Against AI Jurisprudence: Large Language Models and the False Promises of Empirical Judging

Dasha Pruss[*1], **Jessie Allen**[*2]
[1]University of Illinois Chicago
[2]University of Pittsburgh School of Law
dpruss@uic.edu, jallen@pitt.edu

Abstract

As hype around the transformative effects of large language models (LLMs) has taken center stage in popular culture, some judges and legal scholars have suggested that LLMs have the potential to improve the objectivity of judicial decision-making. Proponents argue that using LLMs to find empirical 'evidence' of legal text's meaning can reduce the role of judges' subjective choices, ensuring that judicial rulings faithfully reflect the people's understanding of legal rules, and grounding legal interpretation in a sophisticated empirical investigation of real language use in social context. To the contrary, we argue that LLM jurisprudence underscores the discretionary decisions required to infer ordinary meaning; highlights the inescapable reality that the meaning and application of legal terms is inherently normative; and demonstrates the lack of democratic legitimacy of crowdsourcing legal meaning. We argue that the feature of LLMs that makes them so seductive for legal interpretation – their potential ability to approximate 'ordinary' people's understanding of legal text – reveals the political illegitimacy of empirical judging. We conclude with recommendations and warnings for practitioners in this space.

1 Introduction

In 2024, US Circuit Judge Kevin Newsom made headlines in a self-proclaimed "unusual" concurrence in *Snell v. United Specialty Insurance Company*, a case that hinged on whether James Snell's installation of an in-ground trampoline was covered under the term 'landscaping' in the policy he had purchased from United Specialty Insurance Company (Snell v. United Specialty Insurance Company 2024).

> Here's the proposal, which I suspect many will reflexively condemn as heresy, but which I promise to unpack if given the chance: Those, like me, who believe that 'ordinary meaning' is the foundational rule for the evaluation of legal texts should consider—*consider*—whether and how AI-powered large language models like OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude might—*might*—inform the interpretive analysis. There, having thought the unthinkable, I've said the unsayable.

Newsom goes on to list the advantages for judges of using large language models (LLMs), next-word prediction machines trained on a vast database of language samples, to interpret legal text. Among the strengths of LLMs, he lists their training on "ordinary-language" input ("from so many different sources—e.g., professional webpages, DIY sites, news stories, advertisements, government records, blog posts, and general online chatter"), their ability to "understand" the context of words and "detect language patterns at a granular level," and their accessibility (Snell v. United Specialty Insurance Company 2024).

As hype around the transformative effects of LLMs has taken center stage in popular culture, few areas of intellectual inquiry have remained insulated from hyperbolic claims about AI's ability to solve previously intractable human problems (Angwin 2024; Edwards 2025; Grant and Metz 2022; Markelius et al. 2024). Legal practice is no exception. It is unsurprising that LLMs would be used by lawyers to analyze relevant legal precedents and by laypeople to access free legal advice (Cheong et al. 2024), and it is equally unsurprising that legal briefs have sometimes been infected with 'hallucinated' cases and rulings (Merken 2025; Neumeister 2023). However, the idea that LLMs could be used by judges to interpret authoritative legal text – and thus effectively decide what the law requires or forbids in a given situation – raises new conceptual and political issues.

Judge Newsom is part of a growing chorus of voices advocating the use of LLMs to interpret legal language in contracts, constitutional provisions, and statutes (written laws enacted by legislatures) (Choi 2024; Engel and McAdams 2024; Hoffman and Arbel 2024; Raymond 2024a). Proponents argue that LLMs can improve the objectivity and legitimacy of judicial rulings by introducing a data-driven approach to determining legal meaning. The idea is that querying LLMs about the meaning of key legal texts can replace a judge's intuition with empirical investigation, and, in the process, eliminate or at least minimize the role of subjective bias.

In some ways, using LLMs to determine legal meaning looks like a logical outgrowth of an increasingly dominant style of legal interpretation known as 'textualism'. While there are many methodological and substantive variations associated with different styles of textualism, the core idea is that legal words should be understood and applied accord-

[*]These authors contributed equally.

ing to their "ordinary meaning," that is, "what they conveyed to reasonable people at the time they were written" (Scalia and Garner 2012). But ordinary meaning is not so easy to identify. Along with their own intuitive understanding of language, judges often rely on dictionary definitions to decide whether a legal term applies to a situation in a legal claim. More recently, judges have conducted word searches in online libraries and used corpus linguistics methods to determine the ordinary meaning of terms. However, critics argue that subjective methodological choices enable textualist judges to engage in politically motivated reasoning, threatening the separation between legislators who are tasked with making law and judges who are tasked with interpreting it.

Take a classic example: suppose your local park has a sign reading "no vehicles in the park." Clearly, cars are prohibited from entering the park. But what about bicycles, ambulances, and baby strollers? For a textualist to interpret this rule, they would need to know the ordinary meaning of "vehicles." Relying on a dictionary definition, some textualists might well insist that a baby stroller is prohibited because it satisfies the definition of a "vehicle" as a "means of carrying or transporting something" (per Websters Dictionary). But others would say that the meaning of language is always contextual, and these approaches fail to take into consideration the context of the term. They might argue from a 'common sense' understanding of their native language and their knowledge of usual park activities that few people would use the word "vehicle" to describe baby strollers in this context. Others would advocate considering how ordinary readers likely would understand what harms the law aims to prevent, making it arguably even less likely that strollers are prohibited "vehicles." If only there was a method that could somehow combine all these approaches and predict what an ordinary person actually would be most likely to think.

Enter AI. LLMs, trained on a vast database of diverse language usage, far beyond what is available to even the most ambitious corpus linguist, appear at first blush to answer some of the earlier critiques leveled against textualist interpretation. Proponents say that querying LLMs could produce a methodologically superior way of accurately capturing what the 'typical' person would think about the meaning of a term like 'landscaping' or 'vehicle', akin to recent proposals to use LLMs to generate synthetic approximations, or 'silicon samples', of public opinion and behavior, without actually querying any real people (Aher, Arriaga, and Kalai 2023; Argyle et al. 2023; Lee et al. 2024; Park et al. 2023). Proponents claim that using LLMs to interpret legal text reduces judges' individual discretion and so produces less biased and more democratic results that more closely approximate the objective rule of law ideal and people's understanding of legal rules.

We argue, however, that enthusiasm for AI jurisprudence underestimates the importance of the widely recognized problem that LLMs incorporate and manifest human subjective biases. More significantly, we contend that appealing to LLMs for the meaning of legal text amplifies the false claim that legal interpretation can be reduced to empirical questions. Moreover, even if it were possible to use LLMs to entirely crowdsource legal meaning, the results would not increase law's democratic legitimacy. To the contrary, we argue that the feature of LLMs that makes them so seductive for legal interpretation – their potential ability to approximate 'ordinary' people's understanding of legal text – reveals the political illegitimacy of empirical judging. If LLMs improve the legitimacy of judicial interpretation, it is by illuminating the impossibility of curing the value-laden subjectivity and human fallibility of law. We conclude by discussing five recommendations and warnings regarding this conceptually fraught project for practitioners in this space.

2 Background and Related Work

2.1 Recent Trends in Empirical Legal Interpretation

Judges have always faced the problem of interpreting legal terms because they must apply general laws to particular cases (Blackstone 1765). The basic idea that legal text should be read according to its typical meaning is likewise nothing new.[1] Textualism as a distinctive interpretive mode arose in the 1990s in opposition to the traditional, more pragmatic multifaceted approach to legal interpretation. By setting out to discover 'ordinary meaning', textualism claims to transform legal interpretation from judges' freewheeling projection of their own values to a rigorous investigation, focused on objective social and historical facts. The conservative US Supreme Court Justice Antonin Scalia was textualism's most prominent theoretician and practitioner, and it remains primarily associated with ideologically conservative judges. But over the past several decades, textualist reasoning has come to be seen as the preeminent style of statutory interpretation in US courts (Tobia 2020). Few US judges today would interpret legal text without at least gesturing to a textualist approach.

