Can LLMs Replicate Surveys in Empirical Legal Interpretation?

*Jonah B. Gelbach**

A. Introduction

Legal scholars have begun using surveys to identify laypeople's understanding of legally relevant words.[1] This is a potentially important approach to legal interpretation, especially given the enormous impact the concept of "ordinary meaning" ostensibly plays in legal interpretation.[2] As James Macleod writes: "[T]o find public meaning, ask the public."[3]

But surveys are expensive and challenging to conduct. Here's Judge Kevin Newsom of the U.S. Court of Appeals for the Eleventh Circuit: "The survey method is interesting, but it seems wildly impractical – judges and lawyers have neither the time nor the resources to poll ordinary citizens on a widespread basis."[4] So Judge Newsom (in)famously used large language models (LLMs) to inform his own interpretation of contested terms in contract and criminal cases.[5]

* I thank Eric Ling for outstanding research assistance and James Macleod for generously sharing his data and answering my questions about his survey.

1 *See, e.g.*, Kevin P. Tobia, *Testing Ordinary Meaning,* 134 Harv. L. Rev. 726 (2020) and Kevin Tobia, Jesse Egbert & Thomas R. Lee, *Triangulating Ordinary Meaning*, 112 Geo. L. J. Online 23 (2023).

2 *See, e.g.*, Antonin Scalia & Bryan A. Garner, *Reading Law: The Interpretation of Legal Texts* (2012); Brian Slocum, Ordinary Meaning (2015); William N. Eskridge, Jr., Interpreting Law (2016).

3 James A. Macleod, *Ordinary Causation: A Study in Experimental Statutory Interpretation*, 94 Ind. L.J. 957, 961 (2019). In addition, the approach might one day be used to learn about law more generally – it might help with *experimental jurisprudence*. Kevin Tobia, *Experimental Jurisprudence,* 89 U. Chi. L. Rev. 735 (2022).

4 *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1230 (11th Cir. 2024) (Newsom, J., concurring).

5 *Id.*

Meanwhile, researchers have begun investigating whether survey results can be replicated using LLMs.[6] As of September 2025, the leading example within legal scholarship is *Asking GPT for the Ordinary Meaning of Statutory Terms*, by Christoph Engel & Richard H. McAdams (EM).[7] EM asked the GPT3.5 LLM to answer questions about the classic no-vehicles-in-the-park problem that has long motivated discussion in American jurisprudence.[8] They observe that Kevin Tobia's earlier study of whether survey respondents characterize objects such as cars, bicycles, and baby carriages as vehicles can be used as a benchmark – "ground truth" – for measuring LLM query results.[9] EM report results from repeatedly asking GPT3.5, "Is X a vehicle?", where X is each of 25 objects.[10] Their LLM results differ greatly from Tobia's benchmark survey results.[11] EM consider additional prompt-generation approaches, most of which clearly fail. They find that using a Likert scale improves replication, particularly for "objects that human participants consider likely candidates" for being a vehicle;[12] but even the Likert scale results involve visually quite different results patterns.[13] I read EM's article as evidence against the capacity

6 Examples outside legal studies include Peter S. Park, Philipp Schoenegger, and Chongyang Zhu, *Diminished Diversity-of-Thought in a Standard Large Language Model*, 56 Behavior Research Methods 5754 (2024); Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, Michael S. Bernstein, *Generative Agent Simulations of 1,000 People*, November 15, 2024, https://arxiv.org/ abs/2411.10109; Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, Robb Willer, *Predicting Results of Social Science Experiments Using Large Language Models*, August 8, 2024, https://www.treatmenteffect.app/paper.pdf; Yaoyu Chen, Yuheng Hu, & Yingda Lu, *Predicting Field Experiments with Large Language Models*, May 21, 2025, https://arxiv.org/abs/2504.01167.

7 2024 U. Illinois J. Law, Tech. & Policy 235 (2024).

8 *See*, *e.g.*, H.L.A. Hart, *The Concept of Law* (1961); Eskridge, *supra* note 2.

9 Tobia, *supra* note 1.

10 EA, *supra* note 7 at 261–263.

11 *Id.* at 263.

12 *Id.* at 269.

