MAX PLANCK INSTITUTE
FOR RESEARCH ON
COLLECTIVE GOODS

Christoph Engel
Richard H. Mcadams

Discussion Paper
2024/5

ASKING GPT FOR THE ORDINARY MEANING OF STATUTORY TERMS

Asking GPT for the Ordinary Meaning of Statutory Terms

Christoph Engel[*] and Richard H. McAdams[**]

Abstract

We report on our test of the Large Language Model (LLM) ChatGPT (GPT) as a tool for generating evidence of the ordinary meaning of statutory terms. We explain why the most useful evidence for interpretation involves a distribution of replies rather than only what GPT regards as the single "best" reply. That motivates our decision to use Chat 3.5 Turbo instead of Chat 4 and to run each prompt we use 100 times. Asking GPT whether the statutory term "vehicle" includes a list of candidate objects (e.g., bus, bicycle, skateboard) allows us to test it against a benchmark, the results of a high-quality experimental survey (Tobia 2000) that asked over 2,800 English speakers the same questions. After learning what prompts fail and which one works best (a belief prompt combined with a Likert scale reply), we use the successful prompt to test the effects of "informing" GPT that the term appears in a particular rule (one of five possible) or that the legal rule using the term has a particular purpose (one of six possible). Finally, we explore GPT's sensitivity to meaning at a particular moment in the past (the 1950s) and its ability to distinguish extensional from intensional meaning. To our knowledge, these are the first tests of GPT as a tool for generating empirical data on the ordinary meaning of statutory terms. Legal actors have good reason to be cautious, but LLMs have the potential to radically facilitate and improve legal tasks, including the interpretation of statutes.

[*] Director, Max Planck Institute for Research on Collective Goods, Professor, University of Bonn, Professor Emeritus, Erasmus University Law School.

[**] Bernard D. Meltzer Professor, University of Chicago Law School. The authors thank Elliott Ash, Tom Ginsburg, Dan Klerman, Jonathan Masur, Kevin Tobia, and participants of the workshop at the Center for Law and Economics at ETH Zurich for excellent comments on an earlier draft.

Contents

Introduction

Chief Justice John Roberts devoted his 2023 year-end report to the subject of legal technology, culminating with a discussion of "the latest . . . frontier," Artificial Intelligence.[1] Roberts implicitly referred to large language models (LLMs) such as ChatGPT (GPT) when he said that AI has already shown itself capable of passing some parts of the bar exam and "provid[ing] answers to basic [legal] questions" for those who cannot afford a lawyer. Despite some well-publicized "hallucinations,"[2] where lawyers relied on non-existent cases made up by LLMs, the existing scholarship demonstrates the ability of GPT to identify and summarize cases, write the first draft of briefs and memos, and interpret contract terms.[3] We extend that work here by providing the first test of LLMs as a tool for interpretating statutes.

Given its convenience, we think it inevitable that lawyers and judges will use GPT for this purpose, and yet we agree with the Chief Justice that "any use of AI requires caution and humility." Accordingly, we make only the modest claim that, with the right prompting techniques, GPT very cheaply provides useful data for the empirical assessment of the ordinary meaning of statutory terms. GPT has certain advantages over existing empirical methods for assessing ordinary meaning, as it enables quicker, richer, and more differentiated investigations. At the least, one can use GPT to "triangulate" meaning by using it in combination with other methods.[4] And yet, the wrong methods produce misleading information, a form of junk science that will distract rather than advance the interpretive task.

The value of GPT to statutory interpretation arises despite contentious theoretical questions that divide judges and legal academics about statutory interpretation. In the end, most textualists and non-textualists alike place at least *some* value on an empirical assessment of the ordinary meaning of statutory terms.[5] On the one hand, textualists famously prioritize ordinary

[1] John Roberts, 2023 Year-End Report on the Federal Judiciary (Dec. 31, 2023), at https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf.

[2] *See* Benjamin Weiser, *Here's What Happens When Your Lawyer Uses ChatGPT*, NY TIMES (27 May 2023) (reporting on lawyer who cited non-existent cases in brief based on GPT); Benjamin Weiser & Jonah E. Bromwich, *Michael Cohen Used Artificial Intelligence in Feeding Lawyer Bogus Cases*, NY TIMES (29 Dec. 2023) (reporting the same problem with Google Bard). On the general ethical issues of using GPT in law practice, *see* Amy B. Cyphert, *A Human Being Wrote This Law Review Article: GPT-3 and the Practice of Law*, 55 U. CAL.-DAVIS L. REV. 401, 423-37 (2021). *See also infra* note 28.

[3] *See, e.g.*, Matt Reynolds, *Words with Bots: How ChatGPT and Other AI Platforms Could Dramatically Reshape the Legal Industry*, 109 ABA J. 34, 36 (2023) ("Suffolk Law School Dean Andrew Perlman used ChatGPT to help write a 24-page law review article, draft a U.S. Supreme Court brief on same-sex marriage, craft deposition questions and work on a real estate contract."); Daniel Schwarcz & Jonathan H. Choi, *AI Tools for Lawyers: A Practical Guide*, 108 MINN. L. REV. HEADNOTES 1 (2023) (describing how lawyers can exploit LLMs to produce high-quality legal writing); Neel Ghua, *et al.*, *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models* (Sept. 26, 2023) (reporting on studies of LLM success at various lawyering tasks), at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4583531).

[4] *See* Kevin Tobia, Jesse Egbert & Thomas R. Lee, *Triangulating Ordinary Meaning*, 112 GEO. L.J. ONLINE 23 (2023).

[5] *See* James Macleod, *Finding Original Public Meaning*, 56 GA. L. REV. 1, 4-6 (2021) (explaining that textualists view their inquiry into "original public meaning" as "factual and empirical, not normative"); Kevin Tobia & John Mikhail, *Two Types of Empirical Textualism*, 86 BROOKLYN L. REV. 461, 461 (2021) ("There is significant debate about the meaning of 'ordinary meaning,' but there is general agreement that it is an *empirical* notion, closely connected to facts about how ordinary people understand language."). The term "textualism" itself refers to a set of related interpretive theories rather than a single method. *See, e.g.*, Tara Leigh Grove, *Which Textualism?*, 134 HARV. L. REV. 265 (2020); William N. Eskridge, Jr., Brian Slocum & Kevin Tobia, *Textualism's Defining Moment*, 123 COLUM. L. REV. 1611 (2023).

meaning – "how the ordinary English speaker . . . would understand the words of a statute,"[6] which naturally demands an empirical understanding of how ordinary people use the terms to be interpreted. As Gary Lawson puts it, "Meaning is an empirical fact."[7] That is why textualists have recently shown interest in moving beyond dictionaries to find evidence of meaning in corpus linguistics – the systematic exploration of large corpora of written English.[8]

On the other hand, even non-textualists usually begin with and always consider the text, and usually consider its ordinary meaning.[9] Even if non-textualist interpretation sometimes *also* requires normative reasoning, or positive reasoning about technical or legalistic meaning, ordinary meaning still matters. As Dan Farber explains: "[E]very legal system recognizes the importance of ordinary meaning. . . What method of statutory interpretation would view the ordinary meaning of words as completely irrelevant?"[10] Ordinary meaning refers in some way to how real people ordinarily use the terms being interpreted, which is an empirical issue.

GPT builds on billions of words human speakers have written down, more than any individual human being could ever read or write in her entire lifetime. Arguably, its unprecedented

[6] Amy Coney Barrett, *Congressional Insiders and Outsiders*, 84 U. CHI. L. REV. 2193, 2194 (2017).

[7] Gary Lawson, *Reflections of an Empirical Reader (Or: Could Fleming Be Right this Time?)*, 96 BOSTON U. LAW REV. 1457, 1475 (2016). *See also id*., at 1460 ("[T]o figure out what the document actually says . . . one must be an empirical reader"). Even when their focus is constitutional originalism, scholars like Lawson make the point in an intentionally general way that applies to statutory interpretation as well. On the link between the two, *see* ANTONIN SCALIA, A MATTER OF INTERPRETATION: FEDERAL COURTS AND THE LAW 38 (Amy Gutmann ed., 1997) ("What I look for in the Constitution is precisely what I look for in a statute: the original meaning of the text ...."); Steven G. Calabresi & Hannah M. Begley, *Originalism and Same-Sex Marriage*, 70 U. MIA. L. REV. 648, 649 (2016) ("[A]ll modern originalists ... are original public meaning textualists"). Other originalists echo Lawson's transcendent point about empiricism. *See* Larry Alexander, *Connecting the Rule of Recognition and Intentionalist Interpretation: An Essay in Honor of Richard Kay*, 52 CONN. LAW REV. 1513, 1525 (2021) ("Interpretation of legal texts is an empirical, not a normative, endeavor."); Lawence B. Solum, *Originalist Methodology*, 84 U. CHI. L. REV. 269, 278 (2017) (explaining that "interpretation is a factual inquiry"); Randy Barnett, *Interpretation and Construction*, 34 HARV. J.L. & PUB. POL'Y 65, 66-67 (2011) ("It cannot be overstressed that the activity of determining semantic meaning at the time of enactment required by the first proposition is empirical, not normative.").

[8] *See* Stefan Th. Gries & Brian Slocum, *Ordinary Meaning and Corpus Linguistics*, 2017 BYU L. REV. 1417 (2017); Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788 (2018); James C. Phillips, Daniel M. Ortner & Thomas R. Lee, *Corpus Linguistics & Original Public Meaning: A New Tool to Make Originalism More Empirical*, 126 YALE L.J.F. 21, 21 (2016); Lawrence M. Solan, *Can Corpus Linguistics Help Make Originalism Scientific?*, 126 YALE L.J.F. 57, 57 (2016); Lee J. Strang, *How Big Data Can Increase Originalism's Methodological Rigor: Using Corpus Linguistics to Reveal Original Language Conventions*, 50 U.C. DAVIS L. REV. 1181, 1202 (2017).

[9] See Jonathan T. Molot, *The Rise and Fall of Textualism*, 106 COLUM. L REV 1, 3 (2006) ("[T]extualism has so succeeded in discrediting strong purposivism that it has led even nonadherents to give great weight to statutory text."). For case examples, *see* Sebelius v. Cloer, 569 U.S. 369 (2013) (Sotomayor, J., writing for the Court) ("As in any statutory construction case, '[w]e start, of course, with the statutory text,' and proceed with the understanding that '[u]nless otherwise defined, statutory terms are generally interpreted in accordance with their ordinary meaning") (*citing* BP American Production Co. v. Burton, 549 U.S. 84, 91 (2006)); United States ex rel. Hartpence v. Kinetic Concepts, Inc., 792 F.3d 1121, 1128 (9th Cir. 2015) (en banc) (When interpreting a statute, "our inquiry begins with the statutory text, and ends there as well if the [statute's] text is unambiguous.") (*quoting* BedRoc Ltd. v. United States, 541 U.S. 176, 183 (2004). *See also* Abbe R. Gluck & Richard A. Posner, *Statutory Interpretation on the Bench: A Survey of Forty-Two Judges on the Federal Court of Appeals*, 131 HARV. L. REV. 1298, 1303-04 (2018) (finding that even the nineteen "legal process institutionalists," the older generation of federal appellate judges interviewed, "do not ignore statutory text and indeed many emphasize it."); Valerie C. Brannon, Cong. Rsch. Serv., R45153, STATUTORY INTERPRETATION: THEORIES, TOOLS, AND TRENDS 21 (2018); WILLIAM N. ESKRIDGE JR., INTERPRETING LAW: A PRIMER ON HOW TO READ STATUTES AND THE CONSTITUTION 33-55 (Robert C. Clark et al. eds., 2016).

[10] Dan Farber, *The Hermeneutic Tourist: Statutory Interpretation in Comparative Perspective*, 81 CORNELL LAW REV. 513, 516 (1996).

knowledge base, combined with the power of the algorithm that taps into this mass of language, turns it into a new and powerful source of empirical evidence for how people use words. We explore its potential through three Parts. Part I offers essential background. We begin with the importance of empirical evidence of ordinary meaning and then describe LLMs and their potential for providing such evidence. Finally, Part I asks "which LLM" one should use and explains why we conducted our testing on ChatGPT 3.5 Turbo.

Part II offers the first assessment of GPT as a source of evidence for the ordinary meaning of statutory terms. This Part exploits the fact that Kevin Tobia has run an experimental survey on the meaning of "vehicle," which asked a large number of respondents to identify whether particular objects (e.g., buses, bicycles, roller skates) were a vehicle.[11] We use Tobia's results as a basic benchmark for evaluating GPT. We evaluate four prompting techniques and demonstrate they do or do not create results reasonably similar to Tobia's.

Part III moves beyond benchmarking. In III.A, we test the effects of "telling" GPT that we want to know the meaning of "vehicle" because there is a ban on vehicles in the park. In III.B, we expand this approach by stating that we want to know the meaning of "vehicle" because it appears in some other kind of rule. We test a total of five statutory rules. In III.C, we shift from testing differences in rules to testing differences in the stated purpose of the rule. We report on results of six different statutory purposes. Finally, III.D tests GPT's ability to distinguish between intensional and extension types of meaning, and also its ability to provide evidence of meaning for some historic moment (the 1950s).

In Part IV, we reflect on what we have discovered and offer five tentative lessons for the use of GPT to generate empirical evidence of the ordinary meaning, including some important cautions.

I. The Value of Empirical Information About Statutory Meaning

A. The Value of Empiricism for Ordinary Meaning

As stated in the introduction, textualists clearly believe that meaning is empirical. We pause briefly to consider and reject some arguments that empirical evidence of meaning should *not* matter to non-textualists.

Suppose a new municipal ordinance declares, as in H.L.A. Hart's classic hypothetical, "No vehicles in the park."[12] The question arises whether "vehicle" includes a bicycle. Imagine that you are the interpreter (a judge or enforcer), and you discover a recent empirical study on the subject. You are persuaded as to the high quality of the study's methodology. In particular, you are impressed by the fact that the study focused on the meaning of "vehicle" among the precise population of individuals who are subject to the ordinance, the residents of the municipality that enacted and enforces it. But mysteriously, you are missing some printed pages of the

[11] *See* Kevin P. Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726 (2020).

[12] H.L.A. Hart, *Positivism and the Separation of Law and Morals*, 71 HARV. L. REV. 593, 607 (1958). Many scholars have continued to use this example, which is why we explored it using GPT. See *infra* Part III.

study and are, for the moment, left with this frustrating uncertainty: the study concluded *either* that (a) ninety-nine percent or (b) one percent of the population believe that the term "vehicle" includes a bicycle. Does your theory of statutory interpretation tell you not to bother going back to the website for the missing pages? Is empirical evidence of meaning *that* irrelevant?

We think not. As non-textualists ourselves, we believe the textualists are right that empirical evidence of this sort is at least relevant to statutory interpretation. Language is a practice and when meaning is contested, it is an empirical question who has the superior understanding of linguistic practice at issue. When judges consult a dictionary, they are trying to find a stronger empirical basis for their interpretation than their personal expertise as (native) speakers. Even when they consult nothing but "common sense" or how they could use words at a cocktail party without getting a funny look,[13] they advert to their own intuitive empirical assessment, as someone part of the American culture who communicates in the same language as the statutory text.

To be sure, there is room for disagreement about (1) exactly what empirical evidence has the most direct value, as well as (2) the best methodology for acquiring that evidence. One of many examples of the first issue is whether we should want to know how people in the relevant community *use* the word "vehicle" or *understand* it when others use the word. As for methodology, one of many examples is whether researchers should merely observe people in the community use language (spoken and written) including the word "vehicle," ask them open-ended questions about the word's meaning, or engage them in a careful experimental survey in which they apply their knowledge in various ways, or something else. Without engaging these questions, our point remains simple: statutory meaning turns, in part, on empirical facts about how people use and understand language. We shall see that GPT offers some evidence of these relevant facts.

Nonetheless, let us pause again and briefly engage the putative dissenters, those who *seem* to resist the relevance of empirical evidence to statutory meaning. First, some of the criticism of empiricism in statutory interpretation derives from the fact that judges have sometimes been very bad at it, as Anya Bernstein has trenchantly demonstrated. Bernstein observed judicial opinions citing nineteenth century English novels for the meaning of twentieth century

[13] *See, e.g.,* Johnson v. United States*, 529 U.S. 694, 718 (2000) (Scalia, J.,* dissenting) ("[T]he acid test of whether a word can reasonably bear a particular meaning is whether you could use the word in that sense at a cocktail party without having people look at you funny."); Biden v. Nebraska, 600 U.S. 477, 511 (2023) (Barrett, J., concurring) (referring to the controversial major questions doctrine as "reflecting 'common sense'") (*quoting* FDA v. Brown & Williamson Tobacco Corp., 529 U.S. 120 (2000). At one point, Justice Barrett invokes this hypothetical to prove the common sense of the doctrine:

Consider a parent who hires a babysitter to watch her young children over the weekend. As she walks out the door, the parent hands the babysitter her credit card and says: 'Make sure the kids have fun.' Emboldened the babysitter takes the kids on a road trip to an amusement park, where they spend two days on rollercoasters and one night in a hotel. Was the babysitter's trip consistent with the parent's instruction?