Even avowed textualists do not necessarily agree on every aspect of the interpretive process, let alone on its outcomes, but they do share some basic claims and methods. Textualists aim to interpret "a statute in accord with the ordinary public meaning of its terms at the time of its enactment" (Bostock v. Clayton County 2020).[2] Of course, words must be understood in context (Barrett 2017). The object is to read and apply the statute "the same way any reasonable English speaker would read it" at its original time of enactment, giving it "the ring the words would have had to a skilled reader of words at the time" (Barrett 2017; Easterbrook 1988).

[1] As Sir William Blackstone put it (in 1765), statutory words "are generally to be understood in their usual and most known signification: not so much regarding the propriety of grammar, as their general and popular use" (Blackstone 1765).

[2] The focus of our discussion here is textualism, and our arguments therefore apply to textualist originalism as well. Textualism and originalism are overlapping but not identical interpretive approaches. Generally, originalism is a method of constitutional interpretation and textualism covers all sorts of authoritative legal text. Most, but not all, textualists seek to determine ordinary meaning at the time of enactment, that is, at the time of the text's origin. Some originalists seek to determine the intention of the original enactors, but these days, most are textualist originalists, and seek ordinary meaning at the time of enactment.

In practice, what really distinguishes textualism today from previous interpretive styles is its reliance on certain sources for arguments about the meaning of statutes, and an insistence that those sources are being used to empirically investigate language meaning as a factual reality. Judges frequently employ dictionaries, linguistic "canons" of interpretation "designed to capture the speech patterns" of ordinary language use, and intuitive comparisons with "common sense" usages of the statutory terms (Biden v. Nebraska 2023). These sources are said to provide "evidence" of how ordinary people subject to the statute would understand it (Barrett 2017), though critics often point out that judges exercise significant discretion in gathering this evidence, such as by "shopping" for their preferred dictionary definition (Aprill 1998).

Since the late 2010s, corpus linguistics – a method of quantitatively analyzing patterns of real-world language usage across a large database of corpora – has become vogue in aiding the determination of ordinary meaning (Ehrett 2019; Jennejohn, Nelson, and Nunez 2020). Interrogating corpora is said to be more "empirical" and "scientific" than relying on subjective intuitions and dictionary definitions (Cunningham and Egbert 2019; Solan and Gales 2017; Vogel, Hamann, and Gauer 2018), allowing the quantification of word use frequencies, clusters, and relationships in context. But critics maintain that the method is less accessible than dictionaries in that it requires significant time and technical expertise (Snell v. United Specialty Insurance Company 2024; Ehrett 2019), merely replaces the problem of shopping for the best dictionary definition with deciding which corpora are the most authoritative guides to ordinary meaning (Tobia 2021), and has the added problem of inheriting implicit gender and racial biases in legal corpora (Ehrett 2019; Jennejohn, Nelson, and Nunez 2020). Moreover, the complexity of corpus linguistics presents even more subjective decision points than selecting a dictionary definition.

Regardless of which methods textualists use, their critics charge that beneath the veneer of rigorous objectivity, textualism is every bit as flexible, discretionary and subject to manipulation as the pragmatic interpretive methods it replaces. Focusing on text does not do away with value-laden choices about, for instance, which part of the text is most central, which of multiple definitions to pick, whether to take an expansive or narrow approach to meaning, and what should be included or excluded from relevant context (Aprill 1998; Eskridge, Slocum, and Tobia 2023). In 2022, for example, the Biden administration's public transportation mask mandate was struck down by a Florida district judge, who cherry picked dictionary definitions to argue that the ordinary meaning of "sanitation" did not include wearing a mask during a pandemic (Health Freedom Def. Fund v. Biden 2022; Choi 2024; Gries et al. 2022). The judge found that "sanitation" had two dictionary meanings: (a) active, in "the sense of cleaning" and (b) passive, in "the sense of preserving cleanliness" (Health Freedom Def. Fund v. Biden 2022; Choi 2024). In her view only (b) would include a mask mandate, but according to her corpus linguistics analysis, the more commonly used definition was (a).[3] Insofar as textualism is intended to reign in judicial discretion and not simply serve as "a rhetorical smokescreen for extremely conservative results" (Buchanan and Dorf 2020), the ability to pick and choose ordinary meaning spells trouble.

Lately, some judges and legal academics have proposed that LLMs, trained on vast databases of diverse language use, might help address these criticisms by providing another appropriate, even ideal, source for a legal empirical investigation of meaning.

2.2 LLMs and Ordinary Meaning: A Match Made in Heaven?

With the meteoric rise of generative AI models since 2022, empirically inclined judges have turned to LLMs as a plausible improvement over dictionaries and corpus linguistics (Choi 2024; Engel and McAdams 2024; Hoffman and Arbel 2024; Raymond 2024b). LLMs compactly represent the statistical distribution of words in a vast dataset of textual corpora (Kalyan 2024; Naveed et al. 2025). These models represent words as multidimensional numerical vectors known as word embeddings, where each of the vector's numerous dimensions reflects some aspect of a word's semantic meaning (Choi 2024). LLMs 'learn' word embeddings by exposing a deep neural network to a vast amount of textual data, which adjusts these vectors according to the frequency at which words appear in different contexts. After this initial unsupervised training, LLMs are sometimes fine-tuned with specific instructions and via reinforcement learning with human feedback (RLHF), which employs human workers to give the LLM feedback on some normative dimension (e.g., 'rewarding' coherent outputs and 'punishing' outputs that contain offensive language) (Naveed et al. 2025; Suresh et al. 2024).

Unlike other empirical methods of interpreting legal meaning, such as corpus linguistics, using LLMs requires no technical expertise – anyone can ask OpenAI's ChatGPT or Anthropic's Claude a question about the ordinary meaning of a legal term. Word embeddings, with their hundreds of dimensions, also capture a much richer and more contextual sense of the meaning of words than mere word frequencies (Choi 2024). Moreover, because LLMs are trained on billions of words of everyday language from numerous sources – news pages, Reddit posts, books, and other publicly available sources on the internet (Baack 2024; Engel and McAdams 2024) – they could be especially well-suited to predicting what the average person, as opposed to an elite-trained lawyer, would think a term or statute meant.

Judge Newsom argued as much in his *Snell v. United Specialty Insurance Company* concurrence, a case that hinged on whether installing an in-ground trampoline constitutes "landscaping" (Snell v. United Specialty Insurance Company 2024). Newsom asks, "Is it absurd to think that ChatGPT might be able to shed some light on what the term 'landscaping' means?" He concludes that not only is it not absurd but that LLMs in fact ought to become "one imple-

[3] Gries et al. (2022) and Choi (2024) conducted their own textualist analysis of the case and came to the opposite conclusion.

ment among several in the textualist toolkit" when analyzing ordinary meaning. Among the advantages of LLMs, he lists their training on everyday contextual language, their accessibility, and their transparency (at least relative to the behind-closed-doors construction of dictionaries). He even postulates that LLMs are "probably less vulnerable to manipulation than dictionaries and canons," at least when one's "research process" is well-documented (Snell v. United Specialty Insurance Company 2024).

