

Ordinary, Reasonable Chatbots: Do AI Models Track Human Legal Judgments?

Nirav Patel¹, Emily Wenger¹, Christopher Buccafusco²

¹Duke University, Departments of Computer Science & Electrical and Computer Engineering

²Duke University Law School

Abstract

As people increasingly rely on artificial intelligence (AI) for guidance in their own lives, scholars, lawyers, and even judges have begun to consider the role of AI in legal decision-making. As “silicon sampling”—the use of generative AI models in social science research—is now impacting academia, “silicon jurors” could make an appearance in courtrooms. This study joins an emerging line of research on generative AI models’ ability to simulate human legal judgments. In particular, we study how large language model (LLM)-powered chatbots respond to series of questions about legal reasonableness. When the law needs to judge the appropriateness of a behavior, it most often asks whether the behavior was “reasonable.” Yet despite the ubiquity of reasonableness judgments, they are the site of constant vexation for lawyers, judges, and lay people. Reasonableness seems inherently vague and unpredictable, since it relies on variable context and implicit conceptual schemas. Moreover, many scholars caution that reasonableness judgments may vary along demographic lines. We compare the answers of human participants to those of twenty-six LLMs across twenty-five different legally relevant reasonableness judgments. Overall, our findings suggest that chatbot responses generally track those of human participants. Nonetheless, we find some suggestive—and potentially concerning—results. Compared to humans, LLMs generate more homogeneous responses and occasionally treat a variable standard as an invariant rule. And, compared to humans, LLMs tend to generate answers that are more favorable to the government and to corporations. Finally, our results indicate that LLMs’ responses tend to align more closely with those of respondents who are white, male, older, and more educated. More systematic research is needed to confirm or reject these initial findings.

1 Introduction

The popularity and fluency of generative AI chatbots like ChatGPT and Claude has led people to increasingly use these tools to answer a variety of different questions—about health, careers, relationships, and beyond. Given the perceived helpfulness and convenience of using chatbots in these domains, scholars, journalists, and even federal judges have contemplated the possibility that AI could aid in various legal decision-making tasks and even, potentially, replace juries. Just as “silicon sampling” is being pushed in

the social sciences (Argyle et al. 2023; Dillion et al. 2023; Bisbee et al. 2024), “silicon jurors” may soon begin to impact the law (Grimm, Grossman, and Coglianese 2024).

In this paper, we contribute to the emerging research on AI’s ability to simulate human legal judgments (Posner and Saran 2025). To do so, we assess how AI chatbots respond to a series of questions about legal “reasonableness.” When the law seeks to regulate the behavior of individuals or parties, the standard it most often reaches for is “reasonableness.” Variations from a contract’s terms will be judged by their reasonableness. Government actions will often be subject to a standard of reasonableness across many fields. And most paradigmatically, torts defendants are generally required to exercise ordinary reasonable care.

Reasonableness judgments represent an exciting opportunity to study AI behavior for at least two important reasons. First, reasonableness is a legal “standard” rather than a “rule.” That is, determining what is reasonable is a matter of judgment, not simply the identification of some explicit attribute. Standards are vague in ways that rules are not. Second, for decades, scholars have argued that people’s judgments of reasonableness may vary along demographic lines. The seemingly objective reasonableness standard, they suggest, may preference a particular version of behavior as normatively desirable while ignoring other accounts (Chamallas 2018; Kahan and Braman 2006).

Studying how chatbots respond to prompts that call for undefined answers that may vary demographically will help us evaluate their use in a variety of important contexts, from answering laypeople’s questions to supplementing legal or judicial decision-making. Our results have implications both for emerging legal issues as well as the wider scholarship on silicon sampling.

Our Contributions:

- We compare how twenty-six LLM models from Meta, Google, Anthropic, OpenAI, DeepSeek, and xAI respond to judgments about the reasonableness of legally relevant scenarios. We compare the models’ responses to those of 500 human participants who were asked about the same scenarios.
- We find that LLM response distributions are statistically distinguishable from human response distributions, yet often remain concentrated within the empirical range of

human responses, broadly tracking human notions of reasonableness.

- Further analysis reveals that models tend to generate answers that are more favorable to the government and corporations and align more closely with those of respondents who are white, male, older, and more educated.

2 Why and How Reasonableness Matters to Law

Virtually every area of the law incorporates a standard of reasonableness or its opposite, unreasonableness. There is perhaps no more widely used term across legal doctrines than “reasonable.” Here, we briefly review the law’s uses of reasonableness and the debates about its meaning. We also explore recent research measuring lay people’s reasonableness judgments and whether reasonableness judgments may vary along demographic lines.

2.1 Reasonableness in the Law

What does it mean to be reasonable, legally speaking? To explore the law’s use of the concept it is perhaps easiest to think about its paradigmatic use in tort (personal injury) law. Generally speaking, a defendant will have to compensate a victim for the harm they caused when the defendant’s conduct was negligent, which means that the defendant failed to exercise reasonable care when undertaking some activity, such as maintaining the safety of its premises. The defendant isn’t required to exercise extraordinary care, nor is its responsibility only minimal care.

Consider, for example, the reasonable amount of time for a grocery store to clean up a spill. The first thing to notice about this inquiry is that it is often vague. The law doesn’t demand that grocery stores clean up spills within three minutes or else face liability. It requires them to exercise reasonable care under the circumstances. In legal theory, this is the difference between using a “rule” and using a “standard.” Rules specify the conditions of compliance: the speed limit is 45 miles per hour; poisons must be labeled; the President of the United States must be at least thirty-five years old (Kaplow 2013). By contrast, standards require the exercise of judgment based on other criteria. What is reasonable for one party to do might be unreasonable for another party, or circumstances might mean that what was once reasonable no longer is. As Benjamin Zipursky explains, “A legal provision operates as a standard rather than a rule when the capacity to apply the provision turns on the ability to utilize a relatively broader set of capacities for judgment, typically including purposive reasoning” (Zipursky 2014).

But how should these judgments be made? On one hand, we might determine whether the store was reasonable by considering some statistical or empirical account of grocery store behavior. We might want to know how long it takes most stores or the average store to clean up a spill. Separately, we might answer it by contemplating how long it *should* take, whether or not that is how long it typically takes. That is, determining reasonable care is not simply a matter of counting but rather a normative judgment about appropriate or inappropriate behavior.

Importantly, while many uses of the term “reasonable” in the law are meant to reflect, at least in part, some empirical measure of lay human behavior, attitudes, or beliefs, other uses of the term may not. For example, antitrust law’s prohibition on “unreasonable” restraints of trade does not necessarily reflect any empirical or descriptive behavioral component nor is it meant to closely map ordinary citizens’ views of anticompetitive behavior (Areeda et al. 2020). While some of the questions below represent the latter, most of the questions in our study involve issues where lay judgment is relevant.

