



Alabama Law Scholarly Commons

Articles

Faculty Scholarship

5-2024

Generative Interpretation

Yonathan Arbel

David A. Hoffman

Follow this and additional works at: https://scholarship.law.ua.edu/fac_articles

NEW YORK UNIVERSITY LAW REVIEW

VOLUME 99

MAY 2024

NUMBER 2

ARTICLES

GENERATIVE INTERPRETATION

YONATHAN ARBEL[†] & DAVID A. HOFFMAN[‡]

We introduce generative interpretation, a new approach to estimating contractual meaning using large language models. As AI triumphalism is the order of the day, we proceed by way of grounded case studies, each illustrating the capabilities of these novel tools in distinct ways. Taking well-known contracts opinions, and sourcing the actual agreements that they adjudicated, we show that AI models can help factfinders ascertain ordinary meaning in context, quantify ambiguity, and fill gaps in parties' agreements. We also illustrate how models can calculate the probative value of individual pieces of extrinsic evidence.

After offering best practices for the use of these models given their limitations, we consider their implications for judicial practice and contract theory. Using large language models permits courts to estimate what the parties intended cheaply and accurately, and as such generative interpretation unsettles the current interpretative stalemate. Their use responds to efficiency-minded textualists and justice-oriented contextualists, who argue about whether parties will prefer cost and certainty or accuracy and fairness. Parties—and courts—would prefer a middle path, in which adjudicators strive to predict what the contract really meant, admitting just enough context to approximate reality while avoiding unguided and biased assimilation of evidence. As generative interpretation offers this possibility, we argue it can become the new workhorse of contractual interpretation.

[†] Silver Associate Professor, University of Alabama Law.

[‡] William A. Schnader Professor, University of Pennsylvania Carey School of Law. We thank participants at faculty workshops at Berkeley, Cornell, Minnesota, Penn, Texas A&M, William & Mary and Yale, and Aaron-Andrew Bruhl, Omri Ben-Shahar, Vince Buccola, Jon Choi, James Grimmelmann, Greg Klass, Erik Knutset, Alexandra Lahav, Jeff Lipshaw, Stephen Mouritsen, Daniel Schwarcz, Jed Stiglitz, Frank Pasquale, David Stein, Kevin Tobia, Polk Wagner. We thank Michael Hurley, Elizabeth Meeker and JD Uglum for helpful research assistance. Copyright © 2024 by Yonathan Arbel & David A. Hoffman.

INTRODUCTION	452
I. CONTRACT INTERPRETATION AS PREDICTION	461
II. GENERATIVE INTERPRETATION	473
A. <i>A Gentle Introduction to Large Language Models</i> . . .	476
B. <i>LLMs as a Source of Contractual Meaning</i>	483
C. <i>The Ambiguity Problem</i>	485
D. <i>Filling Gaps</i>	492
E. <i>From Text to Context</i>	495
III. THE FUTURE OF CONTRACT INTERPRETATION	497
A. <i>Interpretation for the 99%?</i>	499
B. <i>Beyond the Textualist/Contextualist Divide</i>	510
CONCLUSION	513

INTRODUCTION

When New Orleans’ levees broke during Hurricane Katrina, devastation, both human and economic, swept the city. And then came the lawyers. In mass contract litigation by policyholders against their insurance companies, advocates fighting over tens of billions of dollars of potential liability ultimately contested the meaning of a single word, representing a concept the companies had excluded from coverage: *Flood*.¹ Plaintiffs labored first to convince judges that *flood* might not mean water damage caused by humans, so they could then prove to a factfinder that their insurance policies didn’t contemplate damage resulting from negligence by the Army’s Corps of Engineers.² Lawyers for the defense argued that the word was unambiguous in context, covering rising waters no matter their cause, and therefore no further

¹ *In re Katrina Canal Breaches Litig.*, 495 F.3d 191, 199 (5th Cir. 2007) (“We will not pay for loss or damage caused directly or indirectly by any of the following. Such loss is excluded regardless of any other cause or event contributing concurrently or in any sequence to the loss. . . . Water[.] . . . [f]lood, surface water, waves, tides, tidal waves, overflow of any body of water, or their spray, all whether driven by wind or not . . .”).

² *In re Katrina Canal Breaches Litig.*, 495 F.3d at 197, 199, 200–01, 203–04; Brief for Appellee-Cross Appellant Humphreys at 16–18, *In re Katrina Canal Breaches Litig.*, 495 F.3d 191 (No. 07-30119), 2007 WL 4266576; Brief for Plaintiff-Appellee Xavier Univ. of La. at 17–40, *In re Katrina Canal Breaches Litig.*, 495 F.3d 191 (No. 07-30119), 2007 WL 4266583; Brief of the Chehardy Representative Policyholders in Response at 14–41, *In re Katrina Canal Breaches Litig.*, 495 F.3d 191 (No. 07-30119), 2007 WL 4266578. On the scope, source, and allocation of negligence, see ANDY HOROWITZ, *KATRINA: A HISTORY, 1915–2015*, at 1–12, 128–33 (2020); see also Campbell Robertson & John Schwartz, *Decade After Katrina, Pointing Finger More Firmly at Army Corps*, N.Y. TIMES (May 23, 2015), <https://www.nytimes.com/2015/05/24/us/decade-after-katrina-pointing-finger-more-firmly-at-army-corps.html> [https://perma.cc/ZA5Z-X6X5].

factfinding was necessary.³ Here, as so often in real court proceedings, though rarely in law school classrooms, expensive, cumbersome and unsatisfactory processes of contract interpretation took center stage.⁴

After years of litigation, the Fifth Circuit—in the best-known and most consequential contracts case of the last generation⁵—held that *flood* was unambiguous: It meant any inundation, regardless of cause.⁶ To get to that outcome, it engaged in the most artisanal and articulated form of textualism available in late-stage capitalism. The court consulted four dictionaries, one encyclopedia, two treatises, a medley of for-and-against, in-and-out-of-jurisdiction cases, and two linguistic, latinized interpretative canons.⁷ That’s on top of the four dictionaries and twenty reporter pages of caselaw analyzing the same problem in the district court.⁸

Notwithstanding such expensive and extensive efforts, the court’s interpretation has come under attack: its dictionary analysis was misleading,⁹ its canons badly deployed,¹⁰ and some of the relevant legal authorities were in fact pro-plaintiff.¹¹ Rather than reach a decision that

³ *In re Katrina Canal Breaches Litig.*, 495 F.3d at 208; Brief of Appellee State Farm Fire & Cas. Co. at 14–26, *In re Katrina Canal Breaches Litig.*, 495 F.3d 191 (No. 07-30119), 2007 WL 2466572; Brief of Appellee Allstate Ins. Co. & Allstate Indem. Co. at 16–37, *In re Katrina Canal Breaches Litig.*, 495 F.3d 191 (No. 07-30119), 2007 WL 4266556.

⁴ Benjamin E. Hermalin, Avery W. Katz & Richard Craswell, *Contract Law*, in 1 HANDBOOK OF LAW AND ECONOMICS 3, 68 (A. Mitchell Polinsky & Steven Shavell eds., 2007) (noting that interpretation is the most litigated type of contract dispute).

⁵ The opinion has been cited nearly 7,000 times over fifteen years, discussed in almost 2,000 secondary sources, and is taught to 1Ls. See, e.g., IAN AYRES & GREGORY KLASS, STUDIES IN CONTRACT LAW 701 (9th ed. 2017).

⁶ *In re Katrina Canal Breaches Litig.*, 495 F.3d at 214–19 (“The distinction between natural and non-natural causes in this context would . . . lead to absurd results and would essentially eviscerate flood exclusions whenever a levee is involved.”).

⁷ *Id.* at 210–19.

⁸ *In re Katrina Canal Breaches Consolidated Litig.*, 466 F. Supp. 2d 729, 747–63 (E.D. La. 2006).

⁹ See Natasha Fossett, *What Does Flood Mean to You? The Louisiana Courts’ Struggle to Define in Sher v. Lafayette Insurance Company*, 37 S.U. L. REV. 289, 303–06 (2010) (arguing that flood as defined in Louisiana Law had a narrower meaning than either the Fifth Circuit or the later Louisiana Supreme Court decision implied).

¹⁰ See Rachel Lisotta, *In Over Our Heads: The Inefficiencies of the National Flood Insurance Program and the Institution of Federal Tax Incentives*, 10 LOY. MAR. L. J. 511, 523 (2012) (criticizing the court for not focusing on the intent of the parties); Fossett, *supra* note 9, at 309–11 (arguing for use of the absurdity canon); Mark R. Patterson, *Standardization of Standard-Form Contracts: Competition and Contract Implications*, 52 WM. & MARY L. REV. 327, 356 (2010) (critiquing the Fifth Circuit for failing to address the significance of the relevant policy being drafted by the Insurance Service Office); Eyal Zamir, *Contract Law and Theory: Three Views of the Cathedral*, 81 U. CHI. L. REV. 2077, 2096 (2014) (critiquing the limited tools used by American courts to regulate standard form contracts, as evidenced by the court’s narrow approach in the Katrina case).

¹¹ See, e.g., *Sher v. Lafayette Ins. Co.*, 2007-CA-0757, 2007 WL 4247708 (La. App. 4th Cir. Nov. 19, 2007) (finding flood ambiguous), *rev’d*, *Sher v. Lafayette Ins. Co.*, 07-2441 (La.

followed from a constraining method, the Fifth Circuit (says its critics) merely affirmed its pro-business priors.¹² If textualism looks like another infinitely malleable and justificatory practice in high stakes cases, what good is it? But textualism's competitor, kitchen-sink contextualism, has been in bad odor for two generations, at least for the sorts of contracts that generally get litigated.¹³ Thus, contract jurists muddle along, looking for a better, more convenient path.¹⁴

In this article we offer a new approach to determining contracting parties' meaning, which we'll call *generative interpretation*.¹⁵ The idea

4/8/08), 988 So. 2d 186; *Ebbing v. State Farm Fire & Cas. Co.*, 1 S.W.3d 459, 462 (Ark. Ct. App. 1999) (holding flood excluded manmade causes); *cf. M & M Corp. of S.C. v. Auto-Owners Ins. Co.*, 701 S.E.2d 33, 36 (S.C. 2010) (finding that rainwater deliberately channeled on insured's land was not flood water).

¹² See Willy E. Rice, *The Court of Appeals for the Fifth Circuit: A Review of 2007–2008 Insurance Decisions*, 41 TEX. TECH. L. REV. 1013, 1039 (2009) (“[T]he Fifth Circuit has received some highly negative coverage in newspapers for its pro-insurer, Katrina-related decisions. . . . Without doubt, for those who believe the Fifth Circuit is a ‘pro-insurer court,’ the discussions of the outcomes and opinions in those cases will do very little to dispel that perception.”); Kenneth S. Abraham & Tom Baker, *What History Can Tell Us About the Future of Insurance and Litigation After Covid-19*, 71 DEPAUL L. REV. 169, 189 (2022) (arguing that homeowners’ unwillingness to buy federal flood insurance helped motivate strict construction of their private contracts); Thomas A. McCann, *5th Circuit Ruling: A Tough Pill to Swallow for Katrina Policyholders*, 20 LOY. CONSUMER L. REV. 100 (2007); Becky Yerak, *Insurers Win Key Katrina Ruling*, CHI. TRIBUNE (Aug. 3, 2007, 12:00 AM), <https://www.chicagotribune.com/news/ct-xpm-2007-08-03-0708020805-story.html> [https://perma.cc/X9ZM-VZXM] (noting the effect on homeowners). To be clear, the earlier ruling came under even more scrutiny. See, e.g., Walter J. Andrews, Michael S. Levine, Rhett E. Petcher & Steven W. McNutt, Essay, *A “Flood of Uncertainty”: Contractual Erosion in the Wake of Hurricane Katrina and the Eastern District of Louisiana’s Ruling in In Re Katrina Canal Breaches Consolidated Litigation*, 81 TUL. L. REV. 1277 (2007) (arguing that the district court’s finding that flood was ambiguous was wrong); Michelle E. Boardman, *The Unpredictability of Insurance Interpretation*, 82 L. & CONTEMP. PROBS. 27, 41 n.45 (2019) (calling the District Court infamous and arguing that the Fifth Circuit ruling was correct); Edward P. Richards, Essay, *The Hurricane Katrina Levee Breach Litigation: Getting the First Geoengineering Liability Case Right*, 160 U. PA. L. REV. PENNUMBRA 267 (2012) (arguing in support of the Fifth Circuit ruling).

¹³ See Lawrence A. Cunningham, *Contract Interpretation 2.0: Not Winner-Take-All but Best-Tool-for-the-Job*, 85 GEO. WASH. L. REV. 1625, 1628–31 (2018) (offering the history of contextualism versus textualism and noting a rise in the latter starting in the early 1990s). *But cf.* 5 CORBIN ON CONTRACTS § 24.7 (2023) (noting a “trend” toward abandoning plain meaning in some states).

¹⁴ Cunningham, *supra* note 13, at 1633–44 (noting proposals to compromise between the two approaches).

¹⁵ For previous discussions of the use of large language models in contracts, see Ryan Catterwell, *Automation in Contract Interpretation*, 12 L. INNOVATION & TECH. 81, 100 (2020) (showing an early version of how information can be extracted from contractual texts); Yonathan A. Arbel & Shmuel I. Becher, *Contracts in the Age of Smart Readers*, 90 GEO. WASH. L. REV. 83 (2022) (arguing that language models could serve as “smart readers” of consumer contracts); Noam Kolt, *Predicting Consumer Contracts*, 37 BERKELEY TECH. L.J. 71 (2022) (arguing that ChatGPT might be useful in helping consumers to understand their contracts and providing examples).

is simple: applying large language models (LLMs) to contractual texts and extrinsic evidence to predict what the parties said at contracting about what they meant.¹⁶ Our goal is to convince you that generative interpretation avoids some of the problems that bedeviled the Fifth Circuit in its Katrina litigation, while being materially more accessible and transparent. Giving courts a convenient way to commit to a cheap and predictable contract interpretation methodology would be a major advance in contract law, and parties may start to include them in their choice-of-law repertoire. We argue that even today's freshly-minted LLMs can be of service, although—as we shall make clear throughout—the tools used to query them still await a process of development, refinement, and validation.

Convincing judges to forgo dictionaries and canons and adopt a chat tool best known today for encouraging lawyers to submit fake authorities will be a tall order.¹⁷ We'll largely proceed by way of demonstrative case studies.¹⁸ Let's start with the word *flood*. In the Katrina case, the question was really whether the widely shared meaning of *flood* reasonably excluded manmade disasters. To answer that question you could, as the court did, turn to the traditional tools of High Textualism. Or you could survey insured citizens (if you could identify them and avoid motivated answers).¹⁹ And you might even, if you were technically sophisticated and patient enough, query a few relatively small databases and ask which words in English generally

¹⁶ Cf. Jonathan H. Choi, *Measuring Clarity in Legal Texts*, 91 U. CHI. L. REV. 1. Choi's excellent paper, though not focused on contract interpretation particularly, significantly advanced understanding of how automated interpretative methods can aid factfinders. We build on his work technically by developing new ways of interacting with large language models and incorporating context and attention mechanisms.

¹⁷ See *infra* text accompanying notes 231–33 (discussing *Mata v. Avianca, Inc.*, No. 22-cv-1461, 2023 WL 4114965 (S.D.N.Y. June 22, 2023)); see also *Ex parte Lee*, 673 S.W.3d. 755, 757 n.2 (Tex. App. July 19, 2023) (explaining the court's suspicion that counsel had filed briefs using ChatGPT and had made up cases and citations).

¹⁸ We mean this demonstration to illustrate the claim that today's models are capable of results that are sensible, predictable, accessible, and cheap. "Sensible" requires a baseline, whether the reader's linguistic priors or the court's own interpretation. As we'll repeatedly say in the text (but want to emphasize again in the notes!) sensibility is distinct from proving that a given result is *correct* (on average or in any specific case). That endeavor requires information about, and a proper definition of, the "ground truth" of the matter. That said, LLMs are meticulously tested on their response accuracy across a large variety of tasks. See, e.g., *Open LLM Leaderboard*, https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard [<https://perma.cc/V72H-9UJ4>] (ranking a number of open LLMs and chatbots across a range of metrics).

¹⁹ See Omri Ben-Shahar & Lior Jacob Strahilevitz, *Interpreting Contracts via Surveys and Experiments*, 92 N.Y.U. L. REV. 1753 (2017) (proposing using surveys to interpret certain mass contracts).

tend to occur, or collocate, with flood in newspapers, books, and the like.²⁰

But we instead turned to a convenient, free, open-source LLM tools resting on databases of trillions of words and asked them to transform words into complex vectors in a process called *embedding*.²¹ As a first cut, this process can be thought of as trying to quantify how much a word or phrase is related to a given category, or dimension. Thus, if there is a dimension for the word *water*, *fish* will score higher than *dogs*. Using an interface we developed, we queried several models about the relation of the policy exclusion term relative to words and phrases describing other potential sources of damage.²²

²⁰ See Stephen C. Mouritsen, *Contract Interpretation with Corpus Linguistics*, 94 WASH. L. REV. 1337, 1378–79 (2019) (proposing using corpus linguistics to interpret contracts).

²¹ For a survey of embedding methods, see MOHAMMAD TAHER PILEHVAR & JOSE CAMACHO-COLLADOS, *EMBEDDINGS IN NATURAL LANGUAGE PROCESSING* 27–110 (2021).

²² All of the code necessary to replicate these results, and the remaining ones in the paper, can be found at: GITHUB, <https://github.com/yonathanarbel/generativeinterpretation/tree/main>. The exclusion term is the language contained at footnote 1, *supra*. Because embeddings are vectors in high-dimensional space, we can measure the distance between them. This method has been used extensively in the literature. See Choi, *supra* note 16, at 24–26 (using method and reporting its usage and limitations). For a non-legal example, see Nitika Mathur, Timothy Baldwin & Trevor Cohn, *Putting Evaluation in Context: Contextual Embeddings Improve Machine Translation Evaluation*, in *PROCEEDINGS OF THE 57TH ANNUAL MEETING OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS* 2799 (2019). We found that while results using this method seem sensible, they are also fragile. To create a more robust measure, we relied on the embeddings of the ten top performing models today (found at <https://huggingface.co/spaces/mteb/leaderboard> on pair classification tasks) and used similar sentence structures. This approach is partly inspired by Maria Antoniak & David Mimno, *Evaluating the Stability of Embedding-Based Word Similarities*, 6 *TRANSACTIONS OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS* 107 (2018). We then calculated the cosine distance, normalized it, and reported the results in the figure below. For an elaboration on the limitations of these techniques, see *infra* notes 242–43 and accompanying text.

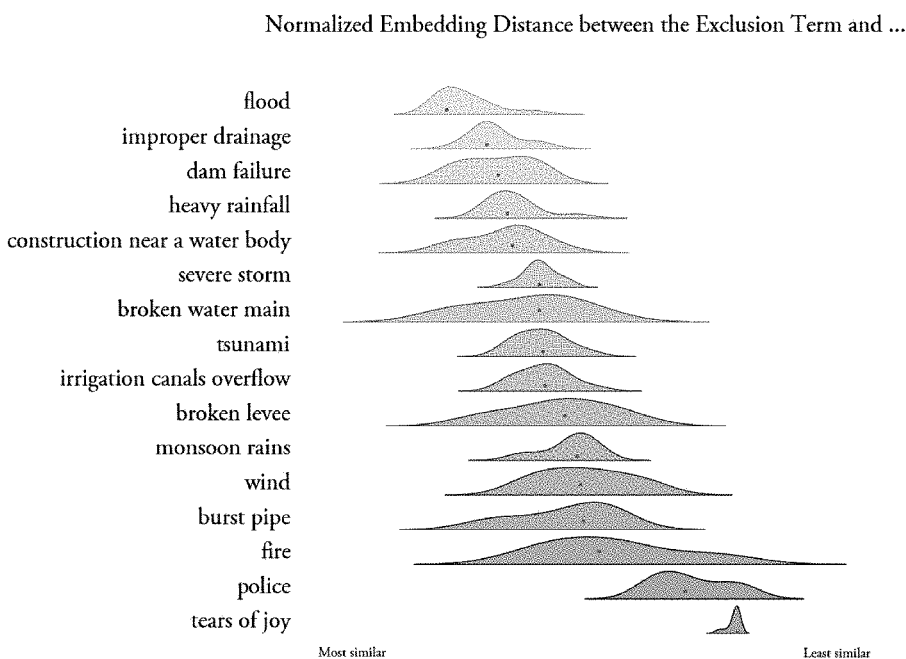


Figure 1: Analysis of the cosine distance—a measure of distance between the numerical representation of terms (embeddings)—between the exclusion clause (“We will not pay for loss or damage caused directly or indirectly by Water . . . Flood . . . all whether driven by wind or not . . .”) and various terms and phrases, as calculated by ten embedding models.

To read Figure 1, focus on the location of the red markers. The further they are from the origin, the more distant the models (on average) consider the semantic relationship between the phrases.²³ In our view, the Figure offers immediately available, objective, cheap support for the court’s judgment that floods can be unnaturally caused. Common sentences regarding floods do not distinguish between the type of cause, but seem more focused on their typicality. Our quality check terms—*tears of joy* and *police*—indeed appear farther to the right than *heavy rainfall* or *severe storm*, indicating that they are less typically associated

²³ The models we use here specialize in creating embeddings that can measure the semantic textual similarity of sentences and words. For technical background, see generally Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer & Yinfei Yang, *Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models*, FINDINGS OF THE ASS’N FOR COMPUTATIONAL LINGUISTICS: ACL 2022, 1864, 1864–65 (2022) (providing background on sentence embeddings).

with floods.²⁴ And while our experiment supports this decision of the court, it challenges others. Louisiana courts refused to exclude water main floods, even though linguistically they appear to be as much of a flooding event as any other.²⁵

Now, the model doesn't provide (nor could it) a scientific answer to the question of whether certain words are sufficiently close to make the plain meaning of *flood* unambiguous. That choice is ultimately a normative one which judges must make. But there is a bit of difference between an informed conclusion based on a statistical analysis of billions of texts and a judgment by a few dictionary editors. And there is an ocean of difference between the baroque and expensive textualism the court used and code that is cheap, replicable, quick, and most importantly, extremely straightforward to use. Simply put, generative interpretation is good enough for many cases that currently employ more expensive, and arguably less certain, methodologies. It's a workable, workmanlike method for a resource-constrained contract litigation world.