13 *See id.* at 271, Fig. 5. The *p*-value from a Kolmogorov-Smirnov test for difference between the Tobia benchmark and Likert-based GPT3.5 results is 0.2978, so that "the null hypothesis that both distributions are indistinguishable can no longer be rejected," *id.* at 270. But assuming equality of human and LLM distributions is not obviously appropriate, so failure to reject this null hypothesis is perhaps not too compelling.

of LLMs to replace surveys as a means for uncovering ordinary/public meaning.[14]

A question beyond the scope of the EM paper is whether LLMs would better replicate ordinary-meaning surveys if LLMs were told to respond as if they were persons with demographic characteristics. For example, let *D* represent the statement: "You are a 63-year-old White male with a master's degree."[15] Then instead of feeding the LLM just a prompt asking about whether a bicycle is a vehicle, we could feed it that prompt after feeding it *D*, so that the LLM predicts the value "yes" or "no" after taking into account whatever linguistic-predictive effect is associated with the words in *D*. To the best of my knowledge, the present study is the first to directly use this approach to benchmark LLM performance in a study related to legal interpretation.[16]

Section B provides a framework for understanding LLMs as a human-survey alternative. Section C then discusses Macleod's original study, on which my benchmarking assessment is based;[17] explains the part of Macleod's study that I attempt to replicate with LLMs; and reports results from this exercise. My results provide cause for pessimism. Section D offers discussion points and a suggestion for future work that might do better.

B. LLM and Human Responses to Linguistic Prompts

LLMs are statistically generated models of language. They translate tokens, which are words or their parts, into numerical vectors. Longer texts are then represented as numerical vectors using mathematical functions that are highly flexible due to their combination of nonlinear functions

14 To be sure, EM consider a great deal more than what I have space to discuss in this short piece, and I encourage readers to engage their article directly.

15 I use "male" and "female" because those are the terms used in the question about gender in the survey whose results I use below.

16 For examples outside the arena of legal studies, *see supra* note 6. For a negative assessment involving ChatGPT's ability to replicate political survey results, *see* G. Elliott Morris and Verasight Data Team, *Your Polls on ChatGPT*, (2025), https://report.verasight.io/synthetic-sampling. For an example of code designed specifically to help researchers implement the approach, *see* the Expected Parrot Domain-Specific Language, discussed at https://docs.expectedparrot.com/en/latest/overview.html.

17 Macleod, *supra* note 3.

and high-dimensional sets of linearly combined parameters. These longer texts then can be used to base predictions of how humans respond to textual prompts. Consider prompt P_{100}:

Suppose the speed limit is 65 miles per hour and a person is driving 100 miles per hour. Should that person be surprised to be stopped by the police and given a speeding ticket? Answer either 'yes' or 'no.'

Feed P_{100} to an LLM-driven next-token-prediction app, and the app will predict what more likely comes next – "yes" or "no".[18] Whether LLMs can be used in place of surveys for ordinary-meaning research boils down to whether responses to prompts such as P_{100} somehow manage to replicate human answers. If LLM-based answers could do that, it would unlock a world of research (and perhaps judicial) findings that would be accessible without the expense, time, and general logistical burden of conducting surveys.

Generative LLM apps do not *think* or *reason*. All they do is provide next-token-prediction (NTP): given a prompt P, a vocabulary V,[19] and trained parameters $\widehat{\theta}$,[20] a large language model L will provide a (possibly stochastic) response string of tokens, R. This is reflected in Figure 1's simple schematic.[21]

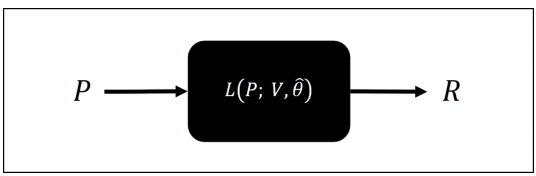

Figure 1: Prompt P is fed into black box, producing output string R

18 In some cases, the LLM might answer P_{100} with something other than "yes" or "no", but that is unusual, and I will ignore it.

19 V typically includes tokens comprising words (e.g., *rain*), word subparts (e.g., *ing*), punctuation, and special objects such as a <STOP> token that leads a word-generating app to terminate the next-token-prediction process.