Advocates of using LLMs for textualism have centered on two main approaches: using LLMs as a synthetic stand-in for what real human populations would think (without actually querying those populations) (Snell v. United Specialty Insurance Company 2024; Engel and McAdams 2024), and using LLMs' word embeddings to estimate the semantic similarity between terms (Choi 2024; Hoffman and Arbel 2024). Returning to the "no vehicles in the park" example: one could determine the ordinary meaning of "vehicle" by prompting an LLM to estimate what percentage of the public would consider each of a set of given objects to be a vehicle (Engel and McAdams 2024), or one could calculate the distances between the word embeddings of 'vehicle' and words like 'bicycle' and 'stroller' using a measure like cosine similarity (Choi 2024; Hoffman and Arbel 2024). Both methods have been shown to be capable of producing results that correlate significantly with an empirical benchmark – in this case, an experimental survey that prompted 2800 English speakers on Amazon Mechanical Turk to determine whether buses, bicycles, ambulances, etc. are a kind of vehicle (Tobia 2020). Based on these results, advocates have concluded that LLMs "hold considerable promise as a way to quickly provide objective answers to legal problems" (Choi 2024).

The use of LLMs for legal interpretation has not been without its critics, however. Responding to Judge Newsom, legal scholars Thomas R. Lee and Jesse Egbert argue that LLMs are "in no position to produce reliable datapoints on questions like the one in *Snell*" (Lee and Egbert 2024) because we know too little about their training data to know whether they are truly representative of ordinary speech, and because RLHF, much of which is done by non-native English speakers, may have unpredictable effects on representativeness. Other critics maintain that using LLMs to infer ordinary meaning is even more value-laden, more discretionary, and less transparent than other empirical methods – particularly corpus linguistics – while facing "substantially the same theoretical issues that confront human interpreters" (Coan and Surden 2025; Lee and Egbert 2024; Tobia 2024). Commenting on the use of cosine similarity between word embeddings, legal scholar Kevin Tobia argues that "judges should not rely on such outputs of algorithmic tools to settle interpretation" (Tobia 2024), because *context* provides the answer about words' meanings, not similarity scores.

Like proponents of LLMs as a cheap stand-in for public opinion research, advocates for using LLMs to interpret legal language sometimes assume their methodological viability depends on fidelity between the views of actual people and their 'silicon' surrogates. Rather than expounding at length on the potential fit between LLM outputs and empirical measures of popular opinion, we will assume that a sufficiently high degree of fidelity is theoretically possible. We focus for the duration of the paper on the conceptual problems revealed by using LLM approximations of ordinary meaning for judicial interpretation.

3 What LLMs Reveal About Empirical Jurisprudence

Proponents argue that using LLMs to find 'evidence' of legal text's ordinary meaning can augment the relative benefits of textualism over other legal interpretive methods by reducing the role of subjective choices, grounding legal interpretation in a sophisticated empirical investigation of real language use, and ensuring that judicial rulings faithfully reflect the people's understanding of legal rules. To the contrary, we argue that LLM jurisprudence underscores the discretionary decisions required to infer ordinary meaning; highlights the inescapable reality that determining the meaning and application of legal terms is an inherently normative project; and demonstrates the lack of democratic legitimacy of crowdsourcing legal meaning.

3.1 LLMs' False Promise of Legal Objectivity

Critics of textualism point out that the methods it relies on – dictionaries, linguistic canons, 'common sense' examples, and corpus linguistics – entail subjective judgments that allow judges to sculpt legal interpretation that fits their political goals (Aprill 1998; Eskridge, Slocum, and Tobia 2023). (Recall the Florida judge's tendentious use of dictionary definitions of "sanitation.") Advocates of incorporating LLMs in legal interpretation say that, with the right precautions or prompting methods in place, LLMs could be the key to helping "reduce opportunities for subjective judgments" (Choi 2024) available with other empirical textualist methods. This echoes familiar ambitions about the objectivity supposedly imparted by algorithmic methods in other contexts (Daston and Galison 2007; Porter 1995). Let's examine this claim.

First, it's important to note that the architecture of LLMs presumes that words do not have a single static meaning but rather that meaning emerges dynamically from complex patterns of context-dependence. Behind the scenes, LLMs use a deep learning architecture called a transformer to process text by breaking it into tokens and converting these into multidimensional vectors of numerical values, or word embeddings. Relationships and semantic similarities with other words are expressed through the numerical relationships between these vectors. Importantly, however, words do not have one single, static embedding; rather, LLMs use what is called a self-attention mechanism to decide how much each word in the input should influence the processing of every other word. Through multiple layers of processing, these initial embeddings are continuously transformed, enabling the model to build contextual representations that capture not just individual word meanings but complex relationships across the entire text. This allows LLMs to distinguish, for example, whether a given use of the word "bat" is most likely to refer to a flying mammal, a long wooden object, or a swatting motion.

Word embeddings confirm the intuition that legal words, isolated from their context, have an indeterminate meaning (Choi 2024). As Tobia puts it, "there is a limit to what we can learn from analysis of individual words, stripped in whole or part from their context" (Tobia 2024). Linguists and philosophers of language have long argued that language meaning is contextual and usage-based (Tomasello 2005; Wittgenstein 2010). This is why LLM outputs are so sensitive to minor changes in the input prompt's wording. When you type a prompt into an LLM like ChatGPT, it converts all the individual words into embeddings and uses this information to predict which words are most likely to come next. Different wording choices in the initial prompt create different attention patterns that compound through the model's layers as it predicts the most likely next tokens.

This sensitivity to context is what makes LLMs so attractive to interpreters of legal language. But it also means that LLMs cannot solve a central problem that dogs all legal interpretation: judges must decide which relevant context to take into account when defining the meaning of words. And it shows how this choice – often not discussed – is likely to be outcome determinative.

Consider Justice Barrett's concurrence in *Biden v. Nebraska*, in which the US Supreme Court overturned the Biden administration's student loan forgiveness program (Biden v. Nebraska 2023). The Court made its decision on the basis of the so-called 'major questions doctrine', an interpretive principle that the US Congress does not delegate questions of 'major' political importance – such as large-scale loan forgiveness – to executive agencies without an explicit statement. Barrett defends the major questions doctrine as a linguistic principle, arguing that it captures the way that a "reasonable interpreter" would understand the "context" of the case (Biden v. Nebraska 2023; Solan 2016). But it is unclear how Barrett distinguishes considering what she considers an objective "common sense" contextual factor she believes is necessary for "discerning ... the text's most natural interpretation" from a "substantive" consideration that skews interpretation by advancing "values external to the statute" (Biden v. Nebraska 2023) (see also Vermeule, 2023). Like shopping for dictionary definitions, judges can shop for relevant context.

Deciding which context is relevant to interpreting a particular legal term is a normative choice. And it is unlikely that there would ever be a one-size-fits all way to define context for every legal text. In *Bostock v. Clayton County*, a landmark US Supreme Court case considered to be a major win for LGBTQIA+ rights, avowed textualists on the US Supreme Court reached opposite conclusions. The case revolved around the meaning of the 1964 Civil Rights Act's prohibition on discriminating against an employee "because of that individual's ... sex." Justice Gorsuch, writing for the majority, reasoned that based on the ordinary meaning in 1964 of the terms "sex" and "because of," the law prohibits employment discrimination based on sexual orientation or gender identity (Bostock v. Clayton County 2020):

> An employer who fires an individual for being homosexual or transgender fires that person for traits or actions it would not have questioned in members of different sex. Sex plays a necessary and undisguisable role in the decision.