In Part I, we introduce the methodologies of contract interpretation and argue that they badly fail at their core purposes of unbiased, accessible ascertainment of what the parties would have wanted. In practice, interpretation operates as a kludgy prediction engine. Both textualism and contextualism strive to estimate what the parties would have said on a matter, accounting for realistic constraints of evidence and cost. But those constraints impose real tradeoffs and can't avoid legitimacy problems generated by courts' motivated reasoning.²⁶ We describe some modern proposed improvements on interpretation's normal science and suggest that, however promising they are, concerns about usability and cost impair their real-world utility.²⁷

²⁴ It is telling that *fire*, while having a wide distribution across models, is nearer to the origin (the embedding of the exclusion clause) than *tears of joy*. One possible explanation is that fire is a form of a commonly insured hazard, and so is conceptually nearer to the origin than the figurative expression. This demonstrates a deep point about the use of embedding distances. What gets measured is word "relatedness," and just as in everyday language, words can be related in many different ways (meaning, length, sound, analogies, register, etc). Despite that, embedding distances is a commonly used technique today and it is consistently shown to produce results that accord with human expectations. See Tomas Mikolov, Kai Chen, Greg Corrado, & Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*, ARXIV (Jan. 16, 2013), <https://arxiv.org/pdf/1301.3781.pdf> [<https://perma.cc/X3NC-EGAX>]; Jeffrey Pennington, Richard Socher, & Christopher D. Manning, *Glove: Global Vectors for Word Representation*, PROCS. OF THE 2014 CONF. ON EMPIRICAL METHODS IN NAT. LANGUAGE PROCESSING (EMNLP), 1532, 1532 (2014).

²⁵ See *Sher v. Lafayette Ins. Co.*, 988 So. 2d 186, 194 n.3 (La. 2008) ("[I]nundation of property due to broken water mains . . . would not be excluded as a 'flood' . . ."); *In re Katrina Canal Breaches Litig.*, 495 F.3d 191, 216 (5th Cir. 2007) ("[U]nlike a canal, a water main is not a body of water or watercourse.").

²⁶ See Rice, *supra* note 12, at 1039 (charging the Fifth Circuit with being pro-business).

²⁷ See *infra* Part I.

Part II is the heart of the Article. Here, we look at several types of interpretative problems generated by real contracts that produced contracts opinions. These range from the easy (what is the predicted meaning of a particular word?), to the hard (is there an ambiguity?), to the metaphysical (what did the parties mean when they clearly hadn't considered the issue?). In each example, we showcase new ways to use large language models to sharpen intuitions about the parties' presumed intent, to illuminate how transparent and objective interpretative methodologies have advantages over intuitive ones, and to suggest that generative interpretation has real promise as a judicial adjunct. The cases we run through include casebook staples, like *Trident Center v. Connecticut General Life Insurance Co.*²⁸ and *C & J Fertilizer, Inc. v. Allied Mutual Insurance Co.*,²⁹ as well as some that should be, like *Famiglio v. Famiglio*,³⁰ *Haines v. City of New York*,³¹ and *Stewart v. Newbury*.³² For many of these cases, our work is based on archival research identifying original contract materials, until now obscured by the judicial opinions that purportedly interpret them.

These case studies show how generative interpretation might be deployed in practice. As we will explore, the technology underlying large language models can do more than merely help us see if *flood* is closer to *levee* than it is to *joy*. Dictionaries, encyclopedias, or corpus linguistics can do that. What makes large language models powerful is the vastness of the data they incorporate; what makes them unique is that they wield an internal mechanism known as "attention" that allows them to account for context. And by becoming context-sensitive, these models can parse the effects of contract text from the marginal value of relevant extrinsic evidence.

Ideally, we would show you that these methods are just as correct, and just as robust, as a judge consulting a dictionary or listening to motivated testimony. We can't quite do that. In most contract cases

²⁸ 847 F.2d 564 (9th Cir. 1988), in RANDY E. BARNETT & NATHAN B. OMAN, *CONTRACTS: CASES AND DOCTRINE* 483 (7th ed. 2021); 847 F.2d 564 (9th Cir. 1988), in E. ALLAN FARNSWORTH, CAROL SANGER, NEIL B. COHEN, RICHARD R.W. BROOKS & LARRY T. GARVIN, *CONTRACTS: CASES AND MATERIALS* 560 (10th ed. 2023).

²⁹ 227 N.W.2d 169 (Iowa 1975); see also Brian Bix, *The Role of Contract: Stewart Macaulay's Lessons from Practice*, in *REVISITING THE CONTRACTS SCHOLARSHIP OF STEWART MACAULAY: ON THE EMPIRICAL AND THE LYRICAL* 241, 252 (Jean Braucher, John Kidwell & William C. Whitford eds., 2013) (describing *C & J Fertilizer* and noting that it is often assigned in casebooks, including Stewart Macaulay's and Charles Knapp's).

³⁰ 279 So. 3d 736 (Fla. Dist. Ct. App. 2019).

³¹ 364 N.E.2d 820 (N.Y. 1977), in ROBERT S. SUMMERS, ROBERT A. HILLMAN & DAVID A. HOFFMAN, *CONTRACT AND RELATED OBLIGATION: THEORY, DOCTRINE, AND PRACTICE* 834 (8th ed. 2021) (reprinting case as example of contract interpretation).

³² 115 N.E. 984 (N.Y. 1917), in SUMMERS ET AL., *supra* note 31, at 948 (reprinting case).

there is no ground truth at hand—we can't really know what the parties intended at contracting and have to make instead our best guess. “Correctness” then has to bow to “good enough.” Are LLMs able to approximate the results of courts, at lower costs, in ways that are fairly replicable and somewhat transparent? And can we offer ways by which courts might reduce parties’ abilities to game inputs into the interpretation machine, so as to reduce barriers to justice?

Current practices as to LLMs and their future uses are contingent: Lawyers tend to use tools before they are theoretically sharp.³³ In Part III, we develop a theory to justify and constrain generative interpretation going forward, as the technology that enables it continues to rapidly develop, the tools used to interact with it advance, and its use by lawyers and judges grows explosively. We make two claims.

First, the method fills a glaring need for a simple, transparent, and convenient way to commit to an interpretative method that helps predict the parties’ intent. If courts follow the evolving best practices, and we provide an initial list here, they will avoid certain access-to-justice and legitimacy problems that have beset the modern contract litigation machine. *Second*, rather than simply a marginal improvement over dictionary-and-canon textualism, or its negation as a form of 1960s-California contextualism,³⁴ use of artificial intelligence (AI) should prompt a top-to-bottom reexamination of the assumptions justifying these approaches to interpretation. As more courts commit to generative interpretation, parties may come to prefer contextual evaluation of meaning when their deals are evaluated, thus flipping a longstanding default rule in contract law.³⁵

We do consider some of the developing objections to the use of large language models, including their hallucinatory errors, biases, black-box methods, and the tension between the rapidity of their deployment and stately needs of precedential decisionmaking. As we show, generative

³³ Consider originalism.

³⁴ For defenses of contextualism, see Jeffrey W. Stempel & Erik S. Knutsen, *Rejecting Word Worship: An Integrative Approach to Judicial Construction of Insurance Policies*, 90 U. CIN. L. REV. 561, 600–01 (2021) (noting the malleability of textualist approaches and advocating for the inclusion of contextual factors in insurance contract interpretation); Jeffrey W. Stempel, *Unmet Expectations: Undue Restriction of the Reasonable Expectations Approach and the Misleading Mythology of Judicial Role*, 5 CONN. INS. L.J. 181, 183–84 (1998) (explaining how the reasonable expectations doctrine is consistent with an ethos of judicial restraint).

³⁵ In some industries, the evidence that parties would prefer that later decisionmakers incorporate context is robust. See William Hoffman, *On the Use and Abuse of Custom and Usage in Reinsurance Contracts*, 33 TORT & INS. L.J. 1, 3 (1997) (describing the origin of nonintegrated contracts in the reinsurance context); William Hoffman, *Facultative Reinsurance Contract Formation, Documentation, and Integration*, 38 TORT TRIAL & INS. PRAC. L.J. 763, 836–37 (2003) (explaining why parties in reinsurance contracts prefer custom).

interpretation's dangers illustrate its limits: Judges will have to use these engines as *tools* to excavate the normative judgments on which all interpretative and adjudicatory exercises rest. Large language models aren't robot judges. What they will do (and maybe are already doing) is help judges illuminate the degree to which we want to give the parties what they really bargained for, as best as we can.

I

CONTRACT INTERPRETATION AS PREDICTION

Jurists interpreting contracts start with a simple question: “What would the parties have said about the meaning of a disputed phrase at the time they entered the contract?”³⁶ That is, to “ascertain the parties’ intention at the time they made their contract.”³⁷ As Alan Schwartz and Robert E. Scott noted in their canonical article *Contract Theory and the Limits of Contract Law*, this question in theory has a “correct answer.”³⁸ In practice, however, it is not always easy or possible to know what it is. Lacking a time machine, adjudicators traditionally have stitched together an answer using imperfect evidence—a mix of the contract’s text, the parties’ statements about the deal (whether from before, during, or after its formation),³⁹ market data,⁴⁰ and some hunches about fairness and efficiency under the circumstances.⁴¹

³⁶ See *Bruce v. Blalock*, 127 S.E.2d 439, 442 (S.C. 1962) (“In construing the contract the Court will ascertain the intention of the parties . . . as well as the purposes had in view at the time the contract was made.”).

³⁷ STEVEN J. BURTON, *ELEMENTS OF CONTRACT INTERPRETATION* § 1.1, at 1 (2008).

³⁸ Alan Schwartz & Robert E. Scott, *Contract Theory and the Limits of Contract Law*, 113 *YALE L.J.* 541, 568 (2003) (“There is a consensus among courts and commentators that the appropriate goal of contract interpretation is to have the enforcing court find the ‘correct answer.’”) [hereinafter Schwartz & Scott, *Contract Theory*]. For criticisms, see Adam B. Badawi, *Interpretive Preferences and the Limits of the New Formalism*, 6 *BERKELEY BUS. L.J.* 1 (2009) (arguing that the frequency and uncertainty of transactions affect the usefulness of formal interpretation methods); Shawn J. Bayern, *Rational Ignorance, Rational Closed-Mindedness, and Modern Economic Formalism in Contract Law*, 97 *CALIF. L. REV.* 943 (2009) (exploring flaws in arguments that formalist approaches cost less without affecting expected results); Robin Bradley Kar & Margaret Jane Radin, *Pseudo-Contract and Shared Meaning Analysis*, 132 *HARV. L. REV.* 1135, 1182–92 (2019) (arguing that sophisticated parties would not and do not prefer acontextual readings). For Schwartz & Scott’s responses to critics, see Alan Schwartz & Robert E. Scott, *Contract Interpretation Redux*, 119 *YALE L.J.* 926 (2010) [hereinafter Schwartz & Scott, *Redux*].

³⁹ Stephen F. Ross & Daniel Trannen, *The Modern Parol Evidence Rule and Its Implications for New Textualist Statutory Interpretation*, 87 *GEO. L.J.* 195, 196–97 (1995) (noting disagreement between Williston and Corbin on parol evidence).

⁴⁰ JOHN BOURDEAU ET AL., *Course of Dealing or Usage of Trade*, in *AM. JURIS.* § 219 (2d ed. 2023) (“Under the Uniform Commercial Code, a course of dealing between the parties . . . may give particular meaning to, and supplement or qualify, terms of an agreement.”).

⁴¹ Omri Ben-Shahar, David A. Hoffman & Cathy Hwang, *Nonparty Interests in Contract Law*, 171 *U. PA. L. REV.* 1095, 1117–30 (2023) (describing courts’ use of public interests in

To put it another way, almost all jurists agree that the goal of contract interpretation—its real ambition—is to be a prediction machine.⁴² That is, to look backward and predict what the parties meant.⁴³ This seems straightforward, akin to the retrospective intent-based inquiries we see in criminal law and tort. Nonetheless, interpretation is “the least settled, most contentious area of contemporary contract doctrine and scholarship.”⁴⁴ That’s because of the many problems it seeks to solve. As Greg Klass puts it, jurists ask (1) whose meaning counts, (2) what type of meaning matters (local/majoritarian, semantic/pragmatic), and (3) what facts determine the legally relevant meaning.⁴⁵ These questions map, imperfectly, onto distinctions between textualists and contextualists. And, at the bottom of the well, contractual interpretation resolves questions of claims to judicial power, and thus legitimates violence.⁴⁶ The result is that parties contesting how to interpret contracts are sometimes arguing about what outcomes are just, not merely which are more likely to lead to parties getting what they want.

But putting aside normative questions, even basic operational empirics about interpretation—the prediction questions everyone agrees are at the core—are hard. Prediction is difficult, and mistakes are inevitable. Accuracy—in the sense of thinking that we really got as close as we could to knowing what the parties would have said—trades off against cost and certainty. Efficiency-minded scholars have repeatedly

interpreting contracts).

⁴² Schwartz & Scott, *supra* note 38, at 568 (noting “consensus” about the “appropriate goal”). There are exceptions. Eyal Zamir, for example, argues that interpretation should adhere to moral and social norms, partly because they are more likely to reflect the parties’ true intent, and partly because only those contracts are worth enforcing. Cf. Eyal Zamir, *The Inverted Hierarchy of Contract Interpretation and Supplementation*, 97 COLUM. L. REV. 1710, 1777–88 (1997). Other common reasons to deviate from the parties’ intentions include attempts to incent clearer drafting, to share valuable information, and to facilitate standardization. See, e.g., Ian Ayres, *Default Rules for Incomplete Contracts*, in 1 THE NEW PALGRAVE DICTIONARY OF ECONOMICS AND THE LAW 585 (Peter Newman ed., 1998) (reviewing the economic theories for the design of default rules). It is inevitable that the parties at times will choose not to think about a relevant possibility to minimize transaction costs or permit a deal. Therefore, when we say that the goal is prediction, consider it the beginning, rather than the end, of interpretation.

⁴³ In recent work one of us elaborates on the idea developed here of interpretation-as-prediction. See Yonathan A. Arbel, *Time and Contract Interpretation: Lessons from Machine Learning*, in RESEARCH HANDBOOK ON LAW AND TIME (forthcoming 2024, Frank Fagan & Saul Levmore eds.).

⁴⁴ Ronald J. Gilson, Charles F. Sabel & Robert E. Scott, *Text and Context: Contract Interpretation as Contract Design*, 100 CORNELL L. REV. 23, 25 (2014); Schwartz & Scott, *Redux*, *supra* note 38, at 928.

⁴⁵ See Gregory Klass, *Contracts, Constitutions and Getting the Interpretation-Construction Distinction Right*, 18 GEO. J. L. & PUB. POL’Y 13, 24–28 (2020) [hereinafter Klass, *Contracts*].

⁴⁶ See Robert M. Cover, *Essay, Violence and the Word*, 95 YALE L.J. 1601, 1601 (1986) (highlighting the relationship between legal interpretation and acts of violence).

argued that as the amount of evidence offered to prove the parties' contemporaneous-to-contracting meaning increased, so does expense across several domains.⁴⁷

As a first cut at that cost, consider that when parties are permitted to adduce additional sources of interpretative evidence, they also increase the range of defensible answers from the tribunal. This means that it becomes harder to know what the factfinder will do—their ability to choose unexpected meanings waxes with the evidentiary inputs.⁴⁸ But worse, both parties and factfinders are motivated in how they offer and process evidence.⁴⁹ In a regime that permits more evidence, parties will offer evidence that favors their view, sometimes unconsciously motivated to avoid presenting data that favors the other side;⁵⁰ factfinders, equally subject to motivated cognition, will process new evidence in biased ways.⁵¹

At the same time, as the types of evidence relevant to contract interpretation become more capacious, parties will seek to introduce more evidence at trial, raising the costs of litigation.⁵² These costs may be significant, even in dispute resolution forums like arbitration that are

⁴⁷ See generally Gregory Klass, *Contract Exposition and Formalism*, GEO. L. FAC. PUBL'NS & OTHER WORKS, at 63 (2017), <https://scholarship.law.georgetown.edu/facpub/1948> [<https://perma.cc/37M3-ZEQZ>] (“The more evidence one allows into interpretation, the less certain the outcome. The costs of such uncertainty in the contractual setting can be especially high.”); Schwartz & Scott, *supra* note 38, at 580 (2003) (“Expanding the evidentiary base is not costless, however. The parties, therefore, face a tradeoff between the efficiency of increased accuracy and the inefficiency of increased contract-enforcement costs.”).

⁴⁸ See Klass, *Contracts*, *supra* note 45, at 63 (“A party that wants to organize its behavior . . . needs to be able to predict how an adjudicator will later interpret that agreement. To the extent thicker interpretive rules reduce predictability, they impose an additional cost . . .”).

⁴⁹ See Christoph Engel, *Judicial Decision-Making: A Survey of the Experimental Evidence*, MAX PLANCK INST. FOR RSCH. ON COLLECTIVE GOODS DISCUSSION PAPER, no. 6, Aug. 2022, at 5 (2022) (noting that even when decision makers are motivated to be impartial, bias has been shown to sneak in inadvertently via race, gender, ideology, and the stereotype that tattoos are typical for criminals); Lawrence Solan, Terri Rosenblatt & Daniel Osherson, Essay, *False Consensus Bias in Contract Interpretation*, 108 COLUM. L. REV. 1268, 1269 (2008) (explaining that “false consensus bias” may cause contracting parties not to recognize different interpretations of their agreement until litigation, at which point judges fall victim to the same bias).

⁵⁰ Schwartz & Scott, *supra* note 38, at 607 (claiming that under standards allowing for recovery of “commercially reasonable” costs and investments, parties would always claim their costs were higher and their investments reasonable).

⁵¹ See, e.g., Solan, Rosenblatt & Osherson, *supra* note 49, at 1298 (“[B]oth laypeople and judges are subject to false consensus bias in deciding whether nonprototypical situations fit within contractual language.”).

⁵² For some evidence on this process in the courts, see generally Lisa Bernstein, *Custom in the Courts*, 110 Nw. U. L. REV. 63 (2015) (showing that courts accept evidence of custom that isn't systematic even in commercial disputes).

built to resolve cases quickly and cheaply.⁵³ The interpretation arms race has led scholars to model when parties would prefer to spend money ex ante on more specified text, rather than spend ex post on litigation.⁵⁴ That is, to pre-commit to methodologies which are less accurate but more efficient.

This is all familiar territory. Now, consider what interpretative methodologies have been on offer to calibrate between predictive accuracy and virtues that center around certainty and efficiency. Like other legal extrapolative enterprises, interpretation has developed two basic methods to solve for the predictive question in the absence of the ability to travel to the time of contracting.⁵⁵ These methods, textualism and contextualism, are represented in the real world by the courts in New York and California, respectively.⁵⁶

New York's textualist judges focus on the contract: They take its words as the canonical source of the parties' meaning and abjure other sources of evidence as predictive grist. Textualists try to use the common sense meaning of words, using dictionaries to obtain the public meaning of the words the parties chose, and grammatical and lexical tools to understand how the words, when collated, create obligation.⁵⁷ Textualism has known advantages, including forcing the parties to think

⁵³ See Richard A. Posner, *The Law and Economics of Contract Interpretation*, 83 TEX. L. REV. 1581, 1605–06 (2005) (arguing that commercial arbitration, where the arbitrator uses commercial common sense to predict intent rather than asking the parties to present evidence, may be preferable when the written contract does not make the parties' intentions immediately clear because it allows the parties to avoid extra expenses).

⁵⁴ Ronald J. Gilson, Charles F. Sabel & Robert E. Scott, *Braiding: The Interaction of Formal and Informal Contracting in Theory, Practice, and Doctrine*, 110 COLUM. L. REV. 1377, 1391 n.35 (2010) (“If conditions are unlikely to change much in the future (the level of uncertainty is low), and thus the ex-ante cost of writing contract rules is low relative to the anticipated gains, the parties' most cost-effective strategy is to write a complex, rule-based contingent contract.”).

⁵⁵ See John F. Manning, *What Divides Textualists from Purposivists?*, 106 COLUM. L. REV. 70, 75 (2006) (arguing that textualism and purposivism remain meaningfully distinct modes of statutory interpretation); see generally Eric A. Posner, Essay, *The Parol Evidence Rule, the Plain Meaning Rule, and the Principles of Contractual Interpretation*, 146 U. PA. L. REV. 533 (1998) (defending textualist approaches in contract law).

⁵⁶ Klass, *Contracts*, *supra* note 45, at 29 (distinguishing New York and California archetypes).

⁵⁷ Joshua M. Silverstein, *Contract Interpretation Enforcement Costs: An Empirical Study of Textualism Versus Contextualism Conducted Via the West Key Number System*, 47 HOFSTRA L. REV. 1011, 1014 (2019) (“Textualist’ judges and commentators argue that the interpretation of contracts should focus primarily on the language contained within the four corners of written agreements.”); Gilson, Sabel & Scott, *supra* note 43, at 40 (“Textualist arguments accordingly focus on the insight that, for legally sophisticated parties who write bespoke contracts, context is endogenous; the parties can embed as much or as little context into a customized agreement as they wish, and they can do so in many different ways.”); see also Uri Benoliel, *The Interpretation of Commercial Contracts: An Empirical Study*, 69 ALA. L. REV. 469, 472–73 (2017) (noting the importance of ambiguity).

carefully about what they mean, and to use contract words in ordinary ways.⁵⁸ This ideological approach to contract interpretation resembles that same concept in statutory and constitutional interpretation;⁵⁹ though it is less politically valenced, it is equally ascendent.⁶⁰

The linguistic textualist project has long been controversial. To begin with, the method of brute sense plain meaning primes judges to overconfidently believe that their beliefs and conclusions are more common than they in fact are.⁶¹ As Arthur Corbin put it long ago, “when a judge reads the words of a contract he may jump to the instant and confident opinion that they have but one reasonable meaning and that he knows what it is.”⁶² Empirical work—experimental⁶³ and sociological⁶⁴—has since found that judges doing plain meaning analysis disagree with each other and with lawyers about things they thought obvious.

Critics also charge textualists with incoherence about ambiguity.⁶⁵ To reach the safe shoals of plain meaning, textualists ask first if the language is unambiguous.⁶⁶ But while textualism provides tools to discover ambiguities, in practice, critics charge, it fails to prioritize one plausible interpretation over the other. It appears to simplify interpretative disputes, but in reality sometimes facilitates expensive, biased battles over extrinsic evidence.⁶⁷

⁵⁸ See Schwartz & Scott, *supra* note 38, at 572.

⁵⁹ For a discussion of the differences between statutory and contract textualism, see William Baude & Ryan D. Doerfler, *The (Not So) Plain Meaning Rule*, 84 U. CHI. L. REV. 539, 563–65 (2017). For an insightful argument that interest in contract interpretation has waned relative to statutory interpretation, see Karen Petroski, *Does It Matter What We Say About Legal Interpretation?*, 43 McGEORGE L. REV. 359, 382 (2012).

⁶⁰ See Ethan J. Leib, *The Textual Canons in Contract Cases: A Preliminary Study*, 2022 WIS. L. REV. 1109 (2022) (studying the use of textualist canons in contract interpretation); Stempel & Knutsen, *supra* note 34, at 565 (“In short, textualism has been resilient and ascendant in the 40 years of the post-Restatement era.”).