Biden, *supra*, at 513. She concludes not: "If a parent were willing to greenlight a trip that big, we would expect much more clarity than a general instruction to 'make sure the kids have fun.'" *Id.* Yet a recent study found that Barrett's intuition was empirically false because the overwhelming majority (92%) of survey respondents said the park trip did not violate the parental instructions. *See* Kevin Tobia, Daniel Walters & Brian G. Slocum, *Major Questions, Common Sense?* 97 S. Cal. L. Rev. (forthcoming June 2024) (manuscript at 41-43), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4520697.

American statutes.[14] As she notes, novels intentionally use words in nonstandard ways, English speakers of different nations use words differently, and meaning can drift considerably from one century to the next.[15] We might add that a single instance of usage is a mere anecdote about what a word can mean, no matter how distinguished the novelist. But the fact that judges have sometimes stumbled badly in the empirical enterprise of interpretation is not a good reason to think that the enterprise is not empirical.

Second, Brian Slocum observes that efforts of empiricism may fail because they are insufficiently sensitive to the contextual nature of language.[16] For example, the fact that the term appears in a legal statute (as opposed, say, to a news article or short story) is vital context, as are the precise constellation of other terms in the statutory text. An empirical methods that ignores such context may lead us astray.[17] To illustrate, gathering data from how people use the term *vehicle* when they are thinking of an insurance policy coverage for vehicles or when they are thinking of criminal liability for operating a vehicle under the influence of alcohol may not predict how they think of vehicles when the context is "no vehicles in the park."

No one really denies that context matters to meaning, but we offer two replies to this concern. First, using the right prompts, it is possible, as we show, to gather evidence via GPT that is sensitive to context. One can then use this data, perhaps in combination with other empirical evidence, to determine the meaning of statutory terms.

Second, at a more general level, we agree that data from the precise context at issue – the ideal data – is definitely better than data from any other context – the non-ideal data. But there remains the *possibility* that non-ideal data is better than nothing. For example, suppose that 99% of respondents in the municipality think that separate ordinances on liability insurance for *vehicles* and on operating a *vehicle* while under the influence do *not* apply to bicycles. In these contexts, a bicycle is, by this evidence, not a vehicle. While that is not itself determinative for the meaning of "no *vehicles* in the park," it does create a presumption in favor of excluding bicycles for that as well.

It may be a weak presumption, one that can be overcome merely by any reason to suppose that the specific context of "no vehicles in the park" will change the meaning in favor of greater breadth, at least to bicycles. But if there is no such argument, or if the argument is met by an

[14] *See* Anya Bernstein, *Democratizing Interpretation*, 60 WM. & MARY L. REV. 435, 442-51 (2018). She refers there to Whitfield v. United States, 574 U.S. 265, 268 (2015), where Justice Scalia cited, among other things, Jane Austen's *Pride and Prejudice* (published 1813) and Charles Dicken's *David Copperfield* (published 1849) for the meaning of "accompany" in a federal statute enacted in 1934. She also points to Muscarello v. United States, 524 U.S. 125, 126 (1998), where Justice Breyer cited, among other things, Herman Melville's *Moby-Dick* (1851) and the *King James Bible* (1611) to determine the meaning of "carry" in a federal statute enacted in the 1960s.

[15] Bernstein, *supra* note 14, at 444-47.

[16] Brian G. Slocum, *Big Data and Accuracy in Statutory Interpretation*, 86 BROOKLYN L. REV. 357 (2021).

[17] *See id.* at 380:

> Statutory interpretation involves consideration of evidence of both general and specific language usage. Corpus linguistics can provide important information about general language usage, but such evidence must be combined with consideration of the specific context of a statute. The latter inquiry is not determined through corpus analysis. The empirical view thus fails to sufficiently account for judicial consideration of the specific context of a statute.

equally strong counterargument that the context of "no vehicles in the park" justifies a narrower reading of vehicles, then the evidence from another context may still tip the balance. In the end, the relevance of context to meaning cannot ultimately be a reason to squarely disregard the empirical evidence from any other context – even "nearby" ones – unless one embraces an unappealing particularism, in which the same word in different contexts not only have different shades of meaning but meanings that are not even correlated with each other.[18]

Finally, Tara Leigh Grove objects to the reduction of statutory interpretation to empirical fact because the claim ignores normative steps in interpretation.[19] We do not contest that there may be normative questions embedded in interpretation alongside empirical ones. For example, we agree with Grove that there are normative issues about "how well-informed the hypothetical reasonable reader" of a statute "should presumptively be."[20] But however one answers these analytic and normative questions, an empirical project remains. One cannot determine how any particular reader, actual or hypothetical, will understand a statute without empirical guidance. Only an empirical understanding of how ordinary people reason about meaning, once informed in a certain way, could illuminate ordinary meaning. As we read her, Grove does not argue otherwise.[21]

There remains the possibility of a theory that says that some normative duty – perhaps grounded in a contentious moral or political theory – compels a particular interpretation regardless of what any particular group of people would imagine the words mean. Perhaps. We will grant that if there are such judges, they will not be interested in empirical evidence of meaning, because they see their role entirely as doing whatever their version of justice requires. But we have now ventured away from the usual meaning of a judge. In the post-legal realist world, the judge may care about normative theory, but the broadest conceptions of judging ordinarily require some fidelity to statutory text, some weight given to the formal legal materials rather than one's preferred normative view of the world. As long as the text matters at all, it also matters not just what the judge thinks the words should mean, but what others think the words do mean. And that inquiry is empirical.

[18] Put differently, we read as Slocum's point, that interpretation is not, in general, empirical, *id.* at 387-88, as really a claim that interpretation is not *merely* empirical. That is consistent with our claim that empirical evidence is relevant and useful to interpretation. *See also* Brian G. Slocum & Stefan Th. Gries, *Judging Corpus Linguistics*, 94 S. CAL. L. REV. POSTSCRIPT 13, 20 (2020) (stating that "statutory interpretation is not empirical in any real sense, even if one or more aspects of an interpretation may have an empirical basis"). The last clause justifies our inquiry into how LLM can improve the empiricism for those aspects that are empirical.

[19] *See* Tara Leigh Grove, *Testing Textualism's 'Ordinary Meaning,'* 90 GEORGE WASH. LAW REV. 1053 (2022).

[20] *Id.* at 1070-71.

[21] *See id.*, at 1073-74 (noting with apparent approval empirical work that does not "go so far as to proclaim that statutory analysis can be *entirely* data-driven") (emphasis added).

B. LLMs as a Source of Ordinary Meaning Empiricism

The empirical turn in legal scholarship in recent years[22] has led to an empirical turn in legal scholarship on statutory meaning. Various scholars have offered different ways of improving on the empiricism embedded in dictionary definitions.[23] Kevin Tobia in particular has authored or coauthored a series of papers using experimental surveys to test questions of ordinary meaning (some results of which we use below),[24] while Jonathan Choi offers computational methods to estimate the cosine similarity of different words.[25]

LLMs such as GPT are a new possible source of empirical information about meaning. To date, no one has explored how GPT might provide empirical evidence to assist lawyers and judges in statutory interpretation.[26] We contend that this new source has great potential and is therefore worth considering for the legal community. At the same time, our efforts to use GPT demonstrates some of the pitfalls to be avoided. We begin with some essential background.

[22] *See, e.g.*, Thomas J. Miles & Cass R. Sunstein, *The New Legal Realism*, 75 U. CHI. L. REV. (2008); Christina L. Boyd, *In Defense of Empirical Legal Studies*, 63 BUFF. L. REV. 363 (2015); Tom Ginsburg & Thomas Miles, *Empiricism and the Rising Incidence of Coauthorship in Law*, 2011 U. ILL. L. REV. 1785 (2011). Regarding the expansion of experimental methods in law, *see, e.g.*, Kevin Tobia, *Experimental Jurisprudence*, 89 U. CHI. L. REV. 735, 735 (2022); Roseanna Sommers, *Experimental Jurisprudence: Psychologists Probe Lay Understandings of Legal Constructs*, 373 SCIENCE 394 (2021). Cornell University Law School published Volume 1, Issue 1 of the *Journal of Empirical Legal Studies* in March 2004. A couple of years later, Law Professors Bernie Black, Jennifer Arlen, Geoffrey Miller, Ted Eisenberg, and Michael Heise organized the first Conference on Empirical Legal Studies, which has met annually since. See https://community.lawschool.cornell.edu/sels/cels-conferences/. Other prominent peer-reviewed journals in law -- *Journal of Legal Studies, Journal of Legal Analysis, Journal of Law and Economics, and Journal of Law, Economics and Organization* – regularly publish empirical papers.

[23] *See, e.g.,* the sources cited supra note 8; William N. Eskridge, Jr., Brian G. Slocum, & Stefan Th. Gries, *The Meaning of Sex: Dynamic Words, Novel Applications, and Original Public Meaning*, 119 MICH. L. REV. 1503 (2021).

[24] *See* Tobia, *supra* note 11; Kevin Tobia, Brian G. Slocum & Victoria Norse, *Ordinary Meanings and Ordinary People*, 171 U. PA. L. REV. 365 (2023); Kevin Tobia, Brian G. Slocum, & Victoria Norse, *Statutory Interpretation from the Outside*, 122 COLUM. L. REV. 213 (2022) (using survey experiments to test whether ordinary people subscribe to traditional canons of interpretation); Tobia, Egbert & Lee, *supra* note 4.

[25] Jonathan H. Choi, *Measuring Clarity in Legal Text*, 91 U. CHI. L. REV. (forthcoming 2024).

[26] We have located two unpublished papers that are closest to our own. First is Andrew Blair-Stanek, Nils Holzenberger, & Benjamin Van Durme, *Can GPT-3 Perform Statutory Reasoning?* (Paper for the June 2023 International Conference on Artificial Intelligence, University of Minho Law School, Portugal), at https://arxiv.org/pdf/2302.06100.pdf. It focuses, however, on asking GPT to provide the legal answers in factual scenarios where the answer requires *statutory reasoning*, *i.e.*, finding relevant statutory provisions and matching the facts of the case to them. For example, the authors ask "how much tax an individual had to pay," *id.* at 2, given some set of facts. Thus, the issue being tested is whether GPT can interpret reasonably clear statutes on its own without assistance by a professionally trained lawyer (and the results are quite mixed). Our inquiry, by contrast, tests whether GPT can give meaningful and reliable responses if the law itself has decided that its interpretation shall not be filtered by legal education, and hence not the interpretation by legal experts matters, but the interpretation by representative members of the general public.
The second paper is Neel Ghua, *et al.*, *supra* note 3 (summarizing existing literature on LLMs including GPT and legal reasoning). Some of the paper concerns interpretation, and some of that concerns statutory interpretation, but only to (1) answer specific legal questions about clear statutory text (as in the prior paper), *id.*, at 103-05, or (2) identify if a judicial opinion used a certain kind of methodology, such as textualism, when interpreting a statute. *Id.* at 116-18. *See also* Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme, *A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering*, In Proceedings of the Natural Legal Language Processing Workshop 2020, at https://ceur-ws.org/Vol-2645/paper5.pdf (testing not GPT but an older machine learning model on a dataset of rules derived from the US Internal Revenue Code).

1. General Background on LLMs

LLMs are prediction engines. Given the prompt they receive from the user and given the large amount of text on which they have been trained, they predict the most likely continuation of the text.[27] Since ChatGPT has been released, ordinary people have seen it with their own eyes. When they type in a question in plain English, they get a meaningful response. The model is not only able to interpret plain language, and to respond in a non-technical manner. It also, *grosso modo*, is able to offer reasons. These models are still far from perfect. In particular their tendency to hallucinate has caught public attention. Not so rarely responses "invent" a reality that does not exist.[28] The responses should therefore be dealt with caution. But for the most part, the responses are grounded and coherent. This includes legal applications, where the responses have been compared with "labelled" data, *i.e.*, lists of responses that are considered correct.[29]

The impressive performance of LLMs results from their architecture, and the data on which they have been trained. An LLM is a neural network characterized by two features: the architecture is layered, and the model is able to work bidirectionally.[30] The first feature makes it possible to have separate passes at the input data, for instance to distinguish basic grammar from the tone of a sentence. The second feature is particularly congenial to language. In a sentence, the same sequence of characters can have very different meaning, depending on the words by which this set of characters is surrounded. The language model can therefore build an expectation when receiving the first words and can revise this expectation in the light of the concluding words.

Technically, LLMs do actually not work with words, they work with vectors of probabilities, called embeddings. Computationally, numbers are much easier to handle than words. More importantly, these vectors of probabilities have a very high dimension, and hence characterize the individual word, the entire sentence, or the entire paragraph, in a high number of respects. The translation of verbal input into such vectors is called a transformer. The use of transformers has revolutionized neural networks. It has not only made them much more efficient, but also much more accurate.[31]

Any algorithm is only as good as the data on which it has been trained. For its latest model GPT-4, its provider OpenAI has not disclosed the composition of the training data.[32] This is

[27] On this core feature of LLMs, *see* Yutian Chen, et al., Token Prediction as Implicit Classification to identify LLM-Generated Text (unpublished manuscript dated 15 Nov. 2023), at https://arxiv.org/pdf/2311.08723.

[28] *See* sources cited *supra* note 2; Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, Weiqiang Jia, *Cognitive Mirage: A Review of Hallucinations in Large Language Models* (unpublished paper dated 13 Sept. 2023), at https://arxiv.org/abs/2309.06794; Yue Zhang, *et al., Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models* (unpublished paper dated 24 Sept. 2023), at https://arxiv.org/abs/2309.01219.

[29] *See* Ghua, *supra* note 3.

[30] For an accessible introduction into the architecture of machine learning models, and neural networks in particular, *see* GARETH JAMES, DANIELA WITTEN, TREVOR HASTIE & ROBERT TIBSHIRANI, AN INTRODUCTION TO STATISTICAL LEARNING WITH APPLICATIONS IN R (2013).

[31] For background, *see* UDAY KAMATH, KENNETH GRAHAM, WAEL EMARA, TRANSFORMERS FOR MACHINE LEARNING: A DEEP DIVE (2022).

[32] The Technical Report is silent on the training data. *See* OpenAI, *GPT-4 Technical Report* (dated 19 Dec. 2023), at https://arxiv.org/abs/2303.08774.

different for the predecessor model GPT-3.5 turbo (which we use in this project for reasons that we explain below). The basic ingredients are a moderated version of the Common Crawl dataset (410 billion words),[33] an expanded version of the WebText dataset (19 billion words),[34] two Internet based corpora of books (12 and 55 billion words), and English language Wikipedia (3 billion words).[35] This huge body of text is much bigger than the amount any human being has a chance to read during her entire life. Indeed, these training materials include vastly more words than the corpora on which corpus linguistics is based, which ranges in the hundreds of millions.[36] In short, LLMs are based on a *large* quantity of data.

Architecture and training data make us confident in the great potential of GPT to assess the ordinary meaning of English words. GPT has not been specifically trained on legal text.[37] For consideration of the "ordinary meaning" of words, this is a feature, not a bug. There may be some contexts in which statutory terms should be interpreted according to the non-ordinary, technical meaning of experts. Yet as we previously discussed, ordinary language is almost always relevant to issues of interpretation.[38]

2. The Potential of LLMs for Statutory Interpretation

The biggest advantage of LLMs is accessibility. Generating the rich empirical evidence we present below did not cost more than some fifty dollars, and, consolidating all the time engaging GPT, did not take longer than a couple of days. Admittedly, preparing the ultimate data generation pipeline was laborious. We had to overcome a series of coding challenges, in particular originating in technical bugs of GPT itself. Analyzing the data requires a certain degree of expertise with data wrangling. We therefore mainly present our own efforts as both a conceptual and a technical proof of concept. But the big providers of LLMs make it increasingly easy (almost as we speak) to build very accessible applications the use of which requires little if any coding expertise.[39] Were the legal community to embrace the method that we propose in this article, it would be possible to build such an interface, and to make it publicly available.[40]

[33] *See* Common Crawl, *Overview*, at https://commoncrawl.org/the-data/.