The fact that the legislators "who adopted the Civil Rights Act might not have anticipated that their work would lead to this particular result" was irrelevant, because under a textualist analysis, "the limits of the drafters' imagination supply no reason to ignore the law's demands." Dissenting, Justice Alito called the majority's reasoning "preposterous," "a pirate ship" that "sails under a textualist flag" but, rather than enforcing the 50-year-old statute's ordinary meaning, amends it to "better reflect the current values of society" (Bostock v. Clayton County 2020). If this dispute is not just about outcome, it reflects a difference of opinion about relevant context and the relationship between social facts and the meaning of statutory text. Does considering what the majority of people in the US thought about homosexuality in 1964 "help to explain what the text was understood to mean when adopted," or does it amount to "departing from the statutory text"? There is no empirical way to settle such a dispute, whether using LLMs or other methods.

Even if there is no objectively *correct* way to determine the appropriate context, however, perhaps LLMs could help with a more modest goal: reducing inconsistencies in which context is considered relevant for interpreting ordinary meaning, an instance of what historians of science Lorraine Daston and Peter Galison call 'mechanical objectivity' – the achievement of consistent results and reduction of human idiosyncratic bias through mechanical methods (Daston and Galison 2007). Suppose that judges could agree on a prompting system to determine ordinary meaning via LLMs,[4] and thus agree on the appropriate scope of context for inferring ordinary meaning.

One way to achieve this goal would be to use a prompt that reproduces as closely as possible patterns in real human survey data. In closely related work, social scientists have recently proposed using LLM responses to survey questions, or 'silicon samples', as a proxy for public opinion. Researchers have demonstrated notable correlations between human subpopulations' responses to survey questions and ChatGPT's responses to the same questions, when prompted to answer as a member of a given subgroup, on a broad range of topics including political candidates, global warming, and consumer preferences (Argyle et al. 2023; Brand, Israeli, and Ngwe 2023; Lee et al. 2024; Park et al. 2023).

Christoph Engel and Richard McAdams apply such an approach to the "no vehicles in the park" thought experiment, testing ChatGPT's ability to replicate an empirical benchmark – an experimental survey on the same topic (Engel and McAdams 2024; Tobia 2020). After testing different prompts, Engel and McAdams settle on asking GPT to estimate, on a 7-point Likert scale from "(almost) none" to "(almost) all," how many participants in the original study would have considered each of the given objects ("bicycle,"

[4]Coan and Surden point out that deciding how to frame the prompt of course "raises most, if not all, of the normative questions that have long bedeviled [legal interpretive] theory" (Coan and Surden 2025).

"ambulance," etc.) to be a "vehicle." They found that this prompting method produced results "good enough" to inspire some confidence in its use for inferring ordinary meaning, adding that "LLMs therefore have the potential to democratize the use of empirical evidence" relative to other textualist methods.

Supposing all judges were limited to using the same prompting method of the same LLM to do their inquiries, we may indeed see an increase in consistency – and a reduction of judges' individual idiosyncratic bias – in the interpretation of ordinary meaning.

The problem is that even if one prompting method is consistently used, LLMs will not remove human subjectivity from the result – they will simply shift the source of that bias to a different part of the process, such as the model developers' design decisions (Coan and Surden 2025). The way that LLMs predict ordinary meaning is itself analogous to what textualist judges already do: making value-laden decisions behind the scenes and presenting the output as objective and reasonable, even though hidden value-laden choices are made in the outputted answer (Birhane et al. 2022). These choices are, however, even more opaque and authoritative than those necessitated by using corpus linguistics or dictionaries – in part because the value-laden choices built into generative AI systems reflect not the preferences of the judge but rather the values embedded in training data, the stochastic element of LLM queries, and seemingly innocuous design decisions by engineers at OpenAI or Google. More broadly, researchers have drawn attention to the ways that LLMs are value-laden in ways that reflect the preferences of their creators (Birhane et al. 2022; Suresh et al. 2024) and biased in ways that reflect the toxic language of people on the internet (Bender et al. 2021; Gadiraju et al. 2023; Gallegos et al. 2024). There is overwhelming evidence that the vast datasets LLMs are trained on – uncurated data from the internet – are rife with hateful language and stereotypes (Bender et al. 2021), with researchers finding that LLMs reflect and amplify broader social biases, including sexism, racism, ableism, islamophobia, and homophobia (Abid, Farooqi, and Zou 2021; Gallegos et al. 2024; Glazko et al. 2024; Kotek, Dockum, and Sun 2023).

If these biases indicate that LLMs are not in fact representative of what many ordinary people think, then their use for inferring ordinary meaning should be called into question. Indeed, critics of silicon sampling have countered that LLMs' biases may be artifacts of the unrepresentative internet-based corpora LLMs are trained on and the particular preferences of human workers employed for RLHF, so LLM outputs do not capture the diversity of perspectives found in actual human populations (Bisbee et al. 2024; Lee and Egbert 2024; Qu and Wang 2024). Santurkar et al., for instance, found "substantial misalignment" between human survey data and LLM responses to questions about a range of political topics, finding left-leaning bias and worse representation for certain subgroups, such as people over the age of 65 (Santurkar et al. 2023; see also Bignotti and Camassa 2024). Lee et al. found that LLMs portray racial minority groups as more homogenous than white people from the US (Lee, Montgomery, and Lai 2024).

On the other hand, insofar as these biases are merely a reflection of society (Resnik 2025), they might be a feature, rather than a bug, of using LLMs to infer what 'ordinary' people think. Argyle et al., for instance, suggest that LLMs' algorithmic biases reflect patterns of association held by actual subgroups of society, and that LLMs can be prompted to reproduce human subpopulations' biases both for and against different perspectives to a high degree of fidelity (Argyle et al. 2023). But that hardly makes their predictions of 'ordinary meaning' objective. Instead it would mean that LLM predictions adopt the predominant subjective biases of the majority, or perhaps some rhetorically influential social minority.

In sum, LLMs may shift the location of, but not eliminate, the role of subjective biases and judgments in interpreting ordinary meaning (Coan and Surden 2025). Much like textualist judges themselves, LLMs require behind-the-scenes decisions about which context is most relevant to determining the meaning of a word, but those decisions are opaque to the user and may introduce artifacts that judges do not notice. In the more likely scenario that LLMs will be used to determine ordinary meaning without a consistent prompting method, using LLMs will simply not reduce the role of discretionary choices at all, making the 'objectivity' justification for using LLMs over other empirical textualist methods a moot point. Regardless, given the prevalence of AI hype, we are concerned that the use of LLMs will exacerbate textualism's 'scientific' quality, suggesting reproducibility and neutrality while obscuring the ever-present normative choices that go into interpretation.

3.2 Ordinary Meaning Is a Normative Phantom

A central claim of empirical jurisprudence is that the meaning of legal rules is an empirical fact that can be objectively discovered (Eskridge, Slocum, and Tobia 2023). In particular, textualists draw a sharp line between legislative purpose as a made-up fiction – an illusory target whose investigation can only mislead interpreters – and ordinary meaning as the one true object of statutory interpretation, a matter of empirically discoverable fact. They observe that for most of the interpretive issues judges must decide, "there *is* no legislative intent" to be empirically discovered because "the majority [of legislators] was blissfully unaware of the *existence* of the issue, much less had any preference as to how it should be resolved" (Scalia 2018). There is, however, an objectively existing text that can be interrogated, and according to textualists, the 'ordinary meaning' of that text can be uncovered through empirical investigation. But using LLMs to do textualist interpretations of legal language reveals a conceptual problem with this claim. It shows that textualist analysis of ordinary meaning is no more empirical or fact-based than the judicial search for the "congressional intent" that textualists disdain as "pure fiction" (Barrett 2017).

LLM determinations of ordinary meaning do not investigate the nature of some preexisting factual entity or event in real life. LLMs use pre-existing (questionably representative) samples of real people's language use to predict how a hypothetical ordinary language speaker would understand the relevant text – in other words, they build an imagined

consensus view of the text they are asked to interpret.