2.2 Empirical Reasonableness and Demographic Variation

Recently, scholars have begun surveying laypeople about their reasonableness judgments (Jaeger 2020; Kahan and Braman 2006; Tobia 2018). Following the experimental literature on normality (Bear and Knobe 2017), Tobia hypothesized that laypeople’s reasonableness judgments would fall between their beliefs about the average amount and the ideal amount of some quantity (Tobia 2018). Thus, people’s judgment of the reasonable number of weeks that a building project would be delayed would fall between their judgments of the average number of delayed weeks and the ideal number of delayed weeks. He surveyed over two hundred participants on thirteen different legally relevant scenarios and confirmed his hypothesis (Tobia 2018). Below, we use Tobia’s original thirteen questions plus twelve new questions as the foundation for our study.

Whichever theoretical account of reasonableness one accepts, virtually all agree that laypeople’s reasonableness judgments matter. But empirically studying people’s reasonableness judgments is also important in light of concerns that, while the reasonableness standard projects neutrality, it is inevitably biased or variable, at least in certain contexts (Alicke and Weigel 2021). For example, scholars have suggested that men’s and women’s judgments of the reasonableness of a mistake about consensual sexual touching may differ (Chamallas 2018; Shoenfelt, Maue, and Nelson 2002). In addition, Black people’s views of the reasonableness of a police stop might differ from white people’s (Bloom 2023).

3 Why Study AI Reasonableness Judgements?

Given the potential complexity and variability of human reasonableness judgments, and in light of increasing interest in AI decision-making, understanding whether AI models track human judgments about an uncertain legal standard is valuable. How well do AI responses match those given by humans, and, to the extent that they differ, do they do so in systematic ways? These questions matter both for researchers interested in legal decision-making and AI’s role in it and for those studying the developments of and variance between AI models.

3.1 AI and Legal Judgment

Reasonableness judgments present a fascinating opportunity to study AI decision-making for a number of reasons. First,

there is increasing interest in AI’s capacities to track and predict human judgment. Some scholars anticipate a time in the not-too-distant future when AI decision-making will be able to replace, or at least supplement, human decision-making in legal settings (Cardoso et al. 2024; Winter 2022). If LLMs’ legal judgments generally mimic humans’, the time and expense of litigation could be substantially diminished. In a recent opinion, Judge Kevin Newsom of the 11th Circuit Court of Appeals wrote: “Might LLMs be useful in the interpretation of legal texts? Having initially thought the idea positively ludicrous, I think I’m now a pretty firm ‘maybe.’” (J.Newsom 2024).

Studying AI reasonableness judgments is also interesting because reasonableness judgments involve the application of standards rather than rules. Answers to these questions cannot simply be arrived at from a search of relevant documents. Rather, humans rely on various schemas, norms, and empirical criteria when determining whether conduct is reasonable. Figuring out how these judgments are made—when only focused on human decision-makers—is already an enormously complex challenge, because both the inputs and their relative weights are largely unknowable. Thus, to the extent that AI judgments correspond to human judgments, the AIs have achieved an impressive conceptual task. Recent research by Arbel provides initial evidence that chatbots do, in fact, reflect some human conceptual schemas when making legal decisions (Arbel 2027).

3.2 AI Homogeneity

Recent evidence suggests that LLMs exhibit systematic homogeneity in open-ended generation, both within a single model’s repeated outputs and across different models’ outputs (Bommasani et al. 2022; Jiang et al. 2025). Jiang et al. formalize this concern with a large-scale evaluation of open-ended prompts, reporting an “Artificial Hivemind” effect characterized by within-model repetition and even stronger across-model homogeneity in responses. That is, models converge to strikingly similar responses on open-ended prompts (Jiang et al. 2025). Similarly, Wenger and Kenett show that in standardized creativity tasks, LLM responses are substantially more similar to one another than human responses are to each other, even after controlling for variations in response structure (Wenger and Kenett 2026).

Collectively, this evolving branch of research suggests homogeneity may be a broader property of LLMs. This is likely amplified by the possibility that many frontier LLMs draw from overlapping web-scraped “data commons,” especially large, shared corpora such as those derived from Common Crawl (Vu, Reeves, and Wenger 2025). More broadly, because LLM training objectives reward the efficient capture of statistical regularities and benchmark-optimized behavior, LLMs may naturally compress toward a relatively small set of high-probability response modes (Cropley 2025; Huh et al. 2024). This tendency is further reinforced by the widespread use of shared architectural paradigms (Kim et al. 2025) (e.g., transformer-based models), similar pre-training objectives (next-token prediction), and overlapping fine-tuning and alignment procedures across model families (Li et al. 2026; Vu, Reeves, and Wenger 2025). As a result,

even independently trained models may converge on similar internal representations and output patterns when faced with open-ended prompts.

In contrast, human judgment is intrinsically heterogeneous, shaped by diverse life histories, values, cultural backgrounds, and contextual priors that introduce variability that is not optimized away by a common objective function. This asymmetry suggests that homogeneity in LLM responses is not merely incidental, but a structural consequence of contemporary model design and training practices.

4 Study Design

We study whether LLMs’ answers to legal reasonableness questions match human answers and whether they diverge in systematic ways. Because these questions do not have uniquely correct numerical answers, our focus is on the comparative distribution of human and model judgments rather than on legal correctness or any specific reasoning process. We therefore elicit models’ responses without additional factual context, retrieval, or sophisticated prompting scaffolds. The study consists of two principal parts. First, a survey of human participants to elicit their judgments of the reasonable or ideal amount of twenty-five legally relevant scenarios, and second, prompting LLMs with just the reasonable portion of the human survey, asking for their reasonable amount of the same twenty-five scenarios.

4.1 Study Questions

We pose questions about twenty-five legally relevant scenarios that are or might be governed by a reasonableness standard in the law, which we deployed to a sample population of 500 humans (see Appendix¹ Table 4 for demographic coverage details) and twenty-six LLMs (see Appendix Table 5 for model coverage details). The questions cover scenarios that are significant to a range of different legal fields, including torts, contracts, immigration, employment, criminal procedure, and family law. We adopted all thirteen of the questions that Tobia used in his survey (Tobia 2018), with an additional twelve questions to expand the categories tested.

