this study. At the same time, the predictions of lay participants by lawyers in the study were systematically less accurate than lawyers' predictions

297. *See, e.g.*, Grove, *supra* note 273, at 1084–85; William N. Eskridge, Jr., Brian G. Slocum & Kevin Tobia, *Textualism's Defining Moment*, 123 COLUM. L. REV. 1611, 1638 (2023) (describing "the tension between textualism as democratic interpretation and the often-esoteric nature of statutory context").

298. *See supra* note 24.

299. *See, e.g.*, Tobia, *supra* note 288, at 741.

300. *See, e.g.*, Felipe Jiménez, *The Limits of Experimental Jurisprudence*, *in* THE CAMBRIDGE HANDBOOK OF EXPERIMENTAL JURISPRUDENCE 79, 83 (Kevin Tobia ed., 2025) ("[T]hose in the driver's seat regarding legal concepts are legal officials and participants, not laypeople.").

301. *See* Holger Spamann & Lars Klöhn, *Can Law Students Replace Judges in Experiments of Judicial Decision-Making?*, 1 J.L. & EMPIRICAL ANALYSIS 149, 158 (2024).

302. *See supra* Section III.B.1.b.

of other lawyers. In particular, the similarities between lawyer and lay participants were far greater than expected by lawyer participants, suggesting that, even considering the aforementioned caveats, the differences between those with and without legal training in making legal judgments may be far lower than commonly assumed by the legal profession.

This leads to point (b), which is that the results of participant accuracy in predicting lawyer and lay consensus reveal which version of textualism and plain meaning better maps on to how judges interpret cases.

Although judges' alignment with lawyer and lay participants was approximately equivalent (albeit a bit higher with lawyers), lawyers' predictions of other lawyer participants were much higher than their predictions of lay consensus. Given that judges are overwhelmingly drawn from the ranks of lawyers, it seems reasonable to infer that their predictions would be substantially similar to those of lawyers. If so, it seems most plausible that when judges are determining the plain meaning of a provision, they are likely to be determining that based on the meaning that a lawyer would understand. By extension, although lawyers and laypeople were aligned at a surprisingly high rate in this study (and perhaps elsewhere), to the extent that lawyers and non-lawyers diverge in their linguistic judgments, these results indicate that judges will likewise diverge from laypeople and will instead interpret a provision in line with how a lawyer would read it.

By further extension, insofar as one subscribes to a Holmesian "law as prediction" theory of law,[303] this would suggest that the "well-informed reader" version of plain meaning is a better account of legal doctrine than the "ordinary reader" version of legal doctrine.

E. Do LLMs Track Plain Meaning?

With the rise of LLMs purportedly equipped with legal and linguistic competence,[304] one open and hotly contested question concerns to what extent these novel computational tools can complement or supplant traditional interpretive tools in judicial decision-making. By comparing the abilities of LLMs to predict linguistic consensus in real-world cases as compared to judges, lawyers and non-lawyers, this Article is the first to formally evaluate this question.

---

303. Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 457 (1897), *reprinted in* DAVID KENNEDY & WILLIAM W. FISHER III, THE CANON OF AMERICAN LEGAL THOUGHT, 31 (2018) ("The prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law.").

304. *See generally* OpenAI, *supra* note 22.

Champions of computational aids — Judge Newsom's cautious "maybe," [305] Engel and McAdams's forecast that LLMs could "radically facilitate . . . the interpretation of statutes,"[306] and Arbel and Hoffman's portrayal of AI as the coming "workhorse of contractual interpretation"[307] — hypothesize that a model steeped in everyday usage should replicate the intuitions lay speakers bring to court. Across 180 real disputes distributed over sixteen doctrinally salient categories, that hypothesis holds: the best model converged with the majority judgments of both lawyers and lay respondents and matched judicial alignment with said consensus. To probe whether this agreement might simply reflect memorization, the study included a contamination check: after each prediction, the LLM was asked to identify the case it had just evaluated. In roughly 90–95% of trials, it could not state the correct case name, and excluding the handful of instances in which it guessed correctly left overall alignment statistics largely unchanged. A further indication that the results are not driven by contamination is that the model's alignment with consensus often diverged from that of the court in key categories — for example, in many rule-of-the-last-antecedent and ejusdem generis cases, the model sided with the modal reader where the judge did not. Taken together, these patterns provide a source of optimism for the view that the model is drawing on generalized semantic patterns rather than merely reproducing cached snippets from its training corpus.

Similarly, these results provide an optimistic empirical response to recent voices of skepticism. Jon Choi warns that "LLM assessments [can] substantially deviate from empirical predictions of language use";[308] Lee and Egbert caution that current AI models "are not up to the task" of ordinary meaning analysis;[309] and Waldon et al. argue that "[j]udges should not rely on direct queries to ChatGPT (or similar chatbots) about the meaning of legal texts."[310] The present results do not eliminate these concerns, but they do complicate them. Across model families, temperature settings, and demographic slices, accuracy hovered at a relatively high and stable rate. While these findings fall short of guaranteeing flawless performance in every future dispute, they indicate that LLMs can, under controlled conditions, often track the consensus, real-world meaning of contested legal terms at rates comparable to the courts. Claims that the technology is categorically incapable of uncovering ordinary usage therefore must be squared with evidence that, in a sizable sample of actual cases, the models frequently

305. *See* Snell, *supra* note 27, at 1225.
306. Engel & McAdams, *supra* note 26, at 296.
307. Arbel & Hoffman, *supra* note 26, at 451.
308. Choi, *supra* note 26, at 1.
309. Lee & Egbert, *supra* note 29, at 1.
310. Waldon et al., *supra* note 25, at 4.

recovered the meanings that human communities — and often the courts themselves — ultimately endorsed. On this view, treating such systems solely as "dialectical partners,"[311] as some have suggested, may be more cautious than strictly required by the present data.

That said, success at pinpointing ordinary meaning is only one piece of the adjudicative puzzle, and these results should not be construed as suggesting that LLMs can or should be used to fully automate the process of legal decision-making, or that they are particularly suitable for aspects of the decision-making process beyond uncovering the ordinary meaning of words in a legal provision. Recall, for example, that the focus of this study was on the interpretation of language. Over the past couple of years, evidence has accumulated suggesting that large language models are good models of human language, both in terms of performance on language-based tasks[312] as well as with regard to their internal representations.[313]

At the same time, evidence has also indicated that (a) language and complex reasoning are distinct cognitive processes,[314] and (b) AI models such as LLMs that are optimized for next-word prediction are not necessarily good models for complex reasoning.[315]

Conversely, break-through AI models that are designed explicitly for complex reasoning, such as OpenAI's o1 and o1-mini models are, according to their own developers, less suitable for most language tasks than general-purpose LLMs.[316]

To the extent that legal decision-making involves a combination of different cognitive processes at different stages of the decision-making process, this would indicate that certain AI models might be better suited for assisting judges with some aspects of the reasoning process, but not others. Indeed, recent neuroimaging work has shown that whereas some modes of legal reasoning (such as reading a contract) strongly engage brain regions associated with the processing language, others (such as evaluating the legal validity of a contract) engage brain

311. *Id*.

312. *See, e.g.*, Kocoń et al., *supra* note 153, at 18.

313. *See* Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf et al., *Driving and Suppressing the Human Language Network Using Large Language Models*, 8 NAT. HUM. BEHAV. 544, 544 (2024).

314. *See, e.g.*, Evelina Fedorenko & Rosemary Varley, *Language and Thought Are Not the Same Thing: Evidence from Neuroimaging and Neurological Patients*, 1369 ANNALS N.Y. ACAD. SCIS. 132, 132 (2016).

315. *See, e.g.*, Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum & Evelina Fedorenko, *Dissociating Language and Thought in Large Language Models*, 28 TRENDS COGNITIVE SCIS. 517, 529 (2024).

316. *Using OpenAI O1 Models and GPT-4o Models on ChatGPT*, OPENAI, https://help.openai.com/en/articles/9824965-using-openai-o1-models-and-gpt-4o-models-on-chatgpt [https://perma.cc/2H7P-9RBY] ("GPT-4o is still the best option for most prompts.").

regions associated with formal and moral reasoning.[317] Moreover, recent benchmarking work has shown that LLMs fail to align with human judges when asked to resolve a legal dispute from start-to-finish using similar materials.[318] Consequently, future benchmarking work evaluating the utility of AI models at different stages of the reasoning process will be critical to ensuring that AI is used in a manner that can replicate the decision-making process of a human judge.

F. (How) Should Judges Use LLMs for Plain Meaning?

In addition to establishing that LLMs can track ordinary meaning in litigated disputes, this study offers pragmatic lessons on how models can be deployed to get the most reliable answers. Below are five operational lessons relevant to judicial adoption.

First, variation in sampling temperature appears to be of negligible consequence. A systematic sweep across nine values, ranging from fully deterministic (0) to maximally stochastic (2), produced no discernible shift in the rate at which models aligned with lay consensus, and even on the highest setting, the model repeated the same modal answer on more than 95 % of trials for a given item. Courts therefore have little to gain from adjusting this parameter and may safely rely on the default. Thus, temperature scarcely matters.

Second, the choice among current flagship systems likewise exerts minimal influence on performance. Alignment rates proved statistically indistinguishable across seven OpenAI variants — including GPT-3.5, GPT-4.1, and the reasoning-oriented o-series — and remained comparable when benchmarked against Anthropic Claude-Sonnet 4.0, Google Gemini Flash Preview 2.5, and DeepSeek-V1. Notably, models explicitly optimized for chain-of-thought reasoning did not outperform their purely linguistic counterparts on this task, despite higher computational cost and latency. For purposes of inferring ordinary meaning, a high-capacity language model therefore appears to offer the most efficient return.

Third, prompt design, by contrast, makes a substantive difference. Repeated yes/no polling caused the model to misestimate the shape of public disagreement, whereas asking directly for probabilities ("What percentage of lay readers would say 'yes'?") cut distribution error roughly in half. Adding a small set of worked examples — just two to eighteen case snippets with the correct answers — reduced prediction

317. Eric Martínez, Jingyuan S. She, David Oluigbo, Edward Gibson, Evelina Fedorenko & Anna A. Ivanova, The Neural Mechanisms of Legal Reasoning (unpublished manuscript) (on file with author).

318. Eric E. Posner & Shivam Saran, Judge AI: Assessing Large Language Models in Judicial Decision-Making 1 (Jan. 17, 2025) (unpublished manuscript) (on file with the Coase-Sandor Institute for Law & Economics at the University of Chicago Law School).

error still further, with the steepest improvements appearing in forecasts of lay usage. Providing representative exemplars and eliciting calibrated probabilities therefore yields markedly more informative outputs than a string of binary votes.

Fourth, when appropriately prompted, LLMs also help curb human overconfidence. Lawyers over-estimated how many peers would share their view in about 70% of cases, while the model did so barely a third of the time. Equally important, the model's prediction error did not increase on the most ambiguous items, suggesting it can flag both the modal interpretation and the depth of consensus. This information can steer judges in deciding whether a statutory clarity threshold has been met.

Finally, the systems' comparative advantage appears to lie in predicting lay consensus, not lawyer consensus. LLMs consistently beat lawyers at forecasting lay responses but did worse at simulating lawyer consensus across the 180 question sample. This pattern suggests that LLMs are most advantageous when the interpretive inquiry focuses on ordinary meaning. Conversely, questions demanding alignment with specialized professional usage may continue to favor the expertise of human practitioners.

G. Computational Canons

Insofar as LLMs can emulate the interpretations of human readers, an additional potential use case of computational tools is not only to help judges determine the plain meaning in a given case alongside their use of canons, but more broadly to refine and discover new linguistic canons beyond those currently in existence. After all, linguistic canons are best conceived as probabilistic heuristics (i.e., rules of thumb that approximate how ordinary speakers encode meaning). If that is right, then the very computational techniques that succeed in modeling ordinary usage can both apply extant maxims and discover previously unrecognized ones. By mining statutory codes, contract databases, and entire runs of the Federal Reporter with LLMs, syntactic parsers, and other machine-learning pipelines, researchers can surface lexical or structural regularities that have not yet been formalized in caselaw. The resulting "emergent canons" would enlarge the textualist toolkit with empirically grounded principles rather than intuition-driven aphorisms.

At the same time, those same tools can shed new light on the venerable canons that already populate the lawyer's handbook. Classic maxims often function as black boxes: For virtually every interpretive "thrust" (for example, *expressio unius*), treatises catalogue a corresponding "parry" (such as exceptions to the negative-implication canon). Yet, judges rarely articulate why one rather than the other

governs the dispute at hand. Feature-attribution methods [319] and counterfactual probing [320] can reverse-engineer the cues — semantic relatedness,[321] syntactic parallelism,[322] word frequency[323] — that lead a model, and by extension a judge, to invoke *ejusdem generis* instead of *noscitur a sociis*. Making those cues explicit would transform the canons from opaque slogans into transparent, testable generalizations about language use.

A parallel evolution has already occurred in linguistics. For decades, the pragmatics literature relied on Grice's conversational maxims — "be relevant," "be informative," "avoid ambiguity" — rendered in fuzzy prose. [324] With the advent of computational approaches such as the rational-speech-act framework, [325] cognitive scientists have translated those dicta into Bayesian models that predict and explain scalar implicatures, reference choice, and a host of other pragmatic phenomena.[326] Empirical work shows that these formalized

---

319. *See, e.g.*, Mukund Sundararajan, Ankur Taly & Qiqi Yan, *Axiomatic Attribution for Deep Networks*, 34 PROC. INT'L CONF. ON MACH. LEARNING 1, 1 (2017); Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim et al., *Learning How to Explain Neural Networks*: *PatternNet and PatternAttribution*, 2017 PROC. INT'L CONF. ON LEARNING REPRESENTATIONS 1, 1 (2017); Luan Luo & Lucia Specia, *From Understanding to Utilization: A Survey on Explainability for Large Language Models* 1, 2 (2024).

320. *See, e.g.*, Anirudh Srinivasan, Venkata S. Govindarajan & Kyle Mahowald, *Counterfactually Probing Language Identity in Multilingual Models*, 1 PROC. MULTILINGUAL REPRESENTATION LEARNING 1, 1 (2023); Venkata S. Govindarajan, David Beaver, Kyle Mahowald, & Junyi Jessy Li, *Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias* 1–2, FINDINGS OF ASS'N FOR COMPUTATIONAL LINGUISTICS (2023).

321. *See, e.g.*, Samira Abnar & Willem Zuidema, *Quantifying Attention Flow in Transformers*, 58 PROC. OF THE ANN. MEETING OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS 4190 (2020).

322. *See, e.g.*, John Hewitt & Christopher D. Manning, *A Structural Probe for Finding Syntax in Word Representations*, 2019 PROC. CONF. N. AM. CH. ASS'N COMPUTATIONAL LINGUISTICS 4129, 4132 (2019).

323. *See, e.g.*, Kenneth Ward Church & Patrick Hanks, *Word Association Norms, Mutual Information, and Lexicography*, 16 COMPUTATIONAL LINGUISTICS 22, 22 (1990); Stefan Th. Gries & Anatol Stefanowitsch, *Extending Collostructional Analysis: A Corpus-Based Perspective on "Alternations"*, 9 INT'L J. CORPUS LINGUISTICS 97, 102 (2009).

324. H. P. Grice, *Logic and Conversation*, 3 SYNTAX & SEMANTICS: SPEECH ACTS 41, 45–51 (Peter Cole & Jerry L. Morgan eds., 1975); *see generally* STEPHEN C. LEVINSON, PRAGMATICS (Cambridge Univ. Press, 1983) (textbook that builds an entire pragmatics curriculum on Grice's principles); Christopher Potts, *Introduction to Pragmatics*, Ling 130a/230a: Introduction to Semantics and Pragmatics 1, 4 (Stanford Univ. lecture notes, 2022) ("Grice's maxims are the backbone of his pragmatic theory.").

325. Michael C. Frank & Noah D. Goodman, *Predicting Pragmatic Reasoning in Language Games*, 336 SCI. 998, 998 (2012); Noah D. Goodman & Michael C. Frank, *Pragmatic Language Interpretation as Probabilistic Inference*, 20 TRENDS IN COGNITIVE SCI. 818, 819 (2016).

326. *See, e.g.*, Noah D. Goodman & Andreas Stuhlmüller, *Knowledge and Implicature: Modeling Language Understanding as Social Cognition*, 5 TOPICS IN COGNITIVE SCI. 173, 173 (2013); Judith Degen & Noah D. Goodman, *Lost Your Marbles? Quantifying Scalar*

maxims both capture and clarify human linguistic behavior.[327] The same logic applies to statutory interpretation: Formal models can predict when a particular canon will be linguistically felicitous and, crucially, explain why it fits the context.

Such explanatory machinery promises additional institutional dividends to the extent that courts attempt to outsource ordinary-meaning judgments to computational agents. Transparent interfaces that expose the linguistic features and weights driving a model's recommendation allow judges to scrutinize, adopt, or reject those recommendations with an informed eye, preserving judicial independence rather than surrendering it to a black box. Litigants obtain a clearer record for appellate review; scholars gain data for doctrinal refinement; and the public acquires a basis for evaluating whether textualist methods are applied consistently. By coupling discovery pipelines for new heuristics with interpretably-rendered outputs for old ones, computational linguistics can both expand the menu of linguistic canons available to the judiciary and render their deployment auditable, contestable, and ultimately more legitimate.

VI. CONCLUSION

This Article is the first to offer a systematic, empirical assessment of a long-standing question in legal interpretation: when judges say they are following the plain meaning of a text and cite dictionaries or linguistic canons in support, are they in fact doing what they claim — using those tools to uncover a shared, consensus meaning — or are they instead selecting the interpretation that best accords with their policy preferences and deploying the tools as post hoc rhetorical support?

Surveying thousands of lawyers and non-lawyers on nearly 200 real-world plain meaning cases, this Article has demonstrated the extent to which courts invoking canons and dictionaries tend to align with linguistic consensus. The results indicate that judges generally invoke traditional interpretive tools not merely as a political window-dressing mechanism, but to support a sincere, albeit error-prone attempt to uncover the meaning of a term at issue in a given legal dispute.