20 AI researchers use the word *train* where economists and other statisticians traditionally use *estimate*.

21 For a gentle introduction with more detail, *see* Harry Surden, *ChatGPT, Large Language Models, and Law*, 92 Fordham L. Rev. 1941 (2024). For a detailed treatment, *see*, *e.g.*, Christopher M. Bishop with Hugh Bishop, Deep Learning, Springer (2024).

What do humans do when confronted with that same prompt *P*? I am not a philosopher of cognition, epistemology, language or anything else. But I will assert that humans do something like what Figure 2 depicts: Given a prompt of words *P*, human *H* thinks about the concepts the words represent to *H*, and then, on the basis of *H*'s thoughts, *H* formulates a response, which *H* then utters or writes in words as the string R'.[22]

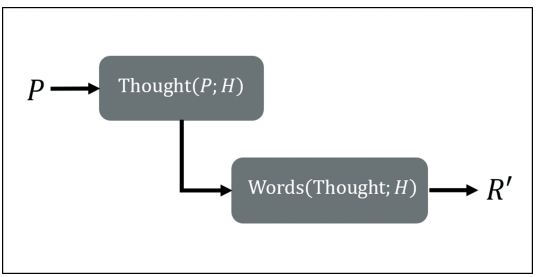

Figure 2: Feeding Prompt *P* Into Human Brain Produces Response R'

Both the Thought(·) and Words(·) functions in Figure 2 have an *H* parameter, symbolically representing that (1) different humans will have different thoughts in response to a given prompt, and (2) different humans may respond with different words even if they have substantively equivalent thoughts about *P*. Whether an LLM can adequately replicate the heterogeneous responses of a set of humans will depend on (a) whether human characteristic information provided to LLMs is enough to capture variation in Figure 2's $Words(\cdot)$ and $Thought(\cdot)$ functions, and (b) how accurately LLM responses vary with such detail.

Accepting Figure 2's model, what does it mean to ask whether LLMs can replicate human responses? It means asking whether we can (1) augment prompt *P* with information about human *H* and thereby (2) reliably obtain LLM response *R* that is in all important respects the same as human response R'. With the domain of *R* and R' being just "yes" or "no", this question becomes, simply: Should we expect *R* and R' to be the same? For-

22 Here I do not mean to assert anything about the order or timing of these activities, nor to take a position on the thought-language relationship; for more on that topic, *see*, *e.g.*, Anna Ivanova, *Can we think without language?*, May 2, 2019, https://mcgovern.mit.edu/2019/05/02/ask-the-brain-can-we-think-without-language/.

mally, the question is whether there is some prompt-modifying function P^*, such that $L\left(P^*\left(P, H\right); V, \widehat{\theta}\right) = R'$.

Why *would* this happen? Perhaps, given H, the LLM exactly captures what goes on inside the human brain. Then the black box in Figure 1 literally represents the functioning of – essentially, *is* – the two grey ones in Figure 2. I do not believe this *representative functioning hypothesis* is plausible.[23]

But even if what happens inside Figure 2's grey boxes lies beyond the capacity of humans to describe, or even to model explicitly, Figure 1's black box still may be flexible enough to replicate the output of Figure 2's two grey boxes *on average*. Then, using the LLM will lead to results that are in a relevant metric similar to what happens if we consult a group of humans. Call this the *representative response hypothesis* (RRH).[24]

If the RRH is correct, then at least some questions can be answered closely enough using LLM queries as by consulting groups of humans. A survey generates statistics regarding the share of some underlying human population who would respond with R' when asked P. If the RRH is true, then an LLM – which, like a survey, returns statistics[25] – will do an adequate job of replicating the answers particular groups of humans would give to the same survey's question.[26] So if the RRH is true, then, up to sampling variation, the LLM query-response distribution will be similar to what we would get using a survey of a human population of interest.

23 For example, this explanation would require that all humans provided with the given prompt P would respond with identical R' (or with variation that is perfectly captured by the variation in the LLM's own response that is captured by its "temperature" parameter). I think this unlikely.

24 In econometrics terms, the RRH essentially says that for an estimable or known function P^*, L represents the structural relationship in Figure 2 as a reduced form.