⁶¹ See *infra* text accompanying notes 116–17.

⁶² 3 ARTHUR LINTON CORBIN, CORBIN ON CONTRACTS: A COMPREHENSIVE TREATISE ON THE WORKING RULES OF CONTRACT LAW 18 (rev. ed. 1960).

⁶³ Solan, Rosenblatt & Osherson, *supra* note 49, at 1285–94 (finding that we overestimate our sense of whether others will agree about contract interpretation).

⁶⁴ See John F. Coyle, *The Canons of Construction for Choice-of-Law Clauses*, 92 WASH. L. REV. 631, 682–87 (2017) (showing that in the absence of a systematic survey, judges can interpret contract language in ways that conflict with the parties’ intentions).

⁶⁵ See Lawrence M. Solan, *Pernicious Ambiguity in Contracts and Statutes*, 79 CHI-KENT L. REV. 859, 859 (2004) (describing problems with the concept of ambiguity).

⁶⁶ 11 SAMUEL WILLISTON, A TREATISE ON THE LAW OF CONTRACTS § 33:43, at 1201–02 (4th ed. 2012) (“When patent ambiguities are found by a court that adheres to the traditional distinctions, they will be resolved by the rules of interpretation or not at all.”). Those supposed rules of interpretation reference § 30:4, where they turn out to combine extrinsic evidence, contract purpose, and rules of construction.

⁶⁷ Ward Farnsworth, Dustin F. Guzior & Anup Malani, *Ambiguity About Ambiguity: An Empirical Inquiry into Legal Interpretation*, 2 J. LEGAL ANALYSIS 257, 271 (2010) (arguing that policy preferences drive statutory ambiguity); Schwartz & Scott, *supra* note 38, at 570 n.55

But even outside of ambiguity, textualism's basic methodological tools are remarkably underdeveloped. Scholars often blame the humble dictionary.⁶⁸ Courts doing textualism are sometimes reversed for failing to use one.⁶⁹ But it's an imprecise tool for discerning the parties' intent at the drafting stage. Selecting between dictionaries is a value-laden act,⁷⁰ and even within a single volume, dictionaries do not provide a single plain, or majoritarian, meaning of words.⁷¹ Critically, dictionary definitions are blind even to internal context, those other parts of the document or statute that textualists do embrace.⁷² As Kevin Tobia demonstrated, definitions can be poor trackers of actual usage, a point well understood by anyone not adding tomatoes to a fruit salad.⁷³

Dictionary-thumping jurists face two opposing critiques: They bind themselves too much,⁷⁴ but also too little.⁷⁵ The first strips the judicial

("Courts seldom distinguish between 'vague' and 'ambiguous' terms . . . More narrowly, however, a word is vague to the extent that it can apply to a wide spectrum of referents, or to referents that cluster around a modal 'best instance,' or to somewhat different referents in different people.").

⁶⁸ Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788, 801–02, 810–11 (2018) (identifying several problems with dictionaries, including their failure to define words in terms of "prototypes" and the inconsistency of definitions across dictionaries); Stephen C. Mouritsen, *The Dictionary Is Not A Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 5 BYU L. REV. 1915, 1919 (2010) (describing "widely shared" false views about dictionaries); Lawrence Solan, *When Judges Use the Dictionary*, 68 AM. SPEECH 50, 50 (1993) ("[W]e commonly ignore the fact that someone sat there and wrote the dictionary . . . and we speak as though there were only one dictionary, whose lexicographer got all the definitions 'right' in some sense that defies analysis."); Samuel A. Thumma & Jeffrey L. Kirchmeier, *The Lexicon Has Become A Fortress: The United States Supreme Court's Use of Dictionaries*, 47 BUFF. L. REV. 227, 276 (1999) ("[A]s with the other steps in the Court's general process of using dictionaries, selecting a specific definition for a term can be problematic, at times appears to lack principled guidance and can determine the outcome of a case.").

⁶⁹ *Lorillard Tobacco Co. v. Am. Legacy Found.*, 903 A.2d 728, 738 (Del. 2006) (reversing for failure to follow dictionary).

⁷⁰ Lee & Mouritsen, *supra* note 68, at 807 ("A common use of a dictionary involves simple cherry-picking.").

⁷¹ *Id.* at 810–11 ("We cannot tell from the opinion whether the *written translator* sense of *interpreter* is less often listed in a real 'survey' of dictionaries because we are not presented with an actual *survey* of dictionaries.").

⁷² WILLISTON, *supra* note 66, at § 32:5, at 692 ("A contract will be read as a whole and every part will be read with reference to the whole."); *see also* Bradley C. Karkkainen, "Plain Meaning": Justice Scalia's Jurisprudence of Strict Statutory Construction, 17 HARV. J.L. & PUB. POL'Y 401, 407 (1994).

⁷³ Kevin P. Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726, 797–99 (2020).

⁷⁴ Nicholas S. Zeppos, *Judicial Review of Agency Action: The Problems of Commitment, Non-Contractability, and the Proper Incentives*, 44 DUKE L.J. 1133, 1143 (1995) (referring to the Court's at times "fanatical" devotion to dictionaries).

⁷⁵ *See* Mouritsen, *supra* note 68, at 1930 (critiquing dictionaries as weak sources of plain meaning and for the absence of context); *see also* *Jordan v. De George*, 341 U.S. 223, 234 (1951) (Jackson, J., dissenting) (calling dictionaries "the last refuge of the baffled judge").

process of its nuanced nature, the latter breeds gamesmanship and bias.⁷⁶ This critique is (to be fair) a little overheated. Sure, judges take dictionaries seriously,⁷⁷ but they also freely admit that dictionaries are not “infallible.”⁷⁸ Even Learned Hand cautioned, “it is one of the surest indexes of a mature and developed jurisprudence not to make a fortress out of the dictionary.”⁷⁹ Dictionaries are normally under-determinative of outcomes, and this is a virtue rather than a vice. As we shall claim, this virtue is equally shared by generative interpretation.

Similarly, the canons of interpretation themselves are difficult to defend empirically.⁸⁰ These canons are traditionally known by their evocative Latin names—in *pari materie*, *expressio unius est exclusio alterius*, *eiusdem generis*, *contra proferentem*, *generalia specialibus non derogant*—and they are used to fill dictionaries’ gaps.⁸¹ They try to address the problem of context by giving heuristics to parse the parties’ proffered meanings.⁸² Popular with judges but absent from the Restatement,⁸³ scholars criticize them as essentially *ad hoc*.⁸⁴ There is no

⁷⁶ Lee & Mouritsen, *supra* note 68, at 798 (“The concern here is that even if we could settle on a theory of ordinary or plain meaning, we are unsure how to assess it.”).

⁷⁷ See, e.g., *In re Liquidation of Am. Mut. Liab. Ins. Co.*, 802 N.E.2d 555, 560 (Mass. 2004) (“Normally, a dictionary definition of a term is strong evidence of its common meaning”); see also *Brigade Leveraged Cap. Structures Fund Ltd. v. PIMCO Income Strategy Fund*, 995 N.E.2d 64, 69 (Mass. 2013).

⁷⁸ *Cyprus Plateau Mining Corp. v. Commonwealth Ins. Co.*, 972 F. Supp. 1379, 1384 (D. Utah 1997) (“Dictionaries, while not infallible (or even consistent), are general guides to common usage.”).

⁷⁹ *Cabell v. Markham*, 148 F.2d 737, 739 (2d Cir. 1945).

⁸⁰ Farshad Ghodoosi & Tal Kastner, *Big Data on Contract Interpretation*, U.C. DAVIS L. REV. (forthcoming 2024) (manuscript at 58) (highlighting the issue of precedent around the use of canons being deployed without regard to the context in which the precedent arose); Leib, *supra* note 60, at 1110 (“Few scholars or lawyers believe they are applied consistently enough to be reliable in predicting case outcomes”).

⁸¹ See generally Edwin W. Patterson, *The Interpretation and Construction of Contracts*, 64 COLUM. L. REV. 833, 852–55 (1964) (identifying canons of contract interpretation).

⁸² The canons of contract interpretation are to be distinguished from the canons of construction in statutory interpretation. As Ryan Doerfler has explored, those canons have been subject to a rehabilitative project over the last generation. See Ryan D. Doerfler, *Late-Stage Textualism*, 2021 SUP. CT. REV. 267, 269 (2022). Of course, the *contra proferentem* doctrine particularly is not necessarily effecting intent, but may motivate clear drafting. See generally Daniel Schwarcz, *The Role of Courts in the Evolution of Form Contracts: An Insurance Case Study*, 46 BYU L. REV. 471 (2021) (making this argument in the context of insurance contracts).

⁸³ Leib, *supra* note 60, at 112 (“Yet the *Restatement* does not treat the textual canons like *expressio unius*, *eiusdem generis*, or *noscitur a sociis* at all.” (emphasis in original)); Ghodoosi & Kastner, *supra* note 80, at 48 (“While substantive canons have remained roughly in equilibrium over time, the chart below demonstrates a trend in which the invocation of textual canons by courts across contract cases is increasing.”).

⁸⁴ 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 (1950) (“[T]here are two opposing canons on almost every point.”).

obvious way to know what to do when different canons lead to different outcomes, meaning that they offer the same kinds of degrees of freedom as dictionaries do.

Nor is it clear that the contractual linguistic canons are rooted in how parties think or write.⁸⁵ The extant empirical work on linguistic canons in statutory interpretation suggests that the answer is: they might be, but only some of the time.⁸⁶ Now, to be sure, some of the canons, like *contra proferentem*, aren't intended to replicate how the parties would have understood the contract at drafting (if that has a stable meaning in contracts deployed to millions of adherents). These normative canons may, or may not, relate to the parties' contemporaneous intentions.⁸⁷ But other canons are intended to reflect ordinary uses of language, and yet have been subject to remarkably little controlled scrutiny.⁸⁸

Notwithstanding its methodological shortcomings, contract textualism is ever more popular.⁸⁹ That's so for a whole host of reasons, but none more so than the weakness of its main conceptual rival: contextualism. This familiar alternative starts with the same premise as textualism: What would the parties have said they meant had we asked them at the time of contracting? But contextualism invites parties to offer extrinsic evidence to build depth into the predictive analysis. By doing so, contextualism seeks to privilege accuracy—the parties' real intended meaning.

⁸⁵ Gregory Klass, *Interpretation and Construction in Contract Law* 48 (Jan. 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2913228 [<https://perma.cc/27YR-G8LN>] (“Rules of construction are only sometimes pragmatically prior to contract interpretation, but not always and not pervasively.”).

⁸⁶ Kevin Tobia, Brian G. Slocum & Victoria Nourse, *Statutory Interpretation from the Outside*, 122 COLUM. L. REV. 213, 241–43, 262 (2022) (finding that some linguistic canons are stated overbroadly or inaccurately but many canons do reflect the intuitive judgment of ordinary people); Brian G. Slocum & Kevin Tobia, *The Linguistic and Substantive Canons*, 137 HARV. L. REV. F. 70, 100 (providing evidence that “some interpretive rules traditionally justified by normative values also have a linguistic basis”); Janet Randall & Lawrence M. Solan, *Legal Ambiguities: What Can Psycholinguistics Tell Us?*, in CAMBRIDGE HANDBOOK OF EXPERIMENTAL JURIS. (Kevin P. Tobia ed.) (forthcoming) (manuscript at 2).

⁸⁷ Christopher J. Walker, *Legislating in the Shadows*, 165 U. PA. L. REV. 1377, 1404 (2017) (arguing that “*contra proferentem*” is not a method by which the true intent of the parties is determined, but rather, is a decision to impose the burden of ambiguity on the drafter).

⁸⁸ Ross & Tranen, *supra* note 39, at 226–27 (“Descriptive canons are based on the way ordinary people express themselves in English.”).

⁸⁹ See Ghodoosi & Kastner, *supra* note 80, at 49 (“[O]ur study provides evidence that textualism is on the rise in contract interpretation.”); Aaron D. Goldstein, *The Public Meaning Rule: Reconciling Meaning, Intent, and Contract Interpretation*, 53 SANTA CLARA L. REV. 73, 77 (2013) (arguing that courts have increasingly moved away from the use of extrinsic evidence to help them understand the parties' intent, leaning instead on “objective” manifestations of intent); Mark L. Movsesian, *Formalism in American Contract Law: Classical and Contemporary*, 12 IUS GENTIUM 115, 115 (2006) (“It is a truth universally acknowledged, that we live in a formalist era. At least when it comes to American contract law.”).

This approach to interpretation, capacious in the types of evidence considered relevant, found its heyday in the 1960s in California and has never been as popular since.⁹⁰ The problem with the approach, according to its critics, is that it does not permit the parties to know what meaning a court will assign to the words they write, since the other side can always offer self-serving meanings *ex post* and, if believable enough, write a new bargain in court to replace the one drafted in the past.⁹¹ Even contextualism's origin story is one of a party suddenly remembering that they actually meant to make the purchase option available only to family members, creditors be damned.⁹² Contextualism makes it difficult to lock down meaning *ex ante*, through merger clauses and the like, which are always subject to later testimonial refutation. Contextualism's consumer protection allure is understandable.⁹³ But even if contextualism could offer more accuracy, critics charge it does so at a high cost.⁹⁴

Indeed, scholars often defend textualism on efficiency grounds.⁹⁵ Though it may be unclear what parties want interpretative rules to be, it's almost certainly the case that lawyer-drafters prefer textualist to contextualist modes of decision. Eric Posner captures the idea well: Parties will often include an explicit merger clause, but few ever bother with an "anti-merger clause."⁹⁶ Thus, from the perspective of

⁹⁰ See *Pac. Gas & Elec. Co. v. G. W. Thomas Drayage & Rigging Co.*, 442 P.2d 641, 643 (Cal. 1968) (finding reversible error when lower court refused to consider extrinsic evidence to demonstrate parties' intent); see also *Masterson v. Sine*, 436 P.2d 561, 562–63 (Cal. 1968) (finding that a trial court erred by refusing to consider extrinsic evidence).

⁹¹ *Masterson*, 436 P.2d at 561 (Burke, J., dissenting).

⁹² *Id.*

⁹³ See *Olah v. Ganley Chevrolet, Inc.*, 946 N.E.2d 771, 774 (Ohio 2010) (holding that buyers of a vehicle are barred from presenting evidence that the car was represented by the dealer as new because the contract says the vehicle is used).

⁹⁴ An admittedly limited survey of enforcement costs did not find meaningful differences between textualist approaches and contextualist ones. See Silverstein, *supra* note 57, at 1021. For an argument that textualism produces higher enforcement costs because the judge-by-judge variation in outcomes produces more litigation, see 6 PETER LINZER, CORBIN ON CONTRACTS 163 (Joseph M. Perillo ed., rev. ed. 2010).

⁹⁵ See Schwartz & Scott, *Redux*, *supra* note 38, at 928 n.1, 941 ("A strong majority of U.S. courts continue to follow the traditional, 'formalist' approach to contract interpretation."). But see Joshua M. Silverstein, *Contract Interpretation and the Parol Evidence Rule: Toward Conceptual Clarification*, 24 CHAP. L. REV. 89, 92 (2020) (arguing that the matter is indeterminate); Silverstein, *supra* note 57, at 1020 ("[C]ontracts scholars can also generally be split into textualist and contextualist camps, with a clear majority falling into the latter group."). There is recent evidence that contract scholars prefer contextualism. Eric Martinez & Kevin Tobia, *What Do Law Professors Believe About Law and the Legal Academy?*, 112 GEO. L.J. 111 (2023).

⁹⁶ Eric A. Posner, *The Parol Evidence Rule, the Plain Meaning Rule, and the Principles of Contractual Interpretation*, 146 U. PA. L. REV. 533, 571 (1998). As Larry Solan later pointed out, limits on "the judicial function" in statutory interpretation "are not easy to find." LAWRENCE M. SOLAN, *THE LANGUAGE OF STATUTES: LAWS AND THEIR INTERPRETATION* 187 (Univ. Chi. Press 2010).

the litigated cases—at least those between rich and lawyered parties—contextualism is simply harder to defend.

And yet, from a certain perspective, contextualism seems well-positioned for a revival. Recall that even contextualism’s critics agree about the first-order goal: to figure out what the parties would have meant at contracting. The problems with contextualism are largely centered around motivated testimony and cost, which persuades the factfinder to ignore the text. But consider: We increasingly live in a world where our thoughts are recorded contemporaneously, whether sent by text, posted on social media, or recorded on TikTok. Such recorded, immutable utterances are cheap to reproduce and appear to courts to be excellent sources of contractual meaning.⁹⁷ Defenders of textualism may argue that permitting their use creates uncertainty, but some of the best arguments against contextualism—that it can be abused *ex post*—are weaker than they used to be.⁹⁸ And yet, we lack a method to know which excited utterances to privilege, and we should worry that courts’ motivated reading will cause them to come to inaccurate or biased understandings.

The debate between textualism and contextualism is old, and scholars have offered various theoretical lenses by which one or the other approach ought to prevail.⁹⁹ Most arguments for or against extrinsic evidence turn on hypotheses about what parties would have wanted (had we asked them) and which methods promote social welfare. These arguments are often theoretically rich but empirically poor.¹⁰⁰

More recently, scholars have offered two new methods, both advancing the certainty values of textualism with a dash of the accuracy interests of contextualism. One school focuses on the use of corpora of words to predict the meaning of phrases in contractual texts—so-called

⁹⁷ See *BrewFab, LLC v. 3 Delta, Inc.*, No. 22-11003, 2022 WL 7214223, at *2–3 (11th Cir. Oct. 13, 2022) (*per curiam*) (affirming that a party’s text message was a personal guaranty that satisfied Florida’s statute of frauds); see also *Cloud Corp. v. Hasbro, Inc.*, 314 F.3d 289, 295–96 (7th Cir. 2002) (finding that a party’s e-mails satisfied the UCC’s statute of frauds and using these as evidence in support of the claim that the contract had been modified); see also *Cosby v. Am. Media, Inc.*, 197 F. Supp. 3d 735, 744 (E.D. Pa. 2016) (holding that tweets may form the basis of a breach of contract claim).

⁹⁸ Cf. Shawn Bayern, *Contract Meta-Interpretation*, 49 U.C. DAVIS L. REV. 1097, 1135 (2016) (pointing out that because text messages are informal, they don’t satisfy some of the deliberation-inducing virtues that textualists would otherwise place in written products).

⁹⁹ See Ross & Trannen, *supra* note 39, at 196–97; see also Joshua M. Silverstein, *The Contract Interpretation Policy Debate: A Primer*, 26 STAN. J.L. BUS. & FIN. 222, 225–26 (2021); see also Mark L. Movsesian, *Severability in Statutes and Contracts*, 30 GA. L. REV. 41, 70 n.184 (1995) (noting that the popularity of the major interpretive approaches “ebbs and flows”).

¹⁰⁰ Silverstein, *supra* note 57, at 1014 (“The textualist/contextualist controversy cannot be resolved in the abstract Unfortunately, empirical evidence bearing on this debate is sorely lacking.”).

corpus linguistics.¹⁰¹ To take the prototypical example, consider the following phrase taken from an insurance contract: “[This insurance] does not apply to ‘bodily injury’ [including death] to any person while practicing for or participating in any sports or athletic contest or exhibition that you sponsor.”¹⁰² An insured dies while snorkeling: Is that a “sports or athletic contest”? As Stephen Mouritsen observes, the question is not easily answerable using the classic dictionary and canon based tools of textualism. And, considering that insurance contracts are drafted by powerful firms, who subject them to regulatory scrutiny, the idea of using extrinsic expressions by either firms or the insured seems hopeless.¹⁰³ Instead, Mouritsen suggests that courts (helped by adversarial presentation by parties) could query language databases to establish whether sports and snorkeling appear relatively close to each other in some number of previous examples. That is, to derive the meaning of the word from its common use in previous texts. (The answer is, more or less, that sports are rule-based competitions, while snorkeling is swimming wearing a goofy mask.)¹⁰⁴

Corpus linguistics is an advance over traditional textualism or contextualism. It provides a methodology that theoretically allows courts to adhere to an objective set of responses when determining the ordinary meaning of words based on their actual usage. Essentially, it’s a form of textualism that doesn’t rely on dictionary definitions or a battery of canons. It mirrors not the static decisions of lexicographers in their secluded, book-filled offices, but rather the public use of words—democratized textualism.¹⁰⁵

But corpus linguistics is inattentive to context.¹⁰⁶ It can only really compare brief snippets of text, rather than whole documents. Thus, although the method has been repeatedly used in statutory interpretation cases—where the stakes are high, parties are commonly

¹⁰¹ See generally Mouritsen, *supra* note 19, at 1360–1407 (making the case for corpus linguistics).

¹⁰² *Id.* at 1340.

¹⁰³ See Christopher C. French, *Insurance Policies: The Grandparents of Contractual Black Holes*, 67 DUKE L.J. ONLINE 40, 44–45 (2017) (discussing the difficulty of interpreting insurance contracts for evidence of real meaning).

¹⁰⁴ Mouritsen, *supra* note 20, at 1371–74 (describing a corpus linguistics approach to the snorkeling example).

¹⁰⁵ For an extended defense, see Jeffrey W. Stempel & Erik S. Knutsen, *Technologically Improving Textualism*, 6 NEV. L.J.F. 10 (2022).

¹⁰⁶ See Choi, *supra* note 16, at 8, 16–17 (arguing that the context “undermines the core claim of corpus linguistics”).

engaged in interpretative battles over short phrases—only one contracts opinion to date has applied the method.¹⁰⁷

A different constraining approach, advanced by Omri Ben-Shahar and Lior Strahilevitz, encourages courts to use survey evidence to decide on the public meaning of certain contractual texts.¹⁰⁸ As they point out, this survey evidence is second best to the predictive ideal we described above:

Contracts should have the meaning that the parties to the transaction assign to the text. [But] it is pointless to ask the actual parties in the litigation what the text meant to them when they formed the contract, because they will bend their answers to fit their litigation goals. So the law should instead ask disinterested people just like them.¹⁰⁹

The authors defend this interesting proposal against various charges.¹¹⁰ Their core survey case is consumer contracts designed for mass audiences.¹¹¹ There, the survey audience and the original adherents are the same people (although separated by time), and we should have fewer worries about the parties intending idiosyncratic meanings.¹¹² But outside of that frame, a problem with the survey approach is that for most litigated contract cases—i.e., commercial cases—the relevant survey audience will be difficult to find, as sophisticated adherents don't take surveys, or will game them, producing the same problems encumbering contextualism.¹¹³

¹⁰⁷ See *Fulkerson v. Unum Life Ins. Co. of Am.*, 36 F.4th 678, 682–83 (6th Cir. 2022) (using corpus linguistic analysis in contract interpretation); see also *Richards v. Cox*, 450 P.3d 1074, 1085–86 (Utah 2019) (Lee, J., concurring) (concurring in majority opinion “to the extent it relies on corpus linguistic analysis” to support constitutional and statutory interpretation). Cf. *Wilson v. Safelite Grp., Inc.*, 930 F.3d 429, 439–40 (6th Cir. 2019) (arguing for use of corpus linguistics in statutory analysis); see also *Caesars Ent. Corp. v. Int’l Union of Operating Eng’rs Loc. 68 Pension Fund*, 932 F.3d 91, 95 (3d Cir. 2019) (using corpus linguistics to interpret “previously”).