[34] Alec Radford, et al., *Language models are Unsupervised Multitask Learners* (2019), at https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[35] Tom B. Brown, et al., *Language Models are Few-Shot Learners*, 8 and Appendix A (22 July 2020), at https://arxiv.org/abs/2005.14165.

[36] The NOW corpus (News on the Web) currently claims to have 18.5 billion words of data. *See* https://www.english-corpora.org/now/. *See also* Tobia, Egbert & Lee, *supra* note 4, at 37 (using the NOW data to explore ordinary meaning); Thomas R. Lee & Stephen C. Mouritsen, *The Corpus and the Critics*, 88 U. CHI. L. REV. 275, 304 (2021) (reporting that the relevant corpora for corpus linguistics "range from hundreds of millions of words to several billion words").

[37] Unlike the recently launched Lexis AI, https://www.lexisnexis.com/en-us/products/lexis-plus-ai.page.

[38] *See supra* notes 4-8 and Part I-A.

[39] For examples, *see* https://cookbook.openai.com.

[40] One of us has already built such an interface for another use case, the implementation of interactive behavioral experiments between multiple instances of LLMs. See Christoph Engel, Max R.P. Grossmann, & Axel Ockenfels, *Integrating Machine Behavior into Human Subject Experiments: A User-Friendly Toolkit and Illustrations*, (MPI Collective Goods Discussion Paper, No.2024/1) (Jan. 3, 2024), at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4682602.

LLMs therefore have the potential to democratize the use of empirical evidence. Not only could judges run experiments of the kind that we report in this article, but so can the parties, or legal scholars observing the dispute. Since our data generation was quick and cheap, the standard excuse for using inferior empirical methods, or just none, vanishes. If it is of the upmost importance for the case at hand which interpretation gets it right, legal practitioners will likely not stop with probing LLMs. But with the help of these models, it will often be possible to constrain the contested area. If, from the perspective of the LLM (or multiple competing LLMs, like GPT on the one hand, and Gemini on the other) one interpretation seems obvious, the burden of argumentation shifts to those who, nonetheless, plead for an alternative interpretation. Much more involved empirical exercises, like the use of computer linguistics, or surveys with human participants, can be reserved for plausibly critical cases, given the evidence from the LLM.

The low cost of LLM use reveals a second advantage. Even well-endowed courts, and wealthy parties, can only afford the generation of so much evidence. For this article, we have given GPT seventeen different tasks (reported in Parts II and III). We have applied each task to twenty-five different candidate objects. We have repeated each of these questions 100 times, to generate an entire distribution and observe the variance. In sum, we made 42,500 requests. Even relatively cheap procedures for generating responses from human subjects, like Amazon Mechanical Turk which researchers like Tobia have used, would not be able to generate that much data. For the courts and the parties, such a data generation exercise would not be affordable. LLMs therefore make it possible to run much more differentiated empirical investigations. It would also be possible to do this iteratively. This opens up the possibility that the judicial users of the evidence come back, once they have seen the first batch of evidence, and probe the empirical basis of their provisional conclusions more closely.

Notwithstanding these advantages, we concede that accuracy is not the only concern. Interpreting statutes is an exercise of power. Democracy values putting power in the hands of the elected representatives in the legislature. The rule of law values giving those who are subject to a law a reasonable chance to foresee what the law demands of them. For these reasons, it would be good to understand exactly what GPT does when responding to a query of the sort that we have given it. Yet the precise architecture of the algorithm is proprietary, as is the exact composition of the training data. GPT is a proverbial black box.[41]

In this respect, using GPT to generate evidence about the ordinary meaning of statutory terms might be compared to using a risk assessment tool like COMPAS to generate evidence about recidivism risk for criminal sentencing.[42] COMPAS – Correctional Offender Management Profiling for Alternative Sanctions – has proved to be controversial. Critics assail COMPAS and

[41] A related normative concern is that many members of the public distrust decisions made by algorithms, and prefer a human decision-maker, so might object to anything that seems like judicial deference to a machine interpretation of legislation. *See, e.g.*, Berkeley J. Dietvorst, Joseph P Simmons & Cade Massey, *Algorithm Aversion. People Erroneously Avoid Algorithms after Seeing Them Err*,144 J. Exp. Psych: General 114 (2015).

[42] *See, e.g.*, State v. Loomis, 881 N.W.2d 749 (Wisc. 2016) (upholding constitutionality of using COMPAS assessments in criminal sentencing as a reason to deny an offender probation). The court explains:
COMPAS is a risk-need assessment tool designed by Northpointe, Inc. . . . A COMPAS report consists of a risk assessment designed to predict recidivism and a separate needs assessment for identifying program needs in areas such as employment, housing and substance abuse. The risk assessment portion of COMPAS generates risk scores displayed in the form of a bar chart, with three bars that represent pretrial recidivism risk, general recidivism risk, and violence recidivism risk . . . on a scale of one to ten. *Id.* at 754.

similar algorithms for the low accuracy of their predictions,[43] for racial bias,[44] and for the fact that the data and algorithm are the property of a politically unaccountable, for-profit firm.[45] The critics argue that trial courts could be overly influenced by the seeming precision of the machine predictions,[46] and lack the expertise to properly assess their probative value.[47] These critiques are part of a broader legal debate over using machine predictions for legal decision making, especially in criminal procedure.

We take these concerns seriously. They might be sufficient to justify avoiding GPT for the purpose of statutory (or other legal) interpretation. Perhaps courts should even tell litigants not to cite to such evidence in legal briefs or court arguments, unless and until the LLMs are open source, and exhibit satisfactory performance.[48] Yet we think that ultimately such use is inevitable in law for reasons we have explained – the overwhelming ease and accessibility of LLMs. Soon enough, we surmise, people will use online LLMs as much as they use online dictionaries, if not more. In our federal system, we predict that quite a few courts will be or already are willing to consider such evidence, so we offer our analysis for how best to use GPT to generate evidence of the ordinary meaning of statutory terms. If GPT is coming or already here, we should make the most of it, harnessing it to provide reliable evidence of statutory meaning, and to avoid the unreliable. Legal scholars need to evaluate its use now rather than later.

We also note some reply to the democratic concern. Unlike an AI tool designed for governmental agents, LLMs are widely available to ordinary people. They can be asked questions by anyone using ordinary language and will reply with ordinary language.[49] GPT is a populist tool in a way that COMPAS is not.

Consider an analogy to a textualist argument for the preeminence of ordinary meaning: the claim that such interpretations are more transparent to ordinary citizens.[50] As the argument goes, the statute will give better notice to those governed by a rule if the textual terms at least presumptively carrying a meaning the ordinary citizen expects. That argument is always limited by the fact that citizens do not usually learn of the content of law by reading statutes. The

[43] *See* Iñigo De Miguel Beriain, *Does the Use of Risk Assessment in Sentences Respect the Right to Due Process? A Critical Analysis of the Wisconsin v. Looming Ruling*, 17 LAW, PROBAB. & RISK 45 (2018).

[44] *See, e.g.,* Aziz Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043 (2019); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014).

[45] *See, e.g.*, Emily Berman, *A Government of Laws and Not of Machines*, 98 B.U. L. REV. 1277 (2018); Alyssa M. Carlson, *The Need for Transparency in the Age of Predictive Sentencing Algorithms*, 103 IOWA L. REV. 303 (2017); Katherine Freeman, *Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis*, 18 N.C. J. LAW & TECH. 75 (2016).

[46] *See* Freeman, *supra* note 45, at 97-98; *Recent Cases, State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing*, 130 HARV. L. REV. 1530, 1536 (2017).

[47] *See* Megan T. Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 303, 306 (2018); *Recent Cases*, *supra* note 46, at 1535.

[48] Prototypes of open-source LLMs are already available: https://ai.meta.com/llama/; https://mistral.ai.

[49] Note that competing empirical methods are not necessarily better controlled democratically. For example, the public does not directly observe dictionary editors decide how to define words or what words to include, the decisions of those collecting corpora of English, nor researchers running experimental surveys on meaning. We do not mean there is no difference from LLMs, which are even more opaque, but we think the transparency issue is one of degree and not kind.

[50] *See, e.g.*, William N. Eskridge, Jr. & Philip P. Frickey, *Statutory Interpretation as Practical Reasoning*, 42 STAN. L. REV. 321, 340 (1990) ("[T]extualism appeals to the rule-of-law value that citizens ought to be able to read the statute books and know their rights and duties."); Bostock v. Clayton County, 140 S. Ct. 1731, 1828 (2020) (Kavanaugh, J., dissenting) (asserting that deviations from ordinary meaning "deprive the citizenry of fair notice of what the law is").

governed mostly get their knowledge of law indirectly, and not always accurately whatever the method of interpretation.[51]

Yet here is where LLMs have unexpected value. As others have noticed, LLMs have the potential to improve access to justice merely by accessing and explaining law to those who ask.[52] We must be vigilant in monitoring GPT in this function, much as we should be concerned that websites offering expert advice – think of WebMD or DIY sites for electrical rewiring work – do not lead people astray. But where the answers are accurate, GPT can potentially lower legal ignorance on a large scale.

A parallel possibility exists for lawyers seeking to support legal arguments of ordinary statutory meaning. Where only the most well-funded lawyers (usually meaning lawyers with the most affluent clients) can afford to conduct experimental survey research or spend the time to learn the best uses of corpus linguistics or cosine similarities, or to hire experts to do the work for them, most lawyers could follow our best prompt method to gather similar information from GPT. Moreover, scholars and others can offer templates for such research.[53] In this respect, the use of the black box of GPT for statutory interpretation might not be ideal for democratic governance, but there is the real prospect of compensating democratic returns.[54]

3. *Which LLM?*

At the time of generating the evidence for this project, we had a choice between the two models provided by OpenAI[55]: GPT-3.5 turbo and GPT-4. On most benchmarks, GPT-4 outperforms GPT-3.5 turbo.[56] Seemingly, we should therefore have used the "better" model. We have, however, decided against GPT-4 as we are chiefly interested in the capability of LLMs to generate distributions of outcomes. We interpret these distributions as the analogue to a sample of human participants. Hardly any behavioral experiment with human participants generates a near uniform set of responses. Rather for a host of reasons, responses vary: the task may be

[51] *See, e.g.,* Benjamin van Rooij, *Do People Know the Law? Empirical Evidence about Legal Knowledge and Its Implications for Compliance*, *In* CAMBRIDGE HANDBOOK OF COMPLIANCE 467 (Benjamin van Rooij & D. Daniel Sokol, eds., 2021).

[52] *See* Roberts, *supra* note 1, at 5 ("AI obviously has great potential to dramatically increase access to key information for lawyers and non-lawyers alike."); Cyphert, supra note [2], at 421-23 ("Scholars have acknowledged that [AI] will not fully solve the justice gap, but have nonetheless predicted it could make a real difference."); Kristen Sonday, *Thomas Reuters Forum: There's Potential for AI Chatbots to Increase Access to Justice*, 25 May 2023 ("Organiations like the Legal Services Corporation and Pro Bono Net have already made great strides in building out content-rich online guides, which will become even more intelligent, accurate, and efficient by using AI"), at https://www.thomsonreuters.com/en-us/posts/legal/forum-spring-2023-ai-chatbots/.

[53] *See* sources cited *supra* notes 39-40.

[54] We thus agree with the general approach of David Engstrom and Daniel Ho, who argue for monitoring, regulating, and improving rather than rejecting AI tools. *See* David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. REG. 800, 854 (2020) ("Given [the significant] stakes, policymakers, agency administrators, judges, lawyers, and technologists should think hard, and concretely, about how to spur, not stymie, government adoption of AI tools while building appropriate accountability mechanisms around their use.").

[55] Gemini Pro has only become publicly available after most of the data generation had already been completed. Originally this model could also only be accessed through chat, via Google Bard, while we need the API to generate data in a fully controlled manner, and to enable multiple repetitions, for generating a complete distribution. As of writing this paper, Gemini Ultra has not been made publicly available.

[56] *See Technical Report*, *supra* note 32, at 4-8.

difficult, and not all participants are equally good at finding the individually optimal solution; in generating the response, several behavioral effects compete, and participants differ in the way how they balance these motives; behavioral regularities are a matter of degree, and different individuals are differently influenced by these regularities. One of us has shown in other work that GPT is subject to similar influences.[57]

The main selling point of GPT-4 is improved accuracy.[58] We were concerned that this improvement comes at the cost of reducing variance. For our purposes, this would be counterproductive: we would no longer see the set of plausible responses that GPT infers from its training data. As we have explained, technically LLMs do next word prediction. Arguably, an increase in prediction accuracy results from increasing the probability that the language model identifies the best possible response, given the prompt.

For many use cases, it is important to get the best possible response. If one is exclusively interested in the model's best guess, one of course wants the model to pick the response that it considers most likely right even if the response is a close call. We are concerned, however, that GPT-4, by discriminating more vigorously between the majoritarian response and minority responses would deny us information about the minoritarian response even though the latter would not have been implausible in the first place. That would defeat the purpose of using an LLM to explore ordinary meaning, where the issue is often whether there is more than one plausible meaning of a term.

To put the idea into numbers: there may be two plausible responses, one with probability 51% (*e.g.*, that *vehicle* includes *bicycle*), the other with probability 49% (that *vehicle* excludes *bicycle*). Or even worse: there may be three plausible responses, one with probability 35%, the next with probability 33%, and the third with probability 32% (*e.g.*, respectively, the context of DUI makes it more likely that *vehicle* includes *bicycle*, the context makes it less likely that *vehicle* includes *bicycle*, and the context makes no difference). Then the most likely response only has the support of a little more than a third, but for all we know – given that the algorithm is not publicly known – it might still constitute the "best reply" according to GPT-4. For our purposes, the ability of LLM to reveal "close calls" is a critical advantage. If we can learn how much the model *had to struggle* with alternative responses, this tells us something about the likely distribution of ordinary meaning in the population.

In principle, one could just ask the model to disclose the responses it has considered, and the probabilities it has assigned to them being the right response. Yet unfortunately GPT-3.5-turbo does not disclose these probabilities. But there is a workaround. GPT makes it possible to define a parameter that it calls "temperature." In the user community, this parameter is often discussed as allowing the model to be more or less "creative." For us it simply is a technology for not only eliciting the one most likely response. Instead in all our data generation, we set this parameter at the high value of one.[59] The resulting distribution of choices enables us to

[57] *See supra* note 40.

[58] *See Technical Report*, *supra* note 32, at 4-8.

[59] A few months ago, GPT has increased the maximum temperature from 1 to 2. We have done a few tests but have gained the impression that temperature > 1 makes the responses very noisy. This is why we have kept the temperature parameter at 1.

measure the probability, given the prompting question, that GPT would give a positive response. This is our proxy for the interpretation of the term in question in the general population.

II. Proof of Concept: Testing GPT Against Benchmark of Statutory Meaning

It is standard in computer science to assess the performance of an algorithm against generally accepted benchmarks. In these tests, one compares the responses produced by the algorithm with "ground truth." The closer the algorithm matches the ground truth on these benchmarks, the more one is willing to trust the algorithm in other domains.[60] This section is written in the same spirit. We have exploited the fact that, in 2020, Kevin Tobia has published results of survey experiments he ran with human participants on the classic hypothetical widely discussed in legal theory: if there is a rule that forbids vehicles in the park, is a certain object to be classified as a vehicle?[61]

In this section, we report on four independent attempts at replicating his results with the help of GPT-3.5 turbo. We start with giving GPT the exact same question that Tobia had asked his participants. The results are mildly impressive. GPT makes a difference between obvious and debatable cases. But overall, it discriminates much less than human participants. If performance could not be improved, one would have reason to be very cautious when introducing GPT data into legal discourse.

Now it has quickly become clear after the introduction of the first LLMs that the way how one asks matters greatly.[62] In our second attempt, we employ a prompting technique that is generally considered to be effective. We no longer confine ourselves to asking for the final assessment (is the object in question a vehicle?). Rather we implement a "chain of thought"[63]: We first ask GPT to define a vehicle, and only thereafter ask it to classify the object in question. Yet for our purposes, this often-helpful prompt does not lead to a substantive improvement. Inspired by a frequent procedure in experiments with human participants,[64] in our third attempt, we replace the original question (is the object of a vehicle?) by the elicitation of a belief. We inform GPT that human participants have been asked this question. We ask GPT to estimate how many of them have given an affirmative response. This too does not substantially

[60] *See, e.g.*, the list of benchmark scores that Google has published when introducing Gemini, https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf (but the scores comparing Gemini with the competition, and in particular with GPT, should be used with care. On many benchmarks, Google has run multiple tests, but only uses the best performing for the comparison).