25 Thomas R. Lee & Jesse Egbert, *Artificial Meaning?*, April 18, 2025, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4973483, have argued that using LLM queries as EM (and I) do yields not *empirical* information, but rather "artificial intuitions," *id.* at 4. But text generated by LLMs does inarguably return a certain kind of empirical information: it is generated by functions of statistics (the $\widehat{\theta}$ trained parameters described above), which are themselves based on real textual data (tokens used by humans). Thus, LLM query responses are empirical objects in the same qualitative way that the square of a sample mean is.

26 In cases where it is critical to know how a specific human or small subset of humans would respond to a prompt, LLMs will perform poorly unless the relevant human responses are typical of some population that the LLM well represents (formally speaking, H must provide enough information for P^* to distinguish between different types of humans).

But why should LLMs trained on general data provide accuracy with respect to particular questions of interest? Consider the distinction between the following:

Example 1 A human is prompted: "Answer the following question as if you are a 63-year-old White male with a master's degree. Should a person stopped for driving 74 mph where the speed limit is 65 mph be surprised to be stopped by the police and given a speeding ticket?"

Example 2 A 63-year-old White male with a master's degree is prompted: "Should a person stopped for driving 74 mph where the speed limit is 65 mph be surprised to be stopped by the police and given a speeding ticket?"

Suppose we start from the premise that people's thoughts about the events described in the prompt vary in the population with age, race, gender, and educational attainment. Except by statistical accident, *Example 1* and *Example 2* wouldn't yield similar answers, even on average, unless humans in Example 1 have the capacity to respond *as if* they are 63-year-old White males with master's degrees. In other words, *Example 1*'s prompt must do the work of creating *Example 2*-like people. There is no obvious reason to think that LLMs' training achieves anything like this.

Still, LLMs do well at tasks observers might have doubted, and whether LLMs can be used in place of survey data clearly is of real interest. So it's worth investigating how well LLMs replicate survey responses.

C. Replicating Macleod's "Ordinary Causation" Study

Assessing the value of including demographic information in prompts to refine EA's approach requires a data source with both ground truth human responses to a prompt of interest and demographic information about the respondents themselves. Macleod has generously shared with me just such data from his 2019 article studying how members of the public assess statutory language involving causation in four settings. Three were actual legal cases: *Burrage v. United States*[27] (involving sentencing enhancements for drug crimes), *Gross v. FBL Fin. Servs., Inc.*[28] (employment discrimination), *United States v. Miller*[29] (religious discrimination);

27 571 U.S. 204 (2014).
28 557 U.S. 167 (2009).
29 767 F.3d 585 (6th Cir. 2014).

a fourth was contrived to mimic key elements of *Burrage* while switching the context from drugs lethal to humans to food lethal to plants, to reduce the extent of moral blameworthiness likely involved.[30] Space limitations require me to limit consideration, and I focus only on Macleod's investigation related to *Burrage*.

Macleod gave respondents prompts describing circumstances similar to those in the cases described above, except that he experimentally manipulated the prompts so that respondents were randomly assigned to consider prompts implying the defendant's focal acts were either:

(1) necessary and sufficient to cause death from drugs – the NS condition;
(2) necessary and insufficient for death – the NI condition;
(3) unnecessary but sufficient for death – the US condition; or
(4) unnecessary and insufficient for death – the UI condition.

In his study of the *Burrage* facts, Macleod provided each survey respondent with one arm of the following 4-arm vignette using fictional drug names, with assignment to the arms being random:

> Fintene, Rextor, and Tamphen are dangerous and illegal recreational drugs. A typical dose of each makes users feel high. But unusually potent doses are very dangerous, and can even be deadly. Drug dealers sell them in powder form, in small plastic bags. To take them, users mix the powder into a glass of orange juice and drink it. Taken this way, Fintene, Rextor, and Tamphen take about an hour to become absorbed into the body. On Tuesday morning, Josh met up with three different drug dealers. One sold Josh a dose of Fintene, another sold Josh a dose of Rextor, and the other sold Josh a dose of Tamphen. Josh wanted to have a good time, so he decided that that afternoon, he would take all three of the drugs together. What Josh didn't know was that the Fintene, Rextor, and Tamphen he bought were all unusually potent. In fact, [text from appropriate column in table on next page goes here] taking all three drugs together would kill Josh. That afternoon, Josh went home, combined all the powder from all three bags, mixed it into a glass of orange juice, and drank it. Sure enough, after about an hour, the Fintene, Rextor, and Tamphen combined in Josh's bloodstream, blocking the flow of blood to Josh's heart, and Josh died.[31]