¹⁰⁸ Ben-Shahar & Strahilevitz, *supra* note 19; Ian Ayres & Alan Schwartz, *The No-Reading Problem in Consumer Contract Law*, 66 STAN. L. REV. 545, 545, 595–96 (2014) (advocating empirical testing to identify surprising and problematic provisions in standard form contracts, against which consumers ought to be warned); Ariel Porat & Lior Jacob Strahilevitz, *Personalizing Default Rules and Disclosure with Big Data*, 112 MICH. L. REV. 1417, 1417, 1419–20 (2014) (advocating the use of surveys to identify the majoritarian preferences for the design of granular default rules).

¹⁰⁹ Ben-Shahar & Strahilevitz, *supra* note 19, at 1802.

¹¹⁰ *Id.* at 1802–13 (making the case).

¹¹¹ *Id.* at 1758 (noting the focus on consumer contracts due to the ease of identifying representative yet disinterested consumers to survey).

¹¹² See *id.* at 1776–77 (noting the utility of surveys for consumer contracts on these grounds).

¹¹³ Cf. *Roberts v. Farmers Ins. Co.*, 1999 WL 1063826, at *4 n.2 (10th Cir. Nov. 23, 1999) (“[W]hat the public expects from an insurance policy is simply not relevant to the legal question of whether the contract is ambiguous.”).

Survey evidence is also an expensive adjudicatory technology. Surveys themselves are difficult to conduct: Judges would need to rely on their adversarial presentation in the ordinary case. And they are increasingly unreliable: Recent work has found that almost a third of online survey respondents use LLMs to complete answers.¹¹⁴ Surveys based on more collated samples face the same sorts of problems that have bedeviled modern polling: nonresponse bias among parts of the population, difficulties of generalization, and inaccuracy. And even here, attention is scarce. It is hard to survey consumers on a twenty-page policy or to expect anyone filling out a survey for a five dollar gift card to attentively consider interdependencies within the contract.

Consequently, though survey methodology is an established technique in trademark cases and could very well be of enormous help in making sense of the meaning of certain consumer contracts, it is unlikely to be a transformative technology in the ordinary contract interpretation case. We are unaware of any cases to date that permit the use of survey evidence to determine contractual meaning.

In summary, notwithstanding broad agreement about the predictive goal of interpretation, there's also a shared sense that there's something amiss in how jurists balance accuracy and efficiency. Textualism promises the latter, but in practice it often merely supercharges the judge's own overconfident priors. Contextualism promises the former, but probably doesn't deliver it, while eroding parties' ability to plan for court outcomes and making litigation prohibitively expensive for all but the wealthiest parties. The two most sophisticated modern improvements on these old technologies—statistical plain meaning and survey evidence—promise to rescue textualism from some of its sins, but haven't been taken up in live cases.

Enter large language models.

II

GENERATIVE INTERPRETATION

The doctrine of reasonable expectations plays a contested role in the regulation of insurance contracts.¹¹⁵ For some courts, the insured's

¹¹⁴ See Veniamin Veselovsky, Manoel Horta Ribeiro & Robert West, *Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks*, ARXIV (June 13, 2023), <https://arxiv.org/pdf/2306.07899.pdf> [<https://perma.cc/CMY2-T3WG>] (noting that 33–46% of mTurk survey workers use LLMs to complete tasks).

¹¹⁵ See generally Stempel, *supra* note 34 (outlining the doctrine's ebbs and flows over time).

reasonable expectations trump the insurance contract's terms, while for many others, the policy's plain language should control.¹¹⁶ Notoriously, these sorts of cases motivate armchair speculation by judges—whose life experience, education, sophistication, and hard-earned cynicism systematically diverge from most lay people. Worse, the interpretations we give words appear very certain in our own minds. Contract interpretation is a prime subject for a phenomenon psychologists call “false consensus effect.”¹¹⁷ To illustrate the effect, Lawrence Solan, Terri Rosenblatt and Daniel Osherson presented contract interpretation questions to both laypeople and judges.¹¹⁸ After giving their opinion, subjects were asked to estimate how many other participants would agree with them.¹¹⁹ This design allows us to compare the actual distribution of answers with how people expected the distribution to look. The results were striking: Both laypeople and judges overestimated how common their chosen interpretations were. Judges even overestimated how much other judges would agree with them.¹²⁰

Thus, one of the risks of introspective interpretation is that its products are very sticky and hard to dislodge. This leads to dissent and reversal, and of course, interpretation that defies parties' expectations. Uncertainty about common interpretation is an appealing case for the use of surveys.¹²¹ And surveys would be of great interpretative use, were it not for the practical difficulties which we've just discussed.

Consider *C & J Fertilizer v. Allied Mutual*.¹²² The president of C&J, a fertilizer firm, purchased a burglary insurance policy from Allied Mutual. The discussions preceding the purchase made it clear that the policy would not cover an inside job.¹²³ The insurance firm in the negotiations tried to insist that to bring a claim, C&J would

¹¹⁶ See RESTATEMENT OF LIAB. INS. § 3 cmt. a (AM. L. INST. 2019) (noting that the plain meaning approach is typically followed instead of contextual approaches). As Dan Schwarcz has explored, the doctrine is unpredictable when applied in real cases. See Daniel Schwarcz, *A Products Liability Theory for the Judicial Regulation of Insurance Policies*, 48 WM. & MARY L. REV. 1389, 1426–30 (2007) (arguing that the doctrine lacks analytical coherence and is inherently vague).

¹¹⁷ Joachim Krueger & Russell W. Clement, *The Truly False Consensus Effect: An Ineradicable and Egocentric Bias in Social Perception*, 67 J. PERSONALITY & SOC. PSYCH. 596, 596 (1994) (defining the effect as “the overuse of self-related knowledge in estimating the prevalence of attributes in a population”); see Brian Mullen, Jennifer L. Atkins, Debbie S. Champion, Cecelia Edwards, Dana Hardy, John E. Story & Mary Vanderklok, *The False Consensus Effect: A Meta-Analysis of 115 Hypothesis Tests*, 21 J. EXPERIMENTAL SOC. PSYCH. 262 (1985) (providing a meta-analysis of false consensus effect).

¹¹⁸ Solan, Rosenblatt & Osherson, *supra* note 49.

¹¹⁹ *Id.* at 1285–94.

¹²⁰ *Id.*

¹²¹ See generally Stempel & Knutsen, *supra* note 33 (describing the worry).

¹²² *C & J Fertilizer, Inc. v. Allied Mut. Ins. Co.*, 227 N.W.2d 169, 176 (Iowa 1975).

¹²³ See *id.* at 171–72.

have to present hard evidence that a theft was made by a stranger.¹²⁴ That idea was embodied in the following promise in the insurance contract:

[Allied will pay for] the felonious abstraction of insured property (1) from within the premises by a person making felonious entry therein by actual force and violence, of which force and violence there are visible marks made by tools, explosives, electricity or chemicals¹²⁵

As it turns out, a burglar robbed the fertilizer plant with style. Leaving some tread marks in the mud, he forced entry into the warehouse and absconded with \$75,000 worth of fertilizer (in today's dollars).¹²⁶ The insurance company, denying the claim, argued that by its plain language, the absence of visible marks made by tools (as opposed to tires) meant that it didn't have to pay.

The Iowa Supreme Court, in a contracts casebook staple, held that the exclusion applied in this way violated the insured's reasonable expectations.¹²⁷ No one could have reasonably expected that burglary would be limited only to those leaving visible forced entry marks.¹²⁸ In reaching that view, the court relied on its own common sense alone. Was it plausibly right?

That question triggers the simplest use cases of LLMs as part of the interpretative process. The judge can simply *ask* the model for an assessment. Fantastical only three years ago, today you might be merelywhelmed by the model's ability to respond coherently and plausibly to this query. Here's the model's response, edited for readability¹²⁹:

An insurance policy reads: “[The insurance company will pay for] the felonious abstraction of insured property (1) from within the premises by a person making felonious entry therein by actual force and violence, of which force and violence there are visible marks made by tools, explosives, electricity or chemicals.” With this in mind, please state your prediction—with the associated numerical

¹²⁴ *Id.* at 171.

¹²⁵ *Id.*

¹²⁶ *See id.*

¹²⁷ *Id.* at 177.

¹²⁸ *Id.* (“[N]othing relating to the negotiations . . . would have led plaintiff to reasonably anticipate defendant would bury within the definition of ‘burglary’ another exclusion denying coverage when, no matter how extensive the proof of a third-party burglary, no marks were left on the exterior of the premises.”).

¹²⁹ For the chat repository, see *Insurance Policy Expectations*, OPENAI (Oct. 7, 2023), <https://chat.openai.com/share/4379b796-cece-4616-b8eb-b6772f13ad37> [https://perma.cc/RAY4-6V84].

level of confidence in parentheses—on the likely expectations of most policyholders under these terms for the following propositions:

Queries and GPT-4 Responses	
Query	GPT-4 Response
Will the policy provide compensation for losses resulting from a substantiated third-party burglary?	Likely Expectation (90%): The policy will compensate for third-party burglary.
Does the policy clearly delineate what is considered an inside job as opposed to a third-party intrusion?	Ambiguous Expectation (50%): The policy doesn't clearly delineate between inside jobs and third-party intrusions.
In instances where a third-party burglary can be definitively established, will the policy necessitate physical evidence of forced entry?	High Expectation (95%): Physical evidence of forced entry is necessary for third-party burglary claims.

Table 1: GPT-4's estimates of propositions regarding the likely content of the gap in the policy.

In other words, the model's results disagreed with the court's majority opinion. They (like a dissenting opinion) predict that policyholders would have expected to be required to provide some evidence of forceful entry to prove that the burglary was not an inside job.¹³⁰

To us these findings are facially plausible: They validate that this cheap and convenient tool could be potentially of use in real cases. But just because the probabilities are reasonable doesn't mean they are accurate. Your intuition should be: prove it! You would want to know more about what the model is doing when it produces percentages, how the choice of the query would have affected the results, and how that methodology fits courts' purposes in interpreting insurance contracts.¹³¹ Let's start there, in Section II.A. We'll then try some more complicated examples in the remainder of this Part.

A. A Gentle Introduction to Large Language Models

When Chat GPT-4 told us that it was 90% likely that the policy would pay in response to a "substantiated third-party burglary," what

¹³⁰ See *C & J Fertilizer, Inc.*, 227 N.W.2d at 184 (LeGrand, J., dissenting) (arguing that the reasonable expectation is to show forceful entry because plaintiff's agent expected that of the insurance contract for his own property).

¹³¹ As we emphasize throughout, model outputs involve a certain degree of randomness. Repeated experimentation, ideally with different prompts, is advisable. See *infra* Section III.A. for discussion of best practices.

happened behind the curtain? We're going to give an explanation a shot here, knowing that doing so is difficult because LLM technology is complex and is rapidly changing. Essentially, LLMs create a statistical model of how words connect by training on torrents of existing texts, some historic and some artificially derived.¹³²

In the common case, LLMs take user input in the form of text and produce an output, also in the form of text. Behind the scenes, the model takes the text and transforms it into numbers. This is essential, because (superficially) computers do not read text. Numbers can encode more information than letters, and they are more valuable in that they allow computers to perform mathematical operations. This is easy to see in the case of ambiguities: Duck is both a verb and a noun. But in a number system, we can use prefixes like 20 for verbs and 10 for nouns, so we can encode the word duck twice. One is, say, 201 and the other 101, to designate the disparate meanings and disambiguate them.¹³³

This simple illustration understates the utility of this process, known as embedding.¹³⁴ Rather than assigning a single number to each word, machine learning models transform them into long lists of numbers—each item on the list capturing some aspects of meaning.¹³⁵ The length of such vectors is very long; one of the latest models in common use employs a vector with 12,288 number-pairs.¹³⁶ For simplicity of exposition, suppose you had a list of common animals and had a two-dimensional vector to describe them. One dimension could be number of feet; another could be if they lived on land or sea. This would produce vectors that we can visualize below¹³⁷:

¹³² Synthetic data is growing in importance, and sometimes may improve model quality. See John Jumper et al., *Highly Accurate Protein Structure Prediction with AlphaFold*, 596 NATURE 583, 587–88 (2021) (noting how training the model using synthetic data improved the model's accuracy significantly).

¹³³ This, in a sense, is what standard English dictionaries do, at least if one were to number the words by order of appearance.

¹³⁴ For a description of embeddings (although without the attention mechanism), see Choi, *supra* note 16, at 20–22.

¹³⁵ What embeddings capture is related to but different from meaning. For a discussion that emphasizes the non-semantic-understanding view, see Lisa Miracchi Titus, *Does ChatGPT Have Semantic Understanding? A Problem with the Statistics-of-Occurrence Strategy*, 83 COGNITIVE SYS. RSCH. 1–2. For sake of exposition, we imprecisely use the word “meaning.”

¹³⁶ Nils Reimers, *OpenAI GPT-3 Text Embeddings – Really a New State-of-the-Art in Dense Text Embeddings?*, MEDIUM (Jan. 28, 2022), https://medium.com/@nils_reimers/openai-gpt-3-text-embeddings-really-a-new-state-of-the-art-in-dense-text-embeddings-6571fe3ec9d9 [<https://perma.cc/5QNE-XGX8>].

¹³⁷ Sea turtles have flippers, not legs. In a more sophisticated representation, we might have adopted a more continuous representation of feet, where flippers are closer to feet than they are to, say, tails.

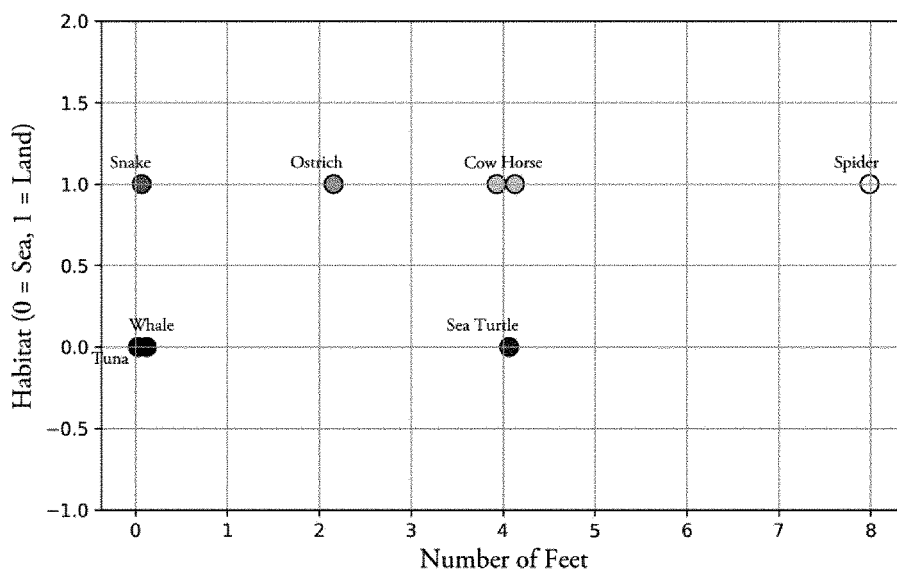


Figure 2: An illustration of the value of encoding meaning via simple embeddings

What makes vectors so powerful is that they allow us to capture not only semantics, but also a syntactic relationship to other words. Horses and cows, in our very simplistic schema, are closer to each other than they are to whales or sea turtles. The snake, always awkward, occupies its own category. If we were to add salamanders, we would spot the emergence of a distinct category of amphibians, alongside the land mammals. Now, suppose you did the same with over 10,000 dimensions.¹³⁸ You can imagine the insights that might result when words are described along such complex dimensions.

Making words dimensional has proved powerful in many machine learning tasks, but was insufficient to power the new LLM revolution. What was needed was the idea of attention.¹³⁹ Read the following sentences:

¹³⁸ A technical clarification: The dimensions in the embedding model do not correspond to clearly defined semantic categories such as “feet” or “habitat.” Rather, they condense information about words in ways that are useful to the attainment of the model’s training objectives. For the best work to date on deciphering the inner working of these complex systems, see Trenton Bricken et al., *Towards Monosemanticity: Decomposing Language Models with Dictionary Learning*, TRANSFORMER CIRCS. THREAD (Oct. 4, 2023) <https://transformer-circuits.pub/2023/monosemantic-features> [<https://perma.cc/4278-PC36>].

¹³⁹ This idea was most powerfully described in a 2017 paper. See Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin, *Attention is All You Need*, ARXIV (June 12, 2017), <https://arxiv.org/pdf/1706.03762>.

“Shohei Ohtani felt the stress. In a desperate attempt, he swung the bat.”

You intuitively grasp that they mean that Ohtani lifted a wooden bat and used it to swing at the baseball. But how do you know that this was right, and not that Ohtani had swung a mammal? As Amelia Bedelia taught us, it’s possible to turn many normal phrases into misadventures if you ignore context.¹⁴⁰ We know that *swung* typically is associated with objects, not animals. And we connect *bat* with Ohtani, a baseball player, which further solidifies our interpretation of the sentence as referring to the object. In other words, our minds naturally pay attention to the context of the word to infer the meaning of any specific word.

An LLM’s *attention mechanism* seeks to achieve the same thing with respect to vectors.¹⁴¹ The model assigns an initial vector to each word in a sentence, which is then enriched by information about its position in the sentence (via positional encoding).¹⁴² Then the attention mechanism assesses which words—say bat or swung—shed light on its meaning.¹⁴³ In the sentence above, words like “stress” and “felt” are not particularly relevant to the meaning of the word “bat”; but both “swung” and “Shohei Ohtani” matter. This allows the model to assign an attention score to each word in the input (relative to the word under analysis) and then reweigh the encoding of the word under analysis relative to the words that are relevant to its interpretation. This means that words do not have stable embedding (as in the older models), but rather, the embedding changes based on the specific context in which they are presented.

These ideas are combined to train a model. A model refers to a collection of parameters (mostly ones called “weights” and “biases”) organized in a specific way whose values are used to transform the input into the model’s output. Modern language models contain tens to hundreds of billions of such parameters, hence their common designation as “Large Language Models.”

pdf [<https://perma.cc/5TXC-AT25>] (introducing a model that can be trained much more quickly relying solely on the concept of attention).

¹⁴⁰ See PEGGY PARISH, AMELIA BEDELIA 20–22 (1963) (pouring dust on furniture after being instructed to “dust the furniture”).

¹⁴¹ For a helpful introduction describing the self-attention mechanism, see SEBASTIAN RASCHKA, YUXI (HAYDEN) LIU & VAHID MIRJALILI, MACHINE LEARNING WITH PYTORCH AND SCIKIT-LEARN 544–61 (2022).

¹⁴² *Id.* at 559–60.

¹⁴³ This is a simplification in several ways. While we discuss words in the text, current models work at the level of a token, which is a part of a word. The model is not directed towards meaning, per se, but rather towards information about other tokens that would help it achieve its training objective. Depending on the architecture, attention may be directed only at preceding tokens. There is more than a single attention mechanism and each one attends to different relationships. There are other subtle simplifications that help the general reader.

Language models are trained with some objective function, a task which they try to achieve and on which they are evaluated. In the context of most popular LLMs, the goal is prediction. The model is presented with the sentence “Shohei Ohtani felt the stress. In a desperate attempt he swung the [?]” and then the model predicts which word would come next. If the model were not calibrated, it might have guessed *lamp* or *materiality*. As these are (probably) incorrect, the model is then led to calibrate toward accuracy through a process called gradient descent.¹⁴⁴ This process repeats itself until the model learns that *bat* follows with 90.14% probability, *ball* with 1.31%, *baseball* with 0.91%, *first* with 0.35%, *club* with 0.29%, and so on.¹⁴⁵

We say the model “learns.” But what does that mean? The simple answer is that during training, the model adjusts the numerical values of billions of parameters such that they would produce predictions that are more likely to achieve its training objectives. It conducts various (fairly simple) algebraic operations to create from a sentence like “Hello, how are you ___” a prediction that the next highest probability word would be “doing.” And yet this simplicity doesn’t capture the process: These parameters are effectively encoded in large, inscrutable matrices whose meaning is wickedly hard to decipher, and whose organization is alien. LLMs do not explain the *why* of their predictions.

You may have read, but would be wrong to conclude, that because the goal is to assign probability to the next word, these models simply replicate text they have seen elsewhere. To effectively predict the next token in a sequence, the models cannot simply memorize what they have seen elsewhere.¹⁴⁶ To predict the continuation of a new sentence

¹⁴⁴ An analogy may capture the intuition behind gradient descent. Suppose you found yourself on a mountain ridge on a pitch-black night, and you are trying to find your way down to the valley below. Feeling with your foot, you sense that going West would lead you upwards, East is levelled, South has a mild declining slope, and North has a steep slope. You head North, and then after a few steps, you test again, to see which direction to take now. This process of finding your way is similar to gradient descent. The model identifies the steepest slope (or gradient) for reducing its deviation from its targets and adjusts its parameters accordingly. It then checks again with more data, iteratively improving until it finds the best or “lowest” configuration. For a helpful and more formal introduction, see Sebastian Ruder, *An Overview of Gradient Descent Optimization Algorithms*, ARXIV 1–3 (June 15, 2017), <https://arxiv.org/pdf/1609.04747.pdf> [<https://perma.cc/3E5Y-TMJV>].

¹⁴⁵ Based on actual predictions of the davinci-002 model with temperature 1, max length = 3 top p = 1, 0 frequency or presence penalty and best of 1. *Playground*, IMGUR, <https://imgur.com/a/37fUjW> [<https://perma.cc/V6DR-BLVP>].

¹⁴⁶ See Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin & Vedant Misra, *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets*, ARXIV 8–10 (Jan. 6, 2022), <https://arxiv.org/pdf/2201.02177.pdf> [<https://perma.cc/E45J-NN4E>] (demonstrating generalization by LLMs occurring after the memorization-like overfitting stage in specific tasks on small, algorithmically generated datasets). *But see* Adam Pearce, Asma Ghandeharioun, Nada Hussein, Nithum Thain, Martin Wattenberg & Lucas Dixon,

like “When they moved to the USA, they set their first home in the state of _____” would require the model to develop a mathematical sense of what are states, what are immigrants, and which states are popular destinations for those who recently arrived.¹⁴⁷ As large as they are, the models are much smaller than the data they are trained on. And so, models necessarily seek deeper representation of the information they train on. This is not unlike how humans read books: they learn from them, but cannot recite them. You can see that model outputs are original because they produce entirely new but responsive text. Of course, this sometimes results in making up facts.