Implicatures, 36 PROC. ANN. MEETING COGNITIVE SCI. SOC'Y 397, 397, 402 (2014); Leon Bergen, Roger Levy, & Noah D. Goodman, *Pragmatic Reasoning Through Semantic Inference*, 9 SEMANTICS & PRAGMATICS 1, 4 (2016).

327. Frank & Goodman, *supra* note 325, at 998, 818 (showing an RSA model quantitatively predicts speakers' and listeners' utterance choices in reference games and reviewing converging behavioral evidence that Bayesian-pragmatic models explain scalar implicatures, reference choice, and other pragmatic phenomena); Degen & Goodman, *supra* note 326, at 397 (demonstrating that RSA-derived production and interpretation probabilities match human judgments of scalar implicature strength); Bergen et al., *supra* note 326, at 1 (extending RSA to more complex conversational queries and validating predictions against experimental data).

This Article is also the first to empirically evaluate a second question of complementary significance — namely, whether novel computational tools, such as LLMs, might offer a useful supplement to judges' use of traditional linguistic canons in plain-meaning cases. Prompting flagship AI models such as GPT-4.1 and o3 on the same plain-meaning cases as those given to human participants, analyses revealed AI's predictions of linguistic consensus to reliably match, though not exceed, those of the court invoking traditional tools. This finding held similarly constant when accounting for demographic variables and potential knowledge of the cases within the AI's training data. These results suggest that some novel computational tools can offer an efficient, if not more effective, supplement to uncover plain meaning.

VII. APPENDIX

This Appendix contains several Parts:

(1) Part 1: Supplemental Information for Canons & Text Analysis

(2) Part 2: Supplemental Information for Study I

(3) Part 3: Supplemental Information for Study II

(4) Part 4: Supplemental Information for Formal Modeling

PART 1: SUPPLEMENTAL INFORMATION FOR TEXT ANALYSIS

A. Materials

The materials for the text analysis consisted of a combination of state and federal cases derived from Harvard Caselaw Access Project.

With respect to the federal cases, the raw sample consisted of every published opinion contained within (a) every volume of Westlaw's federal reporter, up to 935 of the third edition; (b) every volume of Westlaw's federal supplemental reporter, up to volume 392 of the third edition; and (c) every volume of Westlaw's Supreme Court reporter, up to volume 572.

With respect to the state cases, the raw sample consisted of every published opinion contained within (a) the first, second and third editions of Westlaw's Southern reporter, up to volume 275 of the third edition; (b) the second and third editions of Westlaw's Pacific reporter, up to volume 447 of the third edition; (c) the first, second, and third editions of Westlaw's Southwest reporter, up to volume 579 of the third edition; (d) the second and third editions of Westlaw's Atlantic

reporter, up to volume 213 of the third edition; (e) the second and third editions of Westlaw's Northeast reporter, up to volume 130 of the third edition; (f) the second edition of Westlaw's Northwest reporter, up to volume 932; (g) the first five editions of Westlaw's California reporter, up to volume 1 of the fifth edition; and (h) the first five editions of California appellate reporters, up to volume 11 of the fifth edition.

The total number of cases included (a) 1,185,092 in the federal sample; (b) 1,268,391 in the regional state sample; and (c) 138,505 in the California sample.

B. Data Pre-Processing Pipeline

In order to obtain the materials and convert them into analyzable format, a number of extraction and pre-processing steps were conducted.

First, opinions were scraped and read directly from Harvard Caselaw Access Project using a number of scripts from the Python library BeautifulSoup and requests packages from Python.

After scraping the content, separate scripts were used to clean the data, including (a) normalizing Unicode characters; (b) replacing special quotation marks with standard ones; (c) replacing "-" symbols with spaces; (d) converting text to lowercase; (e) tokenizing the text into words; and (f) creating n-grams (bigrams, trigrams, quadgrams and fivegrams) for proximity searches.

Finally, after completing these pre-processing steps, opinions were filtered out from the sample if the number of words in the opinion was fewer than fifty.

For various metrics, words were also lemmatized using the Natural Language Toolkit ("NLTK") package from Python.[328]

C. Metrics

Below is a breakdown of the metrics and the methods for extracting those metrics from the pre-processed data.

1. Plain Meaning

To determine the number of cases referencing plain meaning, for each case, it was determined whether the word "plain" or a synonym thereof occurred within proximity of the word "meaning or a synonym thereof. In particular, a case was determined to have referenced "plain meaning" if (a) the lemma of "plain," "ordinary," "natural,"

328. *See generally* STEVEN BIRD, EDWARD LOPER & EWAN KLEIN, NATURAL LANGUAGE PROCESSING WITH PYTHON (2009).

"common," "sense," "clear," "obvious," "evident," "apparent," "manifest," or "salient" occurred within four words/lemmas of (b) "meaning," "reading," "language," "word," "phrase," "text," or "interpretation."

Conversely, if the above criteria were not satisfied, a case was determined to not have referenced plain meaning.

2. Grammar

To determine the number of cases referencing grammar, for each case, it was determined whether the lemma "grammar" occurred anywhere within the case. If so, and as long as not all instances of the word were followed by the word "school," then a case was determined to have referenced "grammar."

Conversely, if the above criteria were not satisfied, a case was determined to not have referenced grammar.

3. Punctuation

To determine the number of cases referencing punctuation, for each case, it was determined whether the lemma "punctuation" or "comma" occurred anywhere within the case. If so, then a case was determined to have referenced "punctuation."

Conversely, if the above criteria were not satisfied, a case was determined to not have referenced punctuation.

4. *Expressio Unius*

To determine the number of cases referencing expressio unius, for each case, it was determined whether "expressio unius" or a synonym thereof occurred anywhere in the case. In particular, a case was determined to have referenced "expressio unius" if 'expressio unius', 'expresio unius', 'inclusio unius', or 'expressum facit cessare tacitum' occurred anywhere in the case.

Conversely, if the above criteria were not satisfied, a case was determined to not have referenced expresio unius.

5. *Ejusdem Generis*

To determine the number of cases referencing ejusdem generis, for each case, it was determined whether "ejusdem generis" or a synonym thereof occurred anywhere in the case. In particular, a case was determined to have referenced "ejusdem generis" if 'ejusdem generis' or "Lord Tenderton" occurred anywhere in the case.

Conversely, if the above criteria were not satisfied, a case was determined to not have referenced ejusdem generis.

6. *Noscitur a Sociis*

To determine the number of cases referencing noscitur a sociis, for each case, it was determined whether "noscitur" occurred anywhere in the case. If so, a case was determined to have referenced noscitur a sociis.

Conversely, if the above criteria were not satisfied, a case was determined to not have referenced noscitur a sociis.

7. Rule of the Last Antecedent

To determine the number of cases referencing rule of the last antecedent, for each case it was determined whether (a) any of the lemmas "last," "nearest," "closest," or "recent" appeared within three words/lemmas of (b) "antecedent." If so, a case was determined to have referenced the last antecedent rule.

Conversely, if the above criteria were not satisfied, a case was determined not to have referenced the last antecedent rule.

8. Dictionary

To determine the number of cases referencing dictionaries, for each case it was determined whether the lemma "dictionary" appeared anywhere in the case. If so, a case was determined to have referenced dictionaries.

Conversely, if the above criteria were not satisfied, a case was determined not to have referenced dictionaries.

9. Legal Dictionary

To determine the number of cases referencing legal dictionaries, for each case it was determined whether the case referenced one of the three leading law dictionaries, Black's Law Dictionary, Ballentine's Law Dictionary, or Bouvier's Law Dictionary.

In particular, if a case that was determined to have referenced dictionaries also referenced either (a) "black's law" or "blacks law"; (b) "bouvier"; or (c) "ballentine", the case was determined to have referenced a law dictionary.

Conversely, if the above criteria were not satisfied, a case was determined not to have referenced a legal dictionary.

10. Non-Legal Dictionary

To determine the number of cases referencing non-legal dictionaries, for each case it was determined whether the case referenced one of the four leading general-English dictionaries: Merriam Webster, Oxford, Cambridge or American Heritage.

In particular, if a case that was determined to have referenced dictionaries also referenced either (a) "webster," webster's", or "websters"; (b) "oxford," "oxford's," or "oxfords"; (c) "cambridge"; or (d) "heritage," the case was determined to have referenced a non-legal dictionary.

Conversely, if the above criteria were not satisfied, a case was determined not to have referenced a non-legal dictionary.

D. Supplemental Plain Meaning Results

1. Prevalence Within the Federal Judiciary

The prevalence of explicit references to "plain meaning" and relevant synonyms within the federal judiciary over time are visualized in Figure A.1. Descriptively, the prevalence has gone up dramatically as a proportion of all written opinions at all three levels of the federal judiciary over the last several decades.

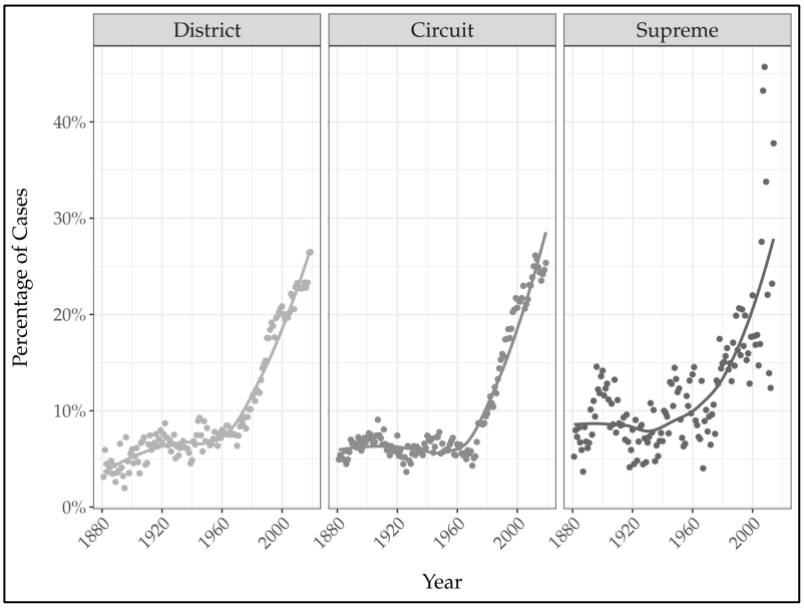

Figure A.1: Proportion of published opinions within the federal judiciary referencing "plain meaning" or relevant synonym over time. Dots represent proportion for a given year. Trend lines represent smoothed LOESS regression lines.

2. Prevalence Within the State Judiciary

References to plain meaning have similarly risen over time within the state courts, albeit the pattern is more mixed and less dramatic than in federal courts. Figure A.2 visualizes this trend at the appellate level broken down by Westlaw regional reporter.

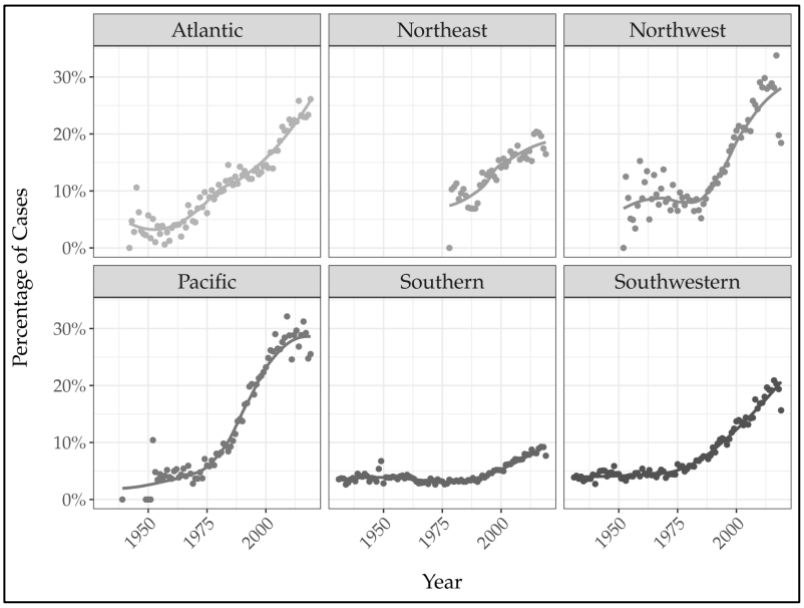

Figure A.2: Proportion of published state appellate opinions referencing "plain meaning" or relevant synonym over time, separated by Westlaw regional reporter. Dots represent proportion for a given year. Trend lines represent smoothed LOESS regression lines.

E. Canon and Dictionary Results

1. Linguistic Canons

As shown in Figure A.3, the prevalence of linguistic canons has risen greatly over the decades with the federal judiciary. As shown in Figure A.4, within the state courts, canon usage has also generally risen over the last several decades, particularly in regions that have seen the largest increases in references to plain meaning.

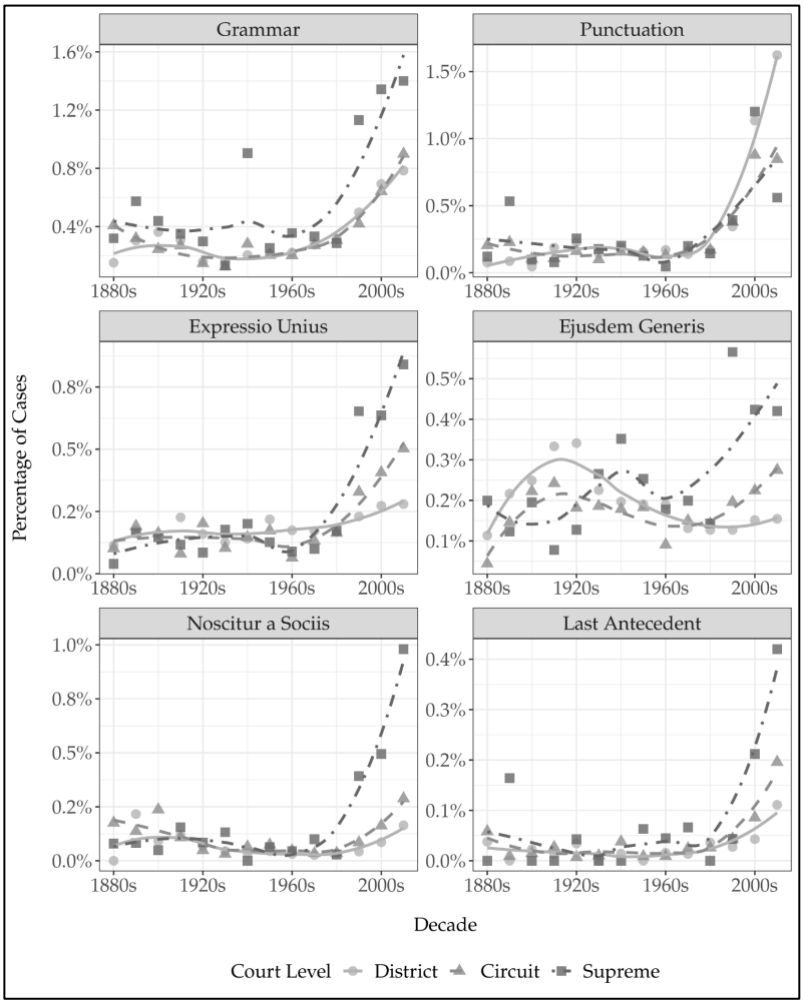

Figure A.3: Proportion of published opinions referencing different linguistic canons over time in the federal judiciary. Shape symbols represent average proportion for a given decade. Trend lines represent smoothed LOESS regression lines.

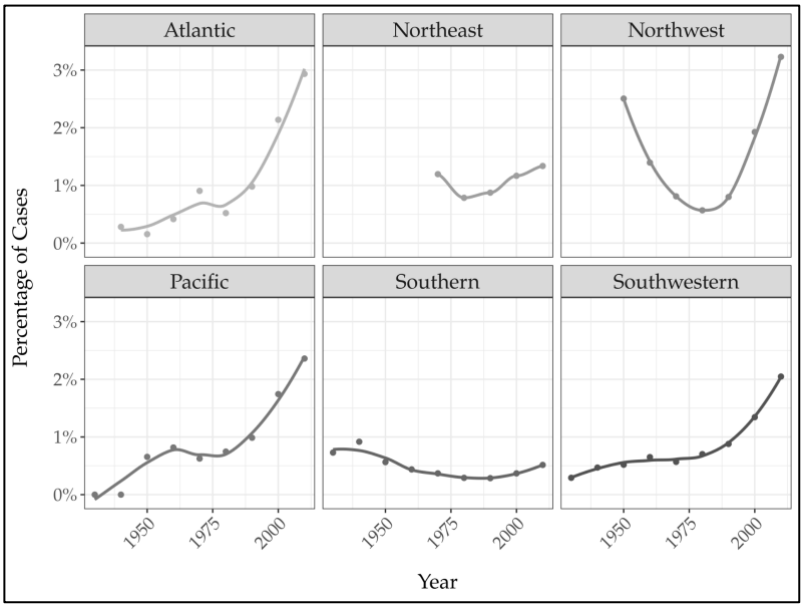

Figure A.4: Proportion of published state appellate opinions referencing at least one linguistic canon over time, separated by Westlaw regional reporter. Dots represent the average proportion for a given decade. Trend lines represent smoothed LOESS regression lines.

2. Dictionaries

As shown in Figures A.5 and A.6, the prevalence of dictionaries is both high and ever-increasing across jurisdictions.

In terms of which type of dictionary courts turn to, Figures A.7 and A.8 confirm that judges reference legal dictionaries (such as Black's, Ballentine's, and Bouvier's) and ordinary non-legal dictionaries (such as Webster's, Oxford, and American Heritage) at remarkably similar rates.

