Generative Interpretation

Yonathan Arbel & David A. Hoffman[*]

Forthcoming: 99 NYU L. REV. ___ (2024)

[***DRAFT*** July 30, 2023]

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 LLMs 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.

---

[*] Associate Professor, University of Alabama Law & William A. Schnader Professor, University of Pennsylvania Carey School of Law. We thank Michael Hurley, Elizabeth Meeker and JD Uglum for helpful research assistance.

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 fact-finder 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, covering rising waters no matter their cause, and therefore no further 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]

---

[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 ... Flood, 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 191, 197, 199, 200-01, 203-04 (5th Cir. 2007); Brief for Appellee-Cross Appellant Humphreys at 16-18, In re Katrina Canal Breaches Litig., 495 F.3d 191 (5th Cir. 2007) (No. 07-30119), 2007 WL 4266576; Brief for Plaintiff-Appellee Xavier Univ. of La. at 17-44, In re Katrina Canal Breaches Litig., 495 F.3d 191 (5th Cir. 2007) (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 (5th Cir. 2007) (No. 07-30119), 2007 WL 4266578. On the scope, source, and allocation of negligence see ANDY HOROWITZ, KATRINA: A HISTORY, 1915-2015, 1-12, 128-33 (2020); *see also* Campbell Robertson & John Schwartz, *Decade After Katrina, Pointing Finger More Firmly at Army Corps*, THE NEW YORK TIMES, May 23, 2015, https://www.nytimes.com/2015/05/24/us/decade-after-katrina-pointing-finger-more-firmly-at-army-corps.html.

[3] In re Katrina Canal Breaches Litig., 495 F.3d at 208. Brief of Appellee State Farm Fire & Casualty Co. at 14-26, In re Katrina Canal Breaches Litig., 495 F.3d 191 (5th Cir. 2007) (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 (5th Cir. 2007) (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).

After years of litigation, the Fifth Circuit—in the best-known and most consequential contracts case of the last generation[5]—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 followed from a constraining method, the Fifth Circuit (says

---

[5] The opinion has been cited nearly 7,000 times over 15 years, discussed in almost 2,000 secondary sources, and is taught to 1Ls. *See, e.g.*, IAN S. AYRES AND GREGORY M. 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, 747-763 (E.D.La. 2006).

[9] 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, 303-306 (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] 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 309-310 (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, 2001) (finding flood ambiguous), reversed by Sher v. Lafayette Ins. Co., 07-2441 (La. 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 (S.C. 2010) (finding that rainwater deliberately channeled on insured's land was not flood water).

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 is simple: applying large language models (LLMs) to contractual texts and extrinsic evidence to predict what the

---

[12] 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*, CHICAGO TRIBUNE, Aug. 3, 2007, at C1 (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, and 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 (2006) (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, *The Hurricane Katrina Levee Breach Litigation: Getting the First Geoengineering Liability Case Right*, 160 U. PA. L. REV. 267 (2012) (arguing in support of the Fifth Circuit ruling).

[13] Lawrence A. Cunningham, *Contract Interpretation 2.0: Not Winner-Take-All but Best-Tool-For-The-Job*, 85 GEO. WASH. U. L. REV. 1625, 1628-31 (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,* at 1633-1643 (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) (early paper showing 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 BERK. TECH. L.J. 71 (2022) (arguing that ChatGPT might be useful in helping consumers to understand their contracts and providing examples).

parties would have 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 we argue that even today's freshly-minted LLMs can be of service.

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. 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).[18] 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 tend to occur, or collocate, with flood in newspapers, books, and the like.[19]

But we instead turned to a convenient, free, open-source LLM tool resting on a database of trillions of words and asked it to transform words into complex vectors in a process called *embedding*.[20] As a first cut, this process can be thought of as trying to quantify how much a word belongs 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

---

[16] *Cf.* Jonathan H. Choi, *Measuring Clarity in Legal Texts,* 91 U. CHI. L. REV. (forthcoming, 2024). 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* at text accompanying notes 181 to 184 (discussing Mata v. Avianca, Inc., __ F.Supp.3d. __, 2023 WL 4114965 (June 22, 2023).); *see also* Ex Parte Allen Michael Lee, __. S.W.3d. __ , 2023 WL 4624777, at *1 n.2 (Ct. App. Tex. July 19, 2023) (explaining the court's suspicion that counsel had filed briefs using ChatGPT and had made up cases and citations).