Finally, consumer-facing chatbots simply invite the user to chat with the model directly. Behind the scenes, however, the model’s behavior is calibrated by settings called “hyperparameters.”¹⁴⁸ The details are quite technical, but one of those hyperparameters is of specific interest. LLMs have “temperature” settings that can be adjusted from low to high. The lower the model’s temperature, the more predictable its output.¹⁴⁹ A very low temperature ensures that the model always outputs the same answer to the same query. A higher one introduces more randomness and outputs that you might think of as “creative.”

So far, so good. Now let’s return to our question: What is the model doing when it assigns a 90% probability to the likelihood that a reasonable person would expect an insurance payment under certain circumstances? The first step for the model is to convert the query we

Do Machine Learning Models Memorize or Generalize?, GOOGLE PEOPLE + AI RESEARCH (Aug. 2023), <https://pair.withgoogle.com/explorables/grokking> [<https://perma.cc/2XAB-RUF7>] (analyzing and discussing limitations of the model behavior described in Power et al., *supra*); Emily M. Bender & Alexander Koller, *Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data*, in PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASS’N FOR COMPUTATIONAL LINGUISTICS 5185, 5186 (2020), <https://aclanthology.org/2020.acl-main.463.pdf> [<https://perma.cc/5XAV-V8CK>] (arguing that systems trained on form alone cannot learn meaning); William Merrill, Yoav Goldberg, Roy Schwartz & Noah A. Smith, *Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?*, 9 TRANSACTIONS ASS’N FOR COMPUTATIONAL LINGUISTICS 1047, 1048 (2021), <https://aclanthology.org/2021.tacl-1.62.pdf> [<https://perma.cc/C3VD-BTQN>] (presenting a mathematical formalization of the limits of model understanding in one experiment proposed in Bender & Koller, *supra*).

¹⁴⁷ For example, the GPT-3.5-turbo-instruct model predicts New York (17.48%), California (10.15%), Michigan (4.84%), Texas (3.76%), Ohio (3.43%), and Illinois (2.71%) as the most likely continuations. *Playground*, IMGUR, <https://imgur.com/a/zB44rZf> [<https://perma.cc/L8UN-AZMW>].

¹⁴⁸ The term “hyperparameter” is necessary to distinguish the model’s own parameters from the parameters that define its training and operation. We reference here the post-training hyperparameters, noting that there are also hyperparameters that dictate the training of the model.

¹⁴⁹ For a friendly technical review, see FRANÇOIS CHOLLET, *DEEP LEARNING WITH PYTHON* 368–76 (2d ed. 2021).

entered to numbers (really, tensors).¹⁵⁰ The next step is crucial: Now the model *attends* to the context of words and uses it to adjust their meaning.¹⁵¹ If the model sees the word “premium” in the current context, it will know to adjust its meaning away from dictionary meanings such as “high quality” and towards “consideration paid for a contract of insurance.”¹⁵²

Armed with a contextual understanding of the query, the model can now run through its vast internal network of parameters and calculate what is the most likely word (really, token) that would follow next. It will assign infinitesimally low probabilities to words that relate to gardening or makeup, but will assign increasingly higher probabilities to words that relate to the insurance context. Once the model determines the most likely continuing words, it orders them by relevancy. In a zero temperature setting, the model will always select the word with the highest probability to follow, but as we increase the temperature it will occasionally pick other words as well. When the model outputs “90%,” it reflects that this number is the most likely continuation of the text preceding it.

This explanation skips over the hardest question, which is *why* the model assigns the highest probability to “90%.” The honest answer is quite unsatisfactory: It picked this number because based on its vast training data and internal statistical model, it found that “90%” is a more likely continuation than “10%.”¹⁵³ This is nothing like an explanation a human would give, where reasons and factual considerations would be provided. It is not the result of an introspective analysis of its internal evaluation. The model outputs brute statistics. It is possible to ask the model to justify itself. And the model will diligently reply with an

¹⁵⁰ See Vaswani et al., *supra* note 139, at 3 (“[A]n attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and outputs are all vectors.”).

¹⁵¹ *Id.*

¹⁵² *Premium*, MERRIAM-WEBSTER DICTIONARY (July 27, 2023), <https://www.merriam-webster.com/dictionary/premium> [<https://perma.cc/H7R8-SGFY>].

¹⁵³ “Verbal elicitation” of large language model confidence, in which a model is given a prompt asking for its internal confidence as in the present experiment, typically produce highly overconfident responses. See, e.g., MIAO XIONG, ZHIYUAN HU, XINYANG LU, JIE FU, JUNXIAN HE & BRYAN HOOI, CAN LLMs EXPRESS THEIR UNCERTAINTY? AN EMPIRICAL EVALUATION OF CONFIDENCE ELICITATION IN LLMs, ARXIV 2306.13063 2 (June 22, 2023), <https://arxiv.org/pdf/2306.13063.pdf> [<https://perma.cc/K2BS-BCK7>] (“[O]ur investigation reveals that LLMs tend to be highly overconfident when verbalizing their confidence.”). See generally Shima Imani, Liang Du & Harsh Shrivastava, *MathPrompter: Mathematical Reasoning Using Large Language Models*, in PROCEEDINGS OF THE 61ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 37 (July 2023) (arguing that large language models display “limited performance” in numerical and arithmetic contexts because “[u]nlike natural language understanding, math problems typically have a single correct answer”).

answer. But it is critical to understand that whatever the model tells you, it is really no explanation at all. It is a *prediction* of what explanation is likely to follow the query. So, working with LLMs admittedly requires a leap of faith, a realization that no better explanation is forthcoming than long inscrutable matrices that produce predictions.

B. LLMs as a Source of Contractual Meaning

With a grasp of the technology in hand, let's work through some more quotidian examples of LLMs' potential use outside of the insurance context. Textualists—as we've described—think that texts have an inherent plain meaning, at least within the context of the written document. The problem is deciding what it is, and whether our intuitions are representative. LLMs may serve as powerful tools to uncover those answers.

We'll start with the divorce of Jennie and Mark Famiglio. Jennie and Mark entered into a prenup before getting married, which committed to a sliding scale of payments from Mark to Jennie if they divorced, tied to the length of their union.¹⁵⁴ Section 5.3.a read:

5.3. JENNIE's Benefits and Obligations. If the marriage ends by dissolution of marriage or an action for dissolution of marriage is pending at the time of MARK's death, then JENNIE shall receive the additional benefits and obligations described in 5.3.a. through d.

a. MARK shall pay to JENNIE, within ninety (90) days of the date either party files a *Petition for Dissolution of Marriage* the amount listed below next to the number of full years they have been married *at the time a Petition for Dissolution of Marriage is filed*.¹⁵⁵

Although Jennie filed a petition for divorce after seven years, she never served the petition and later voluntarily dismissed the action.¹⁵⁶ After ten years, she filed again, and meant it. Under the prenup, seven years of marriage entitled her to \$2.7 million; ten years a whopping \$4.2 million.¹⁵⁷ The parties were left with a consequential but basic interpretative question: When the prenup mentions the number of years at the time “a” petition is filed—did the parties mean the *first* petition or the *ultimate* one?¹⁵⁸

¹⁵⁴ Famiglio v. Famiglio, 279 So.3d 736, 737–38 (Fla. Dist. Ct. App. 2019).

¹⁵⁵ *Id.*

¹⁵⁶ *Id.* at 738.

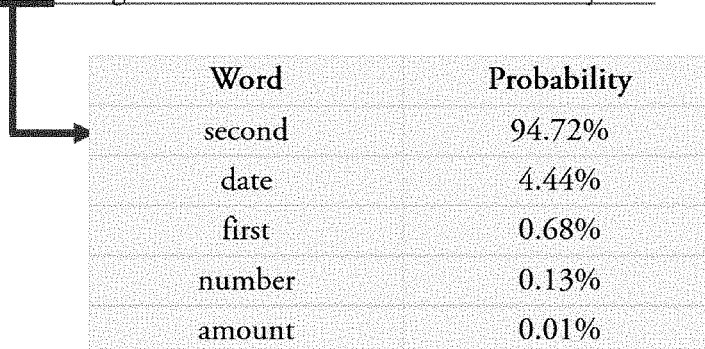
¹⁵⁷ *Id.*

¹⁵⁸ *Id.* at 737 (“In this appeal of a judgment interpreting a prenuptial agreement, the word ‘a,’ the smallest of words in the English language, could mean the difference of a million and a half dollars.”).

Neither party thought witnesses were necessary, as both understood *a Petition* to be unambiguous (and favoring their side).¹⁵⁹ Unfortunately for Jennie, a Florida appellate court ruled against her.¹⁶⁰ Relying in part on dictionaries, it emphasized that “a” is an indefinite article.¹⁶¹ Ordinarily, the court stated, when people predicate a condition on an indefinite event, they mean its first occurrence. Thus, imagine if a golf course posts a rule: “when a thunderstorm approaches, you must end your golf game.”¹⁶² That would be “universally understood to mean the first time a thunderstorm approaches.”¹⁶³ And so, “a” petition filing simply must mean the first one filed. The court’s method of proof seems sensible. But was it right to be so sure of itself?

We presented GPT-4 with the prenuptial agreement and asked it: If one of the parties files a divorce petition, withdraws it, and then a few years later a new petition is filed, what date determines the number of full years of marriage: the first filing or the second one? It produced a sentence that essentially supported Jennie’s view. But to illustrate how the model can help courts be more precise, we can freeze the output in time and take a peek under the hood, as Figure 3 illustrates.

The second filing would determine the number of full years of marriage



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

Figure 3: Davinci-003, temp=1, top-p=1, frequency and repetition penalty=0, best of 1, full spectrum, presented with Famiglio facts and asked “If one of the parties files a divorce petition, withdraws it, and then a few years later a new petition is filed, what date determines the number of full years of marriage: the first filing or the second one?”

¹⁵⁹ *Id.* at 738.

¹⁶⁰ *See id.* at 743.

¹⁶¹ *Id.* at 741.

¹⁶² *Id.* at 743.

¹⁶³ *Id.*

This illustration captures the probabilistic way the model thinks of language and its own process.¹⁶⁴ When it started to produce its answer, it predicted that it ought to start with “The.” Now, neither we nor the model know how it would continue the sentence. It read our question and its partial answer and then made a prediction. Given the context and the vast corpus on which it sits, what should have come next—second or first? It concluded that “second” makes more sense. And once second is produced, the rest of the answer follows.¹⁶⁵

Generative interpretation in this simple case thus offers courts a better sense of the relevant probabilities if the parties were intending to use English in its most public and common sense. And it does so without reference to singular, perhaps idiosyncratic, illustrations pulled from the golf course. Of course, it’s possible that in the context of their deal, extrinsic evidence pointed to a private meaning—or perhaps trade practice could have pushed the court away from the meaning that the model suggests is normal. And, as we’ll discuss, knowing that the court would use the model might have motivated both parties to not so quickly assume that their meaning was unambiguously correct.

C. *The Ambiguity Problem*

As *Famiglio* illustrated, the question of whether a term is ambiguous, permitting extrinsic evidence or not, can be outcome determinative. That’s true for interpretative methods of all stripes. Even the most free-spirited contextualists are not *that* free. They will not waste the parties’ time on a lengthy trial when they think that the language in the contract is simply not “reasonably susceptible” to the interpretation proffered by one of the parties. As a result, a key question in contextualist jurisdictions is which interpretation, exactly, the language is reasonably susceptible to.

Take the well-known case of *Trident v. Connecticut*, often listed as a primary argument against California-style contextualism.¹⁶⁶ A group of lawyers, assisted by other real estate investors, sought to buy commercial real estate to build their law offices. They borrowed \$56 million from Connecticut Insurance, with an agreement to pay it back over fifteen

¹⁶⁴ A cleaner presentation would ask the model to choose a single token from multiple options (as in a multiple choice question where only the number of the answer is requested). This would have the advantage of removing some of the syntactic considerations that inform the probability distribution among tokens, but for the sake of exposition, we retain this simpler presentation.

¹⁶⁵ The usual LLM caveats apply, and the probabilities shouldn’t be interpreted literally. The model could, for example, continue the sentence with “The first filing would not control.”

¹⁶⁶ *Trident Ctr. v. Conn. Gen. Life Ins. Co.*, 847 F.2d 564 (9th Cir. 1988).

years at 12.25% APR. At one point, the agreement stated that the principal could not be prepaid, at least not within the first twelve years of the agreement. However, interest rates fell, and the borrowers sought to prepay the loan with money they would borrow elsewhere.¹⁶⁷ When they were rebuked, they turned to litigation.

The promissory note clearly stated that the borrowers “shall not have the right to prepay the principal amount hereof in whole or in part.”¹⁶⁸ But they pointed to a different clause, creating a 10 percent prepayment penalty for defaulted loans if the lender accelerated.¹⁶⁹ The borrowers’ lawyers relied on the famous statement of California’s contextualism rule, *Pacific Gas*,¹⁷⁰ to argue that they ought to be permitted to offer extrinsic evidence—negotiations, trade usage—in support of their contractual reading.¹⁷¹

In the Ninth Circuit, Judge Kozinski used the case to offer what others have described as a “shrill attack” on the looseness of the California parol evidence rule.¹⁷² He discounted the borrower’s prepayment argument, since it was at the lender’s option. And he concluded that the contract’s “shall not have the right” clause was crystal clear that prepayment was forbidden—standing alone, it was not reasonably susceptible to the borrower’s meaning.¹⁷³ Nonetheless, Judge Kozinski remanded the case. He wrote:

Under *Pacific Gas*, it matters not how clearly a contract is written, nor how completely it is integrated, nor how carefully it is negotiated, nor how squarely it addresses the issue before the court: the contract cannot be rendered impervious to attack by parol evidence. If one side is willing to claim that the parties intended one thing but the agreement provides for another, the court must consider extrinsic evidence of possible ambiguity. If that evidence raises the specter of ambiguity where there was none before, the contract language is displaced and the intention of the parties must be divined

¹⁶⁷ Historic rates had fallen by around 3 percent, meaning an early pre-payment would have meant a saving of ~\$1.1 million over the life of the loan. *Id.* at 6.

¹⁶⁸ Promissory Note at 6, *Trident Ctr. v. Conn. Life Ins. Co.*, 847 F.2d 564 (9th Cir. 1988) (No. 388-030).

¹⁶⁹ See *Trident Ctr.*, 847 F.2d at 567.

¹⁷⁰ *Pacific Gas & Elec. Co. v. G.W. Thomas Drayage & Rigging Co.*, 442 P.2d 641, 645 (Cal. 1968) (holding that “rational interpretation requires at least a preliminary consideration of all credible evidence offered to prove the intention of the parties”).

¹⁷¹ See *Trident Ctr.*, 847 F.2d at 568 (noting reliance on *Pacific Gas*).

¹⁷² Peter Linzer, *The Comfort of Certainty: Plain Meaning and the Parol Evidence Rule*, 71 *FORDHAM L. REV.* 799, 805 n.26 (2002); *Trident Ctr.*, 847 F.2d at 569 (“While we have our doubts about the wisdom of *Pacific Gas*, we have no difficulty understanding its meaning, even without extrinsic evidence to guide us.”).

¹⁷³ *Trident Ctr.*, 847 F.2d at 566.

from self-serving testimony offered by partisan witnesses whose recollection is hazy from passage of time and colored by their conflicting interests.¹⁷⁴

The opinion, written with flair, is in many contracts casebooks, but it is a puzzle in its own right. California's existing rule provided that extrinsic evidence was to be admitted only if the language in the contract was "reasonably susceptible" to the interpretation proffered by the parties.¹⁷⁵ Thus, if Kozinski really had been confident that the language was clear, he should not have remanded.¹⁷⁶ We wondered whether his factual premise was correct and asked LLMs to help.

After obtaining the original promissory note,¹⁷⁷ we introduced the relevant parts to three leading LLMs: GPT-4, Claude 2, and a version of the open source model Llama-2, and then asked for their evaluation.¹⁷⁸ We asked them to read the entire contract and then estimate, as a judge, the likelihood that the parties intended early repayment to be permitted under the agreement. To capture a range of model responses, we repeated the same question many times, while setting the "temperature" at a sufficiently high level to ensure that different responses might be picked.

¹⁷⁴ *Id.* at 569.

¹⁷⁵ *Pacific Gas & Elec. Co.*, 442 P.2d at 644.

¹⁷⁶ See Susan J. Martin-Davidson, *Yes, Judge Kozinski, There Is a Parol Evidence Rule in California—The Lessons of a Pyrrhic Victory*, 25 Sw. U. L. REV. 1, 18–19 (1995) (clarifying California's test for admitting extrinsic evidence). As Prof. Martin-Davidson points out, after remand the defendants won a summary judgment motion and their attorneys' fees. There never was a trial. *Id.* at 4 n.22.

¹⁷⁷ We thank Prof. Todd Rakoff for providing it from his collection.

¹⁷⁸ The 70 billion parameter version of the Llama-2 model was considered the highest performing open source model at the time of writing, and we used the currently highest-performing fine-tuned version of this model, as measured by the HuggingFace LLM Leaderboard. See Riid, *Riid's AI Model Ranks #1 in HuggingFace LLM Leaderboard*, PR NEWSWIRE (Oct. 9, 2023, 9:58 AM), <https://www.prnewswire.com/news-releases/riids-ai-model-ranks-1-in-huggingface-llm-leaderboard-301950871.html> [<https://perma.cc/C39C-V2SG>].

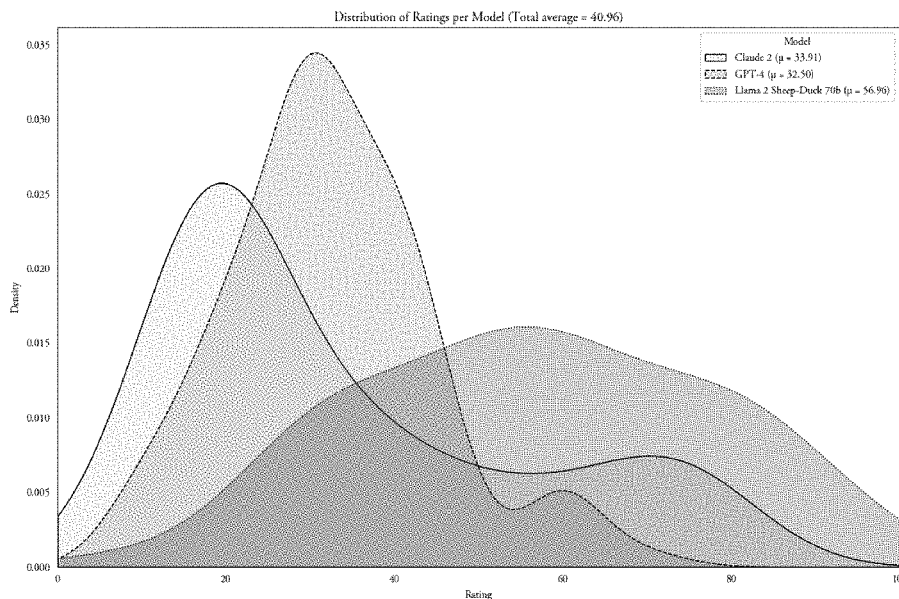


Figure 4: Turbo GPT-4, Claude 2, and Llama-2 70b with set at temperature 1, and fed with the Trident promissory note in full. The models were asked whether the language of the agreement is reasonably susceptible of being read as providing the borrower the right to early repayment. On the x-axis, 0 indicates this interpretation is wrong and 100 is that it is correct.

Figure 4 is suggestive of how generative interpretation can deepen and enrich judicial analysis. Overall, the models roughly agree on average that prepayment is not allowed, with a mean score of ~41. The least powerful model here, Llama-2, was more open to the possibility than the more powerful, proprietary models. But the two most powerful models, Claude 2 and GPT-4, both shared a similar evaluation: they estimated that the majoritarian reading was not that advanced by the Trident group.

One read of this result is that it suggests that Kozinski's intuitive factual premise was wrong, but that he reached the right conclusion. That is, even taking the borrower's argument seriously, the dominant reading rejects a finding of ambiguity. No further extrinsic evidence ought to have been admitted. This would align with common criticisms of the opinion.¹⁷⁹ On the other hand, the models were *not* uniform in their assessment; the probability distribution suggests that at least *some* probabilistic

¹⁷⁹ See Martin-Davidson, *supra* note 176, at 54-55 (arguing that Judge Kozinski erred in admitting extrinsic evidence despite no finding of ambiguity).

readings of the contract permit early repayment. To determine the case, we would want to know more about those minoritarian readings: Are they reflective of discrete linguistic communities, private meanings, or other legally relevant factors? Generative interpretation does not answer the question of whether language is reasonably susceptible of a meaning, it instead helps us visualize a broad spectrum of meaning and quantify how likely a particular result is.¹⁸⁰

Now consider another case turning on ambiguity: *Ellington v. EMI*.¹⁸¹ The issue in this case arose from a 1961 net receipts agreement between the musician Edward Kennedy “Duke” Ellington and his record company, EMI.¹⁸² As was common at the time, the parties agreed on a 50/50 royalty split, after deducting fees charged by third parties that intermediate in foreign markets.¹⁸³ This net receipt agreement bound EMI and its “other affiliate[s].”¹⁸⁴ In the intervening decades, the music industry underwent significant consolidation, and EMI began to use its own affiliates rather than rely on third parties for foreign operations. It sought to deduct those affiliate fees before paying Ellington’s estate.

Feeling blue, Ellington’s grandson sued, arguing that two key phrases in the contract were ambiguous: “(1) the phrase ‘net revenue actually received’ in the royalty provision and (2) the term ‘any other affiliate’ in the definition of Second Party.”¹⁸⁵ The New York Court of Appeals—the country’s preeminent textualist tribunal—rejected the claim. The majority held that the terms were unambiguous: They only reference affiliates that existed at the time of contracting.¹⁸⁶ There is simply no way that they could be read in any other way, given the

¹⁸⁰ Whether a conclusion that is 20 percent likely is legally *reasonable* might turn on several factors we do not explore in the text. Imagine a particular linguistic subcommunity whose understanding of terms correlates with the parties’ (You could think of this as akin to trade usage, but for culture.) In that case, deferring to majoritarian readings would tend to suppress important perspectives. See generally Dan M. Kahan, David A. Hoffman & Donald Braman, *Whose Eyes Are You Going to Believe? Scott v. Harris and the Perils of Cognitive Illiberalism*, 122 HARV. L. REV. 837 (2009) (discussing how simulations can uncover discrete minority perspectives on legally-operative facts that the law should attend); David A. Hoffman, *From Promise to Form: How Contracting Online Changes Consumers*, 91 N.Y.U. L. REV. 1595 (2016) (arguing that younger parties have distinct views of contracting from older ones).