So, if the ordinary meaning of legal terms is a discoverable fact, rather than a construction, one might conclude that this disqualifies LLMs as a useful source for textualist analysis. For instance, Lee and Egbert reject LLM inquiries as a textualist method because LLMs "are engaged in a form of artificial rationalism – not empiricism" (Lee and Egbert 2024). For them, the lynchpin of textualism is treating ordinary meaning as a factual question that can be investigated via empirical methods. In their view, LLMs are not a good textualist tool because "the ideas and opinions they produce are based on a single model's experience with language, not an empirical study of language."

To the contrary, we think LLMs just make more obvious the conflict between textualists' claims to use empirical methods to discover statutory meaning and the object of their interpretive process. We think Lee and Egbert are right that LLMs' textual analysis is not an empirical project – but neither is the determination of a legal text's ordinary meaning.

Textualists often assert the empirical nature of their approach and act as if ordinary meaning is a matter of what real people actually think about something at an actual time and place. For instance, in *Bostock v. Clayton County*, Justice Samuel Alito complains that the Court's decision that the 1964 Civil Rights Act protects homosexual employees from discrimination is wrong, because:

> If every single living American had been surveyed in 1964, it would have been hard to find any who thought that discrimination because of sex meant discrimination because of sexual orientation – not to mention gender identity, a concept that was essentially unknown at the time.

At other times, however, textualists are clear that 'ordinary meaning' is not a factual entity that exists in the real world. Prominent textualists expressly acknowledge that they are not seeking the meaning ascribed to the language at issue by any *real* person or group – that textualism is "not a theory of anyone-in-particular's understanding" (Kesavan and Paulsen 2002). Rather, ordinary meaning is "a construct," the imagined reaction of "hypothetical readers" at the time of its enactment (Barrett 2017).

What can it mean to empirically investigate a hypothetical person's understanding? How is that any more difficult than empirically investigating a legislator's hypothetical intent? And practically speaking, how can we test whether an LLM asked to perform this task is doing it accurately? What is the 'ground truth' against which to compare the LLM's results?

Proponents of LLMs like Engel and McAdams solve this riddle by turning to the experimental results of an unimpeachably empirical investigation – surveys of human subjects (Choi 2024; Engel and McAdams 2024). Once they get the LLM to produce answers that mirror human survey results, they conclude that the LLM is on the track of ordinary meaning. Engel and McAdams therefore conclude that, with "the right prompts" their LLM project "is proof of concept for using GPT to explore the ordinary meaning of statutory terms." But their project does not look like an empirical exploration. What they actually seem to have done is use the LLM to carefully reconstruct earlier results obtained from an empirical survey of actual humans. They chose the outcome they viewed as objectively correct and engineered the LLM inquiries to recreate it.

Once again, this does not so much disqualify LLMs as an authentic textualist method as illuminate a problem with textualist theory, namely the conflict between its supposed empirical approach and the creative, constructed nature of its target.

As Deepa Acevedo observes, the textualist framing of ordinary meaning intentionally avoids identification with any actual human beings (Acevedo 2023). Ordinary meaning cannot be reduced to what specific language actually means to any actual person. The interpretive goal is "something broader," a kind of cultural construction that necessarily incorporates normative standards. Something more like "what a reasonable person would take the author to be conveying by the chosen language" (Gries et al. 2022). Recall Justice Scalia's definition of legal words' ordinary meaning as "what they convey to *reasonable* people" (Scalia and Garner 2012) (our emphasis). And again, a legal text "should be construed *reasonably*, to contain all that it *fairly* means" (Scalia and Garner 2012) (our emphasis). Those are value-laden targets, defined at least in part by the judgment of the interpreter about what is reasonable, and so not susceptible to empirical inquiry (Allen 2018). No doubt this constructed, imaginary, hypothetical ordinary meaning is based partly on the interpreter's observations of how real people have really used language, but it is not any more real or factual or *empirical* than a pragmatic multidimensional interpretation of legal meaning.

3.3 LLM Crowdsourcing of Ordinary Meaning Is Politically Illegitimate

Proponents of LLM jurisprudence have argued that they offer the promise of "democratizing" legal interpretation (Snell v. United Specialty Insurance Company 2024; Engel and McAdams 2024; Hoffman and Arbel 2024) "by leveraging inputs *from* ordinary people and by being available for use *by* ordinary people" (Snell v. United Specialty Insurance Company 2024).

There are good reasons to be skeptical of this claim. In §3.2, we discussed how the constructed, creative nature of LLM outputs is not, strictly speaking, an empirical way to estimate the views of real people. Moreover, as we explained in §3.1, LLMs' biases and variable outputs in response to reworked prompts raise doubts about their reliability as objective predictors of how real people would understand legal text. Recent scholarship has also drawn attention to the undemocratic values encoded in LLMs' design (Birhane et al. 2022; Groves et al. 2023; Kapania et al. 2025; Suresh et al. 2024).[5] LLMs are not a democratic technology, at least

[5]For instance, the "exploitation and erasure" implicit in using LLMs as distorted stand-ins for real human voices (Kapania et al. 2025), and the low ceiling for how participatory LLMs can be made to be, even if ordinary people are included in the design and human feedback process (Groves et al. 2023; Suresh et al. 2024) (but see

as currently manufactured (Bogiatzis-Gibbons 2024). And there is something decidedly undemocratic about allowing the engineering decisions of a few technocratic elites to influence the operation of the courts (though this is arguably happening already).

But there is a more basic problem of democratic legitimacy that arises from using either real, empirical, representative public opinion surveys or their constructed, rationalist, biased machine learning proxies to answer legal questions.

In §3.1, we discussed how the legal survey authors and LLM users struggled to frame questions that would include the right amount of context and right sort of context to get at the 'ordinary meaning' of the text at issue. This might be using a prompt that elicits a response that most closely mimics the results of a public opinion survey about the meaning of a statutory term like "vehicle," a "benchmark" against which to compare the outputs obtained from an LLM using various queries (Choi 2024; Engel and McAdams 2024). But why not, after all, simply ask the human survey subjects or the LLM the ultimate legal interpretive question in the case at hand: does the statutory language apply to the case at hand? Did the girl who rode her bike into the playground violate the legal rule "No vehicles in the park"?

Judges seem to view such a populist approach to jurisprudence as ideal, at least in theory. In oral argument for *Facebook v. Duguid* (Facebook, Inc. v. Duguid 2021), Chief Justice John Roberts mused:

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

The problem with such an approach is that it substitutes the aggregated subjective judgments of the surveyed humans, including whatever biases they harbor – or the LLM's prediction of the collective accumulation of those subjective results – for the ultimate judicial determination of what the law is.

Consider the options available to Judge Newsom when he again used LLMs in a recent criminal case. In *United States v. Deleon*, the defendant challenged his sentence, arguing that he did not "physically restrain" a cashier by pointing a gun at him during a convenience store robbery (United States v. Deleon 2024). Newsom asked three different LLMs, "What is the ordinary meaning of 'physically restrained'?"

After repeating his query multiple times, Newsom found that "the LLMs consistently defined the phrase 'physically restrained' to require the application of tangible force, either through direct bodily contact or some other device or instrument" (Raymond 2024b). Noting that this result "squares comfortably with the results obtained through the traditional, dictionary-driven" method of defining the words 'physically' and 'restrained', as well as with his own intuition, he concluded that a sentence enhancement applied when a robbery victim was "physically restrained" should not have applied to Deleon. Under the LLMs' "plain reading of the text," the cashier was psychologically, but not physically, bound by the gun.