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

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

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

developed, we queried several models about the relation of word flood in its contractual context (attributing flood to water damage) to other potential environments.[21]

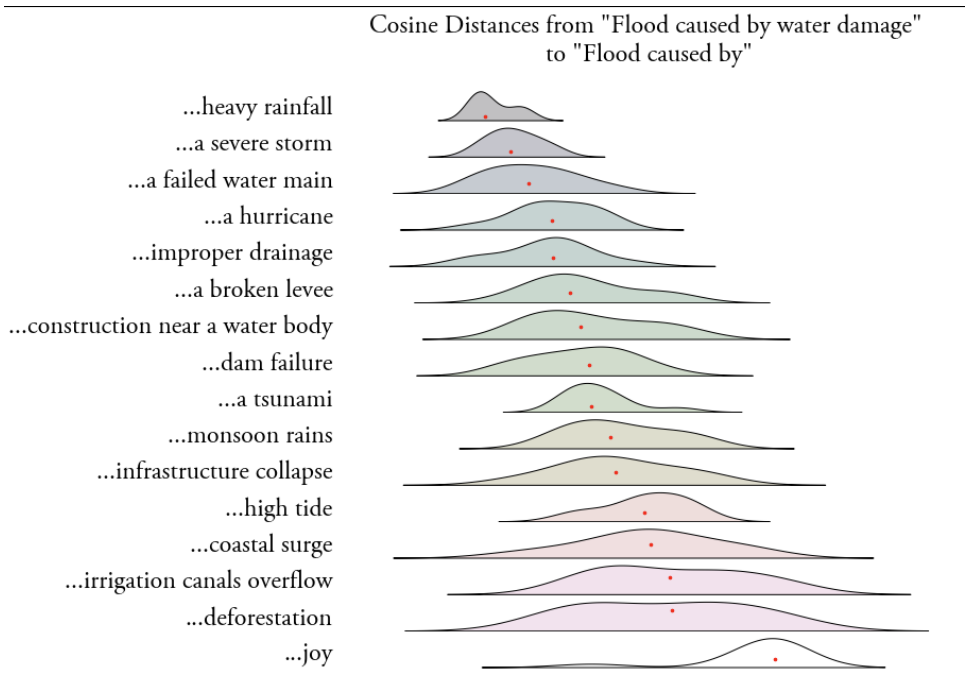

Figure 1: Analysis of the cosine distance—a measure of distance for the numerical representation of terms (embeddings) by language models—between an attention-weighted clause on the cause of flood and other terms.

To read Figure 1, focus on the location of the red markers. The farther they are from the origin, the more distant the term is from the *flood* clause. In our view, the Figure

[21] All of the code necessary to replicate these results, and the remaining ones in the paper, can be found at: https://github.com/yonathanarbel/generativeinterpretation/tree/main. 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, 24-26 (using method and reports its usage and limitations.) For a non-legal example, *see e.g.*, Nitika Mathur, Timothy Baldwin & Trevor Cohn, *Putting Evaluation in Context: Contextual Embeddings Impprove Machine Translation Evaluation*, PROCEEDINGS OF THE 57TH ANNUAL MEETING OF THE ASSOCIATION 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 TRANS. ASS'N FOR COMPUTATIONAL LINGUISTICS 107 (2018). We then calculated the cosine distance, normalized it, and reported the results in the figure below.

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 rather seem more focused on its typicality. Our quality checks, *Flood caused by joy* or *deforestation*, are indeed farther out than flood caused by *rainfall* or a *storm*. And while it supports this decision of the court, it challenges another. Louisiana courts refused to exclude water main floods, even though linguistically they appear to be as much of a flooding event as any other.[22]

Now, the model doesn't provide (nor could it) a scientific way to judge whether words are sufficiently close to make the plain meaning of *flood* unambiguous. 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.

We introduce the methodologies of contract interpretation (in Part I) 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. 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.[23]

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 (the predicted meaning of a particular word) to the hard (whether there is 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

---

[22] Sher v. Lafayette Ins. Co., 2007-2441 (La. 4/8/08), 988 So. 2d 186, 195, on reh'g in part (July 7, 2008) ("inundation of property due to broken water mains . . . would not be excluded as a 'flood'"). *In re Katrina Canal Breaches Litig.*, 495 F.3d at 216 ("Unlike a canal, a water main is not a body of water or watercourse.").

[23] *See infra* at text accompanying notes 32 to 107.