[61] *See* Hart, *supra* note 12; Lee & Mouritsen, *supra* note 8.

[62] For an easily accessible introduction to prompt engineering for lawyers *see* Jonathan H. Choi, *How to Use Large Language Models for Empirical Legal Research*, J. INSTIT. & THEORETICAL ECON. (forthcoming 2024) (part of 39th International Seminar on the New Institutional Economics -- Machine Learning and the Law).

[63] Jason Wei, et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (10 Jan. 2023), at https://arxiv.org/abs/2201.11903.

[64] *See* Mariana Blanco, Dirk Engelmann, Alexander Koch and Hans-Theo Normann, *Belief Elicitation in Experiments. Is There a Hedging Problem?,* 13 EXPER. ECON. 412(2010); Stefan T. Trautmann & Gijs Kuilen, *Belief Elicitation. A Horse Race among Truth Serums*, 125 ECON. J. 2116 (2015).

narrow the gap between Tobia's and our data. We do, however, get much closer in our fourth and final attempt, once we replace the percentage scale by a coarser measure, a seven-point Likert scale[65] running from "(almost) none," "very few", "few", "about half of them", "many", "very money" to "(almost) all".

A. The Benchmark: Tobia 2020

As part of his 1958 debate with Lon Fuller over the nature of law, the philosopher H.L.A. Hart first proposed the hypothetical in which "A legal rule forbids you to take a vehicle into the public park."[66] He used the example to explore the shades of meaning possible in a word like "vehicle," which "[p]lainly . . . forbids an automobile, but what about bicycles, roller skates, toy automobiles? What about airplanes?"[67] The example has been used in jurisprudence ever since.[68] As the prohibition would almost certainly be a statute or ordinance, it has featured prominently in theories of statutory interpretation.[69] For example, three prominent articles study the power of corpus linguistics to shed light on the meaning of "no vehicles in the park."[70]

More relevant for our purposes, Tobia used Hart's hypothetical to test the ability of experimental survey methods to illuminate the empirics of ordinary meaning.[71] Tobia tested dictionary definitions and corpus linguistics against the most direct evidence of ordinary meaning: what some actual Americans – 2,835 of them – thought about whether certain objects were "vehicles."[72] He found that corpus linguistics performed poorly in predicting ordinary meaning

[65] *See* Rensis Likert, *A Technique for the Measurement of Attitudes*, 140 ARCHIVES OF PSYCH 1 (1932); Andrew T. Jebb, Vincent Ng & Louis Tay *A Review of Key Likert Scale Development Advances: 1995–2019*, 12 FRONTIERS IN PSYCH (2021), at https://www.frontiersin.org/articles/10.3389/fpsyg.2021.637547.

[66] *See* Hart, *supra* note 12, at 607. He continues to use the example, with modifications, in H.L.A. HART, THE CONCEPT OF LAW 125-27 (Penelope A. Bulloch & Joseph Raz eds., 2d ed. 1994). Hart's example engaged works such as LON FULLER, THE LAW IN QUEST OF ITSELF 12 (1940), Lon Fuller, *Human Purpose and Natural Law*, 53 J. PHILOS. 697 (1953). Fuller responded in Lon L. Fuller, *Positivism and Fidelity to Law –A Reply to Professor Hart*, 71 HARV. L. REV. 630, 661-69 (1958), and later addressed similar issues in LON L. FULLER, THE MORALITY OF LAW 81-91 (rev. ed. 1969).

[67] Hart, *supra* note 12, at 607.

[68] A recent (23 Jan. 2024) Westlaw search of the Law Reviews & Journals database returned 294 articles to the prompt "vehicle /s park /p hart" and 2602 articles to the prompt "vehicle /s park." The 50th anniversary of the debate did not go unnoticed. *See, e.g.*, THE HART-FULLER DEBATE IN THE TWENTY-FIRST CENTURY (Peter Cane, ed., 2010 (based on a conference held in 2008). The NYU Law Review published a symposium on the anniversary. *See Forward: Fifty Years Later*, 83 NYU L. REV. 993 (2008). *See especially* Frederick Schauer, *A Critical Guide to Vehicles in the Park*, 83 NYU L. REV. 1109 (2008).

[69] *See, e.g.*, ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 36-39 (2012) (discussing "vehicles in the park"); Joshua Kleinfeld, *Textual Rules in Criminal Statutes*, 88 U. CHI. L. REV. 1791 (2021) (referring to the Hart's "classic article" and its "no vehicles in the park" example); William N. Eskridge, Jr. & Judith N. Levi, *Regulatory Variables and Statutory Interpretation*, 73 WASH. U. L.Q. 1103, 1103 (1995) (beginning article with a discussion of vehicles in the park). Schauer recently called the example "tiredly familiar" as he continued to find it useful to a critique of the interpretation-construction distinction. *See* Frederick Schauer, *Constructing Interpretation*, 101 B.U. L. REV. 103, 119 (2021).

[70] *See* Gries & Slocum, *supra* note 8, at 1463-1469; Lee & Mouritsen, *supra* note 8, at 836-45; Daniel Keller & Jesse Egbert, *Hypothesis Testing Ordinary Meaning*, 86 BROOKLYN L. REV. 489, 505, 510-32 (2021) (referring to Hart's hypothetical and using corpus linguistics to resolve whether a "scooter" is a vehicle).

[71] *See* Tobia, *supra* note 11, at 739.

[72] *Id.* at 765. These were "general population" participants from the United States recruited through Amazon's Mechanical Turk or "Mturk." *Id.*

as compared to his experimental survey method. In particular, corpus linguistics tends to reveal a term's "prototypical" uses, but not the full extent of its meaning.[73] We regard Tobia's study as the best extant evidence of ordinary meaning of "vehicles in the park," which is why we use it as benchmark for testing GPT.

For our purposes, it is important to have sufficient variance in the responses from human subjects against which we want to compare the distribution of responses generated by GPT. This is why we focus on the one test in which Tobia has used a large amount of test objects. Specifically, the data reported in his Figure 5 results from asking 2,835 online participants (on MTurk): "is X a vehicle"?[74]

As our Figure 1 shows, [75] the responses he received are nearly uniform only for a few objects. Ninety-seven percent of all participants say that a "truck" is a vehicle. On the other hand, only five and a half percent say that "crutches" are a vehicle. For cars, buses, automobiles and ambulances, a very large majority says they are vehicles. For zip lines and baby carriers, a very large majority says they are not vehicles. Yet for most test objects, the views of human participants diverge. For our purposes, this variance is fortunate, as it gives us a fine-grained benchmark.

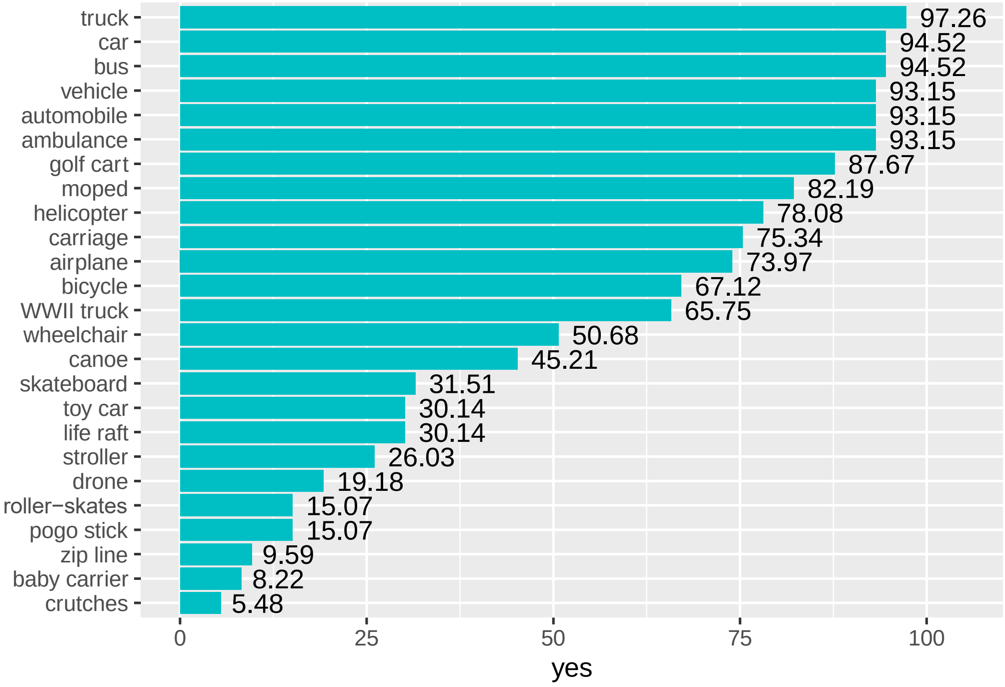

Figure 1
Ground truth: Kevin Tobia's data from MTurk participants

[73] *Id.* at 789-91. More precisely, corpus linguistics does a poor job of demonstrating the "extension" of a term, all the objects that fit the category the term identifies. We discuss the concept of extension, and the related concept of intension, *infra* text accompanying notes 105-110.

[74] *Id.* at 764-766 ("ordinary meaning" condition).

[75] We are grateful to Kevin Tobia for sharing the data required for reconstructing Figure 1 with us.

B. An Attempt at Direct Replication

In our first attempt at replicating Tobia’s results, we ask GPT the exact same question[76]:

> “Is the following a vehicle: <vehicle>?”

We use the exact same twenty-five candidate objects. For each object, we request 100 independent responses.[77] We allow for a rather high degree of variance by setting “temperature” to 1.

As Figure 2 shows, the attempted replication is only mildly successful. GPT apparently has rather strong opinions. For eleven objects, the LLM is perfectly certain that they are vehicles. For three objects, it is perfectly certain that they are not vehicles. Hence with this procedure, the intermediate range shrinks. Moreover, GPT has a pronouncedly different opinion about three objects: with fairly high confidence it classifies a wheelchair, a toy car and a drone as vehicles, whereas human participants are much more hesitant with classifying these objects.

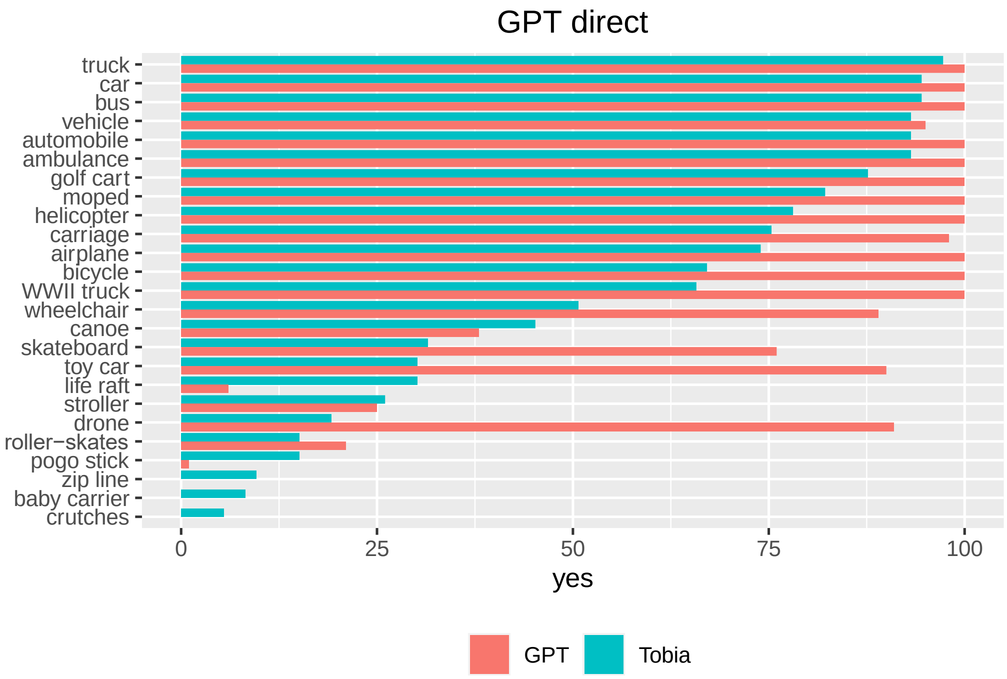
GPT direct

Figure 2

Attempt at directly replicating Tobia’s experiment

We are interested in comparing the responses received from GPT with the responses given by human subjects. As both samples have been tested on twenty-five different objects, and responses of human participants vary considerably across objects, the appropriate statistical

[76] For this first attempt, the system prompt reads simply: ”We want to learn your assessment. Please exclusively respond ‘Yes’ or ‘No.’”

[77] We learned that GPT does not always respect the system prompt “Please exclusively respond ‘Yes’ or ‘No.’” In the interest of always having 100 usable responses, we elicit a larger number (depending on the performance of GPT between 110 and 150). For data analysis we use the 100 first usable responses. This procedure is innocent as GPT does not remember earlier responses unless explicitly instructed to do so (technically: only when using its chat functionality).

procedure for comparing the two data sets is a comparison of the two distributions. A standard test for that purpose is Kolmogorov Smirnov. It turns out highly significant (p = .0054). We can therefore confidently conclude that human responses and these GPT responses are different from each other. This first attempt at replicating Tobia's results is unsuccessful, one of several important cautions we discovered to those seeking to apply GPT to statutory interpretation.

C. Chain of Thought Prompt

For many tasks, the accuracy of LLMs has been improved by what is called a "chain of thought" prompt. Rather than directly asking the question of interest, the LLM is guided towards the equivalent of a mental process for generating the response.78 We have given GPT a classification task. This is why, in this second attempt at replicating the results from human subjects, we have defined the task for GPT as follows:

> We have two related questions for you, one generic and one specific. In generic terms: how do you define a vehicle? Given this definition, do you classify the object that we will mention subsequently to be a vehicle? Hence please give us two responses:
>
> > a) your definition of a vehicle
> > b) the classification of the object.
>
> Please only respond to the second question with "yes" or "no." Here is the specific question:
>
> Answers to the first, generic question are very reasonable. GPT has for instance told us:
>
> "A vehicle is a machine that is designed or used to transport people or goods from one place to another."
>
> "My definition of a vehicle is a machine or device used for transporting people or goods, typically on land, air, or water."
>
> "A vehicle is a man-made object that is designed to transport people or goods from one place to another, typically using wheels, and powered by an engine or some other form of propulsion."
>
> "My definition of a vehicle is a movable object designed to transport people or goods."
>
> "My definition of a vehicle is a mechanical device that is used for transportation, typically on roads or other designated routes."

[78] For background and experiences with this prompting technique, *see* Wei et al., *supra* note 63).

Yet as Figure 3 shows, with this prompting technique, GPT becomes even more opinionated. It now is even 100% sure that 12 candidate objects are vehicles. The Kolmogorov Smirnov test has an even smaller p-value (p = .0018). The results from human subjects do clearly not replicate.

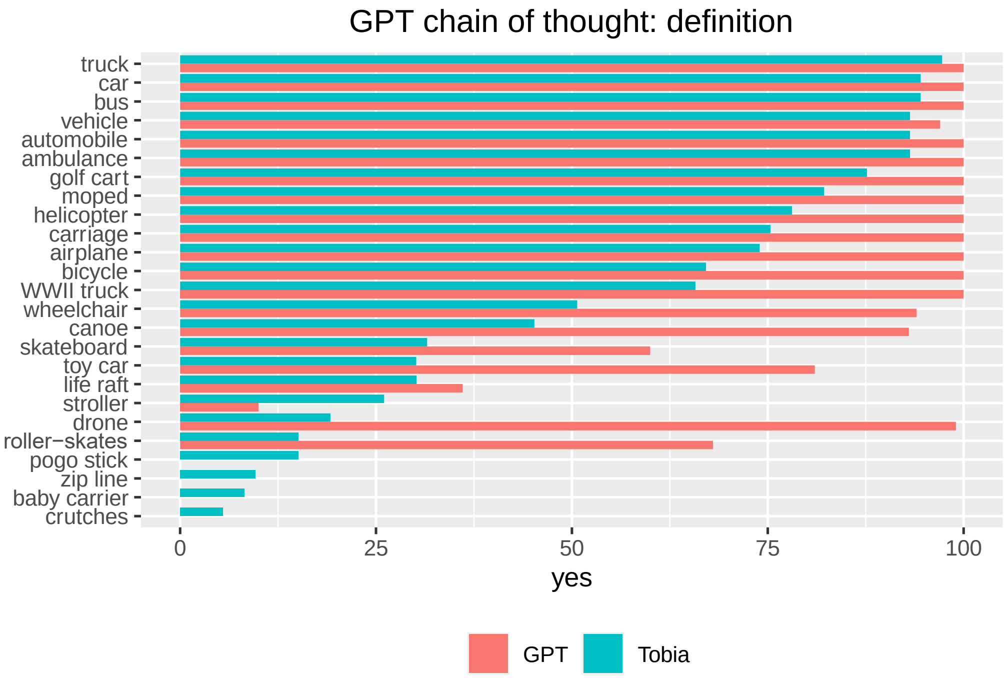

Figure 3
Attempt at replicating Tobia's experiment using a chain of thought prompt

D. Belief Prompt (Asking for a Percentage)

In the experimental literature, it is standard to elicit not only choices, but also beliefs. If the experiment is interactive, beliefs inform the experimenter about the way how one participant has constructed the choices of another participant to which she reacts.[79] If they are concerned that stated beliefs are self-serving and therefore biased, experimenters sometimes invite a new set of participants, explain the design of the original experiment, and ask them for their postdiction of the choices made by participants in the first experiment.[80] In our third attempt at replicating Tobia's results, we leverage this approach.