Whatever you think of its fairness as a mode of decision making, substituting a polling proxy for a judge's individual reasoning is not the process mandated for legal decisions by the text of the US Constitution and by our legal culture, which create courts where judges sit to resolve legal "cases" and "controversies" (U.S. Const., Art. III). A judge who handed over her ultimate decision to either a real panel drawn from the public or an LLM proxy would be violating her duty to decide the case. Just like flipping a coin, such a handover would completely remove the judge's own subjective biases from the legal result. But neither the coin flip nor the poll would satisfy the role requirements of a judge in a rule of law society. They would not be fulfilling their constitutionally provided judicial job to "say what the law is" (Marbury v. Madison 1803). That is why the judges and academics who propose to use LLMs, benchmarked by human surveys or otherwise, are careful to say that they are "not – not, not, not – suggesting that any judge ever query an LLM concerning the ordinary meaning of some word ... and then mechanistically apply it to her facts and render judgment" (Snell v. United Specialty Insurance Company 2024).

Nevertheless, at times the procedures they advocate and the decisions that result seem to produce exactly that mechanistic process. In his concurring opinion detailing his LLM use, for instance, Newsom observed that his approach differed from the other recent LLM textualist approaches (specifically, Hoffman & Arbel (2024) and Engel & McAdams (2024)). "Rather than asking about a word's or phrase's 'ordinary meaning', as I did," Newsom writes, "those studies (at least in part) asked LLMs to apply their predictions about ordinary meaning to a case's facts and then render final decisions – e.g., In the circumstances presented by Deleon's case, 'Was anyone physically restrained?'"

Judges sometimes want to preserve their ability to reject ordinary meaning in cases where that inquiry leads to what they view as absurd or extremely unwelcome social results – for instance, when the original meaning of some law enacted centuries ago conflicts with current widespread social expectations. Critics lambast these 'fainthearted' or 'reasonable' textualists for exposing the unreality of their supposedly neutral fact-based method (MacDonnell 2015). But if judges stick to ordinary meaning as an entirely empirical determination, that is an even bigger problem because it substitutes a populist, opinion-survey mode of decision making for the constitutionally prescribed, culturally accepted method of adjudicating legal conflicts.

Pushed too far, the very aspect of textualism that seems to promise greater democratic legitimacy – decision-making based on an empirical investigation of 'ordinary' people's understanding of legal text – undermines the interpretive method's political legitimacy. This problem is general to all textualist interpretation, but it becomes more visible when we consider using LLMs for textualist analysis. Using LLMs raises the possibility of constructing a scientific, data-based version of ordinary meaning that is superior to the capabil-

(Huang et al. 2024)).

ities of human judges who rely on their intuitive 'common sense' examples, limited textual research, and tendentious dictionary resources (Phillips, Ortner, and Lee 2016; Solan 2016). But recognizing this seductive potential makes plain that the empirical aspect of ordinary meaning has its limits. It's not just that textualists' claims to rely on facts, not values, are sometimes contradicted by their ultimate reasoning. It turns out that were textualist judges to be able to determine the ultimate meaning of legal text in context according to the view of a representative group of ordinary citizens, they could not appropriate that view as the reason for their decisions without abandoning their constitutional and traditional judicial role.

4 Discussion and Concluding Thoughts

The use of LLMs for textualist interpretation is not the first attempt to use computer algorithms in jurisprudence, and it certainly won't be the last (Christin 2016; Daston 2021; Galison 2019; Green and Viljoen 2020; Pruss 2021). Indeed, attempts to conceptualize law as a system of deductive principles and rules, and thereby minimize the role of human discretion, have long drawn comparisons to algorithms and other automatic processes. Critics call such efforts by the pejorative moniker "mechanical jurisprudence" (Pound 1908; Pruss 2021). Textualism – and its embrace of LLMs – is just the latest chapter of a legal philosophy that has intrigued and drawn the ire of US legal scholars for centuries. Although, on reflection, textualist formalism does incorporate a unique and curious feature. It aims to cement its claim to ideal objectivity by appropriating the empirical approach that was pioneered by legal realists as an antidote to formalism (Talesh, Mertz, and Klug 2021).

It may be ripe for criticism, but the motivation for mechanical jurisprudence is resonant, especially in our era of thinly veiled politically motivated court rulings. As Justice Elena Kagan recently commented: "When Courts become extensions of the political process, ... when people see them as trying to impose personal preferences on a society irrespective of the law, that's when there's a problem and that's when there ought to be a problem" (Potomac Watch 2022; Tobia 2022). More fundamentally, it seems that the very idea of a government by the rule of law requires that legal decisions should somehow be produced "by looking outside the decision maker's own subjective will," and yet human judges are subjective creatures (Allen 2018; Bybee 2011; Carter and Burke 2017).

This is the problem textualism claims to solve, and the problem that everyone wants to solve – but so far no one has succeeded. Could rapidly advancing AI technology hold the answer?

No. While LLMs have brought impressive advances in natural language processing, text generation, and translation, they have not eliminated the role of value-laden choices required in the human enterprise of legal interpretation. AI jurisprudence fails much like other historical attempts to remove subjective human contributions via the mechanical objectivity afforded by algorithmic and rule-based methods. Just as Daston and Galison argue that there is no "view from nowhere" in empirical inquiry, even with the application of mechanical methods (Daston and Galison 2007), there is likewise no 'ordinary' meaning of text devoid of normative construction, even with the application of LLMs. If anything, LLMs' architecture, sensitivity to context, and biases make more obvious the imagined, normative quality of the ordinary meaning that textualists claim to be able to discover empirically.

That said, it is important to think through the appropriate role of LLMs in an irreducibly normative legal interpretive enterprise. Given the conceptual, political, and moral problems with using LLMs in adjudication, our own views on AI are rather sour. But whatever one might think about LLMs' methodological viability or the legitimacy of textualism as an interpretive approach, the door to AI use in the courtroom has been opened. "AI is here to stay," Judge Newsom says in his concluding thoughts in his *Snell v. United Insurance Company* concurrence (Snell v. United Specialty Insurance Company 2024). Indeed, Newsom went on to use AI again himself in the decision we discussed involving the meaning of "physically restrained."

As these methods receive more uptake, it will be important to interrogate both the old problems they replicate and reveal and the new issues they present. With this in mind, we present the following warnings and recommendations for practitioners in this space.

1. **Don't believe the hype.** Judges and legal scholars will continue to claim that shifting to LLMs away from dictionaries, intuition, or corpus linguistics will constrain judicial discretion, produce a nuanced, accurate, objective prediction about the meaning of law, and apply it in a reasonable, just manner. But LLMs, like the other tools we have, will nevertheless suffer from the plentiful ways of language. However the legal world ultimately decides to use LLMs, this new technology is not going to solve the fundamental issue of trading off between judicial discretion and more rule-bound consistency.
2. **Jurisprudence must remain a human interpretive enterprise.** We anticipate that LLMs will be increasingly used by judges as an easy and "good enough" (Engel and McAdams 2024) method for inferring ordinary meaning. However, we urge practitioners in this space that LLMs should be thought of as no more definitive than other sources of 'evidence' of ordinary meaning, such as dictionaries, and must only be used with an understanding of their limits, their tendencies to hallucinate, and their sensitivity to context and priming. Moreover, LLMs must never be used to decide the core interpretive questions at issue. In other words, while we fully expect that LLM queries about the ordinary meaning of 'vehicle' will be used as a flawed source of evidence, LLMs should under no circumstances be relied on to answer the question, "did the parent who brought their baby stroller into the park violate the rule 'no vehicles in the park'?"
3. **Reject the bias and opacity that LLMs will introduce to jurisprudence.** While LLMs are susceptible to many of the same issues as other empirical textualist methods, we expect the use of AI to introduce a distinct set of bi-

ases and harms over earlier methods. As we discussed in §3.1, LLMs are compact reflections of biased language on the internet (Bender et al. 2021; Gallegos et al. 2024), shaped in difficult-to-predict ways by the preferences and language patterns of human workers used for RLHF (Lee and Egbert 2024), the values and design decisions of software engineers (Birhane et al. 2022), and Silicon Valley's bottom line. And because LLMs are *large* language models run on supercomputer GPUs inaccessible to the average person and opaque to even the most technologically sophisticated one, the artifacts they introduce are even more difficult to ascertain than those delivered by dictionaries or corpus linguistics tools. Thus, our recommendation is not to use LLMs for inherently normative tasks such as judging. There are always going to be normative values in the legal process, whether LLMs are involved or not. It's better to be explicit and choose values that are socially desirable, rather than accept without question the hidden and often harmful normative values forced on us by technology companies (Nguyen 2024).