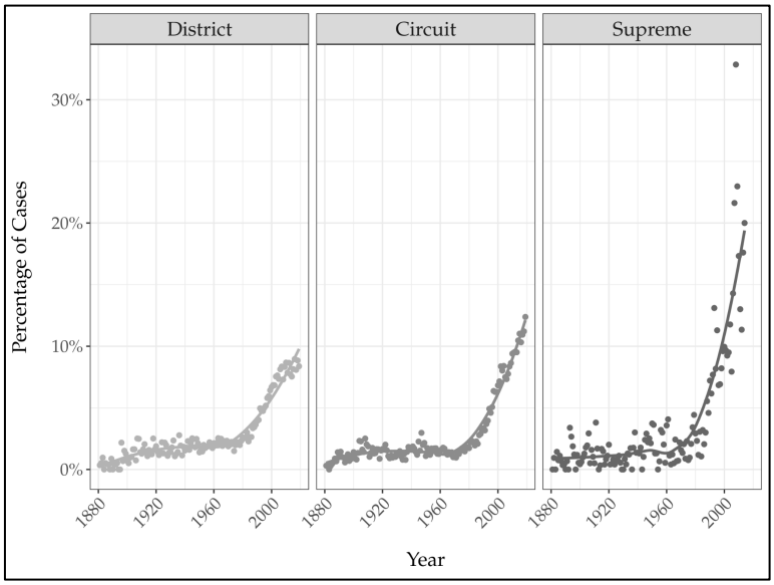

Figure A.5: Proportion of published opinions referencing dictionaries over time in the federal judiciary. Dots represent the proportion for a given year. Trend lines represent smoothed LOESS regression lines.

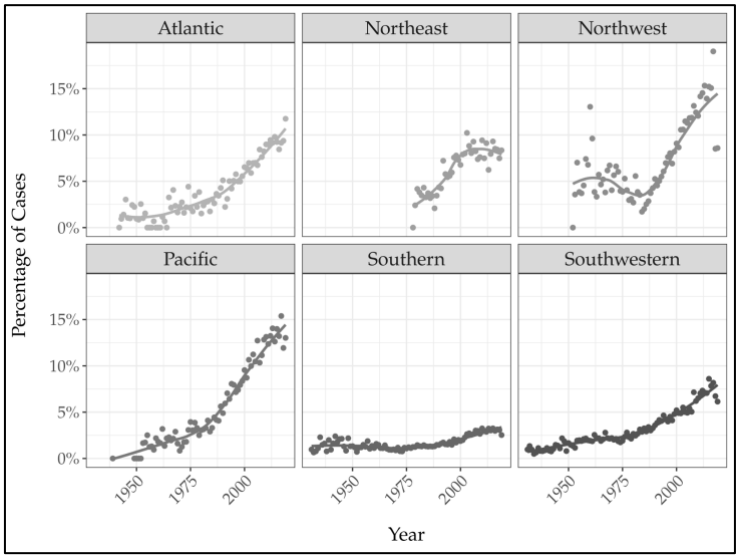

Figure A.6: Proportion of published state appellate opinions referencing dictionaries over time, separated by Westlaw regional reporter. Dots represent the proportion for a given year. Trend lines represent smoothed LOESS regression lines.

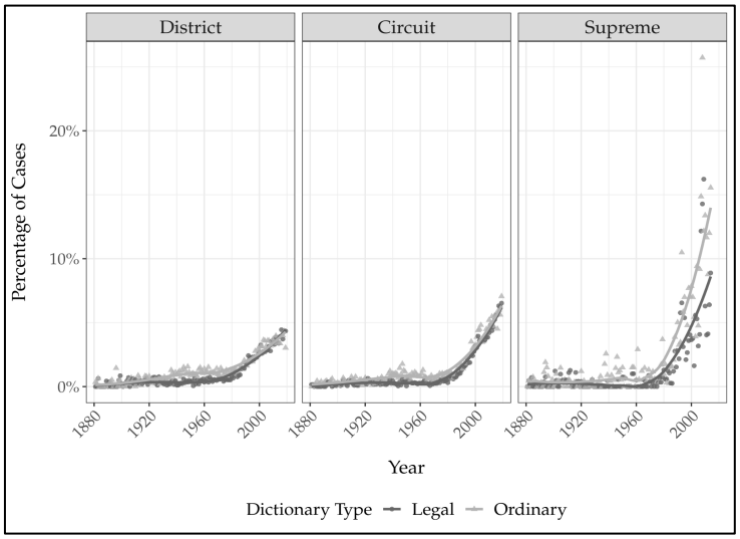

Figure A.7: Proportion of published written opinions referencing legal and ordinary (general English) dictionaries over time in the federal judiciary. Symbols represent the proportion for a given year. Trend lines represent smoothed LOESS regression lines.

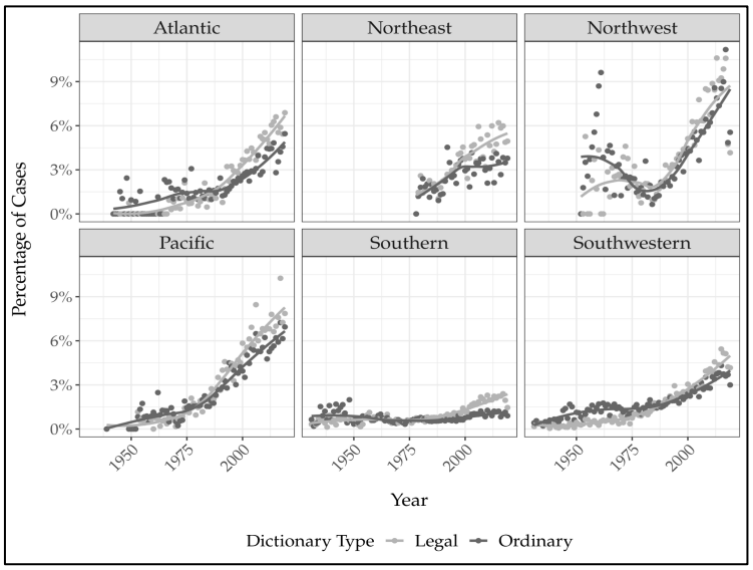

Figure A.8: Proportion of published state appellate opinions referencing legal and ordinary dictionaries over time, separated by Westlaw reporter. Dots represent the proportion for a given year. Trend lines represent smoothed LOESS regression lines.

PART 2: SUPPLEMENTAL INFORMATION FOR STUDY I

A. Supplemental Demographic Results

Table A.1: Lay Participant Demographics

	% of Sample
Gender	
Female	51.7%
Male	48.3%
Age Range	
18–29	22.9%
30–39	20.9%
40–49	16.7%
50–59	19.3%
60–69	15.2%
70–79	4.7%
80 or above	0.3%
Ethnicity	
White	65%
Black	10.8%
Asian	6.4%
Mixed	10.3%
Other	7.5%
Political Affiliation	
Democrat	29.7%
Republican	26.9%
Independent	43.4%
Country of Residence	
United States	100%

Table A.2: Lawyer Participant Demographics

	% of Sample
Gender	
Male	58.2%
Female	40.3%
Non-Binary	0.5%
Transgender	0.2%
Prefer not to answer	0.7%
Write-in	0.2%
Age Range	
18–29	5.4%
30–39	17.4%
40–49	22.9%

50–59	21.5%
60–69	18.4%
70–79	12%
80 or above	2.3%
Race	
White	75.5%
Hispanic or Latino/a/x	9.1%
Asian	6.3%
Black or African American	4.1%
American Indian or Alaska Native	0.5%
Native Hawaiian or Pacific Islander	0.4%
Prefer not to answer	2.5%
Write-in	1.5%
Politics	
Very Liberal	13%
Liberal	28.4%
Somewhat Liberal	18.1%
Middle of the Road	19.7%
Somewhat Conservative	10.8%
Conservative	8.2%
Very Conservative	1.8%
Country	
United States	98.6%
Other	1.4%
Years Licensed	
0–9	21.1%
10–19	24.7%
20–29	20.5%
30–39	17.2%
40–49	12.6%
50 or above	3.9%
Job	
Law firm attorney	33.4%
Solo practitioner	18.8%
Government attorney	10.4%
In-house counsel	8.7%
Public interest / non-profit lawyer	4.4%
Legal academic	2.9%
Other	21.6%

Below is a breakdown of the lawyer participant sample by area of specialization.

Table A.3: Areas of Specialization in Lawyer Sample

Area of Specialization	% of Sample
Administrative Law	8.5%
Admiralty	0.8%
Agency and Partnerships	2.5%
Agricultural Law	0.4%
Alternative Dispute Resolution	6.9%
Animal Law	0.9%
Antitrust	2.3%
Appellate Practice	9.7%
Aviation and Space Law	0.5%
Bioethics	0.3%
Business Associations	9.5%
Civil Procedure	15.8%
Civil Rights	7.1%
Clinical Teaching	1.3%
Commercial Law	13.6%
Communications Law	0.5%
Community Property	1%
Comparative Law	0.5%
Conflict of Laws	0.8%
Constitutional Law	6.7%
Consumer Law	5%
Contracts	23.6%
Corporate Finance	4.5%
Creditors and Debtors Rights	4.3%
Criminal Justice	6.7%
Criminal Law	14.2%
Criminal Procedure	8%
Critical Legal Studies	0.5%
Critical Race Theory	0.5%
Disability Law	2.3%
Education Law	2.1%
Elder Law	2.1%
Election Law	1.2%
Employee Benefit Plans	1.1%
Employment Law	12.7%
Energy Law	1.1%
Entertainment Law	3%
Environmental Law	3.3%
Equity	0.5%
Estate Planning	6.9%
Estate and Gift Tax	2.8%
Estates and Trusts	8.3%

Evidence	7.1%
Family Law	8.1%
Federal Courts	6.8%
Feminist Legal Theory	0.2%
Financial Institutions	1.8%
Forensic Medicine	0.2%
Government Contracts	2.5%
Health Care Law	4.3%
Human Rights	1.2%
Immigration Law	4.1%
Insurance Law	7.6%
Intellectual Property	9%
International Business Transactions	3.1%
International Law	2.4%
International Organizations	0.8%
Judicial Administration	1%
Jurisprudence	1.3%
Juvenile Law	3%
Labor Law	3.8%
Land Use Planning	2.1%
Law Office Management	2.1%
Law and Accounting	1%
Law and Economics	0.8%
Law and Literature	0.3%
Law and Medicine	1.5%
Law and Psychiatry	0.4%
Law and Religion	0.5%
Law and Science	0.8%
Law and Social Science	0.6%
Law and Technology	3.2%
Legal Drafting	8.3%
Legal History	0.8%
Legal Methods	0.9%
Legal Research and Writing	12.8%
Legislation	2.3%
Local Government	4.6%
Military Law	0.6%
National Security	0.6%
Native American Law	0.5%
Natural Resources	1.1%
Nonprofit and Philanthropy	2.6%
Ocean Resources	0.2%
Oil and Gas	0.2%
Payment Systems	0.5%

Poverty Law	1.1%
Products Liability	4.3%
Professional Responsibility	3.7%
Property Law	7.2%
Real Estate Transactions	10.1%
Regulated Industries	1.7%
Remedies	2%
Securities Regulation	2.5%
Sexual Orientation and Gender Identity Issues	0.8%
Sports Law	0.9%
Tax Policy	0.5%
Taxation Corporate	1.2%
Taxation Federal	2.8%
Taxation State and Local	1.9%
Torts	14.4%
Trade Regulation	0.5%
Trial Advocacy	9.8%
Water Rights	0.5%
Welfare Law	0.3%
Women and the Law	0.7%
Workers' Compensation	1.9%
Other	17%

Table A.4: Lawyers' Views on Interpretive Theory

	Yes	**No**	**Other**
Constitutional Interpretation			
Originalism	22.6%	51.8%	25.5%
Living Constitutionalism	62.1%	14.4%	23.6%
Pluralism	28.2	25.1%	46.7%

Statutory Interpretation			
Textualism	71.6%	11.1%	17.3%
Purposivism	58%	20.1%	21.9%
Intentionalism	61%	17.6%	21.4%
Pragmatism	19.1%	62.2%	18.7%
Contract interpretation			
Textualism/Formalism	50.7%	26.3%	22.9
Contextualism/Anti-Formalism	43.1%	27.6%	29.4

Following recent work identifying the distribution of specialties within the legal academy,[329] this study used categories from the AALS (n = 107).[330] The most common area of specialization among lawyers in the sample was Contracts, as 23.6% of lawyers self-identified as specializing in that field. Other common areas included Civil Procedure

329. Martínez & Tobia, *supra* note 208, at 140–45. *See also* Eric Martínez, Measuring Legal Concepts 21 (Feb. 4, 2024) (unpublished manuscript) (on file with the Institute for Law & AI at the University of Chicago Law School) (using the list of specialties to identify the doctrinal areas in which fundamental legal concepts were most likely to pertain to).

330 . *Faculty_Appointments_Register*, ASS'N AM. L. SCHS., https://www.aals.org/recruitment/current-faculty-staff/far/ [https://perma.cc/T9ZV-8EUT]. Areas on the list include areas associated with more traditional "core" 1L courses, such as constitutional law, civil procedure, and torts, as well as a variety of other areas, such as Native American law, poverty law, international law, election law, and elder law.

(15.8%), Torts (14.4%), Criminal Law (14.2%), and Commercial Law (13.6%). 17% of participants self-identified as specializing in something not included in the list.

Results of the interpretive philosophy questionnaire are visualized in Table A.4.

The most widely endorsed approach to interpretation among all theories was textualism, which was also the only theory that was endorsed by a supermajority (71.6%) of participants. Within statutory interpretation, purposivism and intentionalism were also endorsed by a majority of participants, whereas pragmatism — despite being accepted by the majority of legal academics[331] — was rejected by the majority of attorneys in the sample.

Within contract interpretation, preferences of attorneys likewise differed from those of the academy. "Textualism/Formalism" was endorsed by a slight majority of participants (50.7%), compared to just 43.1% for "Contextualism/Anti-Formalism," despite the latter being widely endorsed by United States law professors.[332]

With respect to constitutional interpretation, lawyers in the sample exhibited a clear preference for living constitutionalism over originalism, as the former was endorsed by a majority of participants whereas the latter was rejected by a majority of participants. Pluralism was neither endorsed nor rejected by a majority of participants, with the most common response to this set of views being "Other."

B. Analysis Plan

To formally evaluate the question of whether there is linguistic consensus in plain meaning cases, mixed-effects logistic regressions were conducted using the lme4 package from R.

The outcome variable in these regressions was "agreed_with_majority," which takes each person's yes/no response to a given question and codes it as "1" if the response was in alignment with the majority of responses for that question (i.e., was it "yes" if most people chose "yes"), and "0" if it was not.

In these regressions, "condition" was a fixed effect; and "question" and "participant" were random intercepts.

The pre-registered prediction was that if the lower bound of the confidence interval of the intercept coefficient (after converting from log odds to probability) was above 0.5, this was pre-registered as evidence that participants are consistently choosing one interpretation over another above chance.

---

331. Martínez & Tobia, *supra* note 208, at 152.
332. *Id.*

To formally evaluate the second question of whether judges align with linguistic consensus when invoking certain interpretive tools, separate mixed-effects logistic regressions were conducted.

The outcome variable in these regressions was "agreed_with_judge," which takes each person's yes/no response to a given question and codes it as "1" if the response was in alignment with the court for that question (i.e., was it "yes" if the court said "yes"), and "0" if it was not.

As with the above regressions, "condition" was a fixed effect; and "question" and "participant" were random intercepts.

With respect to confirmatory criteria: If the lower bound of the confidence interval of the intercept coefficient (after converting from log odds to probability) is above 0.5, this was pre-registered as evidence that judges are aligning with consensus above chance.

To assess the robustness of the main findings to demographic and other factors, a series of control analyses were conducted.

C. Fine-Grained Consensus Trends

Consensus data in the main text was reported largely in the aggregate. This Section presents some of the primary consensus results in more fine-grained detail.

1. Case-by-Case Consensus Results

A breakdown of the percentage of lay respondents who aligned with the court in each case is presented in Figure A.9.

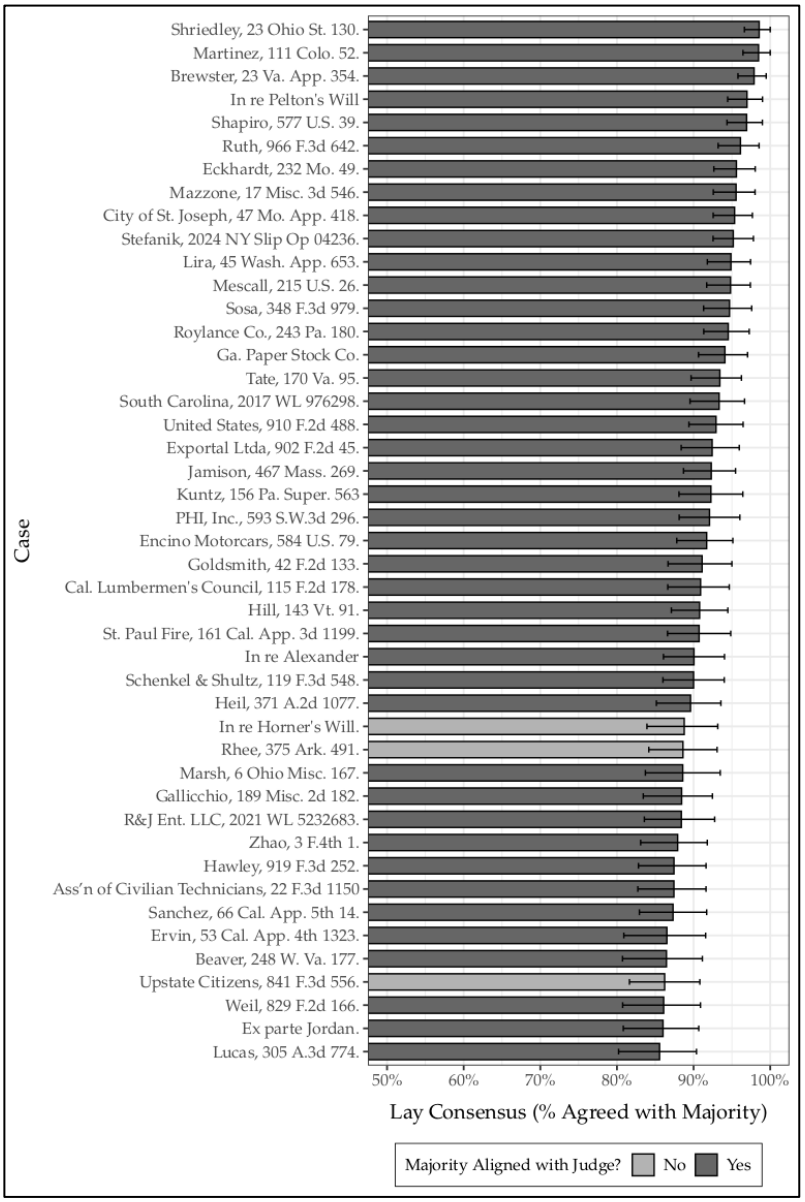

Figure A.9: Lay consensus and judge alignment with consensus, by case. Error bars represent 95% bootstrap confidence intervals of percent agreeing with majority.