Specifically, we define the task as follows:

> In an experiment, 2,835 participants have been asked the question that we will show you below.

[79] For further details, *see* Blanco *et al.*, *supra* note 64; Trautmann et al., *supra* note 64.

[80] For instance, one of us has used that approach in Christoph Engel, Sebastian Kube & Michael Kurschilgen, *Managing Expectations: How Selective Information Affects Cooperation in Social Dilemma Games*, 187 J. ECON. BEHAV. & ORG. 1112021).

> What follows is the question that experimental participants have been asked, not the question we are asking you. From you we want to learn which percentage you believe have responded 'Yes'. Please do not give any explanations. Exclusively respond with a number between 0 and 100.
>
> This has been the question experimental participants have been asked: [“Is the following a vehicle: <vehicle>?”]

As Figure 4 shows, this prompting strategy has a dramatic effect. While in Figure 2 and Figure 3, many of GPT’s responses were extreme, now almost all responses are close to the midpoint. If one inspects individual choices, one sees that this pattern does not result from GPT predominantly giving responses at or near 50%. Rather there is a “regression to the mean”: GPT gives responses all over the range from 0 to 100%.

Unsurprisingly, this response pattern is as clearly distinct from the results received from human participants as with the chain of thought prompt (Kolmogorov Smirnov, p = .0019).

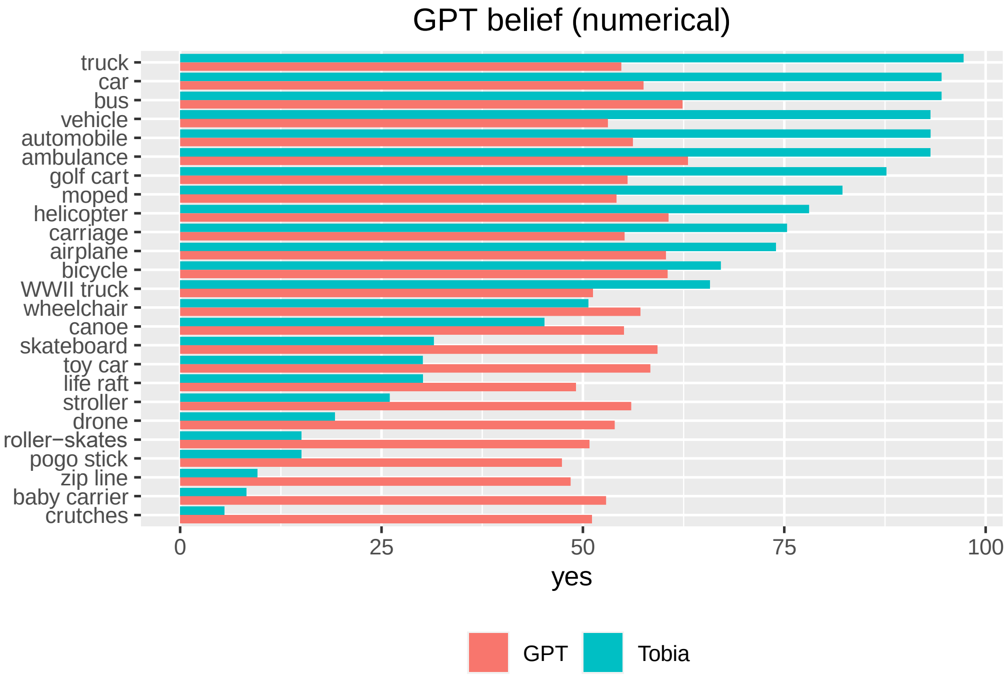

Figure 4

Attempt at replicating Tobia’s experiment asking for percentage beliefs about human choices

E. Belief Prompt (Using a Likert Scale)

It is well known that LLMs are not good at quantitative reasoning.[81] Upon a moment's reflection this is not too surprising. As the name says, large language models have been trained on human language, and have been trained for responding in a way that human recipients can immediately understand. Language is not per se good at quantitative assessments. Actually, a whole branch of developmental psychology has established the distinction between formal and intuitive mathematics: as long as they have not been mathematically trained, and in particular as they are still children, human subjects typically reason about quantitative tasks in a much coarser, qualitative way.[82] This analogy has triggered our fourth attempt at replicating the decisions made by human participants. In this attempt, we replicated the prior effort (a belief prompt) but also tried to bring GPT's measurement of human replies to a more human scale. Rather than asking for a precise percentage, we have introduced a 7-point Likert scale.[83] Hence instead of

> "Exclusively respond with a number between 0 and 100."

we instruct GPT:

> Just respond in one of these seven ways:
> (almost) none
> very few
> few
> about half of them
> many
> very many
> (almost) all

As Figure 5 shows, the mapping is still not perfect, but much improved over all earlier attempts. Visibly, the seven levels on the Likert scale, expressed in ordinary language, are much more congenial to the language model. Now GPT no longer underestimates the probability that objects are vehicles that are relatively clear cases for human participants. This is an improvement over asking for beliefs with a numerical scale (Figure 4). On the other hand, GPT no longer overestimates the probability that objects are vehicles about which human participants are less confident (Figure 2 and Figure 3). Effectively, the correlation between human and GPT

[81] *See* Shima Imani, Liang Du, & Harsh Shrivastava, *MathPrompter: Mathematical Reasoning using Large Language Models* (4 March 2023), at https://arxiv.org/abs/2303.05398.

[82] Key contributions include Elizabeth S. Spelke, *Natural Number and Natural Geometry*, in SPACE, TIME AND NUMBER IN THE BRAIN 287(Stanislas Dehaene & Elizabeth M. Brannon, eds., 2011); HALLARD T. CROFT, KENNETH FALCONER & RICHARD K. GUY, UNSOLVED PROBLEMS IN GEOMETRY: UNSOLVED PROBLEMS IN INTUITIVE MATHEMATICS (Vol. II) – (2012); Moira R. Dillon, Harini Kannan, Joshua T. Dean, Elizabeth S. Spelke, Esther Duflo, *Cognitive Science in the Field: A Preschool Intervention Durably Enhances Intuitive but not Formal Mathematics* – 357 SCIENCE 47 (2017).

[83] Note that this procedure is not inconsistent with putting temperature at the high value of 1. Temperature defines the degree of variance that the language model is allowed in generating responses. Shifting from percentages to the seven-point Likert scale is a change in the definition of the task. To see this difference, consider the results reported in subsection B (the attempt at a direct replication). In that data generating process, following the lead of Tobia, we even had constrained the set of potential responses to only two: yes or no.

responses is quite high for objects that human participants consider likely candidates (for the upper half of the figure). On the lower end, the mapping is less good. GPT ratings remain a bit more inclusive. Yet overall, the mapping is now reasonably good. The null hypothesis that both distributions are indistinguishable can no longer be rejected (Kolmogorov Smirnov, p = .2798).

This is of course not the same as proving that both distributions are indistinguishable. But the data from human participants is also not perfectly representative. The participants on MTurk are not a random draw from a sample that is representative for the population of the United States (as most Americans do not participate in MTurk). Despite the remaining differences between both distributions, we therefore feel entitled to use the belief prompt with a Likert scale as the starting point for the investigations in the following section.

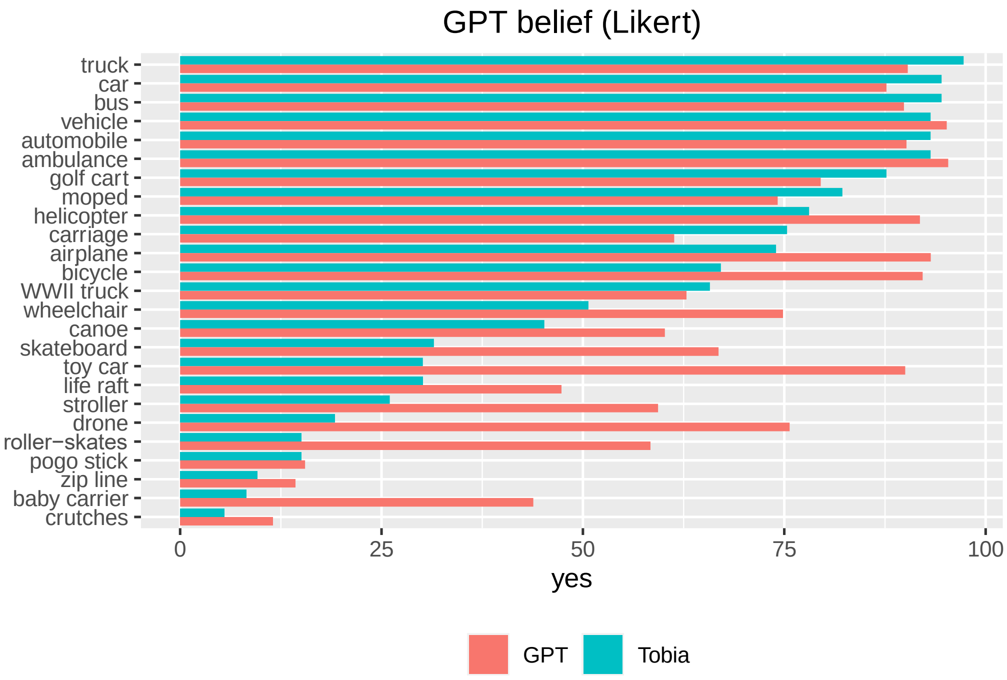

Figure 5

Attempt at replicating Tobia's experiment asking for beliefs about human choices on a Likert scale

In sum, we consider our last effort to "benchmark" GPT to be a success. With the right prompts, GPT gives us empirical data on the meaning of terms that is equivalent to the results of a large and sophisticated experimental survey of English-speaking humans. This is a proof of concept for using GPT to explore the ordinary meaning of statutory terms. Along the way, however, we discovered the bracing and important lesson that three very logical and plausible prompts generated unreliable results.

III. Beyond Replication: Introducing Context to GPT Prompts

In the previous section, we have shown that using a belief prompt, and asking for an assessment on a seven-point Likert scale, brings GPT responses reasonably close to human responses. In the remainder of this paper, we use this prompt as our workhorse to investigate the effect of alternative interpretative techniques. As in the previous section, we always elicit 100 responses, for each of the twenty-five candidate objects. When generating data with GPT, we can introduce context in a very precise manner. Technically, we exploit the possibility to add an "assistant prompt" to the exact same "system prompt" that we have used to generate the context free evidence reported in section II E.

We proceed as follows. In III.A, we "inform" GPT of the context for our inquiry – that we want to know whether an object is a "vehicle" because of a ban on vehicles in the park. In III.B, we broaden this idea by "informing" GPT of multiple alternative statutes in which the term "vehicle" is used. In III.C, we switch from testing different statutory contexts to testing six different purposes for the original no-vehicles-in-the-park rule. Finally, in III.D, we test whether GPT can distinguish intensional and extensional meaning and whether it can provide evidence of meanings from the past.

A. Disclosing the Wording of the Rule

Textualists do not subscribe to mere "literalism."[84] They would grant that textualism is about semantic meaning,[85] and that meaning depends on context.[86] The debate among textualists focuses on what counts as context.[87] As a rule of thumb, the more direct the context, the more likely it is to be considered. Through this contextual window, textualism takes into account which is the effect of subsuming a debated object under a statutory term.

In the first step, we investigate whether, and if so how, informing GPT about the wording of the no-vehicles-in-the-park rule matters to the results on the meaning of the debated term. Specifically, we use the following assistant prompt to instruct GPT:

> You are asking back: Why do you want to know?
>
> I am answering: There is a rule that says: no vehicles in the park.

As Figure 6 shows, if the content of the rule is disclosed, GPT becomes more cautious. Except for the somewhat enigmatic question whether a vehicle is a vehicle, GPT gets less confident than in the neutral setting. One might even wonder whether GPT implicitly assumes that the rule wants to protect visitors of the park from harm. That interpretation might explain why so

[84] *See* Scalia & Garner, *supra* note 69, at 40 ("The soundest legal view seeks to discern literal meaning in context."); John F. Manning, *The Absurdity Doctrine*, 116 Harv. L. Rev. 2387, 2456 (2003) (arguing that textualists are different from "their literalist predecessors in the 'plain meaning' school").

[85] John F. Manning, *What Divides Textualists from Purposivists,* 106 COLUM. L. REV. 70, 70 (2006) (claiming that textualism "gives priority to semantic context").

[86] Lawrence B. Solum, *Communicative Content and Legal Content*, 89 Notre Dame L. Rev. (2013).

[87] *See* Eskridge, Slocum & Tobia, *supra* note 5, at 1660-67.

many responses are positive for mopeds and strollers, and so little responses are positive for drones and wheelchairs.

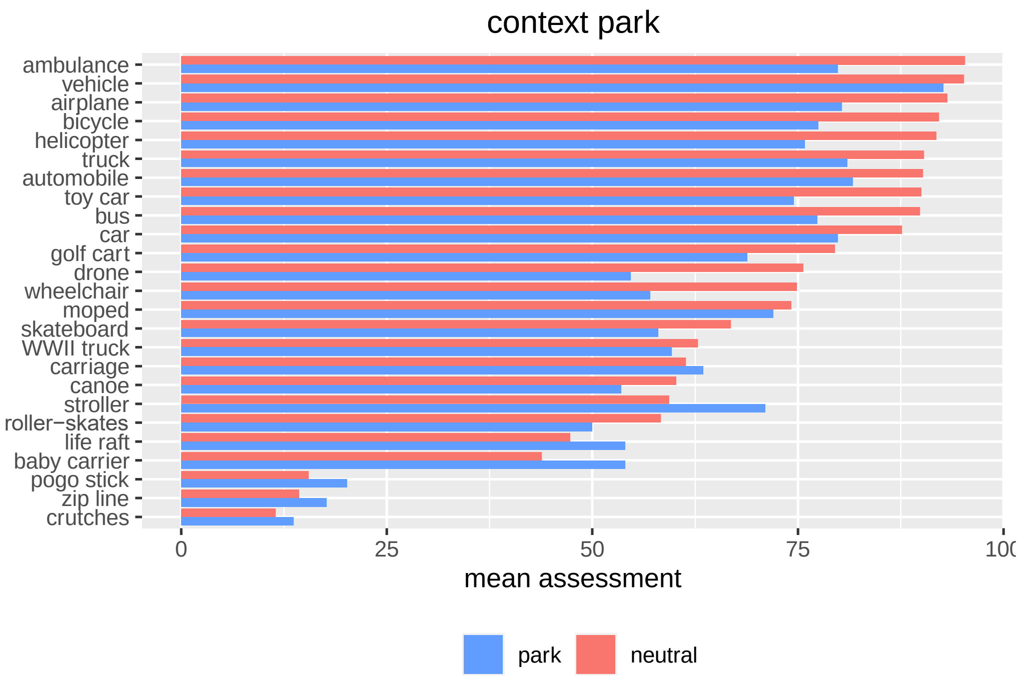

Figure 6
Effect of disclosing the content of the rule

If we reuse the same metric as employed in the previous section, *i.e.*, compare the two distributions with the help of the Kolmogorov Smirnov test, we find a weakly significant difference (p = .0754). This result nicely fits the impression conveyed by Figure 6: taking the content of the rule into account that uses the term "vehicle," interpretations do not radically change. But there is a small discernible shift. When probing GPT, there is some difference between the general meaning and the meaning in a particular context. Context matters.