4. **Beware of 'LLM-shopping'.** As we discussed in §3.1, using LLMs in jurisprudence requires discretion, and choices about the wording of prompts and even which LLM is used can shape the resulting outcome. Human-AI interaction research has shown that human decision-makers – including judges – can use AI decision-making tools in unexpected ways, ranging from confirmation bias to automation bias to algorithm aversion (Kawakami et al. 2022; Pruss 2023; Albright 2025; Stevenson 2018). Empirical research on judges' use of these new interpretive methods is needed to better understand how the introduction of LLMs will interact with the social and discretionary activity of jurisprudence. We anticipate that prompt engineering and 'LLM shopping' will become the new 'dictionary shopping' – in other words, we expect that confirmation bias and politically motivated reasoning is the most likely outcome of increased adoption of LLMs for legal interpretation, and we thus urge legal professionals to take a critical perspective when presented with LLM-derived predictions of legal meaning.
5. **Don't be fooled by the language of 'democratizing' jurisprudence.** As we argued in §3.3, crowdsourcing the meaning and interpretation of statutory terms is at odds with judges' constitutional role. Moreover, the use of methods that are controlled by billionaires to advance populist judges' conservative textualist agenda means that the threat these methods pose for democracy is distinct. As its use by judges is only going to grow, we must continue to be attentive to, and vocally critical of, the ways that the use of LLMs in jurisprudence may shape our democratic institutions.

We conclude by observing that there is a certain irony to claiming to interpret law from the perspective of an 'ordinary person' by using a technology built to provide capacities beyond that of any real human. Of course, law itself is such a technology – something built by humans that claims, or at least aims, to provide a capacity for doing justice beyond that of any real human.

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 298–306. New York, NY, USA: Association for Computing Machinery.

Acevedo, D. D. 2023. The Past as a Colonialist Resource. *Duke Law Journal*, 73(7): 1373–1436.

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*, 337–371. PMLR. ISSN: 2640-3498.

Albright, A. 2025. The Hidden Effects of Algorithmic Recommendations.

Allen, J. 2018. Doctrinal Reasoning as a Disruptive Practice. *Journal of Law and Courts*, 6(2): 215–236.

Angwin, J. 2024. Opinion: Press Pause on the Silicon Valley Hype Machine. The New York Times. https://www.nytimes.com/2024/05/15/opinion/artificial-intelligence-ai-openai-chatgpt-overrated-hype.html.

Aprill, E. P. 1998. The Law of the Word: Dictionary Shopping in the Supreme Court. *Arizona State Law Journal*, 30(2): 275–336.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Baack, S. 2024. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2199–2208. New York, NY, USA: Association for Computing Machinery.

Barrett, A. C. 2017. Congressional Insiders and Outsiders. *The University of Chicago Law Review*, 84: 2193–2211.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery.

Biden v. Nebraska. 2023. 143 S.Ct. 2355.

Bignotti, C.; and Camassa, C. 2024. Legal Minds, Algorithmic Decisions: How LLMs Apply Constitutional Principles in Complex Scenarios. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 120–130.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 173–184. New York, NY, USA: Association for Computing Machinery.

Bisbee, J.; Clinton, J. D.; Dorff, C.; Kenkel, B.; and Larson, J. M. 2024. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4): 401–416.

Blackstone, W. 1765. *Commentaries on the Laws of England*. Clarendon Press.

Bogiatzis-Gibbons, D. J. 2024. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 74–84. New York, NY, USA: Association for Computing Machinery.

Bostock v. Clayton County. 2020. Bostock v. Clayton County, 590 U.S. 644 (2020).

Brand, J.; Israeli, A.; and Ngwe, D. 2023. Using LLMs for market research. *Harvard business school marketing unit working paper*, (23-062).

Buchanan, N. H.; and Dorf, M. C. 2020. A Tale of Two Formalisms: How Law and Economics Mirrors Originalism and Textualism. *Cornell Law Review*, 106(3): 591–676.

Bybee, K. 2011. The Rule of Law is Dead! Long Live the Rule of Law! In *What's Law Got to Do With It?: What Judges Do, Why They Do It, and What's at Stake*. Stanford University Press.

Carter, L.; and Burke, T. 2017. *Reason in Law*. New York: Routledge.

Cheong, I.; Xia, K.; Feng, K. J. K.; Chen, Q. Z.; and Zhang, A. X. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2454–2469. New York, NY, USA: Association for Computing Machinery.

Choi, J. H. 2024. Measuring Clarity in Legal Text. *University of Chicago Law Review*, 91(1): 1–82.

Christin, A. 2016. From daguerreotypes to algorithms: machines, expertise, and three forms of objectivity. *ACM SIGCAS Computers and Society*, 46(1): 27–32.

Coan, A.; and Surden, H. 2025. Artificial intelligence and constitutional interpretation. *U. Colo. L. Rev.*, 96: 413.

Cunningham, C. D.; and Egbert, J. 2019. Scientific Methods for Analyzing Original Meaning: Corpus Linguistics and the Emoluments Clauses. In *Fourth Annual Conference of Law & Corpus Linguistics (2019), Georgia State University College of Law, Legal Studies Research Paper*, 2019-02.

Daston, L. 2021. *Classical Probability in the Enlightenment*. Princeton University Press.

Daston, L.; and Galison, P. 2007. *Objectivity*. New York: Zone Books.

Easterbrook, F. H. 1988. Role of Original Intent in Statutory Construction, The Symposium: The First Annual Federalist Society Lawyers Convention-1987. *Harvard Journal of Law & Public Policy*, 11(1): 59–66.

Edwards, B. 2025. Sam Altman says "we are now confident we know how to build AGI". Ars Tecnica. https://arstechnica.com/information-technology/2025/01/sam-altman-says-we-are-now-confident-we-know-how-to-build-agi/.

Ehrett, J. S. 2019. Against Corpus Linguistics. *Georgetown Law Journal Online*, 108: 50–73.

Engel, C.; and McAdams, R. H. 2024. Asking GPT for the ordinary meaning of statutory terms. *U. Ill. JL Tech. & Pol'y*, 235.

Eskridge, W. N.; Slocum, B. G.; and Tobia, K. 2023. Textualism's Defining Moment. *Columbia Law Review*, 123(6): 1611–1698.

Facebook, Inc. v. Duguid. 2021. 141 S. Ct. 1163, 1174 (2021).

Gadiraju, V.; Kane, S.; Dev, S.; Taylor, A.; Wang, D.; Denton, E.; and Brewer, R. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 205–216. New York, NY, USA: Association for Computing Machinery.

Galison, P. L. 2019. Algorists Dream of Objectivity. In Brockman, J., ed., *Possible Minds: 25 Ways of Looking at AI*. Penguin Publishing Group.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179.