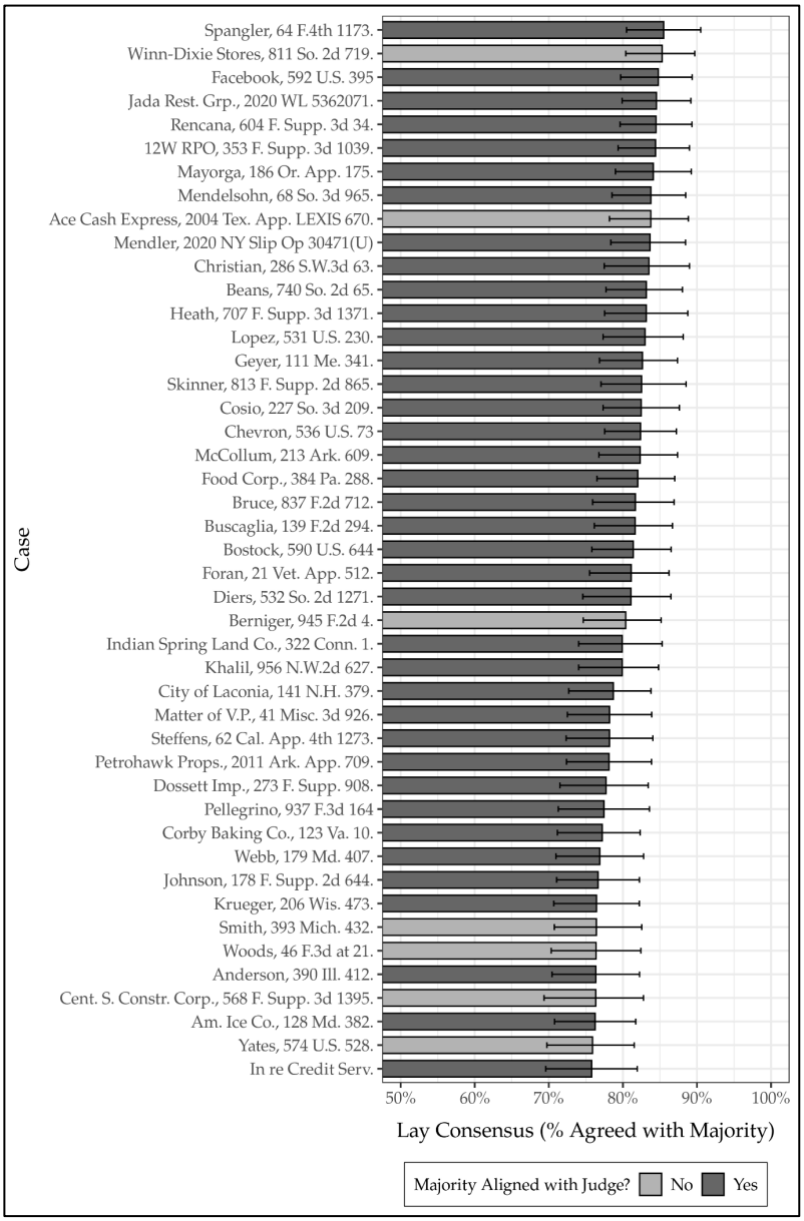

Figure A.9 (cont.): Lay consensus and judge alignment with consensus, by case.

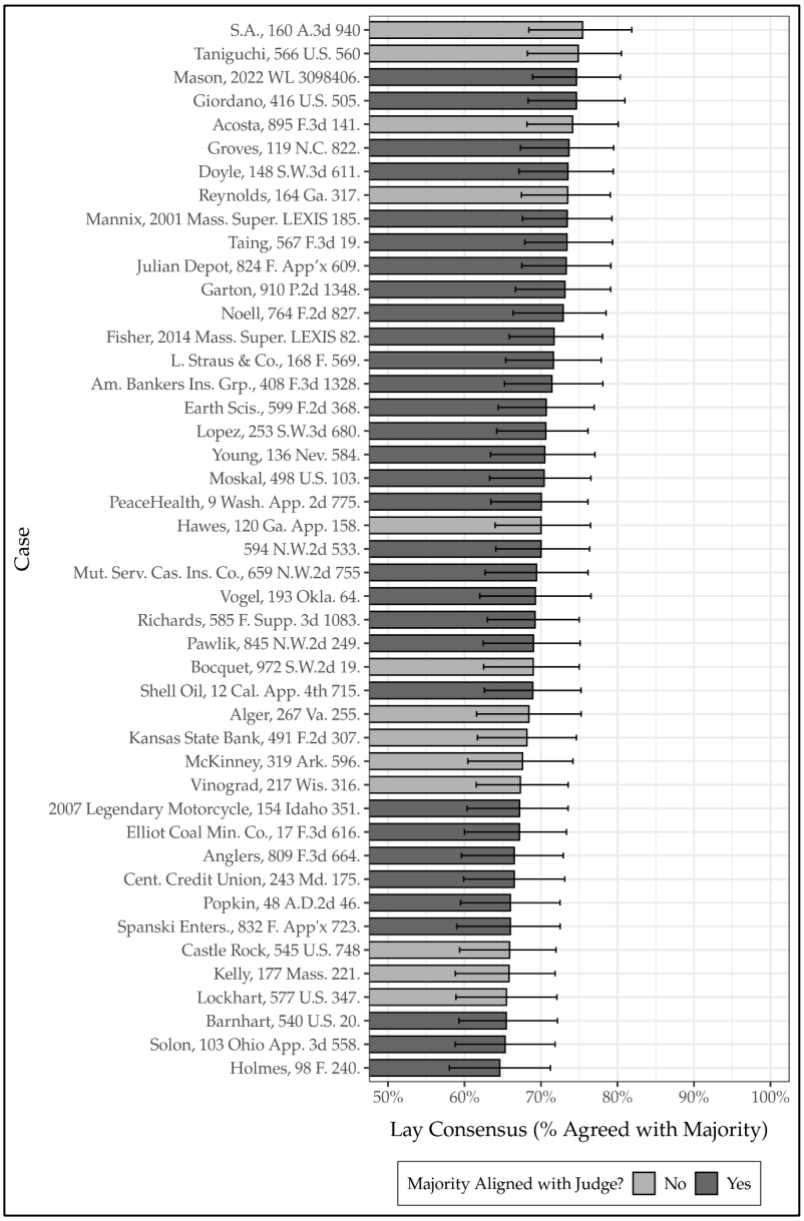

Figure A.9 (cont.): Lay consensus and judge alignment with consensus, by case.

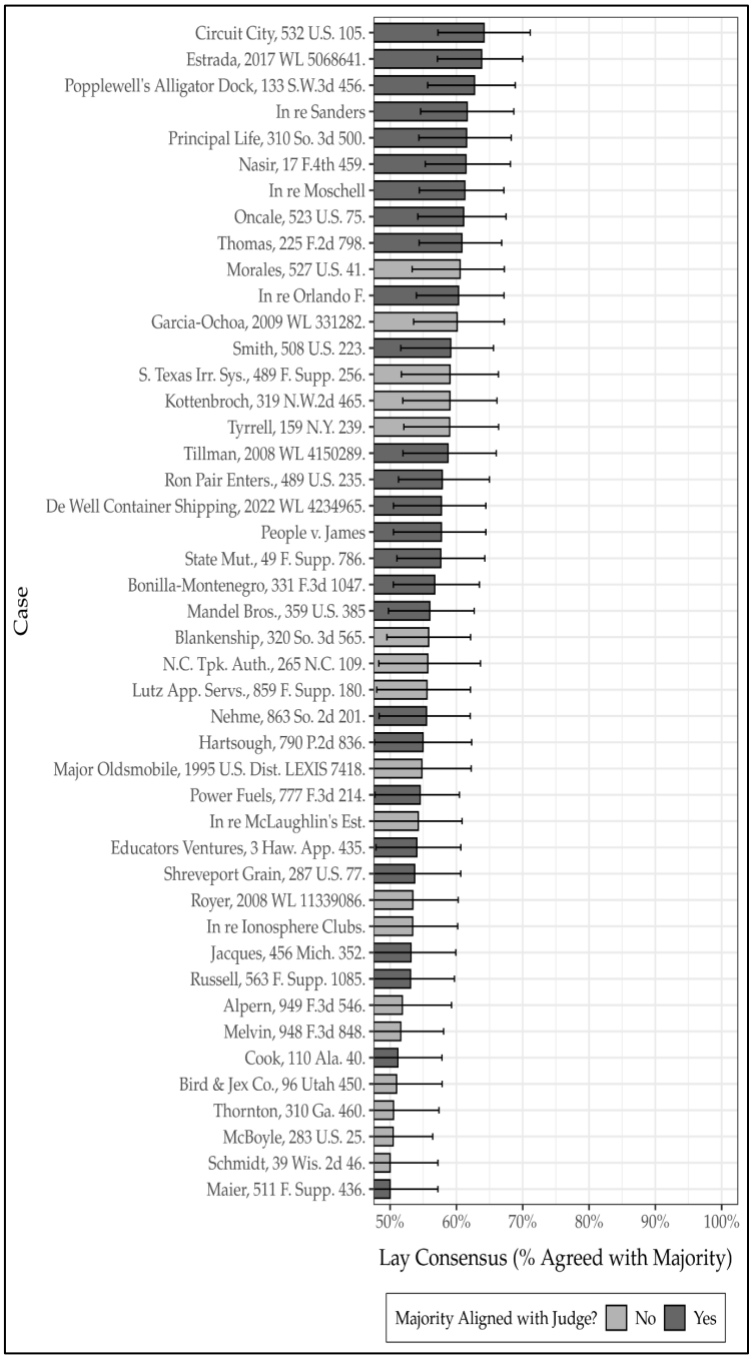

Figure A.9 (cont.): Lay consensus and judge alignment with consensus, by case.

2. Consensus Thresholds

Consensus results by various consensus thresholds (i.e., the percentage of participants adopting the modal interpretation) are visualized in Table A.3 and Figure A.10.

Table A.3. Breakdown of the percentage of cases falling within or above a certain consensus threshold.

	% Cases (Lay)	**% Cases (Lawyer)**	**% Cases Cumulative (Lay)**	**% Cases Cumulative (Lawyer)**
<60%	18.3	16.7	100	100
60–69%	20	18.3	81.7	83.3
70–79%	22.2	15.6	61.7	65
80–89%	23.9	25	39.4	49.4
90%+	15.6	24.4	15.6	24.4

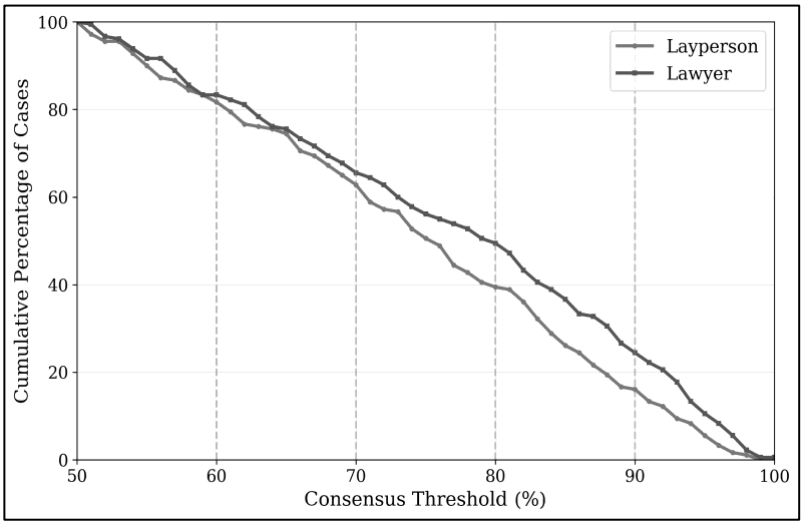

Figure A.10: Cumulative percentage of cases by consensus threshold, as represented by the percentage of a given participant group choosing the modal interpretation.

3. Certainty Results

Below is the distribution of certainty results by case.

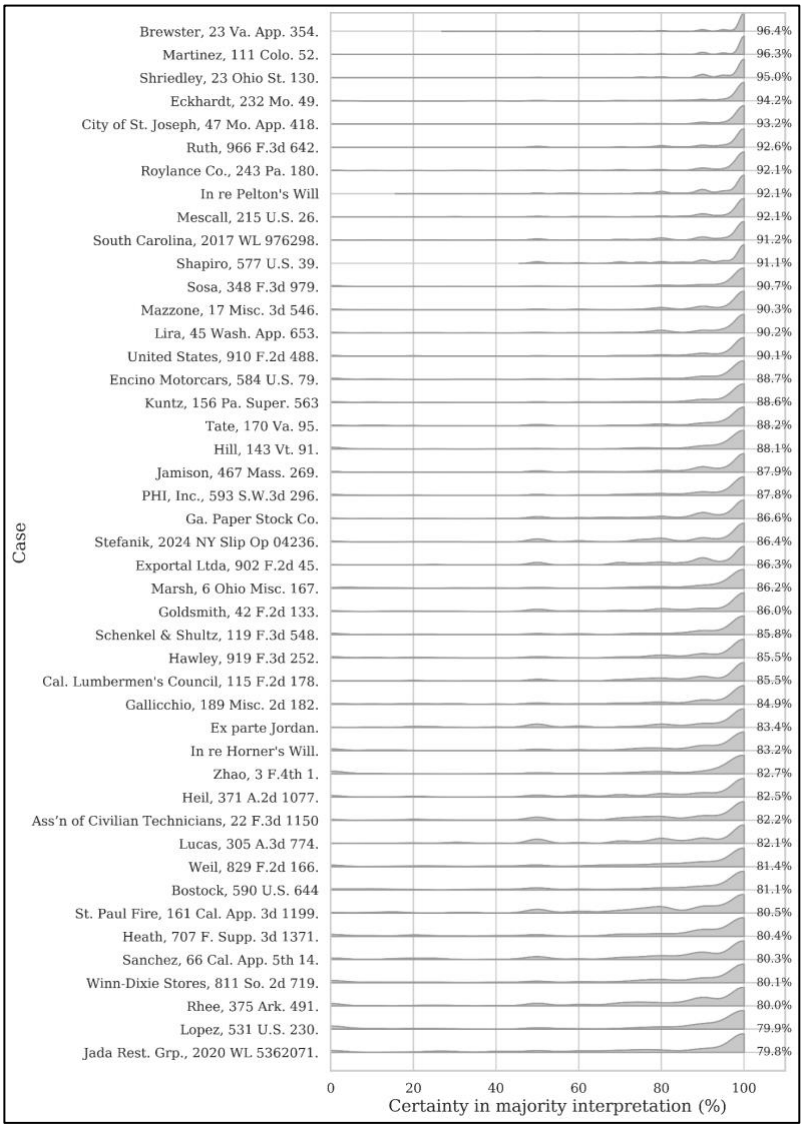

Figure A.11: Certainty in modal interpretation, organized by case.

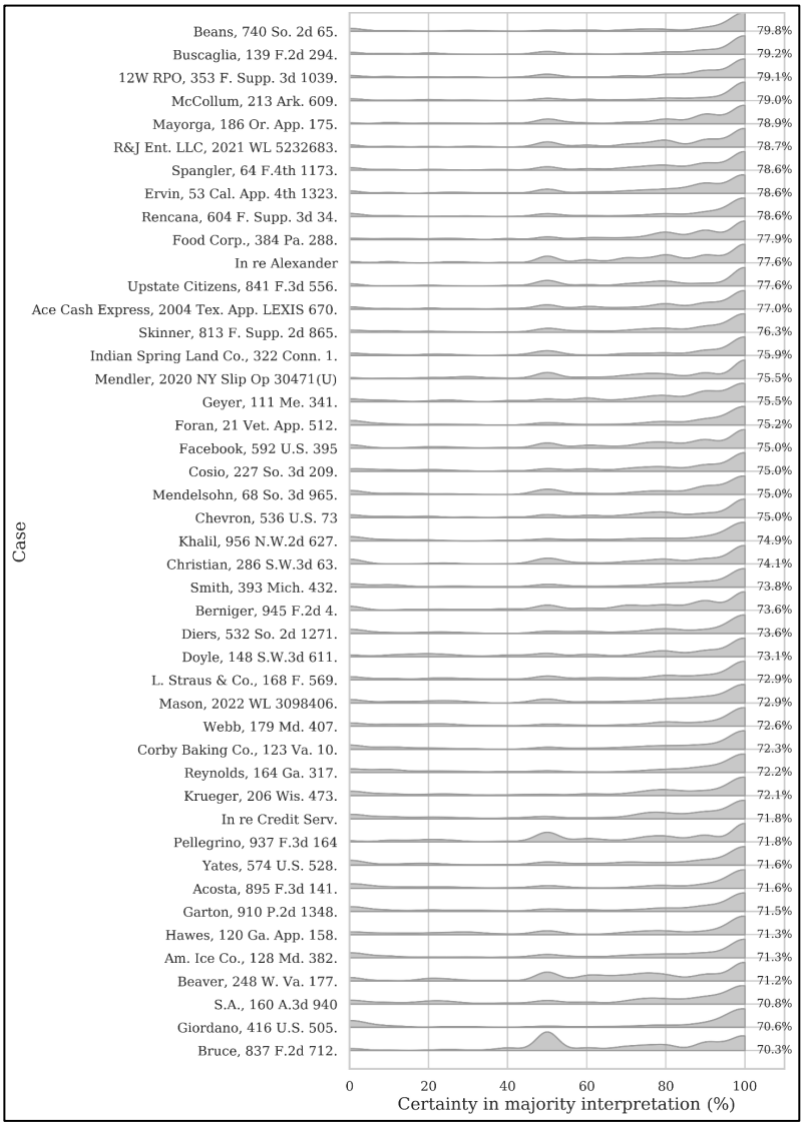

Figure A.11 (cont.): Certainty in modal interpretation, organized by case.