B. Disclosing Alternative Rules

The prior finding invites an obvious extension: In which ways does GPT change its mean response if the classification of an object as a vehicle is relevant for *different* legal rules? We have tested five. We always give GPT the system prompt that has performed best in the comparison with Tobia's data from human participants.[88] Our manipulation is in the additional assistant prompts. We compare the following five prompts:

> **park:** You are asking back: Why do you want to know? I am answering: There is a rule that says: no vehicles in the park.

[88] *See supra* Section II.E.

> **dui:** You are asking back: Why do you want to know? I am answering: There is a rule that says: it is a crime to conduct a vehicle under the influence of drugs or alcohol.
>
> **liab:** You are asking back: Why do you want to know? I am answering: There is a rule that says: if an accident has been caused by a vehicle, the owner is liable even if she has not been negligent.
>
> **enhance:** You are asking back: Why do you want to know? I am answering: There is a rule that says: if a vehicle is used to commit violent crime, punishment is increased by 30%.
>
> **census:** You are asking back: Why do you want to know? I am answering: mandatory census requires that owners list any vehicle they own.

As Figure 7 shows, the dominant determinant is the character of the object, not the context, but context has a secondary significance. When Hart first proposed his hypothetical about vehicles in the park, he asserted that there must be "a core of settled meaning" for a term like "vehicle," in addition to "a penumbra of debatable cases in which words are neither obviously applicable nor obviously ruled out."[89] Indeed, GPT finds a core meaning for "vehicle." Across the legislative contexts we tested, automobiles, trucks, airplanes and cars very likely to be considered vehicles irrespective of the context of the particular rule. Likewise pogo sticks, zip lines and crutches are unlikely to be considered vehicles across all rules.

By contrast, many terms qualify as penumbral, and the legislative context often has some influence. Of greatest interest is the object that has generated the most academic interest ever since Hart proposed the hypothetical: the bicycle. For the bicycle, there is a twelve-point gap in the two legislative contexts in which GPT regards a bicycle to be *most* and *least* likely to be a vehicle (eighty-two percent for a criminal penalty enhancement versus seventy per cent for the DUI crime). Somewhat similar to a bicycle is a "moped," which also has a twelve percent gap between GPT considering it a vehicle in the context of a vehicle-in-the-park ban (seventy-two percent) versus the context of strict civil liability for vehicle accidents (sixty percent). Other penumbral objects with the greatest variation across contexts are the stroller (gap of sixteen percent),[90] carriage (fourteen percent), wheelchair (fourteen percent), and skateboard (thirteen percent). Yet for most other penumbral objects, we fail to detect any significant variation by legislative context.

[89] Hart, *supra* note 12, at 607.

[90] That GPT counts a stroller as included in a ban on vehicles in the park (at seventy-one percent) may be troubling, as at least one Justice has expressed a strongly contrary view. *See* Bostock v. Clayton County, Georgia, 140 S.Ct. 1731, 1822, 1825 (2020) (Kavanaugh, J., dissenting) ("A statutory ban on 'vehicles in the park' would literally encompass a baby stroller. But no good judge would interpret the statute that way because the word 'vehicle,' in its ordinary meaning, does not encompass baby strollers."). Perhaps Justice Kavanaugh is correct and GPT is wrong about the empirical fact of ordinary meaning here, but he does not actually cite any evidence for his intuition. In any event, we restate our view that empirical evidence is merely relevant and not determinative of the proper statutory interpretation of a good judge, given other factors appropriate for consideration.

If we focus on the contexts rather than the objects, we see a possibly disturbing pattern. In a criminal law context, objects are more likely to be classified as vehicles. With not many exceptions, this holds for driving under influence, and for a criminal enhancement if a vehicle has been instrumental in committing the crime. While we object to this result on policy grounds and the interpretive canon of lenity,[91] the result is less surprising if GPT is capturing the way Americans think about legislative context, given American views on criminal punishment.[92] On the other hand, if people in a census are asked to list all vehicles in the household, many objects are less likely to be classified as vehicles.

These remarks highlight only some of the results reported in Figure 7.[93]

[91] *See, e.g.*, United States v. Santos, 553 U.S. 507, 514 (2008) (Scalia, J.) (referring to the "venerable" rule of lenity, which "requires ambiguous criminal laws to be interpreted in favor of the defendants subjected to them."); David S. Romantz, *Reconstructing the Rule of Lenity*, 40 CARDOZO L. REV. 523 (2018) ("[L]enity is a rule of statutory construction that requires a court to resolve statutory ambiguity in favor of a criminal defendant, or to strictly construe the statute against the state."). Of course, in recent years, the Supreme Court has given this canon less weight. *See, e.g.,* Wooden v. United States, 595 U.S. 360, 377 (2022) (Kavanaugh, J., concurring) (noting that, because recent cases indicate a role for lenity only when the statute remains "grievously" ambiguous after considering all other methods of resolving ambiguity, "the rule of lenity therefore rarely if every plays a role.").

[92] *See, e.g.,* John Rappaport, *Some Doubts about 'Democratizing' Criminal Justice*, 87 U. CHI. L. REV. 711, 764-65 ("[W]hile public opinion is certainly less punitive today than it was three decades ago . . . it remains quite harsh."). Our point here is not that Americans always favor harsher punishment but only that the particular context of criminal law triggers greater rather than lesser public concern for the law having a broad scope.

[93] An odd result that shows the importance of caution is the upper left-most cell, where we essentially asked GPT whether a "vehicle" is a "vehicle," and consistently got less than a 100% positive reply (83% to 93% in Figure 7). We note, however, that we are following Tobia in asking this question and, surprisingly enough, he got a similar answer! *See* Figure 1, where only 93.15% of Tobia's respondents identified a "vehicle" as a "vehicle."

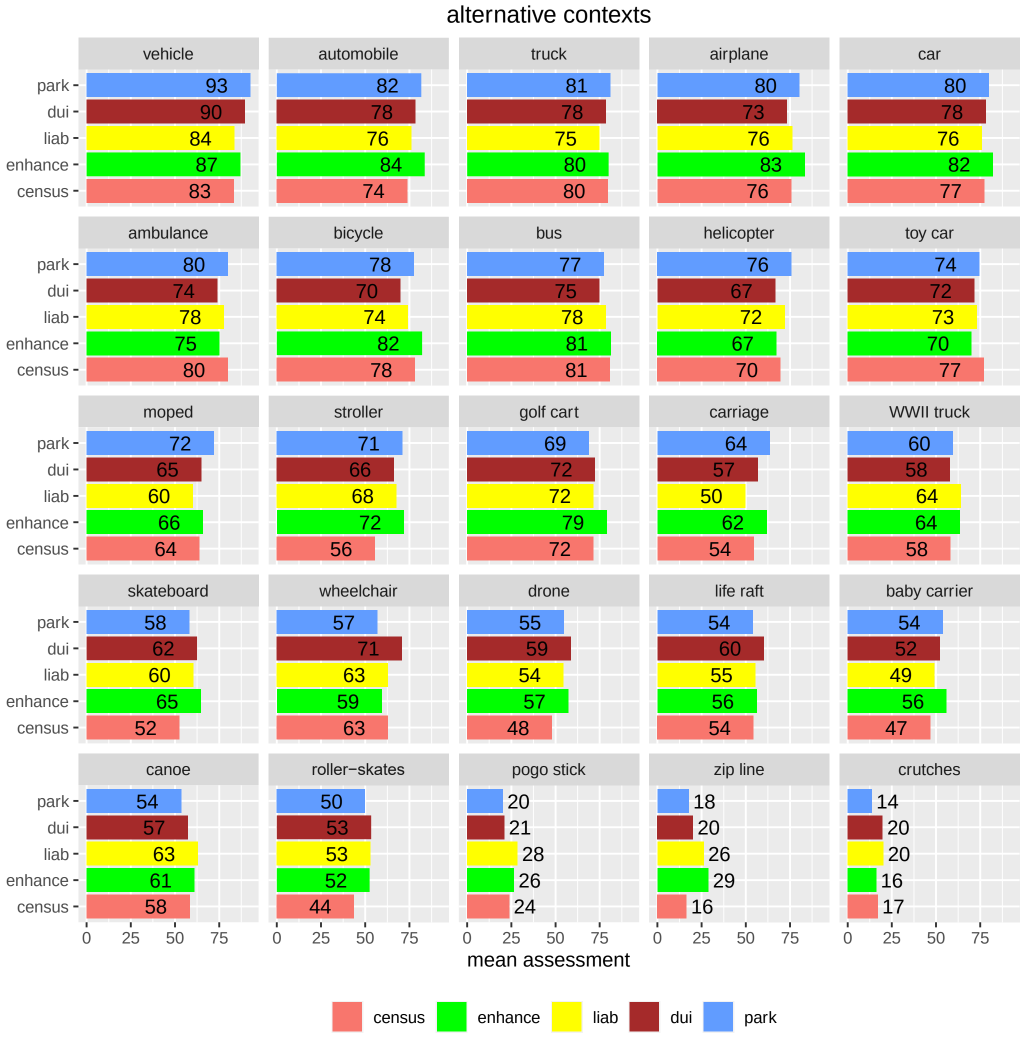

Figure 7
Effect of alternative contexts

C. Disclosing Alternative Purposes

Having varied the rules in which the word "vehicle" appeared, we returned to the original rule – no vehicles in the park – and varied its purpose. That is, we "informed" GPT of the reason for the rule. Obviously, this data is more of interest to non-textualists who primarily focus on the context of legislative purpose.[94] But even though textualists generally reject consideration of the subjective intent of the legislators who voted to enact a bill into law (and therefore reject legislative history evidence[95]), statutory purpose is by no means irrelevant to textualists. They are willing to contemplate a legislative purpose stated or implied in the text of the statute.[96]

[94] *See, e.g.*, Manning, *supra* note 85, at 87 ("[P]urposivism is characterized by the conviction that judges should interpret a statute that carries out its reasonably apparent purpose and fulfills its background justification.").

[95] *Id.* at 84 ("[T]extualists generally forgo reliance on legislative history as an authoritative source of [legislative] purpose").

[96] *See id.* ("Because speakers use language purposively, textualists recognize that the relevant context for a statutory text includes the mischiefs the authors were addressing. Thus, when a statute is ambiguous, textualists think it quite appropriate to resolve that ambiguity in light of the statute's apparent overall purpose."); Anita S. Krishnakumar, *Backdoor Purposivism*, 69 Duke L.J. 1275, 1299, 1305 (2020) (concluding, based on

For that reason, we wondered if GPT might provide a means of testing how statutory purpose influences the meaning of statutory terms. We investigated our ability to probe purpose with different prompts.

We first tried the workhorse prompt that had been reasonably successful with replicating Tobia's data,97 and that had worked well for investigating the effect of context.98 We added purpose to these tasks with the help of the assistant prompts reported below. Yet results were not convincing. GPT again was overinclusive. It had a strong tendency to classify very many objects to be very likely vehicles.

Results became much more plausible with an additional "chain of thought" element. For the investigation of purposivism, we eventually used the following system prompt:

> In this question, we do not ask you about your own assessment. Rather we want to learn your beliefs.
>
> 2,835 human subjects have participated in an experiment. They have been informed about a rule, and its official justification. The experiment has consisted of two stages. In the first stage, participants have been asked to list 5 objects to which the rule, given the justification, is meant to apply. In the second stage, the experimenter has mentioned one object. Participants are asked whether, to their judgement, the object comes under the rubric of the rule.
>
> We are asking you two questions:
>
> 1. which are the 5 objects that you consider most likely participants have listed?
>
> 2. How many participants do you think have responded that the object in question comes under the rubric of the rule, given the justification? [Answers limited to our seven-point Likert scale.99]
>
> We now show you the rule and the justification that participants have seen, and the object that they have been asked to classify.

We tested the following six alternative purposes for not allowing vehicles in the park:

> **annoyance**. The rule says: no vehicles in the park, since people using the park have been annoyed at the loud sounds and air pollution of vehicles in the park.

analysis of 965 opinions from 2005-2016, that "the purposivist Justices on the Roberts Court do *not* appear to have retreated from traditional purposive analysis" and that the textualist justices also "regularly" "traversed into guessing or asserting that Congress had X specific intent or Y specific purpose in mind when it enacted the statute."); Richard M. Re, *The New Holy Trinity*, 18 GREEN BAG 2D 407 (2015) (arguing that purposivism creeps in to textualist analysis because the determination of whether there is textual ambiguity includes consideration of purpose).

[97] *See supra* Part II.

[98] *See supra* Parts II.A and II.B.

[99] Just respond to the second question in one of these seven ways: (almost) none, very few, few, about half of them, many, very many, (almost) all.

accident. The rule says: no vehicles in the park, since there have been a number of accidents in the park involving collisions between inattentive pedestrians and cars or bicycles.

space. The rule says: no vehicles in the park, since some vehicles are taking up too much space, shrinking the space available for enjoying the park.

damage. The rule says: no vehicles in the park, since the grass, gardens, and some small structures in the park have been damaged by vehicles.

local. The rule says: no vehicles in the park, since people who live far away and who don't pay local taxes to support the park are driving to the park and making it crowded.

beauty. The rule says: no vehicles in the park, since vehicles are diminishing the beauty of the park.

We summarize our results in Figure 8. As with disclosing alternative rules, (Figure 7), the dominant effect of disclosing legislative purpose remains the character of the object. Irrespective of the declared purpose, an automobile is far more likely to be classified as a vehicle than a pogo stick. Yet when confronted with an explicit purpose, GPT makes slightly stronger differences within one and the same candidate object depending on the stated purpose of the rule.

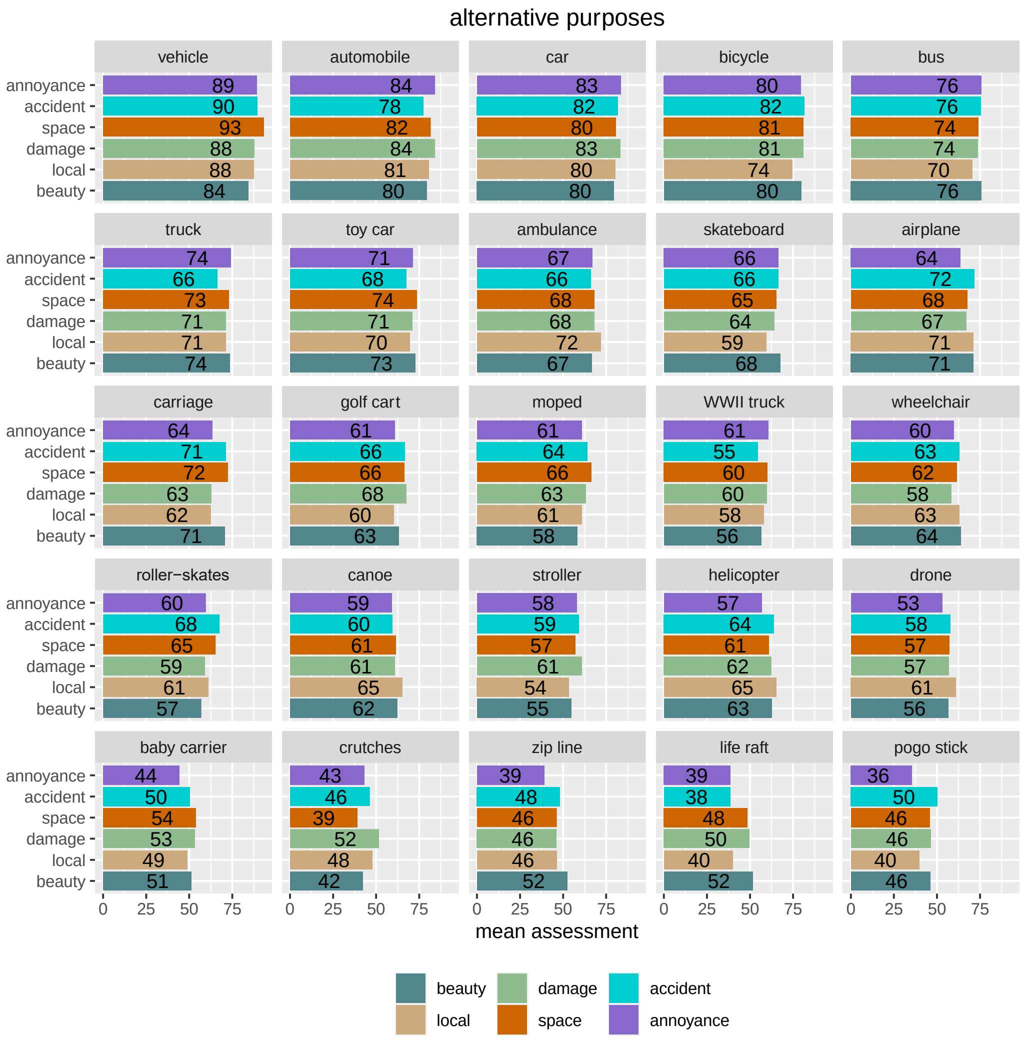

Figure 8

Effect of alternative purposes

Overall, purpose matters. If the stated reason for not admitting vehicles to the park is the annoyance of its visitors, objects are least likely to be classified as vehicles. If the stated purpose is protecting the local community, the mean probability of being rated as a vehicle is very similar. But for all other purposes, the probability is higher, most pronouncedly for a rule motivated with limited space.