¹⁸¹ *Ellington v. EMI Music, Inc.*, 21 N.E.3d 1000, 1001 (N.Y. 2014).

¹⁸² *Id.*

¹⁸³ *Id.* at 1002.

¹⁸⁴ *Id.* at 1005.

¹⁸⁵ *Id.* at 1003.

¹⁸⁶ *Id.* at 1004.

tense that the parties used and the court's aversion to forward-looking language.¹⁸⁷

Again we had access to the original contract. We presented it to the various models for plain language analysis, asking: "Does 'other affiliates' naturally include only the existing affiliates at the time of contract, or does it potentially encompass affiliates that might be created over time?"

Before we describe the model's answer, we should highlight two robustness concerns with generative interpretation. Models are quite sensitive to the prompt used.¹⁸⁸ This opens them to a problem of "leading prompts," queries that lead the model towards a desired answer. And, as we described earlier, models can be set to be hotter (more random) or colder (more deterministic).¹⁸⁹ This allows the user (judge, researcher, policymaker) many degrees of freedom.

To deal with these issues we tried something new. Rather than a single prompt, we used 20 variations of the same legal question, each queried 10 times at a relatively high temperature setting.¹⁹⁰ We presented yes/no questions where yes indicates agreement with the judge's interpretation. Figure 5 below summarizes the results of the experiment among three of the leading models.

¹⁸⁷ See *id.* at 1004–05.

¹⁸⁸ See generally Laria Reynolds & Kyle McDonell, *Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm*, ARXIV (Feb. 15, 2021), <https://arxiv.org/pdf/2102.07350.pdf> [<https://perma.cc/F8WJ-SHW4>] (discussing the effect of prompting techniques on model outputs).

¹⁸⁹ See *supra* note 145 and accompanying text.

¹⁹⁰ Specifically, we set temperature at 1 and top p=1 to encourage a broad range of responses. The twenty prompts were generated by GPT-4, after seeding it with the background of the case and a seed question. See *Ambiguous Contract Interpretation*, OPENAI (July 5, 2023), <https://chat.openai.com/share/e9003c92-5e32-436c-816d-c2add7ac485b> [<https://perma.cc/2D56-39JJ>]. For code, see *supra* note 22.

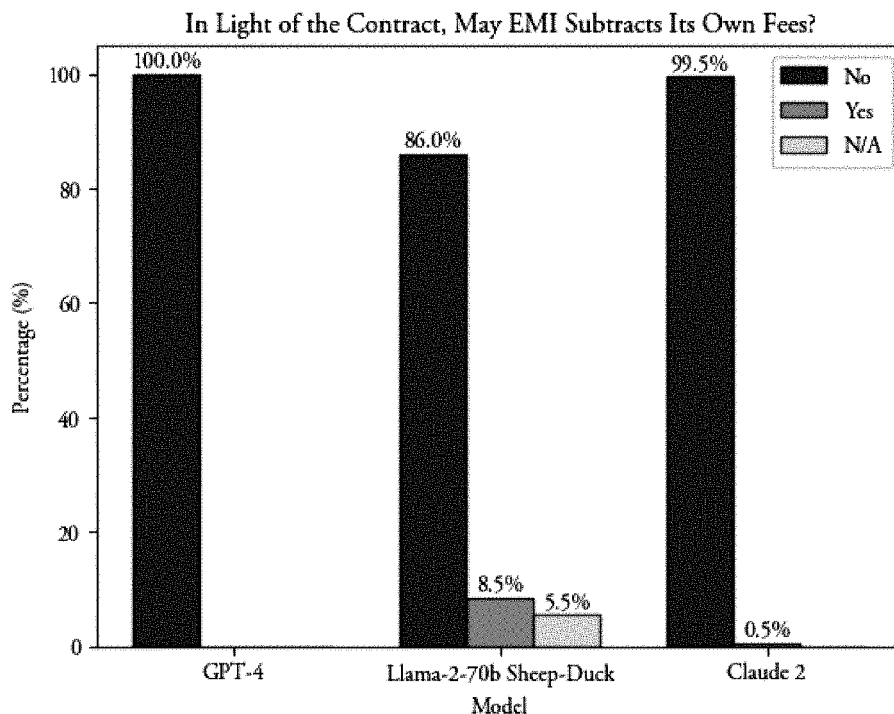


Figure 5: Ellington v. EMI, analyzing the interpretation of “other affiliates” using temperature 1, and responding ten times to twenty prompt variations generated by GPT-4, after seeding it with the background of the case.

As the Figure illustrates, the three models don’t share the New York court’s confidence: The most common interpretation of “other affiliates” includes those that post-date the contract. Llama-2, the open source model, is somewhat open to EMI argument, reflecting that it has some facial plausibility. Of course, even uniformity between powerful models cannot decide cases. The point, rather, is to illustrate the value of LLMs as a convenient check against overconfidence, and a spur to greater reflection. (Though the fact that the dissent thought that the contract was ambiguous might have produced that same introspection.)

Our choice of prompt here and elsewhere is not neutral. But that’s not a unique problem to LLM queries: briefs, jury instructions, testimony, and rules of evidence are all modes of framing that affect, and are often calculated to affect, the judicial output. What we tried to show here is that generative interpretation may offer a means of controlling the inevitable subjectivity of framing. We offered one technique, and future work may measure its debiasing effects, perhaps alongside other novel methods.

D. Filling Gaps

Contracts are incomplete: The parties leave many topics to necessary implication. Such omissions are not always deliberate: Sometimes parties simply have not contemplated a problem—a global pandemic, a supply chain disruption, another peerless ship sailing ex Bombay¹⁹¹—and the court must engage in filling the gaps, rather than merely interpreting words on the page.

Consider the 1977 New York Court of Appeals case, *Haines v. City of New York*.¹⁹² It resolves a dispute about a 1924 contract between the City of New York and an upstate village, in which the City promised to pay the town to process its own sewage so that the city's water supply could be cleaned. (That is, the city paid the village not to pollute.) As the decades passed, the townships grew, and the Federal Government passed environmental regulations. By the early 1970s, facing strong budgetary pressures, New York City refused to continue to pay for the township's expansion of the sewage facilities. A local developer sued, arguing that the contract's absence of a duration term or cabin on the scope of the city's obligation meant that the city was in breach.

The court considered those arguments in a decision that looked only to the written contract. It determined that the parties did not mean for the contract to run forever, in a provision notable for its brevity.

[W]here the parties have not clearly expressed the duration of a contract, the courts will imply that they intended performance to continue for a reasonable time Thus, we hold that it is reasonable to infer from the circumstances of the 1924 agreement that the parties intended the city to maintain the sewage disposal facility until such time as the city no longer needed or desired the water, the purity of which the plant was designed to insure.¹⁹³

The logic here isn't compelling but rests on an empirical prior: By default, parties do not intend contracts to be terminable at will when they write unlimited obligations, and nothing about the language or circumstances of the contract compels a contrary conclusion.

On the related question of whether the city promised (implicitly) to continue to expand the system's capacity, the court was less generous.

By the agreement, the city obligated itself to build a specifically described disposal facility and to extend the lines of that facility to

¹⁹¹ *Raffles v. Wichelhaus*, 159 Eng. Rep. 375 (1864).

¹⁹² *Haines v. City of New York*, 364 N.E.2d 820 (N.Y. 1977).

¹⁹³ *Id.* at 823 (citations omitted).

meet future increased demand. At the present time, the extension of those lines would result in the overloading of the system. Plaintiff claims that the city is required to build a new plant or expand the existing facility to overcome the problem. We disagree. The city should not be required to extend the lines to plaintiffs' property if to do so would overload the system and result in its inability to properly treat sewage. In providing for the extension of sewer lines, the contract does not obligate the city to provide sewage disposal services for properties in areas of the municipalities not presently served or even to new properties in areas which are presently served where to do so could reasonably be expected to significantly increase the demand on present plant facilities.¹⁹⁴

Once more the court alludes to the agreement, but its decision is inattentive to the details. It found an implicit condition to obligation: Extension is required only so long as the system is not overloaded.¹⁹⁵ But this was a gap-filling exercise, informed by the court's judgment about what the parties *should* have said.¹⁹⁶ Such determinations were part of a trend in New York courts in favor of a looser, Cardozoian approach to missing terms.¹⁹⁷

With the cooperation of the New York court system, we obtained the 1924 contract.¹⁹⁸ This contract and the various exhibits are long, especially considering when they were created: about eight pages of Word documents. We entered the text into the two models that can support such long inputs—GPT-4's experimental version and Claude 2—and asked them to assess the validity of several legal arguments given the agreements.¹⁹⁹ Figure 6 illustrates what we found.

¹⁹⁴ *Id.*

¹⁹⁵ Robert M. Jarvis, Phyllis G. Coleman & Gail Levin Richmond, *Contextual Thinking: Why Law Students (and Lawyers) Need to Know History*, 42 WAYNE L. REV. 1603, 1613–14 (1996) (discussing that the City of New York at the time was under severe financial stress and courts rushed to protect it from bankruptcy).

¹⁹⁶ For an argument suggesting that there is no fact-of-the-matter about parties' intent when filling gaps in contracts, see Robert A. Hillman, *More Contract Lore*, 94 TUL. L. REV. 903, 910 (2020); see also Robert A. Hillman, *The Supreme Court's Application of "Ordinary Contract Principles" to the Issue of the Duration of Retiree Healthcare Benefits: Perpetuating the Interpretation/Gap-Filling Quagmire*, 32 ABA J. LAB. & EMP. L. 299, 320 (2017).

¹⁹⁷ Perhaps this part of the opinion responded to the City's financial exigency. See William E. Nelson, *A Man's Word and Making Money: Contract Law in New York, 1920–1960*, 19 MISS. COLL. L. REV. 1, 13 (1998).

¹⁹⁸ E-mail from Marisa Gitto, Reference Services, New York State Library, to Michael Hurley, Research Assistant, University of Pennsylvania Carey Law School (May 22, 2023, 03:01 EST) (on file with authors).

¹⁹⁹ GPT-4, *Talk to GPT-4-32k on Poe*, POE, <https://poe.com/s/Vp9tkyhGnMmHqFvdKp4n> [<https://perma.cc/SPQ4-DMEZ>]; Claude 2, *Talk to Claude-2-100k on Poe*, POE, <https://poe.com/s/DmgexqjOhO6Qx2DADArB> [<https://perma.cc/69HP-ETQY>]. You should take model's self-reported degree of confidence with a grain of salt; it is more meaningful to

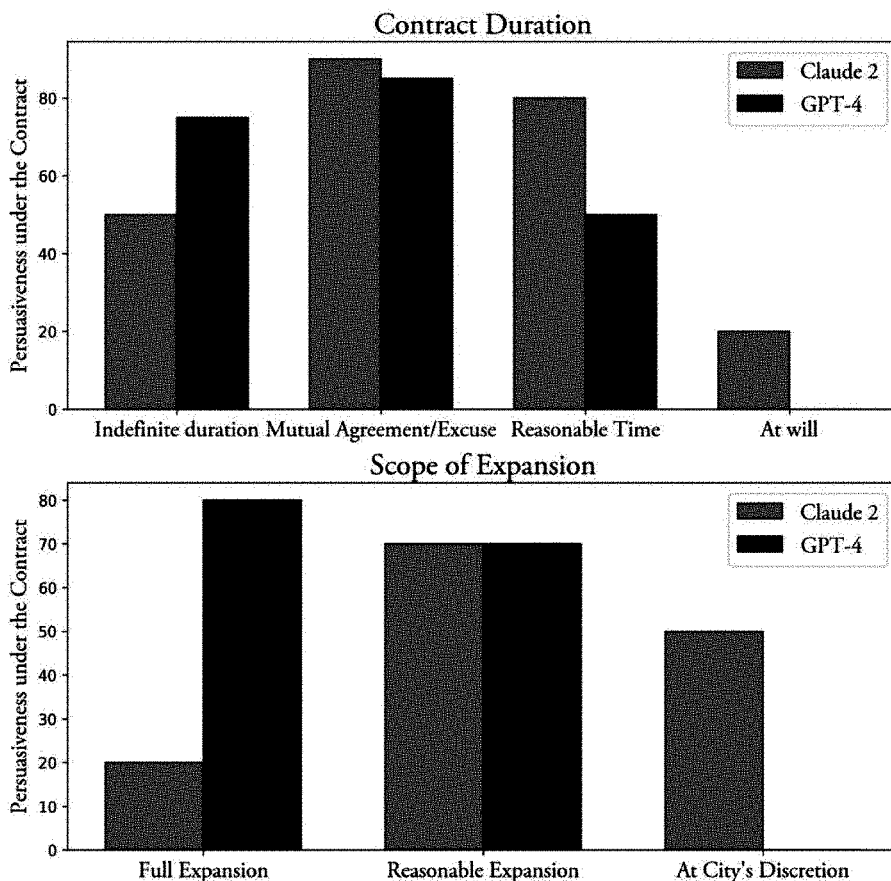


Figure 6: Haines v. City of New York gap filling analysis using Chat GPT-4 (32k context length) and Claude 2 (100k context length).

The first set of questions concerned duration. Neither model's outputs support the city's claim that the parties intended the contract to be terminable at will. And both (with different degrees of confidence) were open to durational gap fillers of an indefinite time, a reasonable time, by joint agreement, or until a time when a legal excuse is present—which is indeed the common law rule for most contracts.²⁰⁰ GPT-4 (like the court) explained, “while a reasonable duration might be inferred

simply compare its expressed confidence with respect to different questions, hence our experiment design here.

²⁰⁰ See *Glacial Plains Coop. v. Chippewa Valley Ethanol Co., LLLP*, 912 N.W.2d 233, 234 (Minn. 2018) (holding that unless otherwise provided, a “contract is of indefinite duration and is terminable at will by either party after a reasonable time and with reasonable notice”).

under common law principles, this argument does not strongly accord with the contract's language."²⁰¹ Overall, the models appear to generally support the court's reading.

The second set of questions involved the scope of the city's obligations. GPT-4 disagreed strongly with the court; it thought that the city's obligation was unbounded. Importantly, it anchored its reasoning in a section of the contract neglected by the court: Section 6. That part obligates the city to extend sewage plans "[w]henver extensions of any of the sewer lines are necessitated by future growth . . . of the respective communities"²⁰² This provision implied the obligation to build additional treatment plants. But Claude 2 was more amenable to the court's interpretation and provided a plausible constraining argument: "The agreement provides for extensions when required by growth, implying a reasonable obligation."²⁰³

E. From Text to Context

So far, we have provided examples that showcase how large language models might power a stronger, cheaper, more robust form of textualism. We now consider how such models can account for contextual evidence such as prior conversations, shared expectations, and industry standards. *Stewart v. Newbury* provides a simple illustration.²⁰⁴ In *Stewart*, a contractor and a business corresponded about the construction of a new foundry. The contractor's offer letter was brief; he offered to do the job and charge either by offering an itemized list or by charging on a cost + 10% basis. This letter was followed by a telephone call where they may have agreed that payment would be made "in the usual manner."²⁰⁵ Finally, the foundry responded in writing that, following the phone conversation, they accepted the bid. As far as we know, that amounts to the entirety of the contracting case file.²⁰⁶

Once the contractor finished the first part of the project, he submitted a bill. The foundry refused to pay. The contractor insisted

²⁰¹ GPT-4, *Talk to GPT-4-32k on Poe*, POE, <https://poe.com/s/Vp9tkyhGnMmHqFvdKp4n> [<https://perma.cc/SPQ4-DMEZ>]. Likewise Claude 2 explained: "A reasonable duration could be implied, though not explicitly stated." Claude 2, *Talk to Claude-2-100k on Poe*, POE, <https://poe.com/s/DmgexqjOhO6Qx2DADArB> [<https://perma.cc/69HP-ETQY>].

²⁰² Agreement Between the Village of Tannersville and the Town of Hunter and the City of New York for Sewage Systems (Aug. 15, 1924) (on file with authors); GPT-4, *Talk to GPT-4-32k on Poe*, POE, <https://poe.com/s/Vp9tkyhGnMmHqFvdKp4n> [<https://perma.cc/SPQ4-DMEZ>].

²⁰³ Claude 2, *Talk to Claude-2-100k on Poe*, POE, <https://poe.com/s/DmgexqjOhO6Qx2DADArB> [<https://perma.cc/69HP-ETQY>].

²⁰⁴ 115 N.E. 984 (N.Y. 1917).

²⁰⁵ *Id.* at 984.

²⁰⁶ *Id.* at 984-85.

that it was customary to pay 85 percent of payments due at the end of every month, but the foundry argued that its payments were only due on (substantial) completion of the project. Seeing no payments made, the contractor stopped work. The parties countersued for breach.

Today, the default rule is that payments in construction contracts are not due until the contract is substantially performed.²⁰⁷ It is unclear that this rule was in place when the parties agreed in 1919. The foundry argued that no payment was due under the contract, and hence, the contractor's refusal to work was wrongful. So now we have an interpretive question: Did the parties agree to a particular payment regime?

The written agreement is too sparse to help, but the phone conversation offers an in. If we believe that the parties indeed agreed to make payments *in the usual manner*, then it is possible to interpret *usual* as referring to an alleged common practice of monthly installment payments. It is also possible, however, that "usual" refers to other standard payment conventions—say, the payment on a cost +10% basis.²⁰⁸

The court left the interpretive question undecided because it remanded for faulty jury instructions. We, however, are not so constricted. We asked today's leading LLMs, GPT-4 and Claude-2, to predict what the parties meant. To do so, we first told the models to assume that the default legal rule would be that payment is conditioned on substantial performance.²⁰⁹ Then, we asked the models to estimate how the parties would have interpreted their deal absent consideration of either extrinsic evidence of the phone conversation or evidence of industry norms. We then added the evidence of the phone conversation, to see how the model's confidence changed, and finally, we added evidence of the custom in the industry. Table 2 summarizes the results²¹⁰:

²⁰⁷ See 22 N.Y. JUR. 2D CONTRACTS § 352; Hillman, *supra* note 196, at 313 (“[C]ourts in construction cases find a duty to pay only after substantial performance.”).

²⁰⁸ See *Stewart*, 115 N.E. at 985.

²⁰⁹ This is not obviously the correct legal rule, then or now, but we had to start somewhere, and we took the court at its word.

²¹⁰ CLAUDE 2.0 POE CONVERSATION, <https://poe.com/s/wLkeCDrPdFpKye3uApSa> [<https://perma.cc/B67E-WJER>]; GPT-4 CONVERSATION, <https://chat.openai.com/share/e05eb214-7e18-46e9-b44a-5a20d6ab3712> [<https://perma.cc/3ML3-R73E>]. Again, you should be skeptical of model's expressed confidence; the direction of change with every new piece of evidence, not its quantification, is informative. For a review of the problem of confidence elicitation, see MIAO XIONG ET AL., *supra* note 153.

Does the Owner have to pay monthly? (Instead of after substantial performance)		
	GPT-4	Claude-2
Letter contract alone	10%	10%
+Phone call	75%	20%
+Industry Norm	95%	90%

Table 2: Expressed confidence in “the duty to pay is monthly” based on legal and transactional context. Presented to GPT-4 (32k context window) and Claude-2 (100k context window).

Table 2 demonstrates how each additional piece of evidence alters the analysis. And for purposes of this case, it shows that, for the models at least, extrinsic evidence was materially important to the outcome.

Illustrating the additional value of each piece of evidence can provide unexpected insight. Judges may fairly worry, when considering potentially unreliable evidence, that mere exposure to the evidence would irreversibly prejudice their decisions. By estimating the probative value of some forms of evidence before closely examining them, the judge can develop a heuristic assessment of probative value with relatively little exposure. The model can thus give structure to the evaluation of extrinsic evidence, making it more attractive to factfinders. And within the limits of its prompts, its conclusions are coherent, cheap, and seemingly plausible.

We traversed ample ground in this Part, seeing various modes of interaction with LLMs and illustrating their *sensibility*. Relative to previous generations of AI, we are struck by how facially reasonable the models’ outputs are. But to feel confident about using these models in the wild, jurists will naturally want to know more about edge cases. How sensitive are the results to the specific prompts (or their multiple variations) we used? How robust are the models themselves, and whether there are opportunities for litigants to manipulate their outputs? Can we quantify the accuracy of these models relative to the *ground truth* intent of the parties? Let’s now consider these problems (and others) in a bit more detail.

III

THE FUTURE OF CONTRACT INTERPRETATION

So convenient are today’s LLMs, and so seductive are their outputs, that it would be genuinely surprising if judges were not using them to

resolve questions of contract interpretation as this article goes to press, about a year after the tools went mainstream. Looking at practical guidance offered to lawyers in the summer of 2023, we see lawyers encouraged to use LLMs to perform legal research, draft deposition questions and contracts, and predict settlement values.²¹¹ And there are hints that judges are already using ChatGPT to answer other kinds of interpretative questions, just as they would use Google.²¹² In one recent survey, one-quarter of judges confessed to using the tool, though many expressed concern about its reliability.²¹³

These models are useful because they offer new tools—fast, cheap, sometimes incorrect ones—in service of old interpretative goals. Courts will soon take a phrase like “dozen” and ask ChatGPT to interpret it, rather than turning to the dictionary or Google; or will ask the model what’s the likely assumption a contract makes when it leaves a gap; or will check if the model thinks an insurance policy contemplated deft burglars. They’ll do so both covertly and overtly, both *sua sponte* and in response to briefing. Almost certainly the first briefs to affirmatively argue for the use of the tool will come from resource-constrained firms. As we illustrated in Part II of this Article, LLMs are already applicable to live problems that courts face every day, and it would be naïve to think they aren’t using them.

Indeed, we’ve seen this story play out many times before. As some readers will recall, when courts first realized that Wikipedia could be used as a source of information,²¹⁴ they were chastised for its use by higher courts,²¹⁵ and then it was eventually folded into the normal set of legal research tools.²¹⁶ But at least in the short run, judges won’t have the tool

²¹¹ Catherine Casey, Ronald J. Hedges, Marissa J. Moran, Stephanie Wilson, Webcast Presentation at A.L.I., Generative Artificial Intelligence (“GAI”) in Practice: What It Is and How Lawyers Can Use It (June 28, 2023) (on file with authors).

²¹² See Luke Taylor, *Colombian Judge Says He Used ChatGPT in Ruling*, THE GUARDIAN (Feb. 2, 2023, 9:53 PM), <https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling> [<https://perma.cc/3AXC-GVN5>] (discussing use by judges of ChatGPT in rulings).

²¹³ Ed Cohen, *Most Judges Haven’t Tried ChatGPT, and They Aren’t Impressed*, THE NAT’L JUD. COLL. (July 21, 2023), <https://www.judges.org/news-and-info/most-judges-havent-tried-chatgpt-and-they-arent-impressed> [<https://perma.cc/3LZU-JD3G>].