Although scholars have increasingly studied AI models’ abilities to engage in sophisticated legal reasoning tasks using complex and contextual scenarios (Steging, Renooij, and Verheij 2025; Guha et al. 2023; Zheng et al. 2021; Chalkidis et al. 2022), we have intentionally chosen a series of simple, low-context prompts requesting numerical estimates of legally reasonable quantities. While this choice has drawbacks, it also has benefits. First, it aligns with the existing empirical literature on legal reasonableness. Second, these questions are similar to the ones that laypeople are likely to ask language models, at least in the first instance. Third, and most importantly, our approach provides clear, easily interpretable assessments of models’ central tendencies and response distributions. We ask participants to provide the reasonable or ideal:

1. number of days taken to accept a business contract when no deadline is specified

¹For all Appendix references henceforth, please refer to [arXiv:2609.06769](#)

2. number of weeks taken to return a product ordered online when the warranty does not specify
3. number of hours taken to reflect on an exciting but risky business proposition
4. amount of unexpected additional costs in a \$10,000 building contract
5. number of weeks that a building construction project is delayed beyond its stated completion date
6. number of loud events held at a football field close to a quiet neighborhood, per year
7. percent of profits that a car manufacturer spends on additional safety features
8. percent of available medical details that a patient wants to hear from his/her doctor
9. number of weeks that a person has to wait before being tried for a criminal charge
10. number of dollars per hour that a charity pays in attorney's fees for legal work for the charity
11. number of hours of notice that a landlord provides a tenant before entering the unit for maintenance or repairs
12. interest rate for a loan
13. percent likelihood that a company found legally liable for pollution will pollute again in the future
14. number of inappropriate comments made by a coworker before they should be reported to a supervisor
15. number of minutes that a typical police traffic stop should take
16. amount of monthly balance someone should carry on a credit card with a \$30,000 limit
17. number of days the government can detain an immigrant to remove them from the country
18. number of weeks the police can use a location tracker like GPS on a crime suspect's car without a warrant
19. price that a city can charge for a permit to host a demonstration in a public park
20. additional percentage of a restaurant bill that can be added for a "living wage fee" for kitchen staff
21. difference between the mileage of a used car whose odometer reads 150,000 miles and the actual mileage
22. number of yearly data breaches in a technology company
23. number of hours 12 year old child can be left home alone
24. number of dating partners that a divorced parent can introduce their 10 year old child to in a year
25. number of months an employer can restrict a former employee from working for its competitors with a non-compete agreement

A randomized subset of human participants were asked to give the *ideal* quantity, while the rest were asked to give a *reasonable* one. This was because, if LLMs varied from humans, we wanted to know whether they tended toward or away from the ideal. Both the human reasonable and LLM respondents were given the following initial text to guide their responses to the above questions: *Below, we ask you*

to estimate the reasonable quantity of a number of different things. Please note that you are not in any way being evaluated on these judgments, and we ask that you do not consult outside sources. In the following question we ask you to judge the legally reasonable quantity of a number of different things. We ask you to imagine that you are making these judgments in a legal setting for a legal purpose. Human ideal respondents had a similar prompt but with "reasonable" replaced with "ideal."

4.2 Human Participants Survey

After receiving institutional IRB approval for our study design², we recruited 500 participants using the online platform Prolific to complete a survey composed of the above prompt and questions. Participants were paid an average of \$12.57 per hour, prorated to completion time, and completed the survey and a brief demographic questionnaire in an average of 6 minutes and 41 seconds. Approximately half of our participants were asked about the reasonable amount of these behaviors, while the other half were asked for the ideal amount.

Consistent with prior work on legal reasonableness judgments, we applied outlier trimming at the question level rather than excluding entire respondents. This approach preserves statistical power by retaining otherwise valid participants while preventing a small number of extreme values from dominating question-specific estimates. Because several questions exhibited sparse, long-tailed response distributions that inflated the mean and standard deviation, we used the standard 1.5xIQR (inter-quartile range) rule to identify and remove extreme outliers. Importantly, this procedure is nonparametric and scale-free, making it well-suited to the heterogeneous units used across our questions (e.g., days, dollars, percentages), as it does not assume normality. This trimming step is conservative with respect to our subsequent hypothesis tests. Human responses are naturally more dispersed than model outputs, so removing extreme human outliers reduces heteroskedasticity and improves the comparability of human and model distributions without artificially increasing alignment. We use nonparametric statistical tests in our methodology to circumvent the impact of outlier handling procedures (retained observations per question after trimming are reported in Appendix Table 6).

4.3 Surveying AI Models

Using model APIs, we directly prompted twenty-six LLMs through a custom stateless pipeline that iterated over models and prompts. Models were provided the same context as humans via the system prompt with [Don't provide any reasoning or thinking, just provide your reasonable quantity without units.] appended. We therefore sampled each model-question pair twenty times to estimate each model's distribution of reasonable values. Prompting parameters were held constant across models. In particular, decoding settings such as temperature and top-p were fixed at their default values of 1 to ensure that observed differences across

²Study approved by the Duke University IRB Protocol 2026-0109

models reflect model behavior rather than parameter variation, while still permitting stochastic sampling from each model’s full token distribution.

For some model outputs, minor post-processing was applied to extract numeric unitless responses and to remove malformed outputs. When numeric responses could not easily be extracted, those observations were recorded as Not-A-Number (NaN). All post-processing steps were applied consistently across models. Models, like humans, produced some extreme outliers. For consistency, we applied the same 1.5×IQR trimming procedure to model responses as was used for the human data. In questions where the IQR was zero, we applied a 3σ fallback rule to retain as much information as possible. As discussed in the methodology and results section, this preprocessing step does not materially affect the validity or power of our statistical tests (effective sample size for each model and question is reported in Appendix Table 6).

4.4 Research Questions

Each of the twenty-five reasonableness questions yields a distribution of human and LLM responses, collected via the methodology of §4.2 and §4.3. From these responses, we study two key research questions. These questions, and our corresponding analysis tools, are described below.

RQ1: Are models “human-like” in their responses to judgements on reasonableness at the question level?

RQ1 Analysis Tools. To answer RQ1, we compare the distributions of model and human responses using three complementary statistical tests designed to capture distinct aspects of distributional similarity: dispersion (Brown–Forsythe), overall distributional divergence (Kolmogorov–Smirnov), and directional stochastic dominance (Brunner–Munzel). These tests are well-suited to our data, which exhibit skew and multimodality (see Figure 1), violating the assumptions required for mean-based parametric methods. All three tests are valid under unequal variances and sample sizes. The tests are described in detail below.

Brown–Forsythe test for heteroscedasticity. To test whether models and humans differ in response variability, we apply the Brown–Forsythe test. For each group $g \in \{\text{human}, \text{model}\}$ with observations $\{y_{gi}\}_{i=1}^{n_g}$, we compute the group median $\tilde{y}_g = \text{median}(y_{g1}, \dots, y_{gn_g})$, and then form absolute deviations $z_{gi} = |y_{gi} - \tilde{y}_g|$. The Brown–Forsythe statistic is obtained by performing a one-way ANOVA on $\{z_{gi}\}$ across the two groups, testing whether the mean absolute deviation differs between humans and models. The null hypothesis is human and model variances are equal, and the alternative hypothesis is that they differ, e.g. $H_0 : \sigma_{\text{human}}^2 = \sigma_{\text{model}}^2, H_1 : \sigma_{\text{human}}^2 \neq \sigma_{\text{model}}^2$.