Descriptively, GPT also makes meaningful differences within objects. Compared with the average rating assuming other purposes, GPT sees less reason to classify a carriage, a golf cart, roller-skates, a baby carrier, a zip line or a pogo stick as a vehicle if the purpose of the rule is said to protect visitors from being annoyed. On the other hand, GPT sees even more reason to classify the object as a vehicle if the purpose is preventing annoyance and the object is either an automobile or a toy car. If the stated purpose is preventing accidents, GPT sees more reason to include wheelchairs, roller-skates, crutches and pogo sticks in the definition. It interestingly sees less reason to include automobiles, trucks, World War II trucks and life rafts, given the intention to prevent accidents. This suggests that GPT is convinced about these objects coming with sufficient safety conveyances.

If the stated purpose is limited space, GPT considers it less necessary to include crutches, and more necessary to include carriages and life rafts. Likewise, if the prohibition has been introduced with the aim of preventing damage to the park, GPT feels less obliged to include crutches, and more obliged to include automobiles, carriages and mopeds. For a fair number of objects, GPT considers it less necessary to prohibit their access to the park if the rule is meant to protect the local community. Specifically, with this purpose, GPT is less likely to classify bicycles, buses, skateboards, life rafts and pogo sticks as vehicles. Finally, if the norm has been introduced to preserve the beauty of the park, GPT is even more concerned about skateboards, carriages, zip lines and life rafts, and it is less concerned about mopeds, roller-skates and crutches.

One may wonder how much trust to put on these results. Differences between objects are pronounced, but differences between purposes are small, and within each object, it often is hard to discern any effect of inducing alternative purposes. As this is standard in quantitative empirical analysis, one may want to use statistical conventions to assess differences. One may want to rely on the finding only if the p-value is below .05.[100] If one regresses the fraction of positive classifications (x is a vehicle) on the object ("vehicle" being the reference category) and purpose ("annoyance" being the reference category), all objects are significantly different from the reference category, as are all purposes except "local."

A straightforward test for individual objects is a chi square test that compares the number of positive classifications across the two purposes that one wants to compare. With the actual data, these tests are never significant at the (conventional) 5% level. Four comparisons are significant at the 10% level ("weakly significant"): pogo stick annoyance vs. accident (36 vs. 50, p = .063); life raft accident vs. beauty (38 vs. 52, p = .065); crutches space vs. damage (39 vs. 52, p = .088); zip line annoyance vs. beauty (39 vs. 52, p = .088). Yet by statistical standards, these results would not be credible, as they rely on multiple testing.

Yet these tests are questionable in the first place for an interesting reason. In the standard case that motivates statistical conventions, the researcher only has access to a limited sample, and wants to make sure this sample is not an atypical draw from the population. We have used one standard test (Kolmogorov Smirnov) for the comparison between Tobia's and our own data. That was appropriate since Tobia's data is limited. We have no chance to increase the number of his observations. Yet this is different for the present question, where we compare different conditions in data that all results from repeatedly asking GPT. The fact that we only use the 100 first complete responses is just a matter of convenience: the number of "yes" responses directly translates into the percentage. Yet at a very affordable cost, we could multiply the number of observations.

[100] Technically, the p-value measures the probability of wrongly concluding that the hypothesized effect is present in the population one wants to understand, given a thought experiment: one draws an infinite number of samples from the population of interest (with replacement), and registers, independently for each sample, whether the null hypothesis (stating that actually the hypothesized effect is not present) is rejected. If this probability is below 5%, one concludes that a false positive result (wrongly accepting the hypothesis) is sufficiently unlikely.

Were we to elicit 1,000 responses instead of 100, results are bound to be very similar: they reflect the degree of certainty GPT has, given its training data and our prompt. Had we ten times more observations, many more comparisons would become significant at conventional levels, for instance for automobiles the comparison between accident and space (78% vs. 82%, p = .029). Other comparisons would remain insignificant, e.g. for automobiles the comparison between local and beauty (81% vs. 82%, p = .604). But even this comparison would become significant if we were to elicit 20,000 responses per condition (p = .011). Were we to elicit 50,000 responses per condition, we would not only find a significant difference for any comparison (unless percentages are identical). We could even apply a (maximally conservative Bonferroni) correction for the fact that we compare six different purposes.[101] For instance, the comparison between local and beauty for automobiles, with the correction applied, would be significant at p = .035.

These examples show: *For assessing the relevance of findings from GPT (alone), significance is not a meaningful criterion.* As long as there is any difference between two conditions, small though it may be, one can always increase the number of requests to GPT until the difference is significant at conventional levels. The important category is what statisticians call the effect size: how big must the gap be to be meaningful? This is not a statistical but a legal question. For some legal problems, it may be possible, or even advisable, to ignore small effects. For other legal problems, even a tiny difference may be critical.

Applying this principle to the present investigation, it is worth noting that GPT makes fairly little difference between alternative motives for banning vehicles from the park. If the object has little resemblance with prototypical vehicles (like a pogo stick or crutches), GPT remains hesitant to bring the object under the rubric of the rule, even if the object might have effects similar to the ones that have motivated the prohibition. And if the object obviously falls under the ordinary understanding of the term, GPT has little inclination to exclude it from the prohibition, even if the stated concern seems rather far-fetched. GPT is, in other words, not very inclined towards reasoning by analogy or by teleological reduction. At least with the rule that we have tested (no vehicle in the park), revealing the intended purpose of the prohibition makes little difference for GPT, and by implication for the way members of the general public are likely to interpret the rule.

D. Using GPT to Explore Historic Meaning: Extensional vs. Intensional

A common way of interpreting statutes is to focus on their meaning at the time of enactment. That is obviously the approach of textualism, which asks for the original public meaning of the statute.[102] But even someone who focuses on the legislative history of the law is emphasizing the meaning at the time of enactment. Even if one embraces dynamic statutory interpretation, in which the meaning of the statute can evolve over time (like the common law),[103] it is usually

[101] And hence would have to multiply calculated p-values by 6! = 720.

[102] *See, e.g.*, SCALIA & GARNER, *supra* note 69, at 41, 83; Victoria F. Nourse, *Textualism 3.0: Statutory Interpretation after Justice Scalia*, 70 ALA. L. REV. 667, 676-80 (2019).

[103] *See* WILLIAM N. ESKRIDGE, JR., DYNAMIC STATUTORY INTERPRETATION (1994).

relevant to ask what the statute meant when first enacted.[104] We therefore thought it useful to explore GPT's ability to identify meaning at a particular time. In the examples below, we prompt GPT to focus on the decade of the 1950s.

At the same time, we combined this exploration with another. We wanted to investigate the fundamental difference in extensional and intensional meaning. Extension meaning refers to "the collection of things that fall within the scope of a term."[105] The extensional meaning of "mammal" would be a list of animals that qualify as mammals. The extension of "planets" is a list of known objects to which the term applies. By contrast, intensional meaning refers to the characteristics or attributes of the term, possibly a set of necessary and sufficient conditions; in short, a definition.[106] The intensional meaning of "mammal" might be any "vertebrate animals in which the young are" (or could be) "nourished with milk from the mammary glands of the mother."[107] The intension of "planet" in our solar system might be "a celestial body" that orbits a star and possesses sufficient mass "to have enough gravity to force it into a spherical shape" and to have "cleared away any other objects of a similar size near its orbit."[108]

One might propose to define statutory terms in either way. Advocates of corpus linguistics are implicitly favoring extension because they look in the corpora for examples of sentences using the term. When Thomas Lee and Stephen Mouritsen used corpus linguistics to ask whether bicycles or airplanes are "vehicles," they looked for sentences in which the term "vehicle" referred to a bicycle or airplane.[109] That would be like looking for sentences that refer to a bat as a mammal or Neptune as a planet. In any case, it is demonstrating that the extension of the larger category includes the specific item listed. What this typical use of corpus linguistics does *not* do is intension. There is no effort to create a definition of "vehicle" from which one could decide what objects belong in the category.

Seeking meaning through intension has certain advantages over extension. As William Eskridge, Brian Slocum, and Stefan Gries explain,

> [I]n 1920 the extension of *airplane* did not include any jets, but its extension in 2021 does. In contrast, even though its extension will change constantly over short periods of time, the intensional meaning of *airplane* might, theoretically, remain stable for long stretches of time.[110]

[104] When Judge Posner disregarded the meaning of Title VII in his concurrence, opting for an updated or "fresh" interpretation, he conceded that "[t]he first and most conventional" approach "is the extraction of the original meaning of the statute," which was the analysis of the judges in the majority of the en banc decision. See Hiveley v. Ivy Tech Community College of Indiana, 853 F.3d 339, 352 (2017) (Posner, J., concurring). We do not mean that this conventional approach is correct. For the best analysis of the different options, *see* Eskridge, Slocum & Gries, *supra* note 23, at 1507-09 (distinguishing societal dynamism, linguistic dynamism, and normative dynamism).

[105] *See* Eskridge, Slocum, & Gries, *supra* note 23, at 1526.

[106] *Id.*

[107] *See* David M. Armstrong, Don E. Wilson, & J. Knox Jones, *Mammal*, in BRITANNICA, last updated Dec. 4, 2023, at https://www.britannica.com/animal/mammal.

[108] This is NASA's gloss on the definition of the International Astronomical Union that controversially excludes Pluto, which is now a dwarf planet. See NASA, "The Definition of a Planet," at https://science.nasa.gov/solar-system/planets/what-is-a-planet/.

[109] *See* Lee & Mouritsen, *supra* note 8.

[110] *See* Eskridge, Slocum, & Gries, *supra* note 23, at 1527.

To take a different example, suppose that in the future we discover a celestial body – call it Astro –that is larger than the dwarf planet Pluto but smaller than the planet Mercury. How do we decide whether Astro is a *planet*? Using corpus linguistics would not be helpful because there are no prior references to the celestial body now known as Astro. But the intensional approach provides criteria that can be applied to new objects. The extension of planet can appropriately expand while the intension remains the same.

Here we explore how GPT operates when asked to determine meaning extensionally and intensionally, both as applied to a specific historical period, a statute or ordinance enacted in the 1950s. We compare the following two system prompts, each with one assistant prompt. The prompts are worth quoting in full:

The Extension Prompt:

> We are interested in the way a term (that we are going to show you below) has been interpreted in the 1950s. To make the question tractable, we propose a thought experiment. Assume that, in that era, 2,835 participants have been asked how they interpret the term.
>
> What follows is the question that these hypothetical participants would have been asked, not the question we are asking you. From you we want to learn how many participants you believe would have responded "Yes." Please note that this is not a knowledge question. Rather we want to learn your belief. Given your world knowledge, which is your best guess?[111]

The Intension Prompt:

> We are interested in the way a term (that we are going to show you below) has been interpreted in the 1950s. To make the question tractable, we propose a thought experiment. Assume that, today, 2,835 participants have been asked how they think the term would have been interpreted in the 1950s. They are admonished not to straight out jump at the response. Rather they are reminded that the meaning of words may change over time. To address the challenge, these hypothetical participants are asked to proceed in two steps. In the first step, they are asked to reflect upon the general scope of the term in the 1950s. In the second step, they are asked whether, given their belief about the general understanding of the term in the 1950s, a specific object would have been brought under the rubric of the term. If the object in question had not existed in the 1950s, they are invited to proceed by analogy.

[111] We followed this prompt with our standard language: "Please do not repeat the task, or the question that human participants have been asked. Do also not give explanations. Just respond in one of these seven ways: [the same Likert scale as used above]. This is the question that hypothetical participants would have been asked: [whether a particular object is a vehicle]." Next, we gave our standard *Assistant prompt*: "You are asking back: Why do you want to know? I am answering: There is a rule that says: no vehicles in the park."

> What follows is the question that these hypothetical participants would have been asked, not the question we are asking you. From you we want to learn how many participants you believe would have responded "Yes" in the second step of the question they have been asked. Please note that this is not a knowledge question. Rather we want to learn your belief. Given your world knowledge, which is your best guess?[112]

As Figure 9 shows, GPT does indeed make a difference between the assessment today (Figure 6) and an attempt at reconstructing the assessment seventy years ago.[113] Overall, in GPT's opinion, objects are more likely to be classified as vehicles today than they would have been classified in the past. More importantly, GPT makes a difference between an intensional and an extensional approach to historical meaning. If GPT reasons from an abstract definition to the application in question (i.e. if it adopts an intensional approach), it is more likely to classify a World War II truck, a carriage, a canoe and a bicycle to be vehicles. These findings suggest that, with the intensional prompt, GPT puts more stress on the question whether the object has already existed in the 1950s. By contrast with the extensional prompt, GPT is more likely to classify a golf cart, a helicopter or a moped as vehicles. Arguably, GPT thinks that these objects are similar enough to objects that were prototypical for vehicles in the 1950s.

.

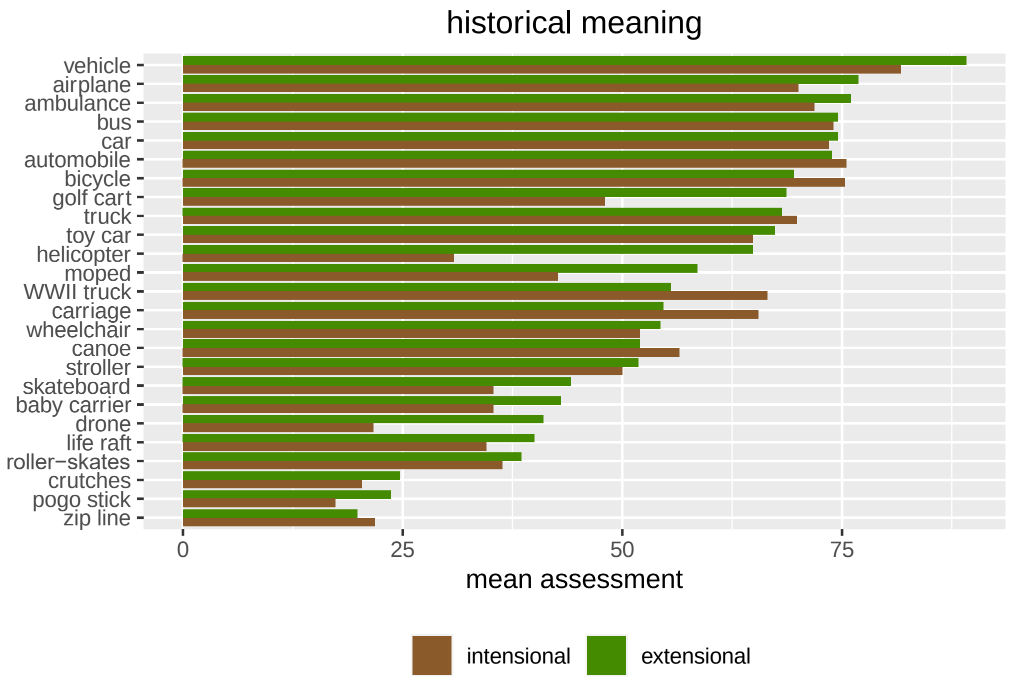

Figure 9

Historical, Intensional & Extensional Meaning

[112] We followed this prompt by the standard language in the prior footnote.

[113] As we have added the assistant prompt "There is a rule that says: no vehicles in the park," the appropriate comparison is data from using this assistant prompt but asking for the beliefs of human participants who have been tested contemporaneously.

IV. GPT And Ordinary Meaning: Some Lessons Learned

As a field of academic study, law became noticeably more empirical a decade or two ago.[114] Today, empirical investigations are no longer confined to specialized areas, like antitrust or patent,[115] but include commercial law,[116] consumer protection,[117] education,[118] criminal law,[119] and comparative law.[120] Yet in the sense of the distinction proposed by H.L.A. Hart, the older empirical investigations predominantly adopted an "external" view to legal issues.[121] The newer trend is to apply empiricism to the "internal" perspective on law, studying statutes and cases with the help of statistical analysis.[122]

Recent articles on statutory interpretation have recognized that law needs empirical ways to test assumptions about ordinary meaning and explored the options of corpus linguistics, experimental surveys, and cosine similarity.[123] With our testing of GPT, we add LLMs to that set of empirical tools. Precisely because LLMs are built on language, and results are formulated in natural language, questions come within the reach of rigorous empirical analysis that would previously have been difficult, if not impossible, to analyze in quantitative terms.