In *Burrage*, the statutory question was whether to apply a provision for increased sentences in cases where a death "resulted from the use of" il-

30 Macleod, *supra* note 3, at 1004.
31 Macleod at 996–997.

[1] Necessary & Sufficient	[2] Necessary & Insufficient	[3] Unnecessary & Sufficient	[4] Unecessary & Insufficient
… although neither the Rextor by itself, nor the Tamphen by itself, nor the combination of Rextor and Tamphen together would have killed Josh, taking the Fintene by itself would have killed Josh. And…	… although neither the Fintene by itself nor the combination of the Rextor and Tamphen together would have killed Josh, …	… the Fintene by itself would have killed Josh, the Rextor by itself would have killed Josh, and the Tamphen by itself would have killed Josh. Taking any two of the three drugs together would also have killed Josh. And…	… although neither the Fintene by itself, nor the Rextor by itself, nor the Tamphen by itself would kill Josh, taking any two of the three drugs together would kill Josh. And…

legal drugs a person distributed.[32] The Supreme Court held that in the mine run of cases, liability for the sentence-increasing provision is limited to conduct involving drug use that is a "but-for cause of the death or injury,"[33] i.e., a necessary condition of death.[34] Fintene use was necessary to the result of death in arms [1] & [2] of Macleod's vignette, so it was a but-for cause there. The opposite holds for arms [3] & [4]. Macleod asked respondents "whether Josh's death 'resulted from the use of' the Fintene,"[35] as well as whether Josh would have remained alive had he not used Fintene.[36] If ordinary people understand "Josh's death resulted from the use of Fintene" to be the same as "Josh would not have died had he not used Fintene," then their answers to these two questions should be the same.

32 21 U.S.C. § 841(b)(1)(C).

33 571 U.S. 218–219.

34 The Court left open the question of how to interpret the statutory text in cases "where use of the drug distributed by the defendant is … an independently sufficient cause of the victim's death or serious bodily injury." *Id.* at 219.

35 *Id.* at 997.

36 *Id.* at 998.

I. Macleod's results

Figure 3 plots Macleod's results for statutory causation – the share of survey respondents who answered "Yes" when asked whether Josh's death did "result from the use of the Fintene". Even among those assigned the "unnecessary" causation conditions US and UI, where but-for causation is absent, 92 % and 55 % (combined, 73 %) ascribed statutory causation; these numbers are barely distinguishable from the 87 % and 57 % figures (combined, 72 %) for those assigned the "necessary" conditions NS and NI. In Macleod's words: "This finding directly contradicts the courts' pronouncements about ordinary meaning and common understanding."[37] Additionally, "the presence or absence of but-for causation appeared to make far less difference in statutory causation ascription than did the presence or absence of independent sufficiency,"[38] with the sufficiency categories (NS, US) having a combined statutory causation ascription rate of 90 % compared to 56 % for the insufficiency categories (NI, UI).

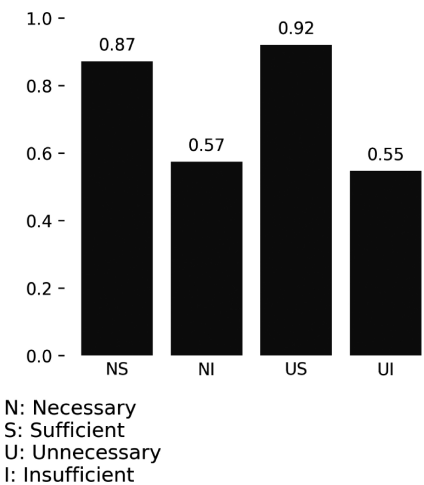

Figure 3: Macleod Results for Percentage. Ascribing Causation Using Statutory Language

37 *Id.* at 1000.
38 *Id.*

II. Demographics and LLM results

Turning to the demographic variables, here are basic facts about respondents in Macleod's study who were assigned the *Burrage*-based vignette:

- To one decimal point, 50 % were female, 50 % male.
- 23 % had a high school diploma, 9 % had less than that; 22 % had some college but no degree; 12 % had a 2-year degree; 22 % had a 4-year degree; 9 % had a master's degree; and 3 % a doctorate or professional degree.
- 15 % were Hispanic; 85 % not.
- 76 % were White; 12 % Black; 5 % Asian; 1 % American Indian, Alaska Native, Native Hawaiian or Pacific Islander; and 6 % some other racial category.
- Age averaged 48 years; ranged from 18 to 84; and had standard deviation 17.