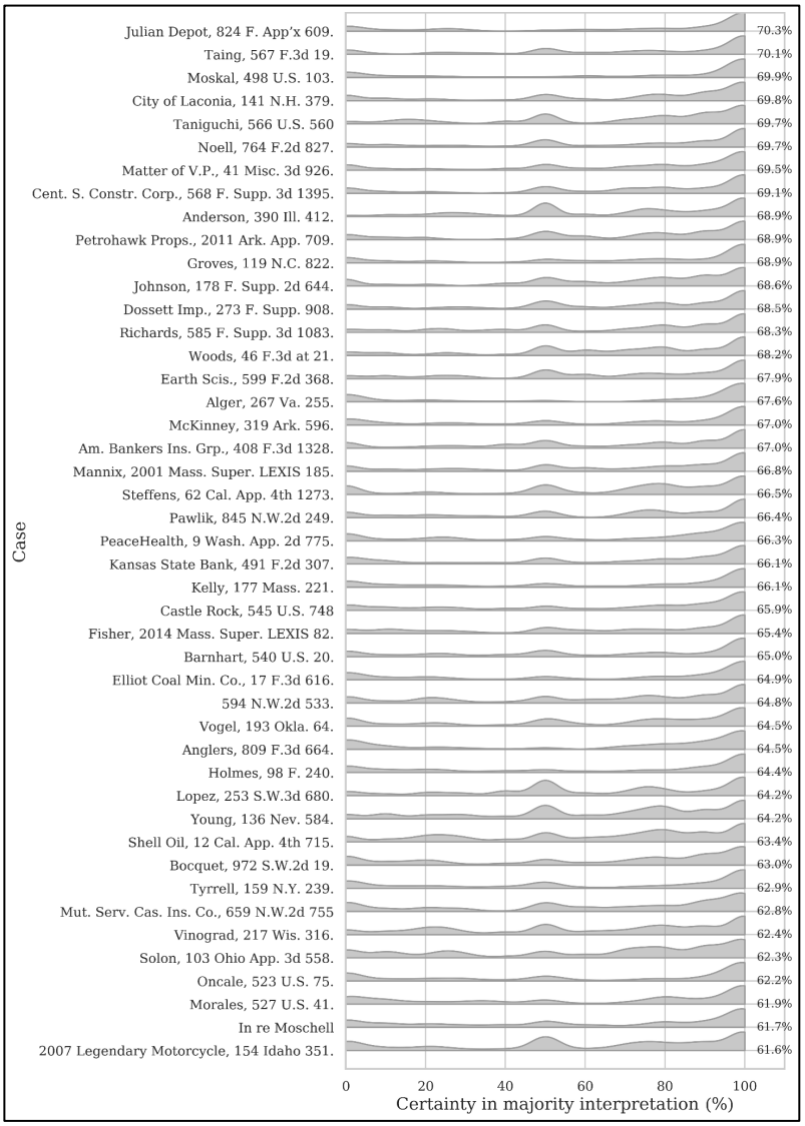

Figure A.11 (cont.): Certainty in modal interpretation, organized by case.

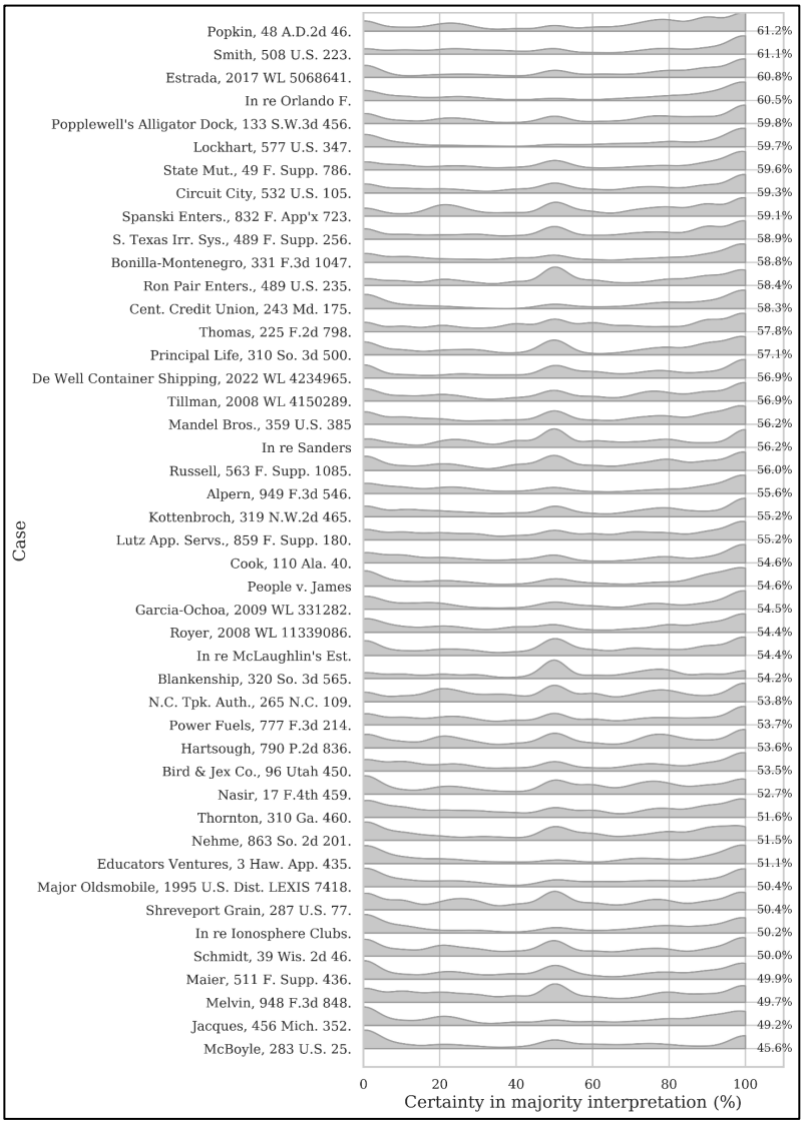

Figure A.11 (cont.): Certainty in modal interpretation, organized by case.

D. Fine-Grained Alignment Trends

Below are some of the alignment results presented at a more fine-grained level.

1. Alignment by Consensus Threshold

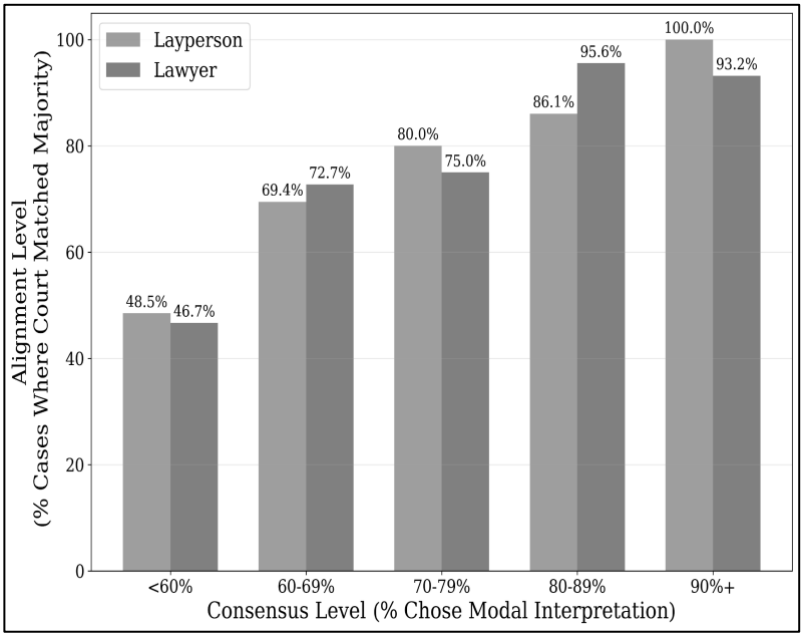

Figure A.12: Judge alignment with consensus-by-consensus level. Bar heights correspond to the percentage of cases in which the judge aligned with the modal interpretation of laypeople (lighter bars) and lawyers (darker bars).

E. Prediction Control Analysis

In the main text, following the pre-registration, participants were excluded if they did not successfully pass a number of attention checks. One might wonder if the accuracy of participants' predictions were influenced by this exclusion criteria.

To assess this, a control analysis was conducted that included those who failed attention checks. The results of this control analysis were consistent with the analysis reported in the main text. With regard to the predictions of lay participants, the mean prediction accuracy was 62.5% (95% CI: 62.0–63.0). When filtering out cases without a clear consensus, the prediction accuracy rose to 65.5% (95% CI: 65.1–66.1).

With respect to the lawyer participants, the mean prediction accuracy of lay consensus was 55.7% (95% CI: 55.0–56.4). When filtering out cases without a clear consensus, the prediction accuracy rose to 59.0% (95% CI: 58.2–59.9).

The mean prediction accuracy of lawyer consensus was 72.4% (95% CI: 71.7–73.0). When filtering out cases without a clear lawyer consensus, the prediction accuracy rose to 76.7% (95% CI: 76.1–77.4).

As with the main analysis, all of these numbers were lower than the accuracy of the court.

F. Alignment Demographic Control Analysis

In order to assess the robustness of the alignment results, additional regression analyses were conducted to determine whether the results held constant when accounting for different demographic variables, such as age, gender, politics and race.

Separate analyses were conducted for the lawyer and layperson alignment data.

With respect to the layperson alignment data, an additional regression analysis was conducted that added age, gender, politics and race as predictor variables. The results of this regression were consistent with the main analysis: the lower bound of the 95% confidence interval of the intercept (after converting from log odds to probability scale) was above 0.5 (0.680-0.746).

The 95% confidence intervals of the different fixed-effect predictors of this regression are displayed in Table A.4.

Table A.4: Confidence Intervals of Coefficient Estimates for Layperson Alignment Control Analysis

	2.5 %	**97.5 %**
(Intercept)	0.679	0.746
condition1	0.538	0.802
condition2	0.394	0.693
condition3	0.256	0.543
condition4	0.444	0.735
condition5	0.110	0.298
condition6	0.659	0.872
condition7	0.131	0.343
condition8	0.379	0.678

condition9	0.245	0.527
condition10	0.429	0.724
condition11	0.267	0.558
condition12	0.402	0.700
condition13	0.570	0.822
condition14	0.221	0.494
condition15	0.472	0.689
age_group1	0.503	0.518
gender_group1	0.494	0.507
race_group1	0.479	0.494
politics_groupNon-Democrat	0.485	0.514

With respect to the lawyer data, an additional regression model was conducted that added age, gender, politics, race and potential familiarity with the case as predictor variables.

The results of this regression were consistent with the main analysis: the lower bound of the 95% confidence interval of the intercept (after converting from log odds to probability scale) was above 0.5 (0.728-0.798).

The 95% confidence intervals of the different fixed-effect predictors of this regression are displayed in Table A.5.

Table A.5: Confidence Intervals of Coefficient Estimates for Lawyer Alignment Control Analysis

	2.5 %	**97.5 %**
(Intercept)	0.728	0.798
condition1	0.559	0.856
condition2	0.418	0.768
condition3	0.199	0.530

condition4	0.370	0.730
condition5	0.111	0.363
condition6	0.456	0.796
condition7	0.129	0.402
condition8	0.338	0.701
condition9	0.289	0.648
condition10	0.133	0.427
condition11	0.299	0.659
condition12	0.202	0.533
condition13	0.826	0.962
condition14	0.260	0.617
condition15	0.418	0.686
age_group1	0.496	0.517
gender_group1	0.487	0.508
race_group1	0.490	0.514
politics_numeric	0.491	0.504
their_jurisdiction1	0.480	0.540

G. Ordinary People Control Study

As stated in the main text, one might still object to the main study design's operationalization of "ordinary people" on the basis that (a) Prolific is skewed towards educated participants, and therefore unrepresentative of "ordinary people"; and (b) participants may have been responding on the basis of their own policy preferences as opposed to providing their own plain meaning judgments.

To account for these possibilities, a control study was conducted. Below are the methods and results of that study.

1. Methods

The methods of this study were identical to the main study, with a few key deviations as outlined below.

a. Materials

The primary materials consisted of a subset of those in the main study as opposed to the full set. In particular, the primary materials consisted of forty items — all of those in the dictionary and judicial authority categories. These two categories were chosen due to the fact that the level of consensus, as well as alignment with consensus, in these two categories were largely representative of the set of materials as a whole.

In order to account for the possibility that participants might respond based on their policy preferences as opposed to their own plain meaning judgments, the attention check questions in the original study were replaced with a different set of questions where the plain meaning of the law would plausibly deviate from most participants' policy preferences.

In particular, the first of these new attention check questions presented participants with a law that prohibited all types of recycling and then asked participants whether, according to the law, it was prohibited to recycle a glass bottle. The exact wording of this provision was as follows:

> Imagine a law states:
>
> > "Recycling of any kind is prohibited."
>
> According to this law, is recycling a glass bottle prohibited?

The second of these new attention check questions presented participants with a law that mandated that state food subsidies be used exclusively to purchase candy and then asked whether, according to the law, it was permissible to use state food subsidies to purchase candy. The exact wording of that provision was as follows:

> Imagine a law states:
>
> > "State food subsidies must be used only to purchase candy."

> According to this law, may state food subsidies be used to purchase fresh vegetables?

Finally, in addition to the primary materials and attention check questions, there was an additional educational demographics questionnaire that asked participants to provide the highest level of educational attainment. The options were modeled directly after those of the United States census categories and consisted of the following:

(1) Less than a high school diploma or equivalent

(2) High school diploma or equivalent

(3) Some college but not a degree

(4) Associate degree

(5) Bachelor's degree

(6) Graduate degree

b. Participants and Procedure

The participants in this study (n = 450) were recruited in the same manner and from the same pool as those from the layperson sample of the original study, via Prolific. Participants were ineligible for the study if they participated in the original study, so as to avoid potential confounds that might arise with overlap between the two samples.

The procedure was largely identical for this study relative to the original study. In particular, participants completed eight experimental trials in random order, along with the two new attention checks randomly interspersed.

In addition, following the completion of the ten trials, participants completed the demographics questionnaire.

c. Analysis Plan

Following the design of the original study, participants were excluded if they failed one or more attention checks.

To account for potentially uneven subsamples in terms of educational attainment, a post-stratification procedure was conducted based on the distribution of educational attainment according to the 2022 United States census.[333] Participant group responses within a given educational attainment group were weighted according to how

333. Press Release, United States Census Bureau, Census Bureau Releases New Educational Attainment Data (Feb. 16, 2023), https://www.census.gov/newsroom/press-releases/2023/educational-attainment-data.html [https://perma.cc/LNR6-A757].

under- or over-represented their group was in the sample relative to the census.

Following this post-stratification procedure, descriptive statistics and parameter estimates of the dependent variables of interest (consensus and alignment with consensus) were computed similarly to the main study.

2. Results

Of the 461 participants who completed the study, 16.2% failed one or both attention checks, leaving 389 (83.8%) participants in the final sample.

Demographic results of the participant sample relative to the United States sample are displayed in Table A.6.

In line with previous estimates of online survey participant demographics, the survey participants possessed higher levels of education attainment, on average, relative to the general population.

Table A.6: Educational Attainment of Survey Participants Relative to General Population

Highest Level of Educational Attainment	**% of Survey Participants**	**% of General Population**
Less than a high school diploma or equivalent	0.5	9
High school diploma or equivalent	9.0	28
Some college but not a degree	22.6	15
Associate degree	9.3	10
Bachelor's degree	32.4	23
Graduate degree	26.2	14

Analyzing the post-stratification-adjusted data reveals that the primary results were robust to these discrepancies.

Across all cases, the mean level of convergence on the modal/majority interpretation (after adjusting for post-stratification) was 77.2% (95% CI: 75.2–79.2).

With respect to alignment, the court's interpretation aligned with that of the majority of participants (after adjusting for post-stratification) in thirty-four out of forty cases (85%).

Results are shown in Figures A.13 and A.14.

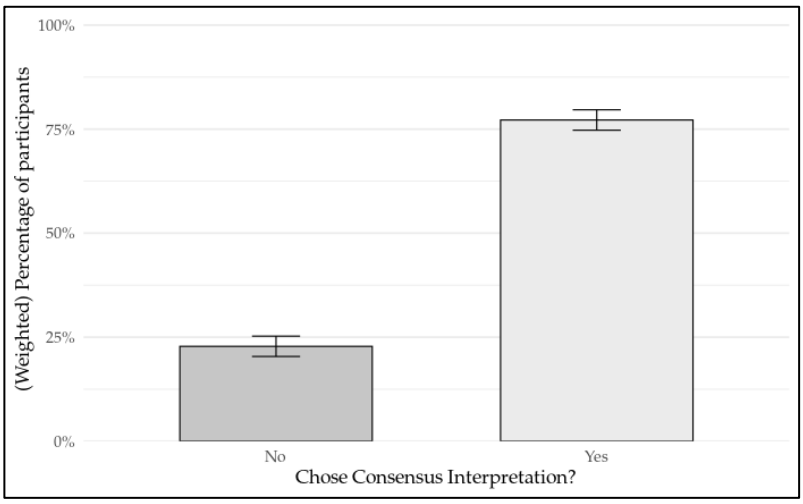

Figure A.13: Degree of linguistic consensus, adjusting for educational attainment. Bar height represents the average weighted proportion of participants who converged on a consensus interpretation for a given item. Error bars represent 95% confidence intervals.