Our effort is merely a first, necessarily exploratory one, but LLMs have enormous potential for revealing ordinary meaning in statutes, as well as some significant potential for unrigorous and poorly motivated prompting that obscures rather than illuminates. The need to understand LLMs as interpretive engines – their strengths and weaknesses – is pressing. No matter what the legal academy says in academic articles on the subject, we can expect LLMs to show up in actual lawyering on statutory issues, as it already has in other respects. The relative ease of GPT makes its use inevitable.

Consider the attractiveness of GPT compared to the empirical alternatives. Experimental surveys are methodologically powerful but expensive and time-consuming. Cosine similarities may require more mathematical comprehension than the typical lawyer and judge possess.

[114] *See supra* note 22.

[115] *See e.g.* Christoph Engel, *Tacit Collusion. The Neglected Experimental Evidence,* 12 J. EMP. LEGAL STUD. 537 (2015); Philippe Aghion, Stefan Bechtold, Lea Cassar & Holger Herz, *The Causal Effects of Competition on Innovation: Experimental Evidence*, 34 J. LAW, ECON. & ORG. 162 (2018).

[116] *See, e.g.*, Eric Talley & Sarath Sanga, *Don't Go Chasing Waterfalls: Fiduciary Duties in Venture Capital Backed Startups*, 52 J. LEGAL STUDIES (forthcoming 2024).

[117] *See, e.g.*, Yannis Bakos, Florencia Marotta-Wurgler & David R Trossen, *Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts*, 43 J. LEGAL STUDIES 1 (2014).

[118] *See, e.g.,* Claudia Cerrone, Yoan Hermstrüwer, Onur Kesten*, School Choice with Consent: An Experiment,* 134 ECON. J. *(forthcoming 2024).*

[119] Tim Friehe, Pascal Langenbach & Murat C. Mungan, *Does the Severity of Sanctions Influence Learning about Enforcement Policy? Experimental Evidence,* 52 J. LEGAL STUD. 83 (2023).

[120] *See, e.g.*, Yun-chien Chang, Nuno Garoupa, Martin T. Wells, *Drawing the Legal Family Tree: An Empirical Comparative Study of 170 Dimensions of Property Law in 129 Jurisdictions*, 13 J. LEGAL ANALYSIS 231 (2021).

[121] HART, supra note 50, at 91.

[122] *See, e.g.*, Curtis Bradley & Eric Posner, *The Real Political Question Doctrine*, 75 STAN. L. REV. 1031 (2023) (reporting on empirical collection of hundreds of 1,200 lower court cases engaging the political question doctrine); Eric Ruben & Joseph Blocher, *From Theory to Doctrine: An Empirical Analysis of the Right to Keep and Bear Arms After Heller*, 67 DUKE L.J. 1433 (2018) (reporting and discussing an analysis of every second amendment opinion -- over 1000 -- from 2008 to 2016). *See generally* William Baude, Adam S. Chilton & Anup Malani, *Making Doctrinal Work More Rigorous: Lessons from Systematic Reviews* 84 U. CHI. L. REV. 37, 57 (2017) (advocating the development of "methodological standards for analyzing case law").

[123] *See, e.g., supra* text accompanying notes __-__ (discussing prior empirical methods of determining ordinary meaning).

GPT is not only more accessible than those two options but is easier to use than the third, corpus linguistics. Or, what is effectively the same, it will *appear* to lawyers to be easier to use well.

Corpus linguistics allows one to seek data about (1) the frequency of a word's appearance in English texts, (2) the collocation of two words (their tendency to be used in the same sentence or passage), and (3) a word in "KWIK"[124] context (seeing many examples of a snippet of text before and after the word of interest).[125] But the lawyer or judge must make a host of decisions before reaching any results, beginning with which one or combination of these three search tools to use. For collocation, one must decide what two words to search for. For example, if one is checking on whether a statutory term (like "vehicle") can refer to a term describing a pivotal object in the facts of the case (like "bicycle"), one must consider whether to seek collocation of some close or exact synonyms of either or both words (like "conveyance" or "transport" for "vehicle," or "bike" or "pedal cycle" for "bicycle").[126] For "KWIK" searches, one must decide how much context to seek and then one has the task of reading through the many examples to see what insights they generate.

To use GPT *well*, one must *also* take time and exercise care, as we have shown. We do not mean to imply otherwise. One message of our paper should be: one cannot take all responses of LLMs at face value. Before GPT evidence can be introduced into prove ordinary meaning, one must understand how LLMs work, and ideally the prompts one uses for asking should have been tested against a benchmark from human subjects.

Yet our caveats will not stop lawyers from using GPT, which will likely prove more popular than corpus linguistics. Where most people and most lawyers have no need to consult corpora for other aspects of their lives, people are learning to use GPT as an all-purpose assistant for a wide array of tasks, which is why many Americans have used GPT and some have downloaded the app onto their phone.[127] If they have not already done so, there can be little doubt then that lawyers who think of GPT as useful for making restaurant recommendations or summarizing cases will soon be citing GPT results in their briefs, and judges may follow suit in their opinions. It is important, therefore, for lawyers and judges to develop a sound methodology for consulting GPT on statutory interpretation, a project we have now begun.

We now summarize and explain what we think are the important lessons we learned from our empirical testing of GPT:

- *First, we should not consider GPT evidence of ordinary meaning unless the prompting method has been separately tested against some reliable benchmark.*

[124] "Key word in context".

[125] *See* Lee & Mouritsen, *supra* note 8, at 831-32.

[126] *Id.* at 847, 875 (discussing the need to check for synonyms).

[127] *See* Jon Porter, *ChatGPT Continues to be One of the Fastest-Growing Services Ever*, THE VERGE (6 Nov. 2023) (noting that, worldwide, "[o]ne hundred million people are using ChatGPT on a weekly basis"); Alyssa Stringer, Kyle Wiggers & Cody Corral, *ChatGPT: Everything You Need to Know about the AI-Powered Chatbot*, TECHCRUNCH+ (30 Jan. 2024) (as of October 2023, OpenAI "amassed 15.6 million downloads"), at https://techcrunch.com/2024/1/30/chatgpt-everything-to-know-about-the-ai-chatbot/.

- *Second, to capture the plausible meanings of a term, rather than just the most common, one must query GPT multiple times.*

- *Third, our most successful method – combining a belief prompt with a Likert scale – is "good enough" for now to justify some confidence in its use, but it requires more testing.*

- *Fourth, we advocate testing of alternative prompts we have not considered; something else could easily prove to be better than our best.*

- *Fifth, pending much more testing, the best use of GPT is in combination with other empirical evidence of meaning.*

We offer some support for these propositions. *First*, from a legal policy perspective, benchmarking is of the utmost importance. We identified and exploited one possible benchmark: Tobia's experimental survey results on the meaning of "vehicle." His data on American users of English enabled us with the precious opportunity to compare the evidence GPT generates with quasi "ground truth." Most of the GPT data we generated deviated substantially from the human data, which casts strong doubt on the reliability of those prompting techniques, at least until other evidence says otherwise. When we simply asked GPT the question "Is the following a vehicle: [object name]," the results we received were significantly different from Tobia's results. The same was true when we used the common "chain of thought" inquiry technique, or an ordinary belief prompt (without limiting the form of the answer to a Likert scale). These were perfectly plausible approaches, but the benchmarking rejects them, which is a strong caution for relying on intuition alone to settle upon a prompting strategy.

Second, for practical assistance in statutory interpretation, one needs to prompt GPT repeatedly to generate a distribution of results. The temptation, of course, is to simply *ask GPT once for the answer* to the interpretive question a statute poses. Hence, a legal practitioner might simply open ChatGPT and ask: *Is a bicycle a vehicle?* There is an alternate method in which this might be a useful beginning.[128] This "single question" approach might be useful if combined with alternative single prompts, along the lines that we have tested in this article: does the response change if one asks for a belief, rather than GPT's own assessment? Does it change if one adds context, in particular the wording of the rule that uses the contested term? Does it change if one adds the moment in time when the rule has been enacted? Does it change if one adds the agreed upon purpose of the rule, or contested definitions of this purpose for that matter? If there is no change, then one has mustered some plausible evidence of meaning.

Yet the "single question" approach has a serious drawback: the absence of a distribution of replies does not allow the researcher to compare the strength of alternative meanings. Any single inquiry reveals nothing about the likely prevalence of a given meaning, but merely GTP's most preferred meaning. Put differently, one does not learn how confident the LLM is in the

[128] If one seeks a single response, one should use the overall most accurate LLM, which at the time of this writing is GPT-4. Unlike our approach, one should *not* set temperature to a high value; that would increase the probability of receiving a minority response. Rather one should set it to zero, and then get GPT's best guess.

given response to a single inquiry. When there may be two plausible responses, we would usually want to know if GPT judges its preferred reply as most likely by a bare majority of 51% or by a near certainty of 99%. In the former case, GPT may assess the "second-best" meaning as 49% likely, an impressively plausible alternative. Remember that our argument is that the GPT data is *relevant*, but given all the other context that is relevant, no one should think that GPT's favored meaning should by itself control and especially not when GPT assesses the competing meaning as being quite strong.

As we said previously,[129] an even stronger case might be where there are three plausible responses, one with probability 35%, the next with probability 33%, and the third with probability 32%. Then the most likely response is still minoritarian. Then GPT's favored answer is not even the most likely meaning. Hence for the contested cases for which empirical evidence may be critical, it is important to generate a complete distribution of responses, rather than the single most likely response.

Third, our belief prompt combined with a Likert scale was reasonably successful, generating results visually similar to and statistically indistinguishable from Tobia's benchmark. This is an important step towards verification. When the match between the results from human participants and from GPT is "good enough," there is room for a radical change in interpretive practice. For GPT does not only generate rigorous evidence but does so at vastly less expense. For financial and for practical reasons, it is not possible to "scale up" Tobia's experimental survey method, but it is possible with GPT. Effectively, GPT could democratize data generation to an unprecedented degree.

Ultimately, the introduction of data generated with the help of LLMs as evidence for ordinary meaning will depend on the perceived benefit. To the prior discussion, we wish to elaborate on two such benefits. First, with the help of LLMs, it may be possible to narrow down the contested domain of complex and expensive legal conflict to the truly critical elements. Take our main example: The rule says: no vehicles in the park. If one asks GPT for its belief about the assessment by experimental participants, and using a seven-point Likert scale, a drone quite clearly qualifies (75.67% yes). If one adds the content of the rule (no vehicles in the park), GPT becomes undecided (54.67% yes). If one focuses on an historic meaning and asks for ordinary meaning at the point in time when the rule has been enacted, instructing GPT to first define five clear applications, GPT becomes skeptical (41% yes). Finally, if one instructs GPT to first develop a definition, from the perspective of the time of enactment, GPT is clearly negative (21.67%). Hence in this contested case, GPT evidence quickly and inexpensively shows what the dispute is actually about: which is the appropriate method for interpreting the rule? Likely the conflict would focus on the relevance of contemporary versus historical ordinary meaning.

This exercise would never be definitive. If one of the parties is not happy with the provisional delineation of the area of conflict, it is up to her to broaden the area. But the more the interpretation suggested by GPT, and possibly probed with a series of alternative prompts, seems unequivocal, the more the burden of argumentation would shift to such a contending party.

[129] *See supra* text accompanying note 58.

The second likely benefit is easiest to explain with an analogy to an established practice in computer science. It originates in the architecture of the most advanced algorithms. These days, most of them are neural networks.130 Neural networks have multiple layers, and often also allow for bidirectionality. Due to both features, output effectively "emerges." It is next to impossible to predict the outcome ex ante, given the architecture, the training data, and the current input. And it is equally difficult to mechanically explain why a certain outcome has emerged. This concern has led to an entire sub-branch of computer science, explainable AI.131 One approach is in the spirit of experimentation: one changes a single element of the input, using a list of alternative inputs, and explores which alternative input would have changed the output. This approach is commonly called reasoning by counterfactual.132

We can again use our running example to illustrate the usefulness of this approach. Let us once more assume that the rule is "no vehicle in the park," and that the contested item is a drone. The defendant argues that drones are a novel, unobtrusive pastime, and should therefore be allowed in. The other party objects that the ordinary meaning of a vehicle includes a drone. If the court considers the methodology appropriate that we have used in our attempt to replicate the Tobia data (Figure 5), it could object: GPT considers drones to be vehicles with 75.67% probability. To be accepted in the park, the object would have to be as different from the ordinary meaning of a vehicle as a pogo stick (15.5%) or a zip line (14.3%).

Fourth, we still do not claim that our single success is sufficient by itself to fully validate even our successful technique (belief prompt with Likert scale). Even this best performing prompt did not yield a perfect match with Tobia's data, though the distributions of responses were no longer statistically distinct. We feel confident to recommend this prompt for tentative use. But before GPT evidence makes it into judicial opinions on statutory interpretation, we recommend that many more exercises along the lines of ours are undertaken. The engineering of LLM prompts is not yet a science, but an art. While a large community, in computer science and beyond, engages in finding more powerful, and more reliable prompts, the debate over prompting is far from closed. We consider it rather likely that, with still different prompts, one could generate discernibly different outcomes, some more consistent with human data than our best.

In particular, there are things we did not test. We did not test ChatGPT 4.0 (for reasons explained), "temperature" settings other than one (for reasons explained), nor repeating the prompt more than 100 times. We did not test more than one benchmark. In particular, our benchmark involved a noun (vehicle), where the question is whether the category the noun defines includes other nouns (e.g., bicycle). We leave for future testing the ability of GPT to generate useful data on other problems of statutory interpretation.

[130] *See supra* notes 30-31.

[131] For an overview of the most prominent approaches, *see* WOJCIECH SAMEK, GRÉGOIRE MONTAVON, ANDREA VEDALDI, LARS KAI HANSEN & KLAUS-ROBERT MÜLLER, EXPLAINABLE AI: INTERPRETING, EXPLAINING AND VISUALIZING DEEP LEARNING (2019).

[132] *See* Ilia Stepin, Jose M Alonso, Alejandro Catala & Martín Pereira-Fariña *A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence*, 9 IEEE Access 11974-12001 (2021); Sahil Verma, *et al., Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review* (2020), arXiv preprint arXiv:2010.10596.

We also sought to test only ordinary meaning, which is the most relevant but not the exclusive way to proceed. Although textualists seem to favor ordinary meaning in every case, there may be contexts where a technical or specialized meaning would be appropriate. Perhaps a statute regulating lawyers, doctors, or hedge fund managers should be interpreted according to the meaning of terms within the regulated industry. Technically, it would be possible to instruct GPT to respond, assuming the role of a trained lawyer, doctor, or hedge fund manager. But before one could trust the results, one would have to carefully investigate how good GPT is at producing such evidence, comparing it to a benchmark of data from the relevant human group. As the language model has not been specifically trained on legal text, or on text from any other group of professional experts for that matter, there is an additional reason for being cautious.

Fifth, although obvious, we observe the important difference between using GPT as one source of empirical data on meaning and using it as the only source. Others have written on the usefulness of combining different empirical approaches together to “triangulate” meaning.[133] Where that suggestion involved the combination of traditional tools (e.g., dictionaries and linguistic canons), corpus linguistics, and experimental surveys, we add that GPT should be considered alongside this mix. Its accessibility may tempt some efforts to use it by itself, but that is a poor idea until there has been much more testing and verification.

Conclusion

Within little more than a year, and despite the persistence of obvious limitations (like hallucinations), LLMs have infiltrated a rich array of social practice. They have already profoundly changed the way how most people search for information. For many purposes, even the production of written text, oral output, and visual stimuli has been entrusted to language models. Language models are here to stay.

Should the responses that LLMs provide to the prompts about the meaning of statutory terms be accepted as empirical evidence of the terms’ ordinary meaning? In this article, we have given a cautiously optimistic response regarding their probative value, or their accuracy, to use the standard term in computer science. Provided that results from test runs come sufficiently close to human responses used as benchmarks, and provided that sufficient care is taken with repetitions, prompting and the representation of the data, results might indeed serve as an easily accessible window into the way how a contested term is interpreted in the wider population. LLMs may provide empirical evidence of ordinary meaning with unparalleled ease. We have explained why legal actors have good reason to be cautious. But language models have the potential to radically facilitate and improve legal tasks, including the interpretation of statutes.

[133] *See* Tobia, Egbert, & Lee, *supra* note 4.