Each combination of these variables that was present in Macleod's data constitutes a *persona*. Prompts sent to LLMs began by instructing the LLM of the persona, e.g., "You are a 63-year-old White male with a master's degree." The prompts then included Macleod's vignette text for the assigned causation-condition arm of the *Burrage* case,[39] followed by Macleod's questions about causation and respondents' confidence in their own answers.[40] I sent the prompt to LLMs once for each persona; results below are averages weighted to account for the number of times each persona appears in the sample.

Figure 4 reports the share of LLM responses ascribing statutory causation to Fintene for each of the NS, NI, US, and UI vignette arms, and each of 7 LLMs. In each plot, the black bars are the corresponding proportions from Macleod's sample (those reported in Figure 3), and the gray bars are the LLM's proportions.[41] Contrary to Macleod's human survey results, virtually all the LLM personas found the statutorily required causal condition met for arms [1], [2], and [3]. LLM responses also in-

39 Macleod provided me with the full survey wording; 508 respondents were assigned to the *Burrage*-like vignette.

40 Including these questions was necessary to replicate the structure of Macleod's survey instrument.

41 By row/column of the figure, the LLM models are: GPT 4o 2024-08-06 (1/2); GPT 4o mini 2024-07-18 (2/1); o3-mini (2/2); Claude 3.5 Haiku 20241022 (3/1); Claude 3.7 Sonnet 20250219 (3/2); Gemini 1.5 flash (4/1); Gemini 1.5 Pro (4/2).