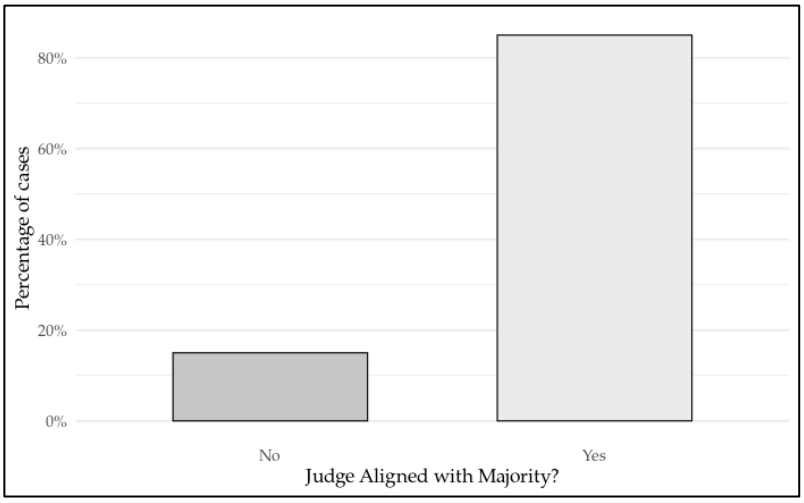

Figure A.14: Judge alignment with linguistic consensus, adjusting for educational attainment. Bar heights represent the percentage of cases in which the majority of participants chose the same interpretation as the court.

H. Easy Case Robustness Check

As stated in the main text, materials were selected so as to span the universe of relevant plain meaning cases, with an emphasis on the exact types of cases in which the different hypotheses being tested were stated to apply. Within this process was a concerted effort to avoid merely focusing on "easy" cases. However, one might still wonder whether the high level of consensus and alignment with consensus was a result of cases that were disproportionately easy, particularly in the sense being less politically charged than most relevant cases (such that judges might have no strong policy preference that might otherwise override the linguistic consensus).

To account for this possibility, an additional robustness check was performed comparing the level of "politically chargedness" between the sample of cases and a random sample of 2,000 cases in the federal and state judiciaries.

1. Methods

To conduct this robustness check, a random sample of 2,000 cases (1,000 federal cases and 1,000 state cases) was drawn from Harvard Caselaw Access Project for comparison with the experimental sample used in Studies 1 and 2. These cases were drawn from the same general sample of cases analyzed in Part II.

To measure how politically charged each of these cases is, an LLM prompting experiment was conducted in which the majority opinion of each of these cases was fed into the context window of GPT-4.1 (prompted through the API), following a prompt that asked the model the following:

> "The following is the text of the majority judicial opinion. Read the text and rate, on a scale of 1-5, how politically charged the case is (1 representing not politically charged at all, 5 representing extremely politically charged)."

For the control sample, all opinions were available in text format, and were therefore fed directly to the model via the user prompt. In the case of the experimental sample, approximately half of the cases were not available in plain text format. In such cases, the opinion was instead fed to the model as a pdf attachment.

After prompting, descriptive statistics were compared both in terms of mean rating, as well as overall distribution.

2. Results

The mean and distribution of politically charged ratings were virtually identical across the two samples. The distribution of the two samples is visualized in Figure A.15.

With respect to mean, the average politically chargedness rating in the experimental sample was 1.54 (SD = 0.887) as compared to 1.51 in the control sample (SD = 0.751). A Mann-Whitney Test revealed no significant differences between the two samples ($p = 0.813$).

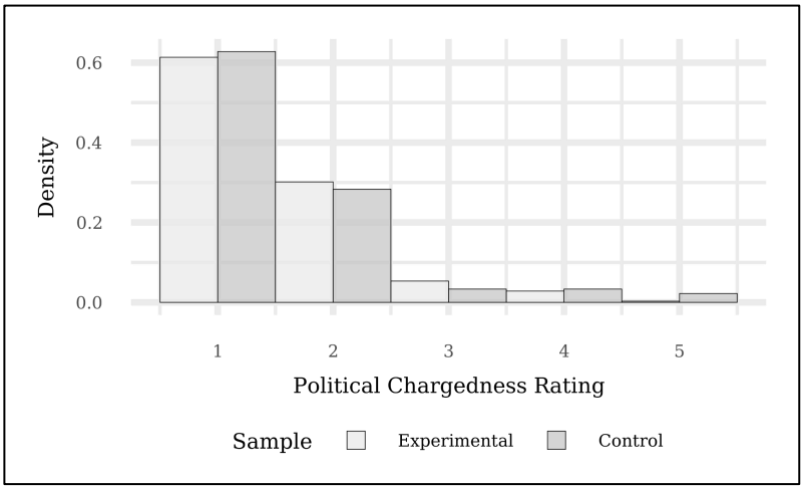

Figure A.15: Distribution of politically charged ratings

I. Smokescreen Robustness Checks

As stated in the main text, one implicit assumption of the study design with regard to testing the smokescreen hypothesis is that, in the aggregate, a judge's policy preferences are independent of linguistic consensus (that is, the judge's preferred interpretation on *policy* grounds is, on average, no more likely than chance to match up with the collective's preferred interpretation on *linguistic* grounds).

The text lays out several reasons for the viability of this assumption. However, one failure mode of this assumption is the possibility that judges might be stacking the linguistic deck in their favor by either (a) filling their docket with cases in which the the text at issue supports their policy preference; or (b) selectively invoking plain meaning in cases where the text supports their policy preference (while failing to invoke it in cases where the text does not support their policy preference).

If so, one would expect to see (a) within the study sample, a high degree of alignment between the political ideology of the judge in a given plain meaning case and the political valence of the case outcome; or (b) outside the study sample, a lack of sensitivity to linguistic clarity in a judge's decision to invoke plain meaning.

A second possibility, which might dictate in favor of a weaker version of the smokescreen hypothesis, is that judges might use canons as a smokescreen in politically charged cases. If so, one would expect to find lower alignment with linguistic consensus in cases with some degree of political chargedness relative to cases with a higher degree of political chargedness.

1. Methods

To test for the possibility of judges stacking the linguistic deck in their favor prior to invoking a traditional linguistic tool, procedures were first conducted to measure, within every case in the study sample: (a) political ideology of the judge authoring the majority opinion in the case; (b) political valence of the case outcome; and (c) political chargedness rating of the case.

To measure political ideology, for every judge, research was conducted to determine (a) for appointed judges, the political party official of the appointing official; (b) for elected judges, the political party whose platform the judge ran under; and (c) for elected and appointed judges, other publicly available information regarding their political affiliation (e.g. status as a politician prior to serving on the bench).

Based on this information, a judge was coded as either being (a) Republican; (b) Democrat; or (c) Unknown. In case of conflicting information between affiliation of appointed officials and that of the judge, precedence was given to the information regarding the judge. In case of political affiliation that did not fall within the two-party system, a judge's ideology was coded as "unknown" unless this party was widely regarded as being aligned with the goals of one side or the other.[334]

Next, to measure the political valence of the case outcome, an LLM prompting experiment was conducted in which the majority opinion of each of these cases was fed into the context window of GPT-4.1 (prompted through the API), following a prompt that asked the model the following:

> "The following is the text of the majority judicial opinion. Read the text and rate, on a scale of 1-5, the political valence of the case outcome (1 representing very conservative, 5 representing very liberal)."

In cases where the opinion was available via text format, the text was fed directly to the model via the user prompt. Approximately half

334. This occurred in just one case: Bruce F. Beilfuss of the Wisconsin Supreme Court in Schmidt v. Dep't of Res. Dev., 158 N.W.2d 306 (Wis. 1968) was affiliated with the Progressive party, which according to sources was widely considered to be left-leaning. *See* Richard M. Valelly, *Wisconsin Progressive Party*, *in* POL. PARTIES AND ELECTIONS IN THE U.S.: AN ENCYCLOPEDIA 1222 (1991) (describing that the Wisconsin Progressive Party grew out of "a liberal faction within the Wisconsin republican Party known as 'the Progressives'"); *Gwin Loser in Clark County*, Marshfield News-Herald, 1 (Sep. 18, 1940) ("In the contested races for nomination to county offices on the Progressive ticket, Bruce Beilfuss, Abottsford attorney, defeated Hugh F. Gwin, incumbent district attorney, by a majority of 195, the unofficial final vote being Beilfuss 1,354 and Gwin 1,159.").

of the cases were not available in plain text format. In such cases, the opinion was instead fed to the model as a pdf attachment.

The same procedure was conducted on a random sample of 2,000 cases (1,000 federal and 1,000 state) as a baseline.

Political chargedness ratings were calculated using a similar procedure, outlined in the Easy Case Robustness Check above.

After collecting these measures, a number of descriptive statistics were computed with respect to the ideological valence of the judges, as well as the political valence of the outcome. In addition, a number of hypothesis tests were conducted to test the robustness of the independence assumption.

2. Judge Ideology

A breakdown of judge ideology is shown in Table A.7. Descriptively, across the 180 case sample, there was a relatively even balance of Democrats (84) and Republican (80) judges, along with 16 judges of an unknown political affiliation (e.g. bankruptcy and magistrate judges).

Table A.7: Breakdown of Political Ideology of Judges in Study Sample

Political Affiliation	**N**	**%**
Democrat	84	46.7
Republican	80	44.4
Unknown	16	8.89

3. Political Valence

A breakdown of average political valence of cases is shown in Table A.8. If judges were stacking the deck towards conservative outcomes by invoking plain meaning, one might expect in the sample for there to be disproportionately conservative outcomes, relative to random samples of cases. Contrary to this prediction, statistical tests (Mann-Whitney) found no evidence that case outcomes in the experimental sample were rated as more conservative than those in the random control sample of 2,000 federal and state cases ($p = 0.622$).

Table A.8: Breakdown of Political Valence of Cases in Experimental Sample v. Control Sample

Sample	Mean Valence	SD	N
Experiment	2.89	0.49	180
Control	2.91	0.511	2000

Even if judges are not stacking the deck towards conservative outcomes, perhaps both conservative and liberal judges are stacking the deck towards their preferred outcome (and with the judicial ideology being split, this might wash out to be ideologically neutral in the aggregate). If so, one would expect more liberal outcomes in cases authored by Democrat judges relative to those authored by Republican judges.

Contrary to this prediction, statistical tests (Mann-Whitney) found no evidence that case outcomes in cases authored by Democrat judges were more liberal than those authored by Republican judges. This held true when coding unknown variables as Democrat ($p = 0.640$), Republican ($p = 0.874$) or excluded from analysis ($p = 0.747$). Descriptively, cases involving Democrat judges had the same political valence score as those involving Republicans (2.89).

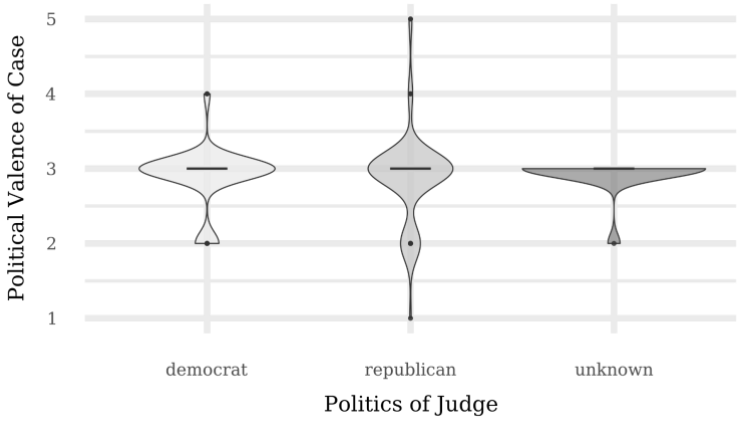

Figure A.16: Political valence of case outcome by politics of judge. Width of violin plots represent the density of cases at a particular valence value. Valence is on a scale of 1-5, with one representing very conservative and five representing very liberal.

4. Political Chargedness

A breakdown of average alignment with consensus by political chargedness is shown in Tables A.9 and A.10. As noted above, under a weaker version of the smokescreen hypothesis, judges might use canons as a smokescreen in politically charged cases. If so, one would expect to find lower alignment with linguistic consensus in cases with some degree of political chargedness relative to cases with a higher degree of political chargedness.

Contrary to this assumption, adding "political chargedness" as a fixed-effect predictor to the alignment regression models did not yield any significant effect of this variable on alignment for either the layperson model ($p = 0.482$) or lawyer model ($p = 0.805$).

Table A.9: Breakdown of Court Alignment with Consensus by Political Chargedness of Case (1-2 v. 3-5)

Chargedness Rating	**N Cases**	**Layperson Alignment**	**Lawyer Alignment**
High (3–5)	16	75%	87.5%
Low (1–2)	164	77.4%	78.7%

Table A.10: Breakdown of Court Alignment with Consensus by Political Chargedness of Case (1 vs. 2-5)

Chargedness Rating	**N Cases**	**Layperson Alignment**	**Lawyer Alignment**
Some (2–5)	67	77.6%	80.6%
None (1)	113	77%	78.8%

5. Linguistic Clarity

As an additional robustness check on whether the invocation of plain meaning (rather than the invocation of the canon in support of plain meaning) is a smokescreen, an exploratory analysis was conducted to measure the relationship between linguistic clarity and the probability of a judge invoking plain meaning.

Results are visualized in Figure A.17. As shown in the figure, the probability of a judge invoking plain meaning increases monotonically as a function of the level of clarity in the text. This result is consistent both with judges being sensitive to clarity when deciding to invoke plain meaning as well as with the notion of individual judges exercising discretion (based on factors aside from clarity) in deciding whether to do so.

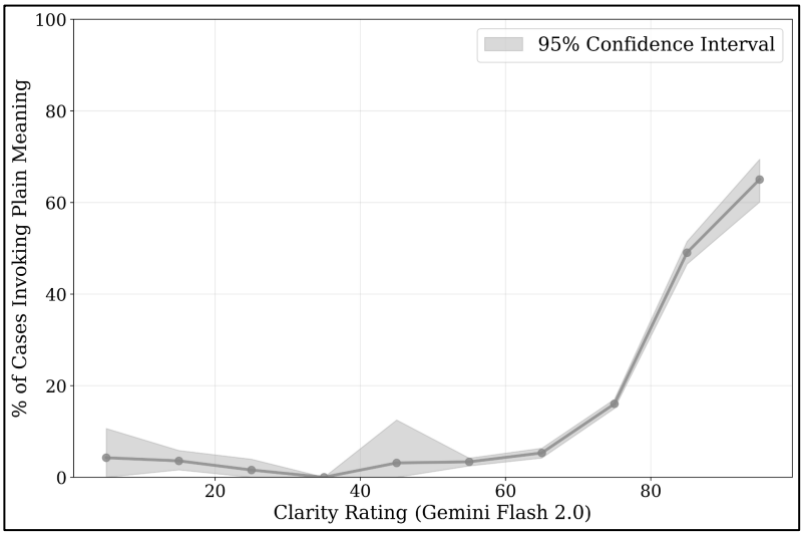

Figure A.17: Likelihood of court invoking plain meaning as a function of clarity rating. X axis represents the average clarity rating of the words at issue of a given interpretation case as judged by Gemini Flash 2.0. Y axis represents the percentage of cases, as judged by Gemini Flash 2.0, where the court judged the meaning of the text to be clear.

PART 3: SUPPLEMENTAL INFORMATION FOR STUDY II

A. Details of Analysis Plan

To evaluate the effectiveness of different models in alignment with consensus, descriptive statistics and parameter estimates were conducted for each model. The outcome variable was "agreed_with_LLM," which took a participant's response and coded it as "1" if it was the same as the LLM; and "0" if it was different. Because the response of an LLM was probabilistic, "agreed_with_LLM" was similarly computed probabilistically — in particular, for every participant's response, "agreed_with_LLM" was computed based on the probability that the LLM gave the same response as that participant (e.g. if the participant gave a "yes" response and the LLM gave "yes" response to that question 95% of the time, then agreed_with_LLM had a 95% chance of being coded as "1" and 5% chance of "0").

To formally evaluate whether novel computational tools can be effective proxies for linguistic consensus relative to judges' use of traditional interpretive tools, mixed-effects logistic regressions were conducted featuring data from both the human experiments and the

LLM prompting experiments — in particular, from the best-performing pre-o1 model from the prompting experiments.

The outcome variable was "agreed_with_LLM_or_judge." The predictor variables were adjudicator and condition as fixed effects; and participant and item as random intercepts.

Similar analyses were conducted with respect to the certainty and prediction data. In addition, a series of control analyses were conducted to assess the robustness of the results to demographic predictors, as well as to account for potential data contamination.

To formally evaluate whether novel computational tools are as effective proxies for linguistic consensus as a "neutral" lawyer or layperson, additional mixed-effects logistic regressions were conducted featuring data from responses to the prediction questions in the human and LLM prompting experiments.

The outcome variable of this regression was "prediction", which was a value between 0 and 100 taken from participant and LLM responses to the prediction questions.

In the case of the LLM condition, a distribution of values was generated that was equivalent in size to the different participant groups and based on the extracted log probabilities of different responses (for example, supposing (a) 200 lay participants gave predictions to item 1 of the *expressio unius* thrust condition, and (b) the extracted probabilities of the LLM responses to this question were 0.8 for a response of "90%" and 0.2 for a response of "100%", a distribution of 200 LLM predictions was generated such that 160 of those responses were "90" and 40 of those responses were "100".)

The predictor variables of these regressions were adjudicator (LLM and lawyer as the levels, contrast coded) and canon (contrast-coded) as fixed effects; and participant and item random intercepts.