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 Ctr. v. Connecticut Gen. Life Ins. Co.*[24] and *C & J Fertilizer, Inc. v. Allied Mut. Ins. Co.*,[25] as well as some that should be, like *Famiglio v. Famiglio*,[26] *Haines v. City of New York*[27] and *Stewart v. Newbury*.[28] 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 helping us to 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" which allows them to attend to context. And by becoming context sensitive, these models can parse out the effects of contract text from the marginal value of relevant extrinsic evidence

But current practices about LLMs and their future uses are contingent: lawyers tend to use tools before they are theoretically sharp.[29] 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 and its use by lawyers and judges grows explosively. We make two claims.

---

[24] 847 F.2d 564 (9th Cir. 1988). *See, e.g.*, RANDY E. BARNETT & NATHAN B. OMAN, CONTRACTS: CASES AND DOCTRINE 483 (7th ed. 2021); E. ALLEN FARNSWORTH, CAROL SANGER, NEIL B. COHEN, RICHARD R.W. BROOKS AND LARRY T. GARVIN, CASES AND MATERIALS ON CONTRACTS __ (10th ed. 2023).

[25] 227 N.W.2d 169 (Iowa 1975). *See* 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 252 (Jean Braucher, John Kidwell & William Whitford eds., Hart Publishing, 2013) (describing C&J and noting that it is often assigned in casebooks, including Stewart Macaulay's and Charles Knapp's).

[26] 279 So.3d 736 (Fla. Dist. Ct. App. 2019).

[27] 41 N.Y.2d 769 (1977). ROBERT S. SUMMERS, ROBERT A. HILLMAN AND DAVID A. HOFFMAN, CONTRACT AND RELATED OBLIGATION: THEORY, DOCTRINE, AND PRACTICE 834 (8th ed. 2021).

[28] 220 N.Y. 379 (1917). SUMMERS ET AL., supra, at 948.

[29] Consider originalism.

First, the method fills a glaring need for a simple, transparent and convenient way to commit to an interpretative method which helps predict the parties' intent. If courts follow the set of best practices we describe, 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,[30] use of artificial intelligence (AI) should prompt a reexamination of the utility of those categories entirely. 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.[31]

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 decision-making. As we show, generative 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?"[32] That is, to "ascertain the parties' intention at the time [the parties] made their contract."[33] As Alan Schwartz and Bob Scott noted in their canonical article, *Contract Theory and the Limits of Contract Law*, this question in theory has a "correct

---

[30] 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, 600-601 (2021); 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).

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

[32] Bruce v. Blalock, 241 S.C. 155, 161, 127 S.E.2d 439, 442 (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.").

[33] STEVEN J. BURTON, ELEMENTS OF CONTRACT INTERPRETATION § 1.1, at 1.

answer."[34] 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),[35] market data,[36] and some hunches about fairness and efficiency under the circumstances.[37]

---

[34] 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.'"); Alan Schwartz & Robert E. Scott, *Contract Interpretation Redux*, 119 YALE L.J. 926 (2010). For criticisms, see Adam B. Badawi, *Interpretive Preferences and the Limits of the New Formalism*, 6 BERKELEY BUS. L.J. 1 (2009); Shawn J. Bayern, *Rational Ignorance, Rational Closed-Mindedness, and Modern Economic Formalism in Contract Law*, 97 CAL. L. REV. 943 (2009); Robin Bradley Kar and Margaret Jane Radin, *Pseudo-Contract and Shared Meaning Analysis*, 132 HARV. L. REV. 1135, 1182-92 (2020) (arguing that sophisticated parties would not and do not prefer acontextual readings).

[35] 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).

[36] JOHN BOURDEAU, PAUL M. COLTOFF, JILL GUSTAFSON, GLENDA K. HARNAD, JANICE HOLBEN, SONJA LARSEN, LUCAS MARTIN, ANNE E. MELLEY, KARL OAKES, KAREN L. SCHULTZ & ERIC C. SURETTE, AMERICAN JURISPRUDENCE § 219 (2nd 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.").

[37] Omri Ben-Shahar, David A. Hoffman and Cathy Hwang, *Nonparty Interests in Contract Law*, 171 U. PA. L. REV. 1095, 1017-1129 (2023) (describing courts use of public interests in interpreting contracts).

[38] Schwartz and Scott, *supra* note 34, 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.

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 is rooted in the public use of words—democratized textualism.