²¹⁴ See Lee F. Peoples, *The Citation of Wikipedia in Judicial Opinions*, 12 YALE J. L. & TECH. 1, 28 (2009) (“Citations to Wikipedia entries in judicial opinions have been steadily increasing since the first citation appeared in 2004.”).

²¹⁵ See *Campbell ex rel. Campbell v. Sec’y of Health & Hum. Servs.*, 69 Fed. Cl. 775, 781 (2006) (rejecting special master’s reliance on Wikipedia, among other online sources, citing several “disturbing” disclaimers on the website and that it could be edited by “virtually anyone”); see also Kenneth H. Ryesky, Letter to the Editor, *Downside of Citing to Wikipedia*, 237 N.Y. L.J. 23, Jan. 18, 2007.

²¹⁶ See Jodi L. Wilson, *Proceed with Extreme Caution: Citation to Wikipedia in Light of Contributor Demographics and Content Policies*, 16 VAND. J. ENT. & TECH. L. 857, 907 (2014)

draft opinions. And why would they? That courts are irreducibly part of the interpretative enterprise—no matter how sophisticated prediction machines get—follows from the obvious point that there are two stages to every contract interpretation problem: figuring out what the parties meant (at contracting), and deciding the “legal significance that should attach to the semantic content.”²¹⁷ The LLM method is simply better for many reference purposes than those currently on offer.

The problem then is not *whether* courts will use LLMs as an aid to interpretation, but *how*. Generative interpretation is a tool and as such, it has strengths, limits, and flaws. To be sure, AI’s most enthusiastic wielders will be its least careful adopters. Thus, our goal in Section III.A is to delimit some principles and limitations for LLM usage by lawyers and judges. With the proper usage of the tool in mind, in Section III.B we suggest that generative interpretation has implications for the continuing vitality of longstanding debates between textualism and contextualism. Or to put it differently, while the uses that we suggest in Section III.A could be thought of as Textualism 2.0—better dictionaries and canons—we don’t think that’s the practical limit of what this method of interpretation can do.

A. Interpretation for the 99%?

As we’ve said, in the coming months and years, we’re sure you will read examples of lawyers and judges using ChatGPT and related tools in perverse, sometimes outright silly ways, and reaching absurd results you think would have been avoided had they just buckled down and done their jobs like careful jurists ought to.²¹⁸ Or, worse, they’ll have these tools generate pedestrian prose that looks like soulless briefing or opinion-writing, but in fact is built on a throne of lies. There’s no question that AI will sometimes be a crutch for lazy or harried lawyers who simply didn’t focus on the details: It might not be ideally pitched at the kinds of people who are reading sentences with care 20,000 words into a law review article.

And yet it’s precisely because LLMs are cheap and workmanlike that they will be attractive to those who want to improve contract interpretation. The biggest single problem with all currently available approaches to contract interpretation isn’t that they are incapable of

(“The advent of Wikipedia and other technological advances has changed legal research. It is unrealistic to believe that the legal community can ignore that reality.”).

²¹⁷ Schwartz & Scott, *Contract Theory and the Limits of Contract Law*, *supra* note 38, at 568 n.50; Edwin W. Patterson, *The Interpretation and Construction of Contracts*, 64 COLUM. L. REV. 833, 833–35 (1964); Klass, *supra* note 45.

²¹⁸ See *Park v. Kim*, 91 F.4th 610, 616 (2d Cir. 2024) (referring a lawyer to the court’s grievance panel for using LLMs to do legal research, badly).

getting correct results some of the time. It's that they are inaccessible to ordinary parties.²¹⁹ Non-wealthy individuals who suffer breach have to lump it,²²⁰ tilt against corporations in internal dispute resolution systems,²²¹ or face financially ruinous fees and prevail in pyrrhic victories.²²² Simply put: There is an access-to-justice problem at the center of contract law as pernicious as the better recognized ones in criminal and constitutional adjudication. The costs and uncertainties of interpreting deals, which form the core of contract litigation, materially contribute to this problem.²²³

Costly interpretation burdens judges too. Chambers are not endowed with reference experts on call for every query. Courts have fewer resources and competencies than the layperson would imagine. This stylized fact alone can explain why dictionaries are popular, and why corpus linguistics is at best experimental; why law office history exists but not law office econometrics; and perhaps even why federal precedent on state issues is more cited than the relevant state law, given that the former is thoroughly indexed in common commercial databases and the latter is not.²²⁴ To substitute for dictionaries and familiar Latin canons, new interpretative tools must be free (or nearly so) and widely available. LLMs satisfy those conditions. Already today, interactions through a chat interface do not require more skill than using a search engine. The deft burglar example offers a proof of concept, and the remaining examples (though not immediately available in your chatbot window) are likely months, not years, away.

²¹⁹ See LEGAL SEVS. CORP., *THE JUSTICE GAP: MEASURING THE UNMET CIVIL LEGAL NEEDS OF LOW-INCOME AMERICANS* 6 (2017) (“86% of the civil legal problems reported by low-income Americans in the past year received inadequate or no legal help.”); E.H. Geiger, *The Price of Progress: Estimating the Funding Needed to Close the Justice Gap*, 28 CARDOZO J. EQUAL RTS. & SOC. JUST. 33, 34–39 (2021) (documenting an array of causes behind the “justice gap”).

²²⁰ Geiger, *supra* note 219, at 37 (“[T]he average household faces 9.3 legal issues per year. 65% of those problems are never resolved; potentially because the claimants cannot afford counsel and do not have the legal literacy to pursue their claims pro se.”).

²²¹ See generally Rory Van Loo, *The Corporation as Courthouse*, 33 YALE J. REG. 547 (2016) (describing internal dispute resolution systems by firms).

²²² See Matthew R. Hamielec, *Class Dismissed: Compelling a Look at Jurisprudence Surrounding Class Arbitration and Proposing Solutions to Asymmetric Bargaining Power Between Parties*, 92 CHI.-KENT L. REV. 1227, 1231 (2017) (arguing that class action waivers and arbitration provisions can result in “negative value suits” where low-resource claimants are pitted against wealthier opponents); see also Gideon Parchomovsky & Alex Stein, *The Relational Contingency of Rights*, 98 VA. L. REV. 1313, 1340 (2012) (noting that class actions can transform individual negative value suits into a single positive value action).

²²³ See Ben-Shahar & Strahilevitz, *supra* note 19, at 1757–58 (discussing interpretation costs); CATHERINE MITCHELL, *INTERPRETATION OF CONTRACTS: CURRENT CONTROVERSIES IN LAW* 110 (2007) (noting expenses associated with contextual approaches to interpretation).

²²⁴ Samuel Issacharoff & Florencia Marotta-Wurgler, *The Hollowed Out Common Law*, 67 UCLA L. REV. 600, 600 (2020) (documenting the “dominance of the federal forum”).

Generative interpretation is a tool which responds to this access-to-justice concern, at several levels.

First, if courts commit to the method, the costs of achieving accuracy in contract interpretation disputes will fall.²²⁵ That's so because the less precise, even if relatively cheap, forms of textualist evidence—dictionaries and canons—will be replaced by better ones. As dispute costs fall and outcomes become more predictable, the returns to opportunistic breach, which generally benefits sophisticated players, will fall.²²⁶ It's true that models may arise to compete in the market, but as we've shown above, more sophisticated models tend to converge on meaning: unlike dictionaries, they are not offering idiosyncratic and curated definitions which differ across people, place, and time.

Second, as outcomes become more certain, and the cost of predicting them falls, there will be *fewer cases* to adjudicate, because parties will likely have a much better sense of what they'll get at verdict, and settle accordingly.²²⁷ LLMs, unlike legal dictionaries, require no specialized legal knowledge to access, and their ease of use will likely improve with time. This implies that there will be a levelling of access to information about law, and a redistribution from more to less repeat players. Further, better calibrated results *ex post* means that parties can spend less time (and money) contracting *ex ante*.²²⁸ A promise of generative interpretation—which it may yet fulfill—is that it will open a form of textualism up to the 99%.²²⁹

The pages of law reviews are littered with proposed technological solutions to supposed problems of excessive legal costs, and unequal access to information about legal outcomes, which turn out to be

²²⁵ Cf. Schwartz & Scott, *Redux*, *supra* note 38, at 930 (noting the primacy of cost in evaluating the correct interpretative rules).

²²⁶ Cf. Eric A. Posner, *A Theory of Contract Law Under Conditions of Radical Judicial Error*, 94 Nw. U. L. REV. 749, 766–69 (2000) (noting that deterministic legal rules discourage opportunistic breach).

²²⁷ Cf. Schwartz & Scott, *Contract Theory and the Limits of Contract Law*, *supra* note 38, at 603 (“When a standard governs, the party who wants to behave strategically must ask what a court will later do if the party is sued. The vaguer the legal standard and the more that is at stake, the more likely the party is to resolve doubts in its own favor.”). This is a partial equilibrium analysis—better adjudication processes invite more commercial activity, which in turn increases contracting.

²²⁸ See Spencer Williams, *Predictive Contracting*, 2019 COLUM. BUS. L. REV. 621 (2019) (arguing that parties could use information about contract outcomes, harnessed through machine learning of large datasets, to change out the contract *ex ante*). But for an insightful discussion of how selection operates to make difficult machine predictions about litigation outcomes, see David Freeman Engstrom & Jonah Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. PA. L. REV. 1001, 1065–67 (2021) (discussing obstacles to prediction).

²²⁹ Schwartz & Scott, *Redux*, *supra* note 38, at 941 (“[T]he more time the court spends on a particular interpretive issue, the less time it can spend on other issues or other cases.”).

either more intractable than the authors thought or ignore virtues that the authors discounted. We should proceed with care, especially when recommending the widespread adoption of a chatbot that sits on matrices whose outputs even its creators do not well-understand. The question is not (in our view) whether generative interpretation offers predictions that are superior in all cases to artisanal, careful, linguistic analysis. It's whether the method is *good enough*, if not today then soon, for resource-deprived courts to adopt in ordinary cases. In evaluating that question of basic competency, it's meaningful, but not dispositive, that even today's unspecialized models can replicate the results of well-considered cases (as Part II explored) and visually illustrate the range of interpretative outcomes that judges might benefit to see.

But Part II offered a curated tour of generative interpretation's greatest hits. It didn't show you where things can go wrong. To make this set of tools perform as well as it can, users should be cognizant of these issues and use it according to newly evolving best practices. To begin, let's start with hallucinatory outputs.²³⁰ In a now-famous case from May 2023, lawyers in a New York federal court turned to ChatGPT for help researching a motion. The tool obliged with helpful cites, but unfortunately had completely made up the opinions in question.²³¹ A sanctions order and plenty of bad press followed.²³² In response to the case, other judges have required lawyers to certify that they had not used any form of Artificial Intelligence in their filings.²³³

False outputs arise from the predictive nature of generative models.²³⁴ Hallucinations are generated texts asserting facts that are

²³⁰ See Sharon D. Nelson, John W. Simek & Michael C. Maschke, *Beware of Ethical Perils When Using Generative AI*, MD. STATE BAR ASS'N (Apr. 19, 2023), <https://www.msba.org/beware-of-ethical-perils-when-using-generative-ai> [<https://perma.cc/CA2W-RZE6>] (“In fact, it can come up with very plausible language that is flatly wrong. It doesn't ‘mean to’ but it makes things up—and that is what AI researchers call a ‘hallucination’ . . .”).

²³¹ Benjamin Weiser, *Here's What Happens When Your Lawyer Uses ChatGPT*, N.Y. TIMES (May 27, 2023), <https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html> [<https://perma.cc/J7A2-TN7X>].

²³² See *Mata v. Avianca, Inc.*, No. 22-cv-1461, 2023 WL 4114965 (S.D.N.Y. June 22, 2023).

²³³ Devin Coledwey, *No ChatGPT in My Court: Judge Orders All AI-Generated Content Must Be Declared and Checked*, TECHCRUNCH (May 30, 2023, 7:32 PM), <https://techcrunch.com/2023/05/30/no-chatgpt-in-my-court-judge-orders-all-ai-generated-content-must-be-declared-and-checked> [<https://perma.cc/82Z5-A6F7>] (explaining the order, which states that “no portion of the filing was drafted by generative artificial intelligence” (such as ChatGPT, Harvey.AI, or Google Bard) or that any language drafted by generative artificial intelligence was checked for accuracy, using print reporters or traditional legal databases, by a human being).

²³⁴ See Benj Edwards, *Why ChatGPT and Bing Chat Are So Good at Making Things Up*, ARS TECHNICA (Apr. 6, 2023, 11:58 AM), <https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them> [<https://perma.cc/PT62-7HBD>] (“[T]he model is fed a large body of text . . . and repeatedly

not quite true.²³⁵ Large language models, remember, are statistical tools optimized to make predictions. But LLMs are not like a helpful librarian that simply pulls out the most relevant book on a topic. Facts are stored in the LLM similar to the way other reasoning and statistical facts are stored, as floating points in a labyrinthian array of vectors. When asked to provide a source on a legal matter, the model employs the same method to elicit both facts and inferences. The output doesn't distinguish facts from inferred facts, and sometimes will predict the world incorrectly.

Recent work has made significant advances in understanding and mitigating hallucination errors, and more powerful models are less susceptible.²³⁶ One solution that is already used in some contexts is connecting the model to a database of facts, so that it can act more like the helpful librarian.²³⁷ Another involves reflective self-evaluation.²³⁸ So while it is appropriate to pay attention to the hallucination problem, we tend to think that this problem will be less salient in the future than it is today. That said, as a best practice, judges would do well to cross-verify

tries to predict the next word in every sequence of words. If the model's prediction is close to the actual next word, the neural network updates its parameters to reinforce the patterns that led to that prediction.""); u/wakka55, REDDIT (Apr. 16, 2023, 2:48 PM), https://www.reddit.com/r/OpenAI/comments/12okltx/openais_whisper_api_sometimes_returns_what_looks [<https://perma.cc/YU4L-KPS6>] (showing that this problem is not limited to textual generation).

²³⁵ See Beren Millidge, *LLMs Confabulate Not Hallucinate*, BEREN'S BLOG (Mar. 19, 2023), <https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate> [<https://perma.cc/5RP8-GNSF>] (describing problem).

²³⁶ See, e.g., Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Zin Zhao, Jian-Yun Nie & Ji-Rong Wen, *The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models*, ARXIV: 2312.10997 (Jan. 5, 2024), <https://arxiv.org/abs/2312.10997> [<https://perma.cc/5DH7-BXZS>] (studying the relationship between model architecture, training, fine-tuning, prompting, and hallucinations); Matt L. Sampson & Peter Melchior, *Spotting Hallucinations in Inverse Problems with Data Driven Priors*, ARXIV: 2306.13272 (June 23, 2023), <https://arxiv.org/pdf/2306.13272.pdf> [<https://perma.cc/4FGM-LMUG>] (arguing that hallucinations can be qualitatively differentiated from fact-based inferences by focusing on activation regions); see also Philip Feldman, James R. Foulds, Shimei Pan, *Trapping LLM Hallucinations Using Tagged Context Prompts*, ARXIV 2306.06085 (June 9, 2023), <https://arxiv.org/abs/2306.06085> [<https://perma.cc/C98S-NS6A>]; Ayush Agrawal, Mirac Suzgun, Lester Mackey & Adam Tauman Kalai, *Do Language Models Know When They're Hallucinating References?*, ARXIV 2305.18248 (Sept. 13, 2023), <https://arxiv.org/abs/2305.18248> [<https://perma.cc/F6KW-N63L>]; Gabriel Poesia, Kanishk Gandhi, Eric Zelikman & Noah D. Goodman, *Certified Reasoning with Language Models*, ARXIV 2306.04031 (June 6, 2023), <https://arxiv.org/pdf/2306.04031.pdf> [<https://perma.cc/H487-9JH7>].

²³⁷ See generally James Briggs & Francisco Ingham, *Fixing Hallucination with Knowledge Bases*, PINECONE, <https://archive.pinecone.io/learn/langchain-retrieval-augmentation> [<https://perma.cc/A9AG-84LV>].

²³⁸ Charlie George & Andreas Stuhlmüller, *Factored Verification: Detecting and Reducing Hallucination in Summaries of Academic Papers*, ARXIV 2310.10627 (Oct. 16, 2023), <https://arxiv.org/abs/2310.10627> [<https://perma.cc/MU88-N42Y>].

the answers that they get from one platform against another, just as in the early days of legal research it would pay to check both Lexis and Westlaw to make sure that your research was complete.²³⁹

Second, models are subject to manipulation. Large language models are malleable; “leading prompts” can lead them to different conclusions. This is roughly analogous to leading questions for witnesses or jury instructions that frame disputes for or against a particular outcome. As anyone who has experience with an LLM chatbot will attest, it is relatively easy to drive conversations toward desired outcomes. In litigation practice, we expect parties to battle over prompts as well as over models, just as they vie to control the framing of the legal questions and forum in litigation today. In response, factfinders can (as we illustrated above) ask the model to itself produce competing prompts, and then, rather than relying on a single query, the factfinder can look at the general trend of responses and share those varying outcomes in their decisions. Such and other innovative methods could potentially control prompt bias. Factfinders will also have to decide whether to defer to the parties’ choice of model, should they make that explicit in their contract. Such choices exist with other interpretative methodologies, but the existence of choices does not mean that all systems are equivalent. Generative interpretation allows participants more control over the process. Instead of a single jury instruction, judges can experiment with multiple variations. Disputes over which models to use can be refined over time, through either contractual choices or emerging industry norms. If manipulation is suspected, parties could litigate it both at trial and on appeal.

The Katrina analysis raises the related problem of model interpretability.²⁴⁰ The way models encode language is not based in semantics. Unlike human-based reasoning, models have a precise sense in which “chocolate” is closer to “bread” than to “nutrition.” This precision can be misleading if interpreted naively. The Katrina example illustrates how distances correspond with a sensible account of meaning. It also shows that the policy exceptions were closer to “fire” than to the arbitrarily chosen word “police.” It is difficult to understand why, precisely, this result followed. Possibly, fire is a category of disaster and in this sense it is closer to the insurance policy. Still, it would be misleading to say that the policy excludes fire damage rather than damage caused by the police. Other terms may lead to more counterintuitive results.

²³⁹ See generally Robert J. Munro, J. A. Bolanos & Jon May, *LEXIS vs. WESTLAW: An Analysis of Automated Education*, 71 L. LIBR. J. 471 (1978) (evaluating platforms against each other).

²⁴⁰ See *supra* notes 1–25 and accompanying text.

This interpretability gap should caution care in the direct translation of model outputs to legal judgments. Yet, it is also the case that, on average, these models predict with great accuracy linguistic distinctions that humans make.²⁴¹ This presents a general tension in language models. They are *generally* extremely good at capturing meaning, but they still make errors and it is not always possible to rationalize or foresee these errors.

It is difficult to talk about interpretability without invoking questions about the rule of law. Surely there is something eerie about ceding control over litigation to poorly understood black boxes. This is why we do not propose this strategy. We see LLMs as performing best in their natural domain, text analysis, not human-critical decisionmaking. Hard choices, like what degree of relatedness is enough to establish that floods are equally understood to be caused by natural and non-natural causes, shouldn't be sloughed off to machines. Exerised in their proper domain, i.e., text analysis, the sort of questions LLMs raise about the rule of law are not qualitatively *that* different than the opaque choices of dictionary editors.

A third consideration focuses on the models' strength: They are naturally inclined to make predictions that maximize probability based on the training text—in other words, they are biased, in a rough sense, towards majoritarian interpretations. Models offer an approximation of general understanding that may simply not be available in any other way, and thus advance long-held goals of contract theory.²⁴² But majoritarian interpretations are just that: They embed and advance the values of the majority. This is doubly problematic. First, courts really ought to be attentive to local, more private, meanings: Public meaning is second best, prioritized because it is efficient and not because it is correct.²⁴³ But more generally, because the linguistic conventions of underrepresented communities are submerged by majoritarian public meanings, they will

²⁴¹ For a discussion of the evaluation metrics, see Niklas Muennighoff, Nouamane Tazi, Loïc Magne & Nils Reimers, *MTEB: Massive Text Embedding Benchmark*, in PROCEEDINGS OF THE 17TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 2014–37 (May 2023), <https://aclanthology.org/2023.eacl-main.148.pdf> [<https://perma.cc/XW3U-E4CZ>].

²⁴² Schwartz & Scott, *Contract Theory and the Limits of Contract Law*, *supra* note 38, at 583–84.

²⁴³ For the foundational work distinguishing local from popular interpretative modes, see 2 SAMUEL WILLISTON, *THE LAW OF CONTRACTS* § 604 (1920). Even textualists understand that strict adherence to the public meaning of words, bereft of any commercial understanding of what the parties could have been doing, will sometimes lead courts astray. *See generally* Stephen J. Choi, Mitu Gulati & Robert E. Scott, *The Black Hole Problem in Commercial Boilerplate*, 67 *DUKE L.J.* 1, 2 (2017) (describing *pari passu* clauses as “a standard provision in sovereign debt contracts that almost no one seems to understand”).

find it more difficult to have their voices surfaced (and thus subsidized) in contract adjudication. Majoritarian interpretative approaches risk silencing entire communities.²⁴⁴

Surely, this is not a problem unique to generative interpretation: dictionaries, canons, and corpora are equally, if not more, vulnerable to the charge.²⁴⁵ And unlike dictionary-and-canon-textualism, it is at least theoretically possible to counter the majoritarian bent of models in several ways. Models trained on curated datasets that reflect the linguistic conventions of distinct communities would bend towards the majoritarian patterns within those communities. Adjustments to the model's hyperparameters elicit more or less majoritarian behaviors from the model. And careful prompt engineering can attune the model to specific contexts.²⁴⁶ This is an active area of research and regulatory scrutiny and should check factfinders.²⁴⁷

Fourth, models may become subject to parties' adversarial attacks or prompt injections, or will be otherwise fragile in unexpected ways.²⁴⁸ By way of illustration, modern AI systems can reliably differentiate between pictures of panda bears and horses or stop signs and yield signs. But if a sophisticated party can imperceptibly change the color of a pixel here and there, that will be enough to make the model erroneously see a horse or a yield sign.²⁴⁹ The same manipulations can be used to "attack" LLM models.²⁵⁰ Slight changes in the wording of a contract—e.g., subtle changes in the presentation of the words—might hack the model logic

²⁴⁴ See, e.g., Majorie Florestal, *Is a Burrito a Sandwich? Exploring Race, Class, and Culture in Contracts*, 14 MICH. J. RACE & L. 1, 36–39 (2008) (discussing role of race and class in an interpretation dispute); Alexandra Buckingham, Note, *Considering Cultural Communities in Contract Interpretation*, 9 DREXEL. L. REV. 129, 156–60 (2016) (arguing for the use of cultural meaning in interpretation); see also *supra* note 172.