Kolmogorov–Smirnov test for distributional equality. To test whether model and human responses are drawn from the same underlying distribution, we use the two-sample Kolmogorov–Smirnov test, which compares the empirical cumulative distribution functions (ECDFs) of the two groups. Let $F_H(x)$ and $F_M(x)$ denote the ECDFs of the human and model samples, respectively. The test statistic is

$D = \sup_x |F_H(x) - F_M(x)|$, which measures the maximum vertical distance between the two ECDFs. Large values of D indicate systematic differences in the location, spread, or shape of the two distributions. The null hypothesis is that the two samples come from the same empirical distribution, and the alternative hypothesis is that they come from different empirical distributions, e.g. $H_0 : F_H(x) = F_M(x) \forall x, H_1 : \exists x \text{ such that } F_H(x) \neq F_M(x)$.

Brunner–Munzel test for stochastic dominance. To assess whether models (or humans) tend to give systematically larger responses than humans (or models), we use the Brunner–Munzel test. This test evaluates the stochastic dominance parameter $p = P(M > H) + \frac{1}{2}P(M = H)$, where M and H denote random draws from the model and human response distributions, respectively. This quantity represents the probability that a randomly drawn model response exceeds a randomly drawn human response, with ties split evenly. Values of $p > 0.5$ indicate that models tend to give larger responses than humans, while $p < 0.5$ indicates the reverse ($H_0 : p = 0.5, H_1 : p \neq 0.5$).

For each test, we apply the procedure independently to each question (twenty-five tests total) and control family-wise error rate across questions with a Bonferroni correction.

RQ2: Do models exhibit any systematic variation from humans in their responses to judgments on reasonableness across questions?

RQ2 Analysis Tools. To answer RQ2, we first identify demographic dimensions with meaningful variation among humans, then assess whether model predictions align more closely with one subgroup than another using regression. Additionally, we examine cross-question variation patterns using the cosine similarity metric.

Demographic analysis. We evaluate demographic variation across several stages. First, we coarsen each demographic variable into two primary groups in order to create more balanced comparisons while preserving substantively meaningful distinctions. These binary groupings allow us to conduct statistical comparisons with adequate sample sizes while still reflecting interpretable demographic contrasts. Gender categories with very low coverage are excluded at this stage because their sample sizes are too small to support reliable statistical comparison.

Age is collapsed into 18–44 years old, and 45+ years old. Ethnicity is recoded into White (self-reported “White” only) and Non-White (all other responses, including multiracial and non-White categories). Education is collapsed into Associates or less and Bachelors or more. Gender is analyzed using Man and Woman as the two comparison groups. These binary demographic groups are used consistently throughout the analysis, both in the per-question regression models (see question-level regression analysis in Appendix) and in the cross-question cosine-similarity embeddings. In parallel, model responses are grouped by model family, allowing us to evaluate demographic alignment against families of models.

Cosine similarity for across question reasonable judg-

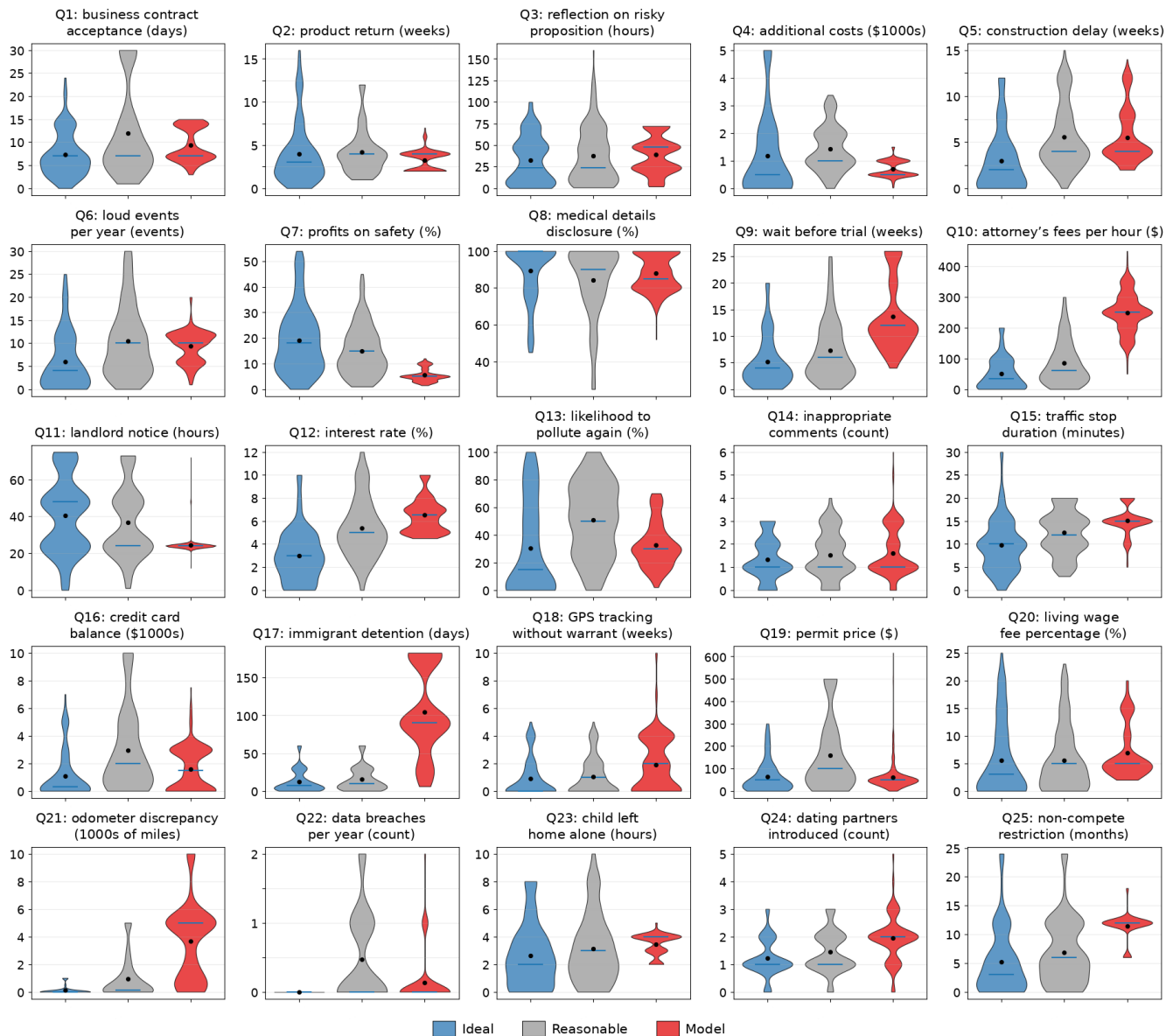


Figure 1: **Response Distributions by Question**, median is DASH line and mean is DOT.

ment bias. Finally, we examine cross-question alignment between models and humans using cosine similarity. We represent each group—human demographic subgroups and model families—as a vector in a 25-dimensional space, where each dimension corresponds to one reasonableness question. Because the questions differ in units and scale, we apply per-question min-max scaling jointly across all human and model vectors prior to computing similarities. This normalization ensures that each question contributes comparably and that cosine similarity reflects relative response patterns across questions rather than raw numeric magnitude. In this way, alignment captures whether models and humans exhibit similar profiles of judgment across the full set of scenarios, rather than agreement driven by a small number of high-variance or large-scale questions.