But corpus linguistics is inattentive to context.[99] 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

[96] *See generally* Mouritsen, *supra* note 19, at 1360-1407 (making case).
[97] Christopher C. French, *Insurance Policies: The Grandparents of Contractual Black Holes*, 67 DUKE L.J. ONLINE 40 (2017) (discussing the difficulty of interpreting insurance contracts for evidence of real meaning).
[98] Mouritsen, *supra* note 19, at 1371-74 (CL approach to snorkeling).
[99] *See* Choi, *supra* note 16, at 8, 16-17 (arguing that the context "undermines the core claim of corpus linguistics").

commonly engaged in interpretative battles over short phrases—only one contracts opinion to date has applied the method to date.[100]

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.[101] As they point out, this survey evidence is a second best to the predictive exercise 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.[102]

The authors defend this interesting proposal against various charges.[103] Their core survey case is consumer contracts designed for mass audiences.[104] There, the survey audience and the original adherents are the same (although separated by time), and we should have fewer worries about the parties intending idiosyncratic meanings.[105] 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.[106]

---

[100] *See* Fulkerson v. Unum Life Insurance Co. of America, 36 F.4th 678 (6th Cir. 2022); *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. Safelinte Group, Inc. 930 F.3d 429, 439 (6th Cir. 2019) (arguing for use of CL in statutory analysis); Caesars Entm't Corp. v. Int'l Union of Operating Eng'rs Local 68 Pension Fund, 932 F.3d 91, 95 n.1 (3d Cir. 2019) (using corpus linguistics to interpret "previously").

[101] Ben-Shahar & Strahilevitz, *supra* note 18; Ian Ayres & Alan Schwartz, *The No-Reading Problem in Consumer Contract Law*, 66 STAN. L. REV. 545 (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, 1419–20 (2014) (advocating the use of surveys to identify the majoritarian preferences for the design of granular default rules).

[102] Ben-Shahar & Strahilevitz, *supra* note 18, at 1802.

[103] *Id*. at 1802-1813 (making the case).

[104] *Id*. at 1758 (noting focus on consumer contracts).

[105] *Id*. at 1776-1777 (articulating the basic consumer contracts case).

[106] *Cf*. Roberts v. Farmers Ins. Co., 201 F.3d 448 (10th Cir. 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.[107] 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 $5 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 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.

---

[107] Veniamin Veselovsky, Manoel Horta Ribeiero, Robert West, *Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks,* ARXIV:2306.07899 (2023) (33-46% of mTurk survey workers use LLMs to complete tasks).

[108] *See generally* 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 (1998).

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.[157] 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." Finally, the foundry responded in writing that, following the phone conversation, they accepted the bid. As far as we know, that amounts to the entirely of the contracting case file.[158]

Once the contractor finished the first part of the project, he submitted a bill. The foundry refused to pay. The contractor insisted that it was customary to pay 85% 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.[159] 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 override this default?

---

[157] 220 N.Y. 379 (1917).
[158] *Id.* at 380-84.
[159] *See* 22 N.Y. JUR. 2D CONTRACTS § 352, Hillman, *supra* note 152, at 313 ("courts in construction cases find a duty to pay only after substantial performance").

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.

The court remanded because of faulty jury instructions, so the interpretative question was left undecided. We, however, are not so constricted. We asked today's leading LLM models, GPT-4 and Claude-4, 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.[160] 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 changes, and finally, we added evidence of the custom in the industry. Table 1 summarizes the results:[161]

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 1 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 to the decisionmaker is useful in part because it permits a real focus on the *quality*, not just the *weight*, of

[160] 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.

[161] CLAUDE 2.0 POE CONSERVATION, https://poe.com/s/wLkeCDrPdFpKye3uApSa (last visited July 30, 2023). Again, you should be skeptical of model's expressed confidence; the direction of change with every new piece of evidence, not its quantification, is reliable.

evidence. Did this conversation really transpire in the way alleged by the contractor? Was sufficient evidence proffered to substantiate the industry custom alleged by the contractor? And, later, whether interpreting the contract in this way is socially beneficial?[162] The model can give structure to the evaluation of extrinsic evidence. And within the limits of its prompts, its conclusions are coherent, cheap, and seemingly plausible.