In terms of confirmatory criteria, the prediction was that if LLMs are as effective as a neutral judge's predictions, then there would either be no main effect of adjudicator/LLM_vs_judge, or positive main effect of LLM condition, such that LLMs would have just as high or significantly higher alignment with consensus than judges.

Results of this analysis are reported in the main text.

B. Alignment Control Analysis

In the main text, LLM alignment analyses were conducted on the basic prompt — that is, how often did the LLM's response to the "yes/no" question coincide with participants' response to that question.

To assess the robustness of this result, an additional control analysis was conducted on the certainty responses of all pre o1 models. In particular, parameter estimates were conducted assessing the percentage of trials in which the LLM and participant's level of

certainty in the "Yes" response were both either (a) at or above fifty; or (b) below fifty.

Results were consistent with the main analyses and are visualized in Tables A.11 and A.12.

Table A.11: Mean Alignment of LLM with Layperson Certainty Responses, along with bootstrapped confidence interval estimates of the mean.

Model	Mean	Lower CI	Upper CI
gpt-4o	0.665	0.660	0.670
gpt-4	0.652	0.647	0.657
gpt-3.5	0.632	0.627	0.637

Table A.12: Mean Alignment of LLM with Lawyer Certainty Responses, along with bootstrapped confidence interval estimates of the mean.

Model	Mean	Lower CI	Upper CI
gpt-4o	0.671	0.664	0.678
gpt-4	0.629	0.622	0.635
gpt-3.5	0.592	0.585	0.599

C. AI Versus AI Analysis

The analyses in the main text focused on evaluating a number of models by OpenAI. One natural question is whether these results would hold when prompting AI models from other labs, such as Anthropic, Google, or Deepseek. To investigate this question, an additional comparison was made analyzing the degree of alignment between different AI models and (a) human participants; and (b) other AI models.

1. Methods

To investigate this, a number of flagship models were given an interpretation prompt using approximately the same protocol as for the OpenAI models. Models included flagship models from DeepSeek

(v3), Google (Gemini Flash 2.5 preview), and Anthropic (Claude 4 Sonnet).

As with GPT models, the interpretation prompt asked the LLM to provide their interpretation of the legal text (yes/no) as applied to the facts of a case.

Temperature settings were kept to a default. Given that log probability extraction is not permitted for these models, these models were prompted five times on each material.

2. Analysis Plan

A modal response was computed for each model and item combination. Next, descriptive statistics were computed to measure the percentage of cases in which a given model's modal interpretation aligned with (a) other models; and (b) lay participants.

3. Results

Results are visualized in Figure A.18. For every given model pair, the modal interpretation across all materials was the same more than 70% of the time. The same was true when comparing models to the modal lay participant judgment.

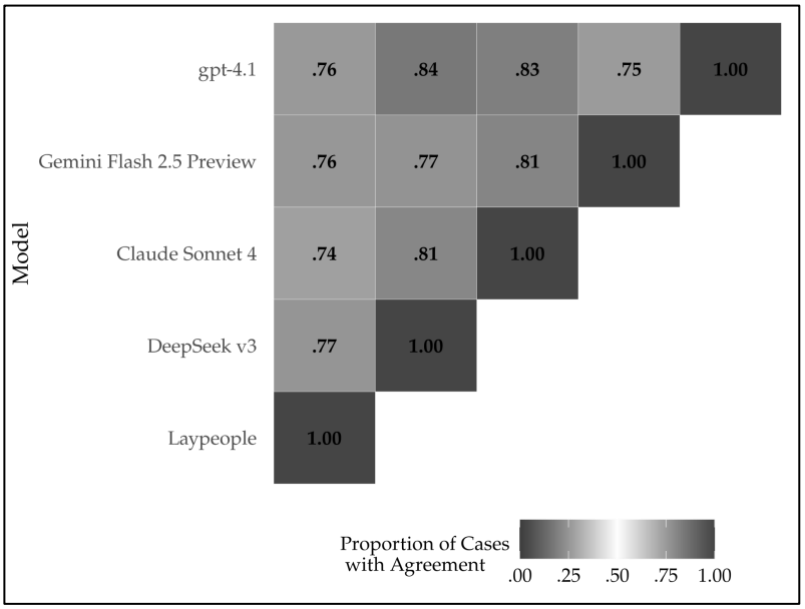

Figure A.18: Alignment of different AI models (a) with each other; and (b) with human participants, as defined by the proportion of cases with the same modal interpretation.

D. Data Contamination Control Analysis

One possible objection to the alignment data is that the cases were already in the LLMs' training dataset, and the LLM was simply regurgitating its knowledge of the case.

To account for this possibility, an additional control analysis was conducted in the highest-performing pre-o1 models that consisted of (a) prompting the LLM to assess whether it could correctly guess the name of the case from which an item was derived; and (b) assessing how well the level of alignment in cases with and without accurate recall of the case.

The results are shown in Tables A.13 and A.14. With respect to how well the models knew the cases, neither GPT-4 nor GPT-4o knew the vast majority of cases. GPT-4 failed to accurately guess 95% of cases. GPT-4o failed to accurately guess 89.4% of cases.

In terms of how well the models performed when accounting for potential recall, both GPT-4 and GPT-4o aligned, on average, with a supermajority of participants across cases when removing items whose case name the LLMs were accurately able to guess.

Table A.13: Proportion of Items in which the LLM accurately recalled the case name, broken down by model

Model	Proportion of Items with Knowledge of Case
gpt-4	0.05
gpt-4o	0.106

Table A.14: Mean Alignment of LLM with Lawyer Certainty Responses, broken down by knowledge of case name

Model	Knew case?	mean	lower_ci	upper_ci
gpt-4	No	0.710	0.703	0.717
gpt-4o	Yes	0.685	0.664	0.708
gpt-4o	No	0.675	0.667	0.681
gpt-4	Yes	0.574	0.540	0.605

E. Exploratory Analyses

To further assess the consistency of the results beyond the above factors, a number of exploratory analyses were conducted based on court level, document type, and jurisdiction (state vs federal).

Results are provided in Tables A.15, A.16, A.17, A.18, A.19 and A.20.

Table A.15: Mean Alignment of LLM with lay interpretations, broken down by jurisdiction

Jurisdiction	Mean Alignment	Lower CI	Upper CI
state	0.696	0.690	0.702
federal	0.662	0.655	0.669

Table A.16: Mean Alignment of LLM with lay interpretations, broken down by legal text genre

Legal Text Genre	Mean Alignment	Lower CI	Upper CI
private	0.702	0.691	0.712
public	0.674	0.668	0.679

Table A.17: Mean Alignment of LLM with lay interpretations, broken down by court level

Court Level	Mean Alignment	Lower CI	Upper CI
circuit	0.704	0.696	0.711
district	0.699	0.688	0.709
supreme	0.647	0.639	0.655

Table A.18: Mean alignment of LLM with lawyer interpretations, broken down by jurisdiction

Jurisdiction	Mean Alignment	Lower CI	Upper CI
state	0.715	0.707	0.724
federal	0.691	0.680	0.700

Table A.19: Mean alignment of LLM with lawyer interpretations, broken down by legal text genre

Legal Text Genre	Mean Alignment	Lower CI	Upper CI
private	0.719	0.704	0.733
public	0.699	0.692	0.706

Table A.20: Mean alignment of LLM with lawyer interpretations, broken down by court level

Court Level	Mean Alignment	Lower CI	Upper CI
circuit	0.719	0.707	0.730
district	0.698	0.683	0.712
supreme	0.692	0.682	0.703

F. Alignment by Consensus

The following plots visualize how well GPT-4.1 aligned with the modal interpretation at various levels of consensus.

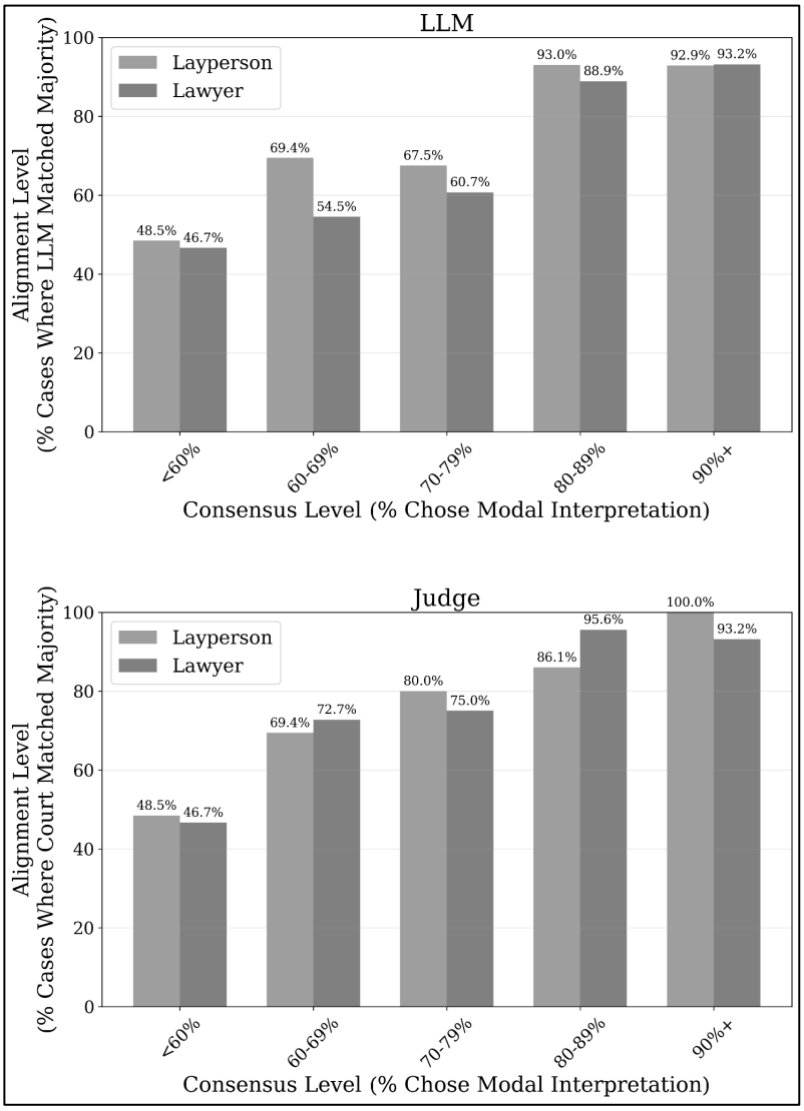

Figure A.19: AI and Judge alignment rates by level of consensus. Bar heights represent the percentage of cases in which the court aligned with the modal interpretation at a particular consensus threshold, defined by the percentage of participants who chose the modal interpretation.

G. Analyzing the Effect of Temperature

In the main text, temperature values were kept to a default of 1, due to previous work showing a null effect of temperature on model

performance. In the case of the basic prompt (i.e. giving the model the same text as that given to humans and asking for its yes/no response), this resulted in models generally giving the same response at an extremely high rate (approximately 97% across all trials).

To test the robustness of the no-temperature effect, as well as to see if higher temperature levels might result in a more human-like distribution of responses, an additional prompting experiment was conducted which prompted a large language model on the study materials at varying temperature levels.

1. Methods

In particular, the model Gemini 2.0 Flash from Google was prompted 100 times on each of the 180 cases at nine temperature levels: 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, and 2.

The Gemini model was chosen both due to its significantly lower cost than OpenAI models and due to its comparable performance in the AI v. AI Analysis above. In addition, given that the temperature function is generally considered to work analogously across LLMs, one can be reasonably confident that these results would generalize to the other models.

Following prompting, the distribution of model responses to each item at each temperature level was computed, and the model's performance was computed in three different ways.

First, as in the main study, the model's performance was measured in terms of the percentage of cases in which the modal response of the LLM matched that of the human.

Second, the model's distribution error rate was computed in terms of the absolute difference between the percentage of yes responses by the model and by that of the human.

Third, the overall consistency of the model's responses was computed in terms of calculating the percentage of trials in which the model returned its modal response.

2. Results

Results are visualized below. Consistent with prior work, performance was largely stable across temperature. The percentage of cases in which the modal response of the LLM matched that of the human was for all temperature combinations either 76.1% or 76.7%, or 137 to 138 cases out of 180.

In terms of consistency of response, even at the highest temperature setting (2), the model returned the modal response (between yes and no) 95.7% of the time for a given prompt, compared to 99.7% of the

time for the lowest temperature setting (0) and 97% of the time at the default temperature setting (1).

In terms of distribution error rate, the model's error rate (i.e., percentage of "yes" responses relative to that of the lay population) was above 30% on average at every temperature setting. It was lowest at the highest temperature setting, with an error rate of 30.7% at temperature setting of 2, compared to 33.6% at a temperature setting of 1.

Table A.21: Exploratory Temperature Analysis Results

Temperature	**Mean Error rate**	**% Cases Aligned With Lay Consensus**	**% Trials LLM Chose Same Response**
0	33.6	76.7	99.7
0.25	33.2	76.7	99.2
0.5	32.8	76.1	98.5
0.75	32.2	76.1	97.7
1	31.6	76.7	97
1.25	31.3	76.1	96.6
1.5	31.2	76.1	97
1.75	30.9	76.1	95.8
2	30.7	76.7	95.7

H. Prediction Accuracy

In the main text, prediction accuracy results are reported in binary terms: That is, for each trial, did the LLM's prediction of how many people would endorse a given interpretation match the majority interpretation (e.g. if the majority of people responded "yes," did the model with at least 51% "yes").

Here we report the prediction accuracy at a more fine-grained level, in terms of the absolute deviation of the LLM's prediction from the true percentage of respondents who adopted the modal interpretation (e.g. if 87% of people responded "yes" for a particular case, how far off was the model's prediction).

Table A.22: Mean Prediction Error Rate of Different Groups by Consensus Level

Consensus Level	LLM→ Lawyers	LLM→ Lay-people	Lawyers → Lawyers	Lawyers → Lay-people	Lay-people → Lay-people
< 60	35.6	19.4	28.8	18.8	24.5
60-69	27.9	18.6	27.8	24.5	26.1
70-79	30.8	19.2	22.8	21.9	24.4
80-89	15.6	12.8	18.3	23.2	21.6
90+	10.9	13.8	12.2	20.6	15.7

Table A.23: Mean Prediction Error Rate of Different Groups

Category	Mean Error	CI Lower	CI Upper
Laypeople → Laypeople	22.7	22.5	22.9
Lawyers → Lawyers	20.9	20.6	21.3
Lawyers → Laypeople	21.9	21.2	23
LLM → Laypeople	16.7	14.7	18.9
LLM → Lawyers	22.4	19.6	25.3

Table A.24: Percentage of Cases in Which Different Groups Overestimated the Percentage of People Who Adopted Their View

Category	% Cases	CI Lower	CI Upper
Laypeople → Laypeople	65	64.5	65.4
Lawyers → Lawyers	70.7	70	71.4
Lawyers → Laypeople	51.1	50.4	51.8
LLM → Laypeople	36.1	29.4	43.3
LLM → Lawyers	38.3	31.1	45.6

I. Few-Shot Prompting

To further investigate whether different hyperparameters might improve model performance, a follow-up prompting experiment was conducted in which an LLM was given, as part of the system prompt, varying amounts of examples of a case, along with the "correct" answer (i.e., the percentage of the relevant readership who responded "yes" to a given question).

The number of examples varied between two and eighteen. All examples sets were drawn from the same thrust/parry combination as that of the question that the model was prompted on in a given trial (one-half from the parry, one-half from the thrust).

Results are visualized below. Descriptively, prediction accuracy in all of the example conditions were higher than in the baseline conditions, both for laypeople and lawyers. In both the baseline and example conditions, error rates were lower in the LLM's predictions of laypeople consensus relative to lawyer consensus.

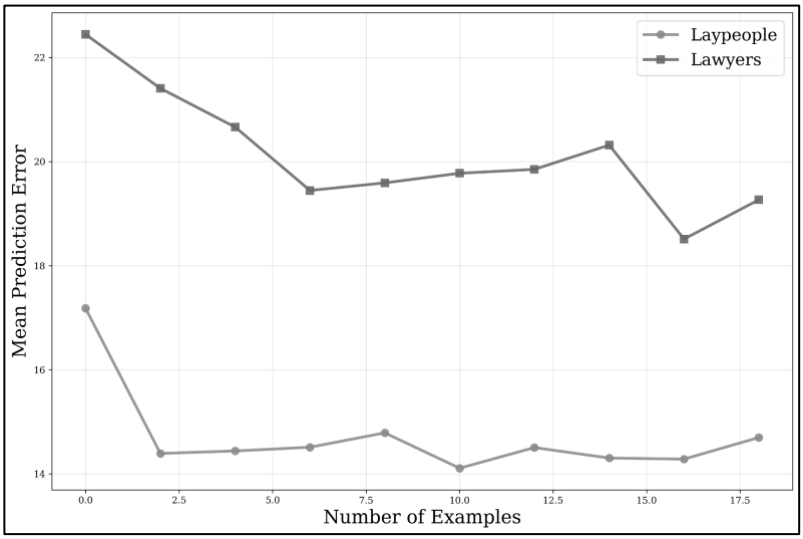

Figure A.20: Mean prediction error rate of LLM by number of examples given to the model in few-shot prompting experiment. Error rates when predicting laypeople consensus are coded in lighter dots. Those of lawyer consensus are coded in darker dots.

PART 4: FORMALIZATION OF CANON (IN)DETERMINANCY, WINDOW-DRESSING, & COGNITIVELY CONSTRAINED FORMALIST MODELS

The purpose of this Part is to formalize four models examined in the main text: (1) the canon indeterminacy model; (2) the canon determinacy model; (3) the window-dressing model; and (4) the cognitively constrained formalist model.

To that end, this Part formally expresses the background assumptions, premises, testable empirical predictions, and falsification criteria for each model.

Note that although definitions and assumptions may be repeated between models, symbols shared across models by default should be assumed to have the same meaning.