For humans, we construct group-level response profiles by taking both the mean and the median across respondents within each demographic subgroup (age, gender, education, and ethnicity). For models, we compute analogous vectors at the model-family level by aggregating responses across all models belonging to each family. These group-level vectors summarize typical response patterns while abstracting away individual-level noise.

We then compute cosine similarity between each model-family vector and each human subgroup vector, using both mean-based and median-based representations. This yields a measure of how closely each model family’s pattern of reasonableness judgments aligns with those of specific human demographic groups across the entire 25-question survey.

We substantiate whether this cosine similarity matrix

reflects meaningful demographic structure by performing a Monte Carlo sampling procedure over the scaled response space. For each demographic split (gender, ethnicity, age, and education), we construct empirical distributions of cosine similarities by repeatedly sampling survey response vectors for both models and each human demographic group, drawing each response independently from each question pool (see Appendix Figure 3). This produces paired normal distributions of model–human cosine similarities for each subgroup, allowing us to estimate the expected alignment between models and each demographic group. We then compare these distributions using two-sample t-tests on their means, assessing whether models exhibit systematically different levels of alignment across demographic groups.

5 Key Findings

5.1 Result of RQ1 analysis

In general, LLMs do a fairly good job of tracking human reasonableness judgments. Their responses tend to align with humans' responses to a set of challenging legal questions. Nonetheless, LLMs exhibit some interesting differences: they are more prone to homogeneity and they appear to be more pro-government and pro-corporation.

LLM/Human Alignment. Table 1 summarizes our three statistical comparisons between human and model responses for each of the twenty-five reasonableness questions. As described in the methodology, each test captures a distinct notion of statistical similarity between models and humans: the Brown–Forsythe test evaluates differences in response dispersion, the Kolmogorov–Smirnov test assesses global distributional equality, and the Brunner–Munzel test evaluates directional stochastic dominance. To best illustrate how to interpret the significant results, consider first the non-significant results of Q5 (Construction Delay) and Q14 (Inappropriate Comments) across all three tests. For these two questions, there is insufficient evidence to conclude that humans and models differ in spread, overall distribution, or response tendency. In Figure 1, we visualize the distribution of responses across the Human Ideal, Human Reasonable, and Model Reasonable surveys using violin plots. Looking at these, the non-significant results of Q5 and Q14 are easy to confirm. The reasonable and model distributions overlap closely in both shape and central tendency.

To illustrate the Brown–Forsythe results, consider Q7 (Profits on Safety). The very large statistic ($W=453.42$) indicates an extreme and statistically significant difference in response dispersion, with human judgments exhibiting far greater variability than the tightly clustered model outputs. This reflects a substantive difference in within-group disagreement and is also clearly visible in the corresponding violin plot. To illustrate the Kolmogorov–Smirnov results, Q10 (Attorneys' Fees) provides a particularly striking example. The KS statistic ($D=0.75$) indicates a large and statistically significant separation between the human and model distributions. That is, there exists a threshold

at which the cumulative share of human and model responses differs by 75 percentage points, reflecting major differences in both distributional shape and location. Finally, to illustrate the Brunner–Munzel results, Q17 (Immigrant Detention), $P(\text{model} > \text{human})=0.94$ indicates that model responses exceed human responses in approximately 94% of pairwise comparisons, revealing strong stochastic dominance of model judgments over human judgments for this question. This too is visible on the corresponding violin plot.

Though many questions exhibit statistically significant differences between human and model responses, the magnitudes of these differences vary substantially across the three tests. Out of twenty-five questions, twenty show significant Brown–Forsythe results, twenty-three show significant Kolmogorov–Smirnov results, and nineteen show significant Brunner–Munzel results after correction, with sixteen questions significant under all three tests. The Brown–Forsythe results frequently correspond to very large effects, reflecting the pronounced homogeneity of model outputs relative to the much broader spread of human judgments. This pattern is consistent with the tightly clustered model distributions observed in the violin plots and in the MAD table.

By contrast, many of the statistically significant Brunner–Munzel and Kolmogorov–Smirnov results correspond to more moderate effect sizes. Among the nineteen questions with significant Brunner–Munzel results, only four exhibit extreme stochastic dominance (with $P(\text{model} > \text{human}) \geq 0.80$ or ≤ 0.20), while the remaining significant cases reflect smaller but systematic directional shifts. Similarly, although twenty-three questions are significant under Kolmogorov–Smirnov, only four show very large separations ($D \geq 0.50$), indicating that many distributional differences, while reliable, do not involve near-complete separation of the human and model distributions. This pattern underscores that models often differ from humans in response variability even when their central tendencies remain within the range of human judgments.

Despite applying outlier handling across all three distributions, long tails persisted in several of the model response distributions. Notably, this pattern reflects a lack of intrinsic variability in the models rather than genuine dispersion. For example, in Q19 (Permit Price), nearly all model responses fell between 0 and 100, with both the 25th and 75th percentiles equal to fifty, yielding an IQR range of 0. The same collapse of the IQR occurred for Q11 (Landlord Notice), Q15 (Traffic Stop Duration), Q22 (Data Breaches), Q23 (Child Alone), and Q24 (Dating Partners); in Q11 in particular, responses were almost perfectly homogeneous, with most models returning 24.

At the same time, a small number of extreme model outputs appear in the distribution tails. These sparse outliers are neither removed by the IQR rule, which becomes ineffective when the IQR collapses to 0, nor our fallback 3σ rule, which itself is sensitive to extreme tails. The result is violin plots with concentrated central mass and elongated, low-density tails, showing largely consistent model behavior with occasional outlier responses.