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 we write this article, only a few months after the tools went mainstream. Looking at practical guidance offered to lawyers in the summer of 2023, we see lawyers are encouraged to use LLMs to perform legal research, draft deposition questions and contracts, and predict settlement values.[163] And there are hints that judges are already using ChatGPT to answer other kinds of interpretative questions, just as they would use Google.[164] In one recent survey, one-quarter of judges confessed to using the tool, though many expressed concern about its reliability.[165]

These models are useful because they offer tools—fast, cheap, sometimes incorrect—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 ask the model what's the likely assumption a contract makes when it leaves a gap; or check if it 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.

---

[162] Williston suggests that it is more difficult to order specific performance against a recalcitrant contractor than to order (effective) specific performance against a non-paying principal. 15 WILLISTON ON CONTRACTS § 44:48 (4th ed.).

[163] Catherine Casey, Reveal Brainspace, Ronald J. Hedges, Ronald J. Hedges LLC, Marissa J. Moran, N.Y.C. Coll. of Tech., Stephanie Wilson, Reed Smith LLP, *Generative Artificial Intelligence in Practice: What It Is and How Lawyers Can Use It* (June 28, 2023) (on file with authors).

[164] 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.

[165] 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/.

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,[166] they were chastised for its use by higher courts,[167] and then it was eventually folded into the normal set of legal research tools.[168] But at least in the short run, judges won't have the tool 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."[169] The 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 A is to delimit some principles and limitations for its usage by lawyers and judges. With the proper usage of the tool in mind, in Section 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 Part 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

---

[166] Lee F. Peoples, *The Citation of Wikipedia in Judicial Opinions*, 12 YALE J. L. & TECH. 1, 28 (2010) ("Citations to Wikipedia entries in judicial opinions have been steadily increasing since the first citation appeared in 2004.").

[167] 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, *Downside of Citing to Wikipedia*, N.Y. L.J., Jan. 18, 2007, at 2.

[168] 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) ("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. . . .").

[169] Schwartz & Scott, *supra* note 34, at 568 n.50; *see generally* Edwin W. Patterson, *The Interpretation and Construction of Contracts*, 64 COLUM. L. REV. 833, 833–35 (1964); Klass, *supra* note 40.

buckled down and done their jobs like careful jurists ought to. 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 of real use in contract interpretation. The biggest single problem with all currently available approaches to contract interpretation isn't that they are incapable of getting correct results some of the time. It's that they are inaccessible to ordinary parties.[170] The result is that non-wealthy individuals who suffer breach have to lump it,[171] tilt against corporations in internal dispute resolution systems,[172] or face financially ruinous fees and prevail in pyrrhic victories.[173] 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 interpretating deals, which form the core of contract litigation, materially contribute to this problem.[174]

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

---

[170] 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").

[171] Geiger, *supra*, at 38 ("[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.").

[172] *See* Rory Van Loo, *The Corporation as Courthouse*, 33 YALE J. REG. 547 (2016).

[173] 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); 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).

[174] Ben-Shahar & Strahilevitz, *supra* note 18, 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).

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.[175] 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.

Generative interpretation is a tool which responds to this access-to-justice concern. If courts commit to the method, the costs of achieving accuracy in contract interpretation disputes will fall.[176] 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.[177] Over time, one possibility is that 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.[178] And better calibrated results *ex post* means that parties can spend less time (and money) contracting *ex*

[175] Samuel Issacharoff & Florencia Marotta-Wurgler, *The Hollowed Out Common Law*, 67 UCLA L. REV. 600 (2020) (documenting the "dominance of the federal forum").
[176] *Cf.* Schwartz & Scott, *Redux*, *supra* note 34, at 930 (noting the primacy of cost in evaluating the correct interpretative rules).
[177] *Cf.* Eric A. Posner, *A Theory of Contract Law Under Conditions of Radical Judicial Error*, 94 N.W. U. L. REV. 749, 766-69 (2000) (noting that deterministic legal rules discourage opportunistic breach).
[178] *Cf.* Schwartz & Scott, *supra* note 34, 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.

ante.[179] A promise of generative interpretation—which it may yet fulfill—is that it will open a form of textualism up to the 99%.[180]

The pages of law reviews are littered with proposed technological solutions to supposed problems of excessive legal costs, which turn out to be 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,* right now or soon, for resource-deprived courts to adopt in ordinary cases. In evaluating that question of basic competency, it's meaningful that even today's unspecialized models can replicate the results of well-considered cases (as Section II explored) and prompt courts to consider their own priors.