²⁴⁵ Steven J. Burton, *A Lesson on Some Limits of Economic Analysis: Schwartz and Scott on Contract Interpretation*, 88 IND. L.J. 339, 350–51 (2013) (arguing that majoritarian readings can privilege certain views).

²⁴⁶ For an illustration, see Arbel & Becher, *supra* note 15, at 99–104.

²⁴⁷ See *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, at 4, COM (2021) 206 final (Apr. 21, 2021) (stating that a goal of the proposal is to "minimize the risk of algorithmic discrimination, in particular in relation to the design and the quality of data sets used for the development of AI systems").

²⁴⁸ For an expanded discussion, see Arbel & Becher, *supra* note 15.

²⁴⁹ Agnieszka M. Zbrzezny & Andrzej E. Grzybowski, *Deceptive Tricks in Artificial Intelligence: Adversarial Attacks in Ophthalmology*, 12 J. CLIN. MED., no. 9:3266, 2023, at 2 ("Suppose we consider even minor perturbations to the image, such as the change in colour of just one pixel. Then, such models are uncertain for small perturbations.").

²⁵⁰ For a formal exploration, see Jindong Wang et al., *On the Robustness of ChatGPT: An Adversarial and Out-of-Distribution Perspective 8* (Int'l Conf. on Learning Representations 2023 Workshop on Trustworthy and Reliable Large-Scale Mach. Learning Models, 2023), <https://openreview.net/pdf?id=uw6H5SkoM29> [<https://perma.cc/57LC-RUET>].

system and alter its interpretation.²⁵¹ There is no known general solution to such issues. But if judges and parties become aware of the possibility of such subtle manipulations, they might develop defenses, like using sanitized versions of the contract in their analyses.²⁵²

Fifth, models are sensitive to time. As your neighborhood originalist will tell you, the meaning of words is embedded in the time they were used. If we want to interpret the meaning of a contract signed in 1924, we should account for the linguistic conventions of the time. Models are trained on data indiscriminately: It is unlikely that they will be able to interpret a term as it was read in a specific period in time. The problem is compounded since the training data may include information that was not available for the contracting parties at the time of contracting. This may well include the decision of a trial court when the appellate court seeks to interpret the contract. We can think about this as pollution of the database: For example, perhaps Hurricane Katrina associated “levee” with “flood” more closely than it was at the time the relevant insurance contracts were signed.²⁵³ Or perhaps the *Stewart* example was confounded by the subsequent decades of linguistic evidence of payment defaults.

This problem is longstanding. Judges’ innate sense of language is also grounded in the linguistic conventions in which they are personally embedded. Dictionaries and corpus linguistics have an advantage here because one could seek a dictionary or a corpus from the relevant time period. But even this advantage is limited, because dictionaries are updated in intervals of decades,²⁵⁴ and corpora cover considerably fewer

²⁵¹ From the model’s perspective, “please” and “please” are not the same word. For an accessible exploration, see Computerphile, *Glitch Tokens - Computerphile*, YouTube (Mar. 7, 2023), <https://www.youtube.com/watch?v=WQ2X3oZEJOA> [<https://perma.cc/35Q5-JP6G>]. Various other examples are esoteric: Certain models act unexpectedly when presented with specific nonsensical words like “SolidGoldMagikarp.” See *Forbidden Tokens Prompting Results*, GOOGLE SHEETS, https://docs.google.com/spreadsheets/d/1PAZNCks11qoUpiojTJpj0odCYQL2_HGQgam8HSwAopQ/edit#gid=0 [<https://perma.cc/6FHG-LS6H>]. But in high stakes settings, such vulnerabilities can be exploited.

²⁵² Courts could require, for example, that texts will be presented in plain text format. This would limit some forms of attacks—especially those that are embedded in the graphical layer of the document. But the bitter lesson from cybersecurity is that security is a process, not a product. For illustration, see Riley Goodside (@goodside), X (Oct. 13, 2023, 9:15 PM), <https://twitter.com/goodside/status/1713000581587976372> [<https://perma.cc/9FJL-B4A9>].

²⁵³ A more far-fetched problem is parties trying to inject meaning into the record, just as they would in a normal interpretation dispute by way of after-action lawyer letters and the like. But because parties expect performance, not breach, and the relevant corpora for LLMs is so vast, jurists should worry less about this problem than the internal-to-the-text adversarial attacks we describe above.

²⁵⁴ See *History of the OED*, OXFORD ENGLISH DICTIONARY, <https://www.oed.com/information/about-the-oed/history-of-the-oed/?tl=true> [<https://perma.cc/JSV4-U64Y>]; *Merriam-Webster’s Ongoing Commitment*, MERRIAM-WEBSTER, <https://www.merriam-webster.com/about-us/ongoing-commitment> [<https://perma.cc/SX6H-YXWN>].

texts when they are sliced to relevant time periods.²⁵⁵ Thus, courts will have to consider whether the use of language has shifted over time, and perhaps restrain the use of generative interpretation in cases where its training data suffers from linguistic drift. Another way to put this is that generative interpretation is likely to be least useful for old contracts, where worries about subsequent judicial opinions interpreting like terms are most severe, unless and until specialized models with time delineated training data come online.

Sixth, generative interpretation will need a language of its own. Although scholars often hype objective, scientific methods of proof and judgment, this way of explaining and justifying the exercise of power is unconvincing, and perhaps repulsive, to the population at large.²⁵⁶ (Which is one reason we've tried to tamp down the statistics and claims to singular answers in this paper.) Juries, after all, aren't presented with simple probabilistic proofs, and judges don't typically justify their decisions by saying they have a 51 percent chance of being right.²⁵⁷ Thus, a real problem for the method—which it shares with corpus linguistics and the survey methodologies discussed in Part I—is how to explain itself to lay audiences in ways that reinforce, rather than diminish, judicial legitimacy.²⁵⁸ It's sociologically normal to say that the word *chicken* takes meaning from the dictionary and trade usage.²⁵⁹ This sociological framework does not yet exist for black box language models.²⁶⁰ Part of the credibility cachet is being built organically, as the public comes to experiment with LLMs and find utility in their responses; another part will be built over time, as generative interpretation tools are built and

²⁵⁵ See Mouritsen, *supra* note 20, at 1378 (“One of the challenges for examining usage in context in a corpus is that the greater the specificity of the search, the fewer examples appear in the corpus.”).

²⁵⁶ See David A. Hoffman & Michael P. O’Shea, *Can Law and Economics Be Both Practical and Principled?*, 53 ALA. L. REV. 335, 339 (2002) (“Most intriguingly, the studies suggest that in certain cases people prefer that legal decisions not be made on an economic basis.”).

²⁵⁷ As Nesson famously argued, the fact-finding system (and juries) exists to achieve legitimacy, not just accuracy. Charles Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357, 1358–59 (1985).

²⁵⁸ Cf. Benjamin Minhao Chen, Alexander Stremitzer & Kevin Tobia, *Having Your Day in Robot Court*, 36 HARV. J. L. & TECH. 127 (2022) (presenting experimental evidence that subjects are not biased against algorithmic decisionmakers).

²⁵⁹ See *Frigalment Importing Co. v. B.N.S. Int’l Sales Corp.*, 190 F. Supp. 116, 121 (S.D.N.Y. 1960) (adopting the broader meaning of the word after contextual inquiry).

²⁶⁰ See Hasala Ariyaratne, *The Impact of ChatGPT on Cybercrime and Why Existing Criminal Laws Are Adequate*, 60 AM. CRIM. L. REV. ONLINE, 2023, at 7 (“Since ChatGPT uses complex deep learning algorithms, it is often a black box with no clear reason why it provided a certain output.”); David S. Rubenstein, *Acquiring Ethical AI*, 73 FLA. L. REV. 747, 766 (2021) (“[D]eep learning neural networks drive some of the most powerful, sophisticated, and functional AI systems, but their complexity renders them inscrutable to humans.”); Nelson et al., *supra* note 230 (“AI is largely a ‘black box’—you cannot see inside the box to see how it works.”).

validated. Ultimately, courts will have to find ways to wrap the results from automated interpretation in packages that help laypeople to see law as engaging in a values-driven, communal, constrained exercise, and not merely the highest probability next-token predictions.²⁶¹

The solution likely lies in a specific type of transparency. Just as much as judges are sociologically committed to certain types of dictionaries, so will it be the case that certain models will emerge as robust and trustworthy. The current practice of interpretation is largely indefensible on this score; because we have no window into the court's processes, we cannot see the dictionaries it did not select or the words it chose not to focus on. But we can know what model a court picks, and from that selection, what probabilities it assessed. We cannot know exactly how the model produced those outcomes, as this knowledge lies in its vast, inscrutable matrices. But so long as a judge discloses not only the version of the model that she employed but also the particular prompts that she used, generative interpretation is more replicable than any other method on offer.²⁶² (We have tried to show how that would work in the notes of this article.) Indeed, courts might go further: They can *capsule* the results of their inquiries and incorporate them as permanent links to their opinions.

In summary, generative interpretation promises an accessible, relatively predictable, tool that will help lawyers and judges interpret contracts. If it's to achieve that promise, courts will need to be careful to use this tool while being mindful of its uses and limitations. To guide what would inevitably be a process of exploration, we offered a series of best practices based on the technical foundations and legal constraints that define the limits of this tool. As a default, judges should disclose the models and prompts they use and try to validate their analyses on different models and with multiple inputs. Ideally, they'd capsule their findings online. They'll want to be careful about parties' manipulative behavior, and they'll want to consider how (and whether) to excavate private, non-majority meanings. By doing so—and by saying what they are doing clearly and with appropriate recognition of LLMs'

²⁶¹ Related to this rhetorical concern is one about attribution and basic fairness that citizens may have about use of LLMs. *See, e.g.*, Sheera Frenkel & Stuart A. Thompson, 'Not for Machines to Harvest': Data Revolts Break Out Against A.I., N.Y. TIMES (July 15, 2023), <https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html> [<https://perma.cc/9TQQ-MSSN>]; Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 748 (2021) ("In this Article, we argue that ML systems should generally be able to use databases for training, whether or not the contents of that database are copyrighted."); *see also* Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley & Percy Liang, *Foundation Models and Fair Use* (Stan. L. & Econ. Olin Working Paper No. 584, 2023).

²⁶² The model disclosure should include the model's hyperparameters, much like judges share the version of the dictionary they consulted.

foibles—courts can fairly experiment with this new technology and achieve a better grasp on the contract’s meaning, without abusing the tool or subjecting themselves to reversals.

B. *Beyond the Textualist/Contextualist Divide*

As we described in Part I, the modern debate about interpretation takes as a given that prediction is the goal. But in deciding how to best accomplish prediction, scholars and courts disagree about an empirical meta-question: How would most parties prefer that courts interpret their deals?²⁶³ Many have argued that sophisticated parties prefer textualism.²⁶⁴ Others assert that contextualism is preferred, especially within longer-term relational contexts.²⁶⁵ Some argue such preferences are, well, contextual.²⁶⁶ Litigated cases appear to be all over the map.²⁶⁷ The views of poorer parties are more rarely studied. True, contextualism promises to protect parties from bait-and-switch maneuvers and opportunistic drafting. But who can afford it?

Generative interpretation challenges the utility of this old binary. Starting with textualism, its proponents have said that it builds a common commercial vocabulary and motivates clear contract drafting.²⁶⁸ But if applied correctly, generative interpretation (as a form of textualism) can predict parties’ intent well even without invocation of specialized language or expensive drafting. And if courts follow our proposed best practices, this method is also predictable *ex ante*. When parties can anticipate in advance the choice of model—and we argue that they should be able to contract for it explicitly—then they can clarify disputes well ahead of litigation. Even if the judge consults a broader evidentiary base than the contract itself, models can incorporate it and produce consistent outputs.

²⁶³ See Bayern, *supra* note 38, at 1101.

²⁶⁴ See, e.g., Schwartz & Scott, *Redux*, *supra* note 37, at 941 (“[P]arties prefer textualist interpretive defaults.”).

²⁶⁵ See Lisa Bernstein, *Merchant Law in a Merchant Court: Rethinking the Code’s Search for Immanent Business Norms*, 144 U. PA. L. REV. 1765, 1769–70 (1996) (discussing the fact that business arbitrators avoid business norms); Benoliel, *supra* note 57, at 493 (concluding that sophisticated parties prefer textualism). For a survey of the scholarly literature, see Silverstein, *supra* note 99, at 278–81; see also U.C.C. § 2-202(a) (AM. L. INST. & UNIF. L. COMM’N 2023) (stating that the meaning of contractual terms may be supplemented by usage of trade in some cases).

²⁶⁶ See Adam B. Badawi, *Interpretive Preferences and the Limits of the New Formalism*, 6 BERKELEY BUS. L.J. 1, 54 (2009).

²⁶⁷ See Silverstein, *supra* note 99, at 259 (noting courts’ mixed approaches in litigated cases).

²⁶⁸ Gilson et al., *supra* note 44, at 40–41.

By contrast, contextualism promises accuracy by integrating all relevant evidence. Its champions think it protects the weak from the powerful and reflects the real premises of relational contracting relationships.²⁶⁹ But as a judicial practice, it encourages gamesmanship,²⁷⁰ exposes decisionmakers to bias-inducing testimonies, increases uncertainty,²⁷¹ and more than anything, is simply very expensive. Generative interpretation can also serve as a form of contextualism. It is cheaper to incorporate context into the process when the model can feed on dozens of pages of evidence. Models are not prejudiced by parcels of evidence like human decisionmakers. And armed with LLMs, judges can assess the incremental probative value of proposed elements of evidence at the summary judgment stage. As we demonstrated with respect to *Stewart*, the judge can weigh in advance whether litigation over, say, the records of a phone conversation would be materially important to the outcome. This kind of prioritization is generally the approach of the Uniform Commercial Code: The models can turn an interpretation ladder into something more objective.²⁷²

All of this suggests a disruption of the traditional impasse. Generative interpretation allows both predictability and restraint, while also offering better linguistic accuracy. And it corrals litigation costs.²⁷³ Or to put it differently, the choice between four corners or no corners at all is a product of its time and of a specific adjudicatory technology. As this technology improves, judges can relax old safeguards towards a more inclusive approach.

To be sure, generative interpretation would be a simple flip in the default: Parties could indicate that their meaning was not to be determined by large language models, just as they can now commit to avoiding certain dictionaries or choosing others.²⁷⁴ Just as using a

²⁶⁹ See *supra* notes 87–96 and accompanying text (discussing contextualism).

²⁷⁰ See Gilson et al., *supra* note 44, at 41 (noting that “[u]nder a contextualist theory, a party for whom a deal has turned out badly has an incentive” to look to the other party’s negotiations, past practices, and trade customs to find enough evidence to force a settlement).

²⁷¹ See Schwartz & Scott, *Contract Theory and the Limits of Contract Law*, *supra* note 38, at 587; Schwartz & Scott, *Redux*, *supra* note 38, at 944–47 (arguing that certain parties prefer textualist defaults in part because of the risk of error).

²⁷² See, e.g., U.C.C. § 1-303(e) (2024) (establishing a hierarchy of evidence in contract interpretation, where express terms prevail over course of performance, which prevails over course of dealing, which prevails over trade usage).

²⁷³ Schwartz & Scott, *Redux*, *supra* note 37, at 946 (suggesting that controlling litigation costs is one reason that sophisticated parties prefer to avoid extrinsic evidence).

²⁷⁴ See 5 MARGARET N. KNIFFIN, CORBIN ON CONTRACTS: INTERPRETATION OF § 24.9 (Joseph M. Perillo ed., rev. ed. 1998) (arguing that courts should enforce private meanings, “however we may marvel at the caprice”); see, e.g., *Smith v. Wilson*, 3 B. & Ad. 728, 728 (1832) (holding that “parol evidence was admissible to sh[ow] that . . . the word thousand, as applied to [the contract], denoted twelve hundred”).

dictionary to interpret a secret cipher is a foolish way to interpret a deal,²⁷⁵ following parties' expressed interpretative preferences is wise. Generally speaking, giving parties the ability to control how contracts are interpreted respects their autonomy and carries efficiency benefits.²⁷⁶ So too here: Generative interpretation expands the kinds of evidence that *most* parties would like courts to consider, but it won't be for everyone.

Even if all generative interpretation does is flip the default on extrinsic evidence surrounding contracting, it still has important distributive effects. Textualism's many virtues can be recast as its elitist faults. Poorer parties, or uncounseled ones, often misunderstand the relationship between contractual disclaimers of reliance and oral sales talk.²⁷⁷ Though the Restatement of Consumer Contracts suggests that courts should be more open to the idea that contracts that disclaim obligation in the face of contrary promises should not be enforced,²⁷⁸ it does little to help with interpretative disputes which are less obviously unjust. And yet there are many examples of parties' proffered meaning being excluded as violative of the parol evidence rule,²⁷⁹ or simply not considered because the meaning is purportedly plain.²⁸⁰ As the *Ellington* example above demonstrates, even on their own terms such decisions may be questioned.²⁸¹ But if the parties have not otherwise indicated, generative interpretation will provide more evidence to courts that extrinsic meaning ought to matter in discerning what the parties contemporaneously would have said they meant.

An exemplary case that generative interpretation could benefit is *Smith v. Citicorp*.²⁸² The Smiths needed to borrow money to repay an old loan and pay for some home improvements. They turned to Citicorp, which purported to create a revolving loan agreement, secured by their home. The key to the dispute was that the interest rate on this loan was 13.99% APR, a rate only permissible for revolving loans, not closed ones. The Smiths argued that closed is exactly what the loan agreement was. Miraculously, the Smiths had signed affidavits from *two* Citicorp employees, attesting that Citicorp never intended to make advances on this loan (which would have defined an open-ended, revolving, loan).

²⁷⁵ Kniffin, *supra* note 274, § 24.13 (noting that courts should and do enforce the parties' vernacular).

²⁷⁶ Schwartz & Scott, *Contract Theory and the Limits of Contract Law*, *supra* note 38, at 569.

²⁷⁷ Lawrence M. Solan, *The Written Contract as Safe Harbor for Dishonest Conduct*, 77 CHI.-KENT L. REV. 87, 92 (2001) (identifying ways in which integrated agreements promote injustice).

²⁷⁸ RESTATEMENT OF THE LAW, CONSUMER CONTRACTS § 7 (AM. L. INST., Tentative Draft, 2023).

²⁷⁹ See, e.g., *Gold Kist, Inc. v. Carr, Jr.*, 886 S.W.2d 425, 430 (Tex. App. 1994).

²⁸⁰ See, e.g., *Greenfield v. Philles Recs., Inc.*, 780 N.E.2d 562, 569–70 (N.Y. 2002).

²⁸¹ See *supra* notes 179–86 and accompanying text.

²⁸² *Smith v. Citicorp Pers.-to-Pers. Fin. Ctrs., Inc.*, 477 So.2d 308, 311 (Ala. 1985).

But the Supreme Court of Alabama ignored that highly probative and rare evidence because it laid outside the four corners of the contract.

We think this result gives too much weight to generalized worries about courts' competency to evaluate extrinsic evidence. It would be a trivial task to incorporate the affidavits into the generative analysis, and, as we've shown, they can be weighted according to the judge's priors. This would not resolve questions of credibility and relevance, but the flexibility of incorporating it at the margins might radically improve the accuracy of the court's analysis.

Because generative interpretation blurs the line between textualism and methods of interpretation that are more capacious in their evidentiary sources, and because it enables a new set of evaluative metrics and socio-legal advantages, we think that it ultimately won't be (just) Textualism 2.0. Rather, it will become a distinctive method of evaluating contractual meaning, marked by its own jargon, normative commitments, and practitioner community. That new methodology will take time to develop. As we said, in the early days, judges will dip in and out of the application, using it as one would a dictionary, or a refresher CLE on the canons of contract interpretation. Only when lawyers start to argue that the tool can provide better answers to interpretative questions will courts ask if that is true, and whether answers from ChatGPT should supplant those from Merriam's or Black's dictionaries.

CONCLUSION

In this Article, we introduced generative interpretation, a method of interpreting legal texts using large language models. Our work follows a rapidly evolving practice: Lawyers and judges are already experimenting with these models in law offices and chambers across the country, some covertly, others less so. We offered a deep dive into the way the technology works (and fails) and explored techniques of using it to better perform interpretative tasks. We demonstrated that the technique can be applied to famous contracts cases, often arriving at the same answers at lower cost and with greater certainty, and sometimes exposing ambiguities, dislodging sticky priors about meaning, and parceling out the marginal effect of new evidence on interpretation.

In our view, generative interpretation is a developing tool with unsettling implications for legal practice and contract theory. Because language models are attentive to context, and because they can voraciously digest long texts, they promise a future of better-calibrated (and more predictable) textualism. The models' complex encoding of language far outstrips that of any dictionary, and extensive training data give them a superior sensitivity to actual usage. All of that promises a

considerably more sensitive way to predict meaning, but it won't replace judges. Attempting to do so would ignore the model's real structural limitations, which include their opacity, hallucinatory nature, latent biases, and susceptibility to adversarial attacks by sophisticated parties. And using today's models for interpretation requires some expertise—creating multiple prompts or running code isn't the day-to-day work of state trial court judges.

Keeping these limitations in mind, we argued that generative interpretation nevertheless paves an important middle ground between too-cold textualism and too-hot contextualism. The traditional tradeoffs between textualism and contextualism take as a given that our textualist inquiry must depend on dictionaries and that extrinsic evidence is necessarily costly and prone to manipulation. Because generative interpretation is easy to deploy, cheap, and accurate, and because it is not prone to those specific biases, it suggests a workable third way. We argue that, given this technology, parties would prefer courts to ascertain meaning using *some* extrinsic evidence. As such, generative interpretation will become a majoritarian default.

With time, these discussions will spill over to even broader debates about statutory and constitutional interpretation, originalism and public meaning, and the relative competencies of courts and agencies to reach unbiased, predictable outcomes. We deferred direct discussion of these issues, not least because we're not competent to resolve them. This work, nonetheless, fits into this broader interpretative project of assigning meaning to legal instruments.

We close by offering a different sort of prediction. If, in fact, these models can ascertain party intent to a close-enough approximation, it seems obvious that courts will (and should) use them to make interpretation better. But if that's really true, we wonder why parties would continue to commit to contracts at all? Formal contracting is expensive. Why not, instead, simply write out jointly held goals at the beginning of the relationship and let models spit out codes of conduct and legal responsibility as problems arise down the line?²⁸³ Or to put it differently, right now, generative AI looks like a promising judicial adjunct. But the future of this technology is more disruptive by far: Formal contracts themselves may be made obsolete. Or, at the very least, jurists should consider the marginal value of contracting if the terms themselves are fairly determinable from the parties' goals.

²⁸³ Cf. Cathy Hwang, *Deal Momentum*, 65 UCLA L. REV. 376, 380 (2018) (describing the use of term sheets as deal motivators).