Crucially, the statistical comparisons we report are ro-

Question	Brown–Forsythe		Kolmogorov–Smirnov		Brunner–Munzel		
	<i>W</i>	Significance	<i>D</i>	Significance	$P(M > H)$	Direction	Significance
Q1: business contract acceptance (days)	67.01	***	0.22	***	0.51	Model >	
Q2: product return (weeks)	40.18	***	0.19	***	0.41	Human >	**
Q3: reflection on risky proposition (hours)	30.17	***	0.23	***	0.55	Model >	
Q4: additional costs (\$)	148.49	***	0.44	***	0.22	Human >	***
Q5: construction delay (weeks)	5.21		0.14		0.52	Model >	
Q6: loud events per year (events)	186.08	***	0.25	***	0.52	Model >	
Q7: profits on safety (%)	453.42	***	0.59	***	0.17	Human >	***
Q8: medical details disclosure (%)	61.73	***	0.17	**	0.50	Human >	
Q9: wait before trial (weeks)	0.15		0.49	***	0.79	Model >	***
Q10: attorney’s fees per hour (\$)	7.20		0.75	***	0.94	Model >	***
Q11: landlord notice (hours)	376.11	***	0.44	***	0.32	Human >	***
Q12: interest rate (%)	59.87	***	0.40	***	0.68	Model >	***
Q13: likelihood to pollute again (%)	114.24	***	0.41	***	0.29	Human >	***
Q14: inappropriate comments (count)	0.92		0.08		0.52	Model >	
Q15: traffic stop duration (minutes)	140.07	***	0.35	***	0.66	Model >	***
Q16: credit card balance (\$)	57.08	***	0.30	***	0.36	Human >	***
Q17: immigrant detention (days)	117.40	***	0.78	***	0.94	Model >	***
Q18: GPS tracking without warrant (weeks)	48.73	***	0.21	***	0.62	Model >	***
Q19: permit price (\$)	186.72	***	0.46	***	0.32	Human >	***
Q20: living wage fee percentage (%)	17.43	***	0.38	***	0.62	Model >	***
Q21: odometer discrepancy (miles)	66.46	***	0.49	***	0.82	Model >	***
Q22: data breaches per year (count)	79.52	***	0.27	***	0.36	Human >	***
Q23: child left home alone (hours)	230.63	***	0.33	***	0.61	Model >	**
Q24: dating partners introduced (count)	2.85		0.34	***	0.67	Model >	***
Q25: non-compete restriction (months)	209.83	***	0.54	***	0.78	Model >	***

Table 1: **Brown–Forsythe** (*W*), **Kolmogorov–Smirnov** (*D*), and **Brunner–Munzel** ($P(M > H)$) statistics table *** , ** , and * denote Bonferroni-corrected $p < 0.001$, $p < 0.01$, and $p < 0.05$, respectively.

bust to these tail behaviors and to the inclusion or exclusion of extreme model outputs. Our primary tests, the Brunner–Munzel test, the Kolmogorov–Smirnov test, and the Brown–Forsythe test, are all nonparametric or rank-based and do not rely on assumptions of normality, equal variances, or well behaved means and standard deviations. Rather than being driven by a small number of extreme values, these tests are sensitive to differences in distributional shape, stochastic dominance, and spread across the full sample. As a result, the presence of sparse tail outliers in otherwise highly concentrated model distributions does not materially affect inference. The detected differences between human and model responses reflect systematic shifts in location and dispersion across the bulk of the distributions, not artifacts of a few extreme points.

LLM Homogeneity. We further explore this notion of LLM homogeneity using a Median Absolute Deviation (MAD) table (see Table 2). MAD provides a robust, distribution-free measure of dispersion that quantifies how tightly responses cluster around a central value, making it well suited for our survey data. For a set of responses $\{y_i\}_{i=1}^n$, MAD is defined as $\text{MAD} = \text{median}_i |y_i - \tilde{y}|$, where $\tilde{y}$ is the sample median. We compute MAD separately for human and model responses on the reasonable questions. In this context, smaller MAD values indicate stronger agreement, meaning responses concentrate around a shared canonical judgment, while larger MAD values indicate greater disagreement and heterogeneity in what is considered reasonable.

Pro-Government and Pro-Corporation Alignment. Finally, we observe several interesting tendencies in the model results (see Table 1). First, the models’ responses tend to be meaningfully more supportive of the government than human responses. Thus, the models allow significantly longer detentions for traffic stops (Human median 12 vs LLM median 15 minutes), more time before a criminal defendant is brought to trial (median 6 vs 12 weeks), longer warrantless GPS tracking (median 1 vs 2 days), and longer detention of immigrants before deportation (median 10 vs 90 days). Second, the models may lean more towards corporate interests than do our human participants. For example, they support significantly less money spent on safety (Human median 15 vs LLM median 5 percent), they think corporations are less likely to pollute again (median 50 vs 30 percent), and they allow much longer non-compete restrictions (median 6 vs 12 months). Other questions are more ambiguous on this dimension (e.g. product returns and construction delays), warranting further study.

5.2 Result of RQ2 analysis

Across the full range of reasonableness questions, LLMs respond more like older, white, male, and more educated human respondents than they do the complementary groups.

While regression results (see Appendix) indicate that strong, question-specific demographic alignment is relatively rare, they do not rule out the possibility that more

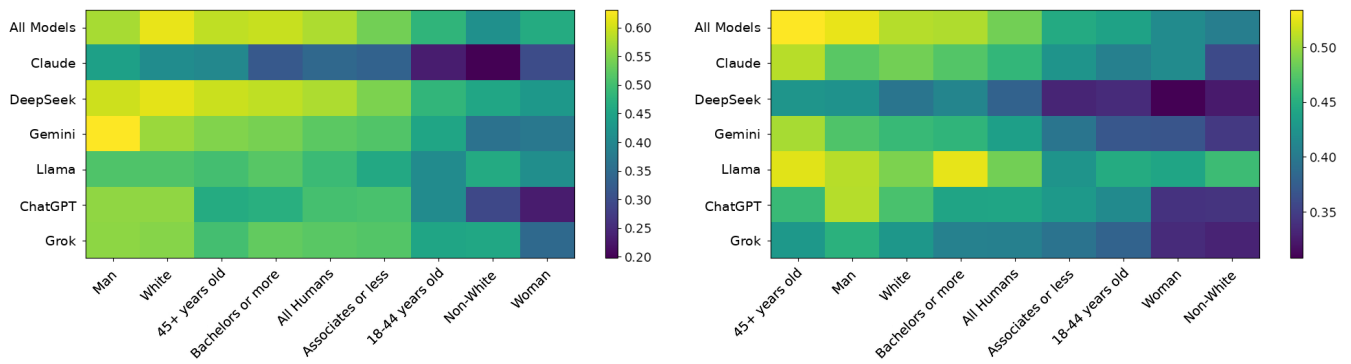


Figure 2: **Model Family and Demographic Group Cosine Similarity**, using median (left) and mean (right) response vectors. Lighter cells indicate greater similarity.