But Section II offered a curated tour of generative interpretation's greatest hits. It didn't show you where things can go wrong. To make this tool perform as well as it can, users should be cognizant of these issues and use it according to evolving best practices. To begin, let's start with hallucinatory outputs.[181] 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.[182] A sanctions order and plenty of bad press followed.[183] In

---

[179] *See* Spencer Williams, *Predictive Contracting*, 2019 COLUM. BUS. L. REV. 621 (arguing that parties could use information about contract outcomes, harnessed through machine learning of large datasets, to change out they contract ex ante). But for an insightful discussion of how selection operates to make difficult machine predictions about litigation outcomes, *see* David Freeman Engstrom and Jonah Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversariliasm*, 169 U. PA. L. REV. 1001, 1065-1067 (2021) (discussing obstacles to prediction).

[180] Schwartz & Scott, *supra* note 34, at 941 (noting "the more time the court spends on a particular interpretive issue, the less time it can spend on other issues or other cases").

[181] Sharon D. Nelson, John W. Simek & Michael C. Maschke, *Beware of Ethical Perils When Using Generative Ai!*, 46-JUN, WYO. LAW., at 28, 30 ("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'. . . .").

[182] Benjamin Weiser, *Here's What Happens When Your Lawyer Uses ChatGPT*, THE NEW YORK TIMES (May 27, 2023),
https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

[183] *See* Mata v. Avianca, Inc., __ F. Supp. 3d. __, 2023 WL 4114965 (June 22, 2023).

response to the case, other judges have required lawyers to certify that they had not used any form of Artificial Intelligence in their filings.[184]

False outputs arise from the predictive nature of generative models.[185] Hallucinations are generated texts asserting facts that are not quite true.[186] 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.[187] One solution that is

---

[184] 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/ (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.").

[185] 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/ ("the model is fed a large body of text . . . and repeatedly 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."); waka55 (u/wakka55), REDDIT (Apr. 16, 2023, 2:48 PM), https://www.reddit.com/r/OpenAI/comments/12okltx/openais_whisper_api_sometimes_returns_what_looks/ (showing that this problem is not limited to textual generation).

[186] Beren Millidge, *LLM's confabulate not hallucinate*, BEREN'S BLOG (Mar. 19, 2023), https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/ (describing problem).

[187] *See e.g.*, 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 (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; *see also* Ayush Agrawal, Lester Mackey, & Adam Tauman Kalai, *Do Language Models Know When They're Hallucinating References?*, ARXIV: 2305.18248 (May 29, 2023), https://arxiv.org/abs/2305.18248; *see also* 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.

already used in some contexts is connecting the model to a database of facts, so that it can act more like the helpful librarian.[188] So while it is appropriate to pay attention to the hallucination problem, it's not obvious that it is a fundamental, persistent, and broad concern. That said, as a best practice, judges would do well to cross-verify 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.[189]

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 chat bot will attest, it is relatively easy to drive conversations toward desired outcomes. In litigation practice, we should expect that the parties themselves will submit competing prompts, just as they vie to control the framing of the legal questions 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.

A third consideration focuses on the models' strength: they are naturally inclined to make predictions that maximize probability—in other words, they are biased 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.[190] 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.[191] But more generally, because the linguistic conventions of underrepresented communities are submerged by majoritarian

---

[188] *See generally* James Briggs & Francisco Ingham, *Fixing Hallucination with Knowledge Bases*, PINECONE, https://archive.pinecone.io/learn/langchain-retrieval-augmentation/.

[189] *See generally* Robert J. Munro, J. A. Bolanos & Jon May, *LEXIS vs. WESTLAW: An Analysis of Automated Education*, 71 LAW LIBR. J. 471 (1978).

[190] Schwartz & Scott, *supra* note 34, at 583-584.

[191] For the foundational work distinguishing local from popular interpretative modes, *see* 2 SAMUEL WILLISTON, THE LAW OF CONTRACTS, § 604, 1162 (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, 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").

public meanings, they will find it more difficult to have their voices surfaced (and thus subsidized) in contract adjudication. Majoritarian interpretative approaches risk silencing entire communities.[192]

Surely, this is not a problem unique to generative interpretation: dictionaries, canons, and corpora are equally, if not more, vulnerable to the charge.[193] And unlike dictionary-and-canon-textualism, it is at least theoretically possible to query models about the linguistic conventions of distinct communities, enabling courts to come closer to understanding what the parties before them intended their language to mean.[194] This is an active area of research and regulatory scrutiny and should check factfinders.[195]

Fourth, models may become subject to parties' adversarial attacks.[196] By way of illustration, modern AI systems can reliably differentiate between pictures of panda bears and horses, and 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 see a horse or a yield sign.[197] The same manipulations can be used to "attack" LLM models.[198] Slight changes in the wording of a contract—e.g., subtle changes in the

---

[192] *See, e.g.,* Majorie Florestal, *Is a Burrito a Sandwich*, 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 (2016) (arguing for the use of cultural meaning in interpretation); *see also supra* note 142.