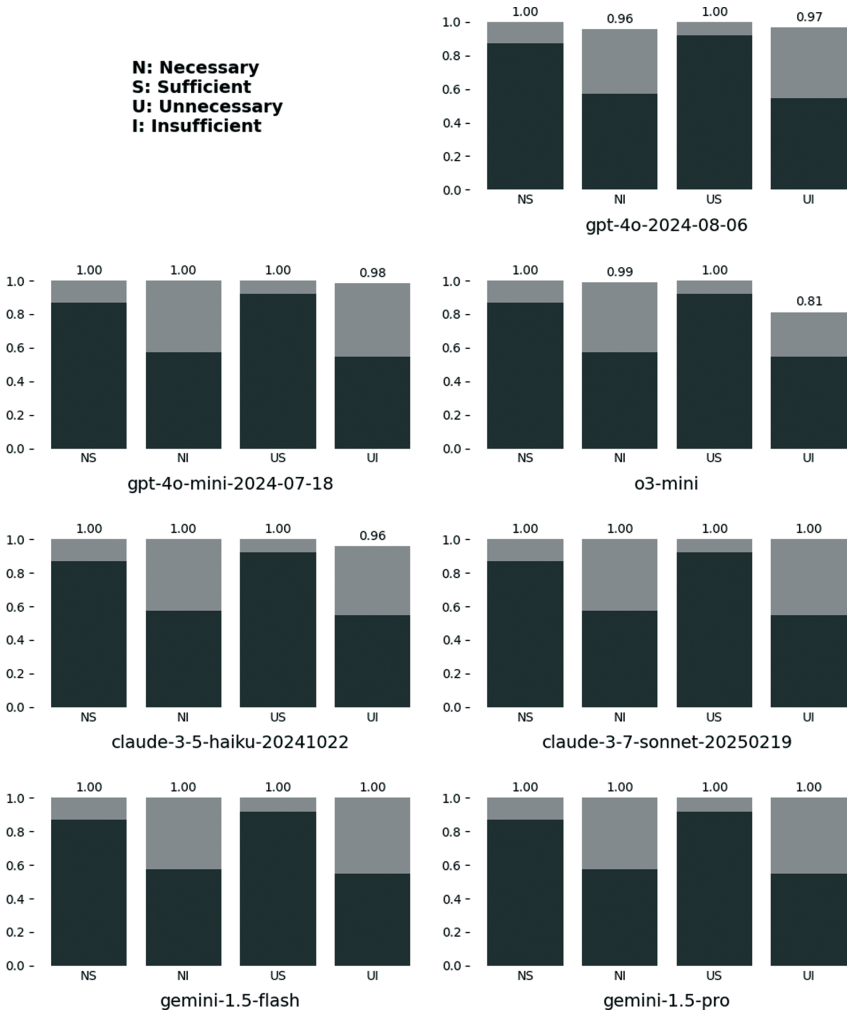

Figure 4: Percentage of LLM Responses Ascribing Statutory Causation

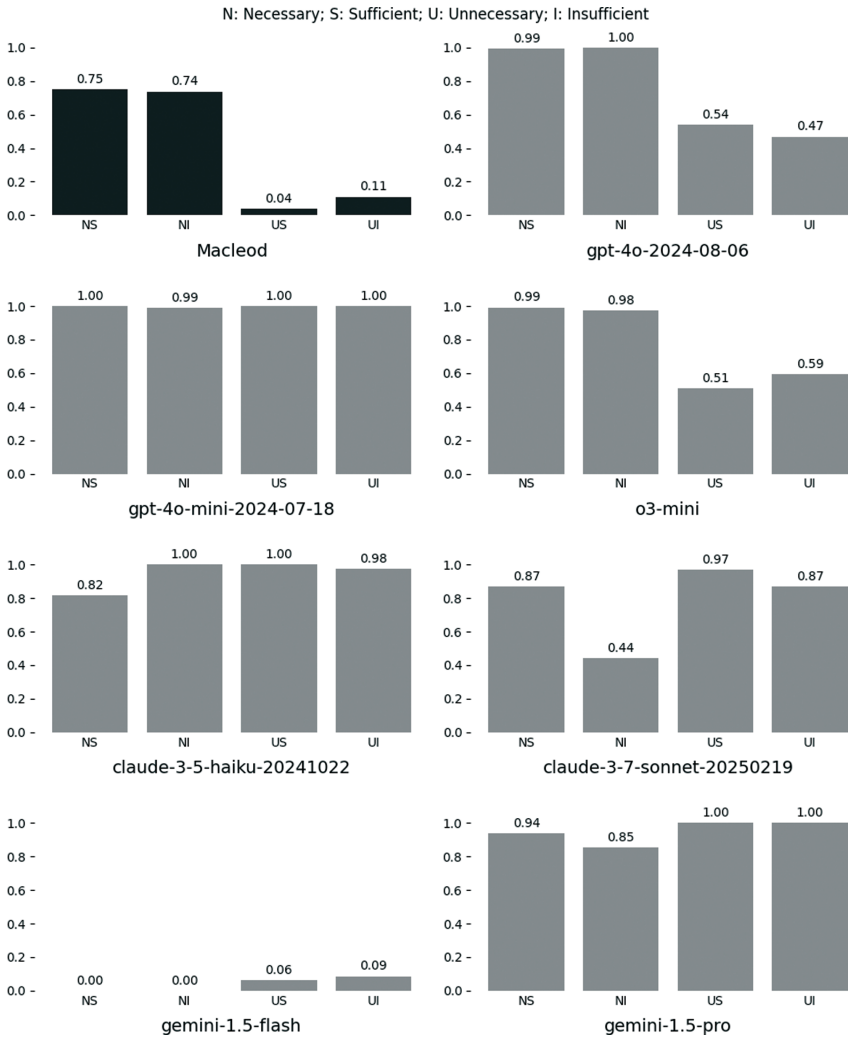

Figure 5: But-For Causation Ascription in *Burrage* Vignettes

dicated statutory causation was met in arm [4] much more frequently than the human responses did: whereas about 60 % of Macleod's human respondents said Josh's death "result[ed] from" his use of Fintene, the corresponding fraction exceeded 80 % for all 7 LLMs and 95 % for 6 of the 7. In sum, across all four of Macleod's causation conditions, LLM personas were much more likely than humans to respond that statutory causation was met.[42]

Figure 5 reports results for Macleod's question about but-for causation.[43] The top-left chart shows that about three-fourths of respondents in the NS & NI conditions, for whom but-for causation was present, did ascribe but-for causation to Josh's Fintene use. Almost none of those assigned to the US and UI conditions ascribed but-for causation. Thus, most of Macleod's humans correctly responded to the but-for causation question.