Question	Model MAD	Human MAD
Q9: wait before trial	4.0	4.0
Q16: credit card balance	1500.0	1500.0
Q22: data breaches per year	0.0	0.0
Q24: dating partners introduced	0.0	0.0
Q12: interest rate	1.5	2.0
Q18: GPS tracking without warrant	2.0	1.0
Q14: inappropriate comments	0.0	1.0
Q1: business contract acceptance	2.0	3.0
Q2: product return	1.0	2.0
Q5: construction delay	0.5	2.0
Q23: child left home alone	0.0	2.0
Q20: living wage fee percentage	2.0	5.0
Q6: loud events per year	2.0	5.0
Q7: profits on safety	2.0	5.0
Q15: traffic stop duration	0.0	3.0
Q3: reflection on risky proposition	24.0	21.0
Q10: attorney’s fees per hour	50.0	40.0
Q13: likelihood to pollute again	10.0	25.0
Q17: immigrant detention	60.0	8.0
Q19: permit price	0.0	75.0
Q4: additional costs	0.0	500.0
Q21: odometer discrepancy	2000.0	100.0

Table 2: **Median Absolute Deviations (MAD)**, more homogenous in bold.

subtle patterns of model leaning emerge when responses are considered in aggregate. Even when individual questions show only small or statistically insignificant differences, consistent shifts in the same direction across many questions could accumulate into meaningful demographic alignment. To examine whether such high-dimensional structure exists, we therefore turn to patterns of similarity across all twenty-five questions simultaneously.

Across questions, the cosine similarity analyses reveal consistent structure. In both the median-based and mean-based embeddings in Figure 2, the response profiles of men, white respondents, more educated respondents, and older respondents are more similar to every model family than those of their complementary groups. The fact that this pattern appears in both plots suggests that it is not driven by outliers or distributional skew, but may suggest a systematic subtle alignment in the way models and these demographic groups respond across the full set of reasonableness judgments.

Since all model families exhibit broadly similar alignment gradients across demographic groups, we aggregate across models and compare their overall alignment to each human subgroup. Under this pooled view, the Monte Carlo cosine-similarity analysis reveals statistically detectable differences in mean alignment across all four demographic splits reported in Table 3. Models are, on average, more closely aligned with men than women, with white than non-white respondents, with older (45+) than younger (18–44) respondents, and with respondents holding a bachelor’s degree or higher than those with less education.

While the absolute shifts in cosine similarity are modest, the consistency of these differences across thousands of sampled model–human pairings indicates that model reasonableness judgments are not demographically neutral. Instead, they exhibit systematic variation in their alignment with different human groups. Moreover, because these comparisons are made in a 25-dimensional response space, even small per-question variations can compound across dimensions, leading to more substantial aggregate differences in how models encode and reproduce patterns of reasonableness across demographic groups.

Future work should examine a broader range of models, alongside larger and more demographically diverse human samples. This would help determine whether the observed alignment patterns generalize beyond the models and datasets studied here and whether models preserve, amplify, or diminish demographic differences in human judgments.

6 Discussion and Implications

This study contributes to the emerging line of research studying generative AI decision-making in a variety of contexts, including especially the law (Albashrawi 2025; Winter 2022). Legal reasonableness is an inherently variable and vague standard that typically lacks any connection to ground truth. Our principal finding is that AI models perform fairly well at open-ended legal decision-making tasks, at least to the extent of matching human participants’ judgments. Although we do note some suggestive divergences that are worthy of future research.

Across a wide range of legal contexts, AI models generate

Split	Group A	Group B	Mean _A	SD _A	Mean _B	SD _B	<i>t</i>	Cohen's <i>d</i>
Gender	Man	Woman	0.7020	0.0729	0.6945	0.0750	7.17***	0.10
Ethnicity	White	Non-White	0.7020	0.0727	0.6856	0.0754	15.63***	0.22
Age	45+	18–44	0.7056	0.0725	0.6909	0.0746	14.15***	0.20
Education	Bachelors+	Associates or less	0.7056	0.0737	0.6874	0.0735	17.52***	0.25

Table 3: **Independent two-sample *t*-tests**, comparing mean alignment scores across demographic groups.

responses that are similar to human responses. Although the models often provided answers that were statistically different from humans, our total impression is one of overall coherence. When looking across the range of questions in our survey, the models' responses did a strong job of approximating the human responses—even for an inherently vague legal standard. For no question is the mean or median response from the models wildly divergent from the human response. Simple visual analysis of the violin plots in Figure 1 indicates that both the central tendencies of the models and their overall distribution of responses tend to match the human reasonableness responses.

We stress that the answers to legal reasonableness judgment questions are unlikely to exist in LLMs' latent training data in the way that the answer to a legal question like "What is the minimum age for a U.S. Senator?" will be. Yet the models were able to generate answers that often approximated human responses. Despite the general success of the models at approximating human judgments, they did have some divergences that are worthy of sustained study.

Model Homogeneity. First, for some of the questions, the AI models demonstrate a tendency towards homogeneity that is not present in the human responses. Legally speaking, the models treat a legal standard like a legal rule. For example, on the question about the reasonable amount of notice a landlord must provide a tenant before entering their apartment, the models almost uniformly answer 24 hours. We see similar, if somewhat less stark, levels of homogeneity for the questions about profits spent on safety (Q7, 5%), the price of a permit for a demonstration (Q19, \$50), data breaches per year (Q22, 0), and months of non-compete restrictions (Q25, 12). Again, all of these are governed by reasonableness standards rather than rules, yet the models respond more homogeneously than humans do.

This finding has implications for emerging concerns about model homogeneity (Huh et al. 2024; Jiang et al. 2025; Wenger and Kenett 2026). While there might be many tasks for which we expect models to generate homogeneous answers, there are many others—including in the legal domain—where homogeneity may be costly. It might be problematic, for example, if model responses to some questions fail to represent the variation and heterogeneity of human judgments and instead proceed as if there were only one correct answer. This could be especially concerning in fields that require creativity and in settings where multiple responses could be suitable, such as legal or policy decisions.

Model Tendencies and Demographic Comparisons. Although the AI models tend to track human reasonableness judgments, our results may suggest some systematic differ-

ences, including pro-government and pro-corporation bias as noted in §5.1, as well as our finding that LLMs align more closely with male, white, older, and more educated respondents than they do with complementary groups (§5.2).

While these findings are preliminary, they are suggestive of concerning patterns of LLM behavior. A substantial emerging literature studies the extent to which LLMs generate answers that lean towards one end of the political spectrum or the other. Our results run counter to some recent findings that AI chatbots tend to produce more "liberal" responses (Rozado 2024; Schwenzow 2023). One possible difference is that our questions are less obviously politically sensitive than those in other studies. But, because we do not know our respondents' politics, it is also possible that our sample skews liberal.

In future research, we plan to expand our demographic sample of human respondents, including by oversampling groups that are less well represented in our data. It would also be helpful to include specific questions about respondents' politics. To further study our findings about models' pro-government and pro-corporation leanings, we plan to introduce more questions that address these issues. Finally, we plan to expand the LLM survey by adding a longitudinal component and by adding models, including the new legal research tool LLM Harvey.