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

[194] For an illustration of this use case, *see* Arbel & Becher, *supra* note 15, at 99-104.

[195] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislation Acts*, at 4, COM (2021) 206 final (Apr. 21, 2021) (stating that a goal of the proposal is to "minimise 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. . . .").

[196] For an expanded discussion, *see* Arbel & Becher, *supra* note 15.

[197] Agnieszka M. Zbrzezny & Andrzej E. Grzybowski, *Deceptive Tricks in Artificial Intelligence: Adversarial Attacks in Ophthalmology*, 12(9) J. CLIN. MED. 3266 (2023) ("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.").

[198] For a formal exploration, *see* Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang, Xing Xie, *On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective*, ARXIV: 2302.12095 (Mar. 29, 2023), https://arxiv.org/pdf/2302.12095.pdf.

presentation of the words—might hack the model logic system and alter its interpretation.[199] 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.

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.[200]

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,[201] and corpora cover considerably fewer texts when they are sliced to relevant time periods.[202] Thus, courts will have to consider whether the

---

[199] 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=WO2X3oZEJOA. Various other examples are esoteric: certain models act unexpectedly when presented with specific nonsensical words like "SolidGoldMakigarp." *See* FORBIDDEN TOKENS PROMPTING RESULTS, https://docs.google.com/spreadsheets/d/1PAZNCks11qoUpiojTJpj0od-CYQL2_HGQgam8HSwAopQ/edit#gid=0 (last visited July 20, 2023). But in high stakes settings, such vulnerabilities can be exploited.

[200] 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.

[201] *See* HISTORY OF THE OED, https://www.oed.com/information/about-the-oed/history-of-the-oed/?tl=true (last visited July 20, 2023); *See* MERRIAM-WEBSTER ABOUT US ONGOING COMMITMENT, https://www.merriam-webster.com/about-us/ongoing-commitment (last visited July 20, 2023).

[202] Mouritsen, *supra* note 19, 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.").

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.

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 uncompelling, and perhaps repulsive, to the population at large.[203] (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% chance of being right.[204] Thus, a real problem for the method—which it shares with corpus linguistics and the survey methodologies discussed above—is how to explain itself to lay audiences in ways that reinforce, rather than diminish, judicial legitimacy.[205] It's sociologically normal to say that the word *chicken* takes meaning from the dictionary and trade usage.[206] This sociological framework does not yet exist for black box language models.[207] 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.[208]

---

[203] 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.").

[204] 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 (1985).

[205] *Cf.* Benjamin Minhao Chen, Alexander Stremitzer and Kevin Tobia, *Having Your Day in Robot Court*, 36 HARV. J. L. & TECH. 1 (2022) (presenting experimental evidence that subjects are not biased against algorithmic decisionmakers).

[206] *Cf.* Frigaliment Importing Co. v. B.N.S. Int'l Sales Corp., 190 F. Supp. 116 (S.D.N.Y. 1960) (adopting the broader meaning of the word after contextual inquiry).

[207] Hasala Ariyaratne, *The Impact of Chatgpt on Cybercrime and Why Existing Criminal Laws Are Adequate*, 60 AM. CRIM. L. REV. ONLINE 1, 7 (2023) ("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) ("deep learning neural networks drive some of the most powerful, sophisticated, and functional AI systems, but their complexity renders them inscrutable to humans."); Nelson, Simek & Maschke, *supra* note 181, at 30 ("AI is largely a 'black box'--you cannot see inside the box to see how it works.").

[208] 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 and Stuart A. Thompson, *'Not for Machines to Harvest': Data Revolts Break Out Against A.I.*, THE NEW YORK TIMES (July 15, 2023), https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html; Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 748 (2021) ("In this Article,

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 he 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 not only discloses 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.[209] (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 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' 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.

---

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*, ARXIV: 2303.15715 (Mar. 28, 2023), https://arxiv.org/pdf/2303.15715.pdf.

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