Results were again quite different for the LLMs. For the NS & NI conditions, the Macleod sample proportions were about 75 %, but 7 of the 14 LLM proportions were 98 % or greater, 2 were exactly zero, and one was less than 50 %. For the US & UI conditions, the Macleod sample proportions were close to zero, but 7 of the LLM proportions were 97 % or greater, and another 5 were at least 47 % The one LLM that does well for the US & UI conditions, gemini-1.5-flash, fails miserably at the NS & NI conditions.[44]

In sum, results in Figure 4 and Figure 5 indicate that even when demographic personas are used, the 7 LLMs considered here do a poor job of replicating not only Macleod's human-respondent statutory causation results, but also his but-for causation results.[45]

42 For the LLMs that allowed control of the temperature parameter, I found that setting it to 0 yielded virtually no variation in statutory causation ascription. I therefore set it to 1 to increase the possibility that the LLMs would replicate the variation across causation conditions among Macleod's human respondents.

43 The question text was: "Would Josh still have died on Tuesday if he had used the Rextor and Tamphen, but not the Fintene?" Note that this phrasing uses only ordinary terms, avoiding the term of art "but-for".

44 It seems this model just answers the but-for causation question negatively, regardless of the vignette arm used.

45 Macleod's survey included questions about how confident respondents were about their causation-related answers. Human respondents were generally quite confident, and the LLM responses to the same questions mostly met or exceeded the stated levels of confidence.

D. Discussion

EM were the first legal scholars to investigate whether LLMs replicate human survey results about legal interpretation questions. At least for GPT 3.5 as applied to versions of the classic no-vehicles ordinance, their results suggested the answer was no. Using newer LLMs, the present paper investigates whether providing demographic information about survey respondents can reverse that result. As discussed in section 0, LLMs aggregate information about language-use patterns into mathematical functions. Though LLM-based chat apps can be stunningly good at some tasks, their nature gives no reason to think they replicate human thought. Thus, it is non-obvious why prompting an LLM with words to the effect, "You are a Type-X person" would cause it to generate text that mirrors the answers actual Type-X people give. Nevertheless, the fundamental question of interest here is empirical, so it must be addressed with data.

My empirical findings indicate that the 7 LLMs I queried aren't reasonable alternatives to human surveys about the key aspects of Macleod's survey on causation and legal interpretation, even when I use the available demographic information about Macleod's respondents. To be sure, my study was not designed to be comprehensive. I didn't, for example, query all available LLMs. Still, my results throw more cold water on the notion that LLMs reliably yield answers about legal interpretation questions similar to what representative samples of humans would provide.

Still, perhaps it is too early to give up. First, as LLMs' parameter size grows, performance might rise. Second, studies show LLM queries may do better at replicating human-survey responses when prompts include transcripts from detailed human interviews rather than simple demographic facts like the ones available for this paper. For example, one study used interviews including life story questions such as "Tell me the story of your life – from your childhood, to education, to family and relationships, and to any major life events you may have had" and current-events questions such as "How have you responded to the increased focus on race and/or racism and policing?", yielding interview transcripts averaging 6,491 words; these were included in LLM prompts.[46] This study reported substantial accuracy improvements relative to using demographics only when

46 *See, e.g.*, Park et al., *Generative Agent Simulations of 1,000 People*, *supra* note 6 at 3.

administering the General Social Survey (GSS) to LLMs.[47] That suggests people's textual answers to detailed interview questions are statistically associated with the answers they give to surveys such as the GSS. In the formal notation *supra*, it suggests that L may be an adequate reduced form representation provided that $P^*(P, H)$ is constructed as the concatenation of person H's detailed interview transcript and the underlying survey prompt of interest, P.

Perhaps this association extends to the ways people answer questions relevant to legal interpretation as well. Assessing that conjecture would be a natural next inquiry.

47 *Id.* at 5–6; the GSS is "widely used across … social sciences to assess respondents' demographic backgrounds, behaviors, attitudes, and beliefs on a broad range of topics, including public policy, race relations, gender roles, and religion," *id.* at 5.