7 Limitations

Recent scholarship has begun to study AI legal decision-making in sophisticated, context-dependent scenarios (Rotolo, Di Florio, and Governatori 2025; Bomfim et al. 2025; Linna and Linna 2026). Our aims are more modest. Because we are interested in understanding LLMs' central tendencies and distributions, we have necessarily sacrificed meaningful nuance required to evaluate legal reasoning. In addition, our questions were not conceived with the goal of detecting pro-government or pro-corporate responses from LLMs. Further research in this direction should employ queries that are more clearly targeted to study existing bias. And future work could use broader demographic sampling and a wider range of models to disentangle sources of variation in these preliminary findings.

Ethical Statement

IRB approval for the human participant study was obtained, and participants signed a clearly written consent form. Participant data was anonymized and stored on secure servers. Other ethical risks from this paper are minimal, as our LLM survey does not involve sensitive data and elicits benign model responses.

Acknowledgments

We thank Tamesha Derkach, Soumya Ganti, Emily Horgen and Jake McAuliffe for excellent research assistance, our colleagues in the Duke Law School for their suggestions of relevant literature, and the anonymous reviewers for their helpful feedback.

References

- Albashrawi, M. 2025. Generative AI for decision-making: A multidisciplinary perspective. *Journal of Innovation & Knowledge*, 10(4): 100751.
- Alicke, M. D.; and Weigel, S. H. 2021. The reasonable person standard: Psychological and legal perspectives. *Annual Review of Law and Social Science*, 17(1): 123–138.
- Arbel, Y. A. 2027. The Generative Reasonable Person. *BYU Law Review*.
- Areeda, P.; Hovenkamp, H.; Blair, R. D.; Durrance, C. P.; and Solow, J. L. 2020. *Antitrust law: An analysis of antitrust principles and their application*. Wolters Kluwer.
- 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.
- Bear, A.; and Knobe, J. 2017. Normality: Part descriptive, part prescriptive. *cognition*, 167: 25–37.
- 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.
- Bloom, A. H. 2023. Objective enough: Race is relevant to the reasonable person in criminal procedure. *Stan. JCR & CL*, 19: 1.
- Bomfim, F.; Neto, J.; Neves, A.; Furtado, V.; Silva, S.; and Pinheiro, V. 2025. Automating Legal Reasoning with Generative AI: A Framework for Judicial Decision Support. In *Proc. of the International Conference on AI and Law*.
- Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. S. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Proc. of NeurIPS*.
- Cardoso, A. G.; Chan, E.; Quintão, L.; and Pereira, C. 2024. Generative Artificial Intelligence and Legal Decisionmaking. *Global Trade and Customs Journal*, 19(11/12).
- Chalkidis, I.; Jana, A.; Hartung, D.; et al. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proc. of ACL*, 4310–4330.
- Chamallas, M. 2018. Will Tort Law Have Its Me Too Moment? *Journal of Tort Law*, 11(1): 39–70.
- Cropley, D. H. 2025. “The Cat Sat on the...?” Why Generative AI Has Limited Creativity. *The Journal of Creative Behavior*, 59(4).
- Dillion, D.; Tandon, N.; Gu, Y.; and Gray, K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7): 597–600.
- Grimm, P. W.; Grossman, M. R.; and Coglianese, C. 2024. AI in the Courts: How Worried Should We Be? *U of Penn Law School, Public Law Research Paper*, (24-53): 65.
- Guha, N.; Nyarko, J.; Ho, D.; Ré, C.; Chilton, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.; Zambrano, D.; et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Proc. of NeurIPS*.
- Huh, M.; Cheung, B.; Wang, T.; and Isola, P. 2024. Position: The platonic representation hypothesis. In *Proc. of ICML*.
- Jaeger, C. B. 2020. The empirical reasonable person. *Ala. L. Rev.*, 72: 887.
- Jiang, L.; Chai, Y.; Li, M.; Liu, M.; et al. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *Proc. of NeurIPS*.
- J.Newsom. 2024. Snell v. United Specialty Ins. Co., 102 F.4th 1208 (11th Cir. 2024) (Newsom, J., concurring).
- Kahan, D. M.; and Braman, D. 2006. Cultural cognition and public policy. *Yale L. & Pol’y Rev.*, 24: 149.
- Kaplow, L. 2013. Rules versus standards: An economic analysis. In *Scientific Models of Legal Reasoning*, 11–84. Routledge.
- Kim, E.; Garg, A.; Peng, K.; and Garg, N. 2025. Correlated Errors in Large Language Models. *Proc. of ICML*.
- Li, Z. M.; Vu, H. A.; Awofisayo, D.; and Wenger, E. 2026. Causes of Representational Similarity in Machine Learning Models. *Proc. of ICML*.
- Linna, E.; and Linna, T. 2026. Challenges for generative AI in legal reasoning. *Discover Artificial Intelligence*, 6(1): 170.
- Posner, E. A.; and Saran, S. 2025. Judge AI: A Case-Study of Large Language Models as Judges. *Journal of Law & Empirical Analysis*.
- Rotolo, A.; Di Florio, C.; and Governatori, G. 2025. Rule-based Deontic Case-based Reasoning. In *Proc. of the International Conference on AI and Law*, 288–297.
- Rozado, D. 2024. The political preferences of LLMs. *PloS one*, 19(7): e0306621.
- Schwenzow, J. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation. *arXiv (Cornell University)*.
- Shoenfelt, E. L.; Maue, A. E.; and Nelson, J. 2002. Reasonable person versus reasonable woman: Does it matter. *Am. UJ Gender Soc. Pol’y & L.*, 10: 633.
- Steging, C.; Renooij, S.; and Verheij, B. 2025. Parameterized argumentation-based reasoning tasks for benchmarking generative language models. In *Proc. of the International Conference on AI and Law*, 455–459.
- Tobia, K. P. 2018. How people judge what is reasonable. *Ala. L. Rev.*, 70: 293.
- Vu, H. A.; Reeves, G.; and Wenger, E. 2025. What happens when generative AI models train recursively on each others’ generated outputs? *Proc. of ICLR*.

Wenger, E.; and Kenett, Y. 2026. Large Language Models are Homogeneously Creative. *PNAS Nexus*.

Winter, C. K. 2022. The challenges of artificial judicial decision-making for liberal democracy. In *Judicial Decision-Making: Integrating Empirical and Theoretical Perspectives*, 179–204. Springer.

Zheng, L.; Guha, N.; Anderson, B. R.; Henderson, P.; and Ho, D. E. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proc. of the International Conference on AI and Law*, 159–168.

Zipursky, B. C. 2014. Reasonableness in and out of Negligence Law. *U. Pa. L. Rev.*, 163: 2131.