Glazko, K.; Mohammed, Y.; Kosa, B.; Potluri, V.; and Mankoff, J. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 687–700. New York, NY, USA: Association for Computing Machinery.

Grant, N.; and Metz, C. 2022. Google Sidelines Engineer Who Claims Its A.I. Is Sentient. The New York Times. https://www.nytimes.com/2022/06/12/technology/google-chatbot-ai-blake-lemoine.html.

Green, B.; and Viljoen, S. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 19–31. Barcelona, Spain: Association for Computing Machinery.

Gries, S. T.; Kranzlein, M.; Schneider, N.; Slocum, B. G.; and Tobia, K. 2022. Unmasking Textualism: Linguistic Misunderstanding in the Transit Mask Order Case and Beyond. *Columbia Law Review*, 122(8): 192–213.

Groves, L.; Peppin, A.; Strait, A.; and Brennan, J. 2023. Going public: the role of public participation approaches in commercial AI labs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 1162–1173. New York, NY, USA: Association for Computing Machinery.

Health Freedom Def. Fund v. Biden. 2022. 599 F. Supp. 3d 1144 (M.D. Fla. 2022).

Hoffman, D.; and Arbel, Y. 2024. Generative Interpretation. *New York University Law Review*, 451.

Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1395–1417. New York, NY, USA: Association for Computing Machinery.

Jennejohn, M.; Nelson, S.; and Nunez, D. C. 2020. Hidden Bias in Empirical Textualism. *Georgetown Law Journal*, 109(4): 767–812.

Kalyan, K. S. 2024. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6: 100048.

Kapania, S.; Agnew, W.; Eslami, M.; Heidari, H.; and Fox, S. E. 2025. 'Simulacrum' of Stories: Examining Large Language Models as Qualitative Research Participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–17.

Kawakami, A.; Sivaraman, V.; Cheng, H.-F.; Stapleton, L.; Cheng, Y.; Qing, D.; Perer, A.; Wu, Z. S.; Zhu, H.; and Holstein, K. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–18. New York, NY, USA: Association for Computing Machinery.

Kesavan, V.; and Paulsen, M. S. 2002. The Interpretive Force of the Constitution's Secret Drafting History. *Georgetown Law Journal*, 91(6): 1113–1214.

Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, 12–24. New York, NY, USA: Association for Computing Machinery.

Lee, M. H.; Montgomery, J. M.; and Lai, C. K. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1321–1340. New York, NY, USA: Association for Computing Machinery.

Lee, S.; Peng, T. Q.; Goldberg, M. H.; Rosenthal, S. A.; Kotcher, J. E.; Maibach, E. W.; and Leiserowitz, A. 2024. Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias. *PLOS Climate*, 3(8): e0000429.

Lee, T. R.; and Egbert, J. 2024. Artificial Meaning? SSRN Scholarly Paper. https://papers.ssrn.com/abstract=4973483.

MacDonnell, T. C. 2015. Justice Scalia's Fourth Amendment: Text, Context, Clarity, and Occasional Faint-Hearted Originalism. *Virginia Journal of Criminal Law*, 3: 175.

Marbury v. Madison. 1803. 5 U.S. 137 (1803).

Markelius, A.; Wright, C.; Kuiper, J.; Delille, N.; and Kuo, Y.-T. 2024. The mechanisms of AI hype and its planetary and social costs. *AI and Ethics*, 4(3): 727–742.

Merken, S. 2025. Trouble with AI 'hallucinations' spreads to big law firms. Reuters. https://www.reuters.com/legal/government/trouble-with-ai-hallucinations-spreads-big-law-firms-2025-05-23/.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; and Mian, A. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5): 1–72.

Neumeister, L. 2023. Lawyers submitted bogus case law created by ChatGPT. A judge fined them $5,000. AP News. https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c.

Nguyen, C. T. 2024. Value Capture. *Journal of Ethics and Social Philosophy*, 27(3): 469–504.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, 1–22. New York, NY, USA: Association for Computing Machinery.

Phillips, J. C.; Ortner, D. M.; and Lee, T. R. 2016. Corpus Linguistics & Original Public Meaning: A New Tool to Make Originalism More Empirical. *Yale Law Journal Forum*, 126: 21–32.

Porter, T. M. 1995. *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton, N.J: Princeton University Press.

Potomac Watch. 2022. Opinion: Potomac Watch. Elena Kagan vs. John Roberts on Supreme Court Legitimacy. Wall Street Journal. https://www.wsj.com/podcasts/opinion-potomac-watch/elena-kagan-vs-john-roberts-on-supreme-court-legitimacy/469a15fe-6c7d-4443-b4be-385d55ac83b1.

Pound, R. 1908. *Mechanical Jurisprudence*. Columbia University Press.

Pruss, D. 2021. Mechanical Jurisprudence and Domain Distortion: How Predictive Algorithms Warp the Law. *Philosophy of Science*, 88(5): 1101–1112.

Pruss, D. 2023. Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 312–323. New York, NY, USA: Association for Computing Machinery.

Qu, Y.; and Wang, J. 2024. Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1): 1–13.

Raymond, N. 2024a. Conservative US judge says AI could strengthen 'originalist' movement. Reuters. https://www.reuters.com/legal/transactional/conservative-us-judge-says-ai-could-strengthen-originalist-movement-2024-04-01/.

Raymond, N. 2024b. US judge runs 'mini-experiment' with AI to help decide case. Reuters. https://www.reuters.com/legal/transactional/us-judge-runs-mini-experiment-with-ai-help-decide-case-2024-09-06/.

Resnik, P. 2025. Large language models are biased because they are large language models. *Computational Linguistics*, 1–21.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, 29971–30004. PMLR.

Scalia, A. 2018. *A Matter of Interpretation: Federal Courts and the Law - New Edition*. Princeton University Press.

Scalia, A.; and Garner, B. A. 2012. *Reading Law: The Interpretation of Legal Texts*. Thomson/West.

Snell v. United Specialty Insurance Company. 2024. 102 F.4th 1208 (11th Cir. 2024).

Solan, L. M. 2016. Can Corpus Linguistics Help Make Originalism Scientific? *Yale Law Journal Forum*, 126: 57–64.

Solan, L. M.; and Gales, T. 2017. Corpus Linguistics as a Tool in Legal Interpretation. *Brigham Young University Law Review*, 2017: 1311.

Stevenson, M. 2018. Assessing Risk Assessment in Action. *Minnesota Law Review*, 103(1): 303–384.

Suresh, H.; Tseng, E.; Young, M.; Gray, M.; Pierson, E.; and Levy, K. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1609–1621. New York, NY, USA: Association for Computing Machinery.

Talesh, S.; Mertz, E.; and Klug, H. 2021. *Research Handbook on Modern Legal Realism*. Edward Elgar Publishing.

Tobia, K. 2021. Dueling Dictionaries and Clashing Corpora. *Duke Law Journal Online*, 71: 146–158.

Tobia, K. 2022. We're Not All Textualists Now. *New York University Annual Survey of American Law*, 78(2): 243–262.

Tobia, K. 2024. Algorithmic Interpretation. *U. Chi. L. Rev. Online*, 1.

Tobia, K. P. 2020. Testing Ordinary Meaning. *Harvard Law Review*, 134(2): 726–807.

Tomasello, M. 2005. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

United States v. Deleon. 2024. 116 F. 4th 1260 (11th Cir. 2024).

U.S. Const., Art. III. 1787.

Vermeule, A. 2023. Text and "Context". *Yale Journal on Regulation*.

Vogel, F.; Hamann, H.; and Gauer, I. 2018. Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies. *Law & Social Inquiry*, 43(4): 1340–1363.

Wittgenstein, L. 2010. *Philosophical Investigations*. John Wiley & Sons.