A. Canon-I

Table A.25: Notation for canon indeterminacy model

Symbol	**Description**	**Range**
$C = \{1, \ldots, N\}$	Universe of cases invoking "plain meaning"	N large/unknown
$S = \{k_1, \ldots, k_s\} \subseteq C$	Survey sample of cases	$s \leq N$
Ω	Relevant readership (ordinary / expert)	large, finite
Φ_k	Words at issue in case k	text string
$\mathcal{I}_k = \{i_{k1}, \ldots, i_{kmk}\}$	Competing interpretations in case k	$m_k = 2$
π_{kj}	Share of Ω choosing interpretation i_{kj}	[0, 1]
$Con(k) = max_j\, \pi_{kj}$	Consensus index for case k	$[1/m_k, 1]$
τ	Consensus / determinacy threshold	.5–.67
$A_k = \mathbb{1}\{Con(k) > \tau\}$	1 iff case k is *determinate*	{0, 1}
$I_k = 1 - A_k$	1 iff case k is *indeterminate*	{0, 1}

$q = P_{k \in C}(I_k = 1)$	Population share of indeterminate cases	[0, 1]
λ	Indeterminate share required by H_{CI}	$0 \leq \lambda \leq 1$
$\mathcal{T}$	Set of linguistic tools	finite
$\mathcal{T}(i_{kj})$	Tools that favor i_{kj} (CU: non-empty)	subset of $\mathcal{T}$
ρ_{kj}	Force of $\mathcal{T}(i_{kj})$	$\mathbb{R}_{\geq} 0$
$\bar{I}_s = (1/s)\sum_{k \in S} I_k$	Sample share of indeterminate cases	[0, 1]

1. Background Assumptions

a. Canon Universality Principle (CUCI)

$$\mathcal{T}(i_{kj}) \neq \emptyset \quad \text{for every } i_{kj} \in \mathcal{J}_k$$

- Definitions.
    - Let C ={1,...,N} be the universe of all cases in which the court invokes "plain meaning."
    - For each case k let $I_k = \{ i_{k1}, \ldots, i_{kmk} \}$ be the rival interpretations of the disputed text Φ_k.
    - Let T denote the master toolbox of interpretive tools (linguistic canons, dictionaries, precedent).
    - For any interpretation i_{kj} the notation $T(i_{kj}) \subseteq T$ denotes the subset that supports that reading.

- Meaning. The axiom states that every candidate interpretation has at least one corresponding interpretive tool that supports it: T (i_{kj})$\models \emptyset \; \forall \; i_{kj}$. This codifies Llewellyn's maxim that "there is always a counter-canon" (though not necessarily one of equal force).

- Practical implication. CU guarantees that a non-negative "canon force" measure ρ_{kj} (defined in CCM_{CI} below) is well-defined for every i_{kj}.

- Scope note. In line with the main text, the notion of "tool" is broad so as to include any source a court routinely cites to support a plain meaning judgment.

b. Canon–Consensus Monotonicity Principle (CCM_{CI})

$$\rho kj = f(\pi kj), \qquad f'(x) > 0$$

- Definitions. Let π_{kj} be the proportion of the relevant readership Ω who select interpretation i_{kj}. Let $\rho_{kj} \geq 0$ quantify the *force* of the supporting tool T (i_{kj}).

- Meaning. A strictly increasing function f links linguistic consensus and canon force: if an interpretation commands a larger reader share it *necessarily* entails a stronger supporting canon. Formally,

$$\pi_{kj} > \pi_{k\ell} \Longrightarrow \rho_{kj} > \rho_{k\ell}, \qquad \pi_{kj} = \pi_{k\ell} \Longrightarrow \rho_{kj} = \rho_{k\ell}$$

- Intuition. Under plain meaning, linguistic tools are treated as applicable insofar as they mirror linguistic understanding; a meaning that resonates more with the relevant readership is backed by a stronger, more applicable linguistic tool.

- Link to indeterminacy. If $\text{Con}(k) = \max_j \pi_{kj} \leq \tau$, CCM_{CI} implies $\rho_{k1} \approx \rho_{k2} \approx \ldots$; no side has a decisive edge. That is, for every canon there is a counter-canon *of equal force*. Conversely, strong linguistic consensus ($\text{Con}(k) > \tau$) entails a dominant toolset.

2. Canon Indeterminacy Model

$$\text{Con}(k) = \max_j \pi_{kj,}$$

$$A_k = \mathbb{1}\{\text{Con}(k) > \tau\}, \quad I_k = 1 - A_k, \quad q = P_{k \in C}(I_k = 1)$$

$$H_{\mathrm{CI}}(\lambda):\quad q \geq \lambda, \qquad 0 \leq \lambda \leq 1$$

- Definitions.
  - π_{kj} – share of the relevant readership that prefers interpretation i_{kj}.
  - Con(k) – modal share in case k; the larger of the π_{kj}'s.
  - τ – consensus cut-off that marks determinacy.
  - A_k – indicator that case k is *determinate* (Con(k) > τ).
  - I_k – indicator that case k is *indeterminate* (Con(k) ≤ τ).
  - q – population fraction of indeterminate cases.
- Meaning.
  - The hypothesis asserts that some large fraction of plain meaning cases are indeterminate, such that neither candidate interpretation commands a sufficiently high level of consensus among ordinary or expert readers.
  - Formally, the hypothesis places a *lower bound* on the proportion of indeterminate disputes in the entire "plain-meaning" universe:

$$q = P_{k \in C}(I_k = 1) \geq \lambda.$$

  - Equivalently, the fraction of determinate cases is bounded above:

$$P_{k \in C}(A_k = 1) = 1 - q \leq 1 - \lambda.$$

- Strong vs. weak. λ and τ specify the strength of the hypothesis. For example, $\lambda = 1$ and $\tau = 0.5$ yield the strongest version of the claim (approximately zero determinate with more than 0.5 consensus), similar to that specified in the preregistered analyses of the main text. $0 < \lambda < 1$ yields a weaker claim that still requires the share of indeterminate cases to be no less than λ.

- Tool-level implication (via $\mathrm{CCM}_{\mathrm{LI}}$). Because $\rho_{kj} = f(\pi_{kj})$ with $f'(x) > 0$, every indeterminate case must satisfy $\rho_{k1} \approx \rho_{k2} \approx \ldots$; hence for at least $\lambda \times 100\%$ of disputes each canon has a counter-canon of roughly equal force.

3. Canon Determinacy Model

$$H_{CD}(\lambda): \quad q < \lambda, \qquad 0 \leq \lambda \leq 1$$

- Definitions. As in the canon indeterminacy model.

- Meaning. The model posits that only a small fraction of cases are linguistically indeterminate: the population share satisfies $q < \lambda$. Equivalently, at least $1 - \lambda$ of disputes are *determinate*, meaning one interpretation commands high consensus.

- Strong vs. weak. Smaller λ makes the claim stronger (more determinacy required), larger λ weakens it.

- Tool-level implication (via CCM_{CI}). Because $Con(k) > \tau$ in most cases, the dominant interpretation enjoys a *stronger* supporting canon set: $\rho_k* \gg \rho_{k' \neq k}*$.

Thus, canons rarely come in perfectly balanced pairs.

4. Predictions

Let

$$\bar{I}_s = (1/s) \sum_{k \in S} I_k \qquad \text{(sample share of indeterminate cases)}$$

$$\bar{I}_s \geq \lambda \implies \text{supports Canon Indeterminacy Model}$$

$$\bar{I}_s < \lambda \implies \text{supports Canon Determinacy Model}$$

Interpretation. The two boxed rules provide a simple decision framework:

- *Support for Indeterminacy.* Whenever the observed indeterminacy rate $\bar{I}_s$ meets or exceeds the pre-specified threshold λ, the data are consistent with the Canon Indeterminacy Model. The larger the gap $\bar{I}_s - \lambda$, the stronger the support.

- *Support for Determinacy.* When $\bar{I}_s < \lambda$, the evidence instead favors the Canon Determinacy Model, with greater departures below λ implying stronger determinacy.

- *Sampling uncertainty.* In practice one should attach sampling error to $\bar{I}_s$ (e.g., a 95% confidence interval). If the entire interval lies above (below) λ the evidence robustly supports the Indeterminacy (Determinacy) model; overlap with λ warrants a more cautious interpretation.

B. Window-Dressing Model

Table A.26: Notation for the window–dressing model

Symbol	**Description**	**Typical range**
$C,\ S$	Universe of "plain-meaning" cases / empirical sample	N large; $s \leq N$
$\mathcal{J}_k$	Competing interpretations in case k	$m_k = 2$ common
i_k^*	Modal interpretation (highest π_{kj})	singleton
J_k	Interpretation adopted in the opinion	element of $\mathcal{J}_k$
$\mathcal{T}$	Toolbox of linguistic canons	finite
$\mathcal{T}(i_{kj})$	Tools supporting i_{kj}	subset of $\mathcal{T}$
T_k	Tool actually cited	element of $\mathcal{T}$

d_{kj}	(Unobserved) policy desirability of i_{kj}	$\mathbb{R}$
B_k	$\mathbb{1}\{J_k = i_k^*\}$ (court matches modal interpretation)	{0, 1}
θ_k	Chance alignment $1/m_k$	.5 when $m_k = 2$
$\bar{B}_s = (1/s)\sum_{k \in S} B_k$	Sample alignment rate	[0, 1]
$\bar{\theta}_s = (1/s)\sum_{k \in S} \theta_k$	Sample chance baseline	[0, 1]
δ	Systematic excess alignment ($E[\bar{B}_s] = \bar{\theta}_s + \delta$)	$\delta > 0$
$E[\cdot]$	Expectation over the joint sampling and model distribution	—

1. Background Assumptions

Below are the background assumptions. Note that the canon universality principle is formally equivalent to that of the canon determinacy model. It is repeated here both for convenience and to clarify its functional role in the window-dressing model.

a. Canon Universality Principle (CUWD)

$$\mathcal{T}(i_{kj}) \neq \emptyset \quad \text{for every } i_{kj} \in \mathcal{J}_k$$

- Definitions.

- $I_k = \{i_{k1}, \ldots, i_{kmk}\}$ – rival interpretations of the words at issue Φ_k.
- $\mathcal{T}$ – judge's toolbox (canons, dictionaries, corpus snippets, etc.).
- $\mathcal{T}(i_{kj}) \subseteq \mathcal{T}$ – tools that doctrinally favor i_{kj}.

- Meaning. Every candidate interpretation is backed by *at least one* tool. In Llewellyn's language: "for every canon, a counter-canon."

- Motivation. Ensures that step (P2) of the mechanistic model — citing a tool to justify the chosen reading — is always feasible, no matter which interpretation the judge prefers for policy reasons.

b. Policy–Meaning Independence Axiom (IAWD)

$$(dk1, ..., dkmk) \perp\!\!\!\perp ik^*, \quad \forall\, k$$

- Definitions.

  - d_{kj} – policy desirability of interpretation i_{kj} to the deciding judge.

  - $\mathbf{d}_k = (d_{k1}, \ldots, d_{kmk})$ – policy vector for case k.

  - i_k^* – modal interpretation among the relevant readership (highest π_{kj}).

- Meaning. Judges' private policy valuations do not systematically align with the linguistic consensus of ordinary (or well-informed) readers. Formally, the joint distribution of the policy vector $\mathbf{d}_k$ and the modal interpretation i_k^* factors:

$$P(\mathbf{d}_k, i_k^*) = P(\mathbf{d}_k)\, P(i_k^*)$$

- Rationale and intuition.

  - *Different dimensions.* Policy preferences live on an ideological axis (e.g. pro-business vs. pro-consumer), whereas i_k^* is determined by linguistic usage.

- *Consensus across political spectrum.* Empirically, groups possessing disparate policy preferences (liberals and conservatives) rarely disagree on the meaning of words. In the present study, there was alignment between the two groups in over 90% of cases.

- *Attention check data.* Empirically, people tend to converge on a common interpretation of a text even when that interpretation goes against their policy preference. In a control experiment for Study I, the vast majority of participants (86%) correctly answered two attention checks that asked them to interpret a law whose clear meaning went against their policy preferences.

2. Mechanistic Premises of the Smokescreen Model

a. Premise P1: Policy-Maximizing Choice

$$J_k = \arg\max_j d_{kj}$$

- Definitions.
  - d_{kj} – policy desirability to the deciding judge of interpretation i_{kj}.
  - J_k – interpretation actually adopted in the majority opinion for case k.
- Meaning. The judge selects the reading that maximises her subjective policy pay-off. Linguistic considerations do not constrain this step; the choice is driven entirely by $\mathbf{d}_k$.
- Rationale.
  - Judges, like other political actors, have ideological priors and constituency concerns.
  - When doctrinal tools are plentiful (TU_{SS}), they face little cost in first choosing on policy grounds.

- Implication. Because of Independence Assumption (IA), the probability that J_k coincides with the modal interpretation i_k^* is no greater than random chance.

b. Premise P2: Post-Hoc Tool Citation

$$\mathrm{Tk} \in \mathcal{T}(\mathrm{Jk})$$

- Definitions.
    - $\mathcal{T}(\mathrm{J_k})$ – subset of tools that doctrinally favor the adopted reading J_k.
    - $\mathrm{T_k}$ – specific canon, dictionary entry, etc. . .
- Meaning. After choosing J_k, the judge selects *any* tool from $\mathcal{T}(\mathrm{J_k})$ to justify that choice.
- Support from CU. Canon Universality guarantees $\mathcal{T}(\mathrm{J_k}) \neq \emptyset$, so the judge can always find at least one respectable authority.
- Rhetorical effect. The cited T_k casts the decision as a neutral exercise in textual interpretation, masking the policy motive embodied in P1.

3. Empirical Prediction

a. Alignment *with modal interpretation occurs only at chance level*

Let

$$\theta_k = 1/m_k, \qquad B_k = \mathbb{1}\{J_k = i_k^*\}$$

so θ_k is the probability of matching the modal interpretation under random choice among the m_k live readings and B_k records whether the court actually matches.

$$P(J_k = i_k^*) = \theta_k, \qquad \forall\, k \in C$$

- Derivation.
    1. *Policy-driven selection* (P1) makes J_k depend only on $\mathbf{d}_k$.

2. *Independence axiom* (IA) severs any statistical link between $\mathbf{d}_k$ and the modal interpretation i_k^*.

3. Therefore the conditional probability of matching i_k^* is exactly the chance rate $\theta_k = 1/m_k$.

- Aggregate implication.

  $E[B_k] = \theta_k$, and $E[(1/|S|) \sum_{k \in S} B_k] = (1/|S|) \sum_{k \in S} \theta_k$

  Across any representative sample, the observed alignment rate should cluster around the chance baseline.

4. Falsification Criteria

$$H_{\neg SS}: \quad P(J_k = i_k^*) = \theta_k + \delta, \qquad \delta > 0$$

A systematic excess δ implies that courts do, in fact, track ordinary meaning (and cannot be relying on tools purely as post-hoc window-dressing).

C. Cognitively Constrained Formalist Model

Below is the alternative model to the window-dressing model. Under this Cognitively Constrained Formalist model, judges use canons to support/justify a sincere attempt to arrive at plain meaning.

This model predicts that judges align with ordinary meaning at an equivalent rate to that of a neutral interpreter facing the same informational constraints.

1. Additional Notation

Note that all symbols in Table A.3 retain their meaning unless re-defined below.

Table A.27: New symbols introduced for the cognitively constrained formalist model

Symbol	Description	Typical range
$\hat{\imath}_k^*$	Judge's *perceived* modal interpretation in case k	element of $\mathcal{J}_k$
ε_k	Mis-perception probability $P(\hat{\imath}_k^* \neq i_k^*)$	[0, 1]
B_k	Alignment indicator $\mathbb{1}\{J_k = i_k^*\}$	{0, 1}

$\bar{\varepsilon_s}$	Sample error rate $1 - (1/s)\sum_{k\in S} B_k$	[0, 1]
ε^{bench}	Benchmark error among neutral readers	[0, 1]

2. Background Assumptions

The background assumptions of this model are the same as those of the window-dressing model.

3. Mechanistic Premises of the Consensus–First Model

a. Premise Q1 : Consensus-Maximising Choice with Noise

$$J_k = \hat{\imath}_k^*, \qquad P(\hat{\imath}_k^* = i_k^*) = 1 - \varepsilon_k$$

- Definitions.
    - i_k^* – true modal interpretation among the relevant readership.
    - $\hat{\imath}_k^*$ – judge's estimate of that modal interpretation.
    - ε_k – probability the estimate is wrong in case k.
- Meaning. The judge chooses the reading she *believes* aligns with linguistic consensus; Misalignment may occur only through measurement error (e.g. false consensus bias). Unlike the smokescreen model, policy preferences do not meaningfully affect the chosen interpretation, conditioning on the judge invoking plain meaning.
- Implication. The expected alignment with true consensus is $E[B_k] = 1 - \varepsilon_k$; any departure from perfect alignment is attributed solely to unavoidable noise, not ideology.
- Agnosticism about Canons. Note that the model is agnostic about how judges arrive at their linguistic judgment and what role, if any, canons play in informing this step.

b. Premise Q2 : Ex Post Tool Citation

$$Tk \in \mathcal{T}(\hat{\imath}k^*)$$

- Definitions.
    - T_k – specific tool cited in the opinion.
- Meaning. After settling on $\hat{\imath}_k^*$, the judge cites any doctrinal tool that endorses that reading.
- Support from CU. Canon Universality ensures at least one such tool is always available, so citation never constrains the decision reached in Q1.

4. Empirical Prediction

Let

$$B_k = 1\{J_k = i_k^*\}, \qquad \varepsilon_k = 1 - P(B_k = 1)$$

- Derivation.
    1. *Choice rule* (Q1) ties B_k solely to the error event $(\hat{\imath}_k^* \neq i_k^*)$.
    2. Therefore $P(B_k = 1) = 1 - \varepsilon_k$ and $E[B_k] = 1 - \varepsilon_k$.
- Aggregate implication. For any representative sample S of s cases,

$$\bar{\varepsilon}_S = 1 - (1/s) \sum_{k \in S} B_k \to E[\varepsilon_k]$$