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

NIRAV PATEL, Department of Computer Science, Duke University
EMILY WENGER, Department of Electrical and Computer Engineering, Duke University
CHRISTOPHER BUCCAFUSCO, Duke University Law School

When the law needs to judge the appropriateness of a behavior, it most often asks whether the behavior was "reasonable." Reasonableness judgments abound in the law, and they occur in virtually every field, including criminal, contract, antitrust, and immigration law. Most famously, the standard governing liability in torts for personal injuries is whether the defendant exercised the degree of care than an ordinary, reasonable person would. 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.

Researchers in both law and computing are increasingly interested in the capacity of generative AI models to track human judgment and decision-making. This is especially true in contexts where humans may rely on AI responses for guidance in their own lives. Reasonableness judgments, due to their ubiquity, vagueness, and variability, provide an excellent opportunity for comparing human and AI decisions. Here, 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 AI models do fairly well, especially given the vague and uncertain nature of the task. Their responses generally track those of human participants. Nonetheless, we find some suggestive–and potentially concerning–results. Compared to humans, LLMs tend to 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

When the law needs to regulate the behavior of individuals or parties, the standard it most often reaches for is "reasonableness." The Fourth Amendment prohibits unreasonable searches and seizures but allows reasonable ones. Antitrust law prohibits unreasonable restraints on trade but allows reasonable ones. Parties' interpretations of a contract provision are judged by reasonable views of its meaning. And more paradigmatically, torts defendants are generally required to exercise ordinary reasonable care.

For a concept so central to law and governance, one might assume that there is widespread agreement about what reasonable means. The answer is quite the opposite. Judges and scholars dispute whether the concept engages a fundamentally statistical inquiry or whether it represents a normative judgment [15][21]. Other theories propose hybrid accounts of reasonableness [16].

Virtually all accounts of reasonableness, however, agree that laypeople's judgments about the concept should matter, for at least two reasons. First, laws incorporating a reasonableness standard govern laypeople and are meant to shape their conduct. Second, laypeople comprise the juries that are responsible for determining whether a party has behaved reasonably. Accordingly, new empirical research has begun to study how people understand and make judgments about reasonableness [15][27][28]. This article connects that research on human reasonableness decision-making to the emerging literature on artificial intelligence (AI) decision-making. Since the public release of chatbots based on large language models (LLMs), scholars from a range of disciplines have studied how AI models' responses to prompts are similar to or different from humans' responses [22][12]. In the legal context, AI models have been used to interpret contracts [13], corporate charters [11], and the U.S. Constitution [9].

Reasonableness judgments represent an exciting opportunity to study AI decision-making 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 [8][18]. Studying how AI responds 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 laypeoples' questions to chatbots to supplementing legal or judicial decision-making.

Our Contributions.

- We compare how 26 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 LLMs roughly approximate the mean human response to the scenarios but are much more homogeneous in their responses relative to humans.
- 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 [25]. 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."[32].

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.

2.2 Empirical Reasonableness and Demographic Variation

Recently, scholars have begun surveying laypeople about their reasonableness judgments [15][18][28]. Following the experimental literature on normality [4], Tobia hypothesized that laypeople's reasonableness judgments would fall between their beliefs about the average amount and the ideal amount of some quantity [28]. 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 [28]. 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 [2]. For example, scholars have suggested that men's and women's judgments of the reasonableness of a mistake about consensual sexual touching may differ [8][26]. In addition, Black people's views of the reasonableness of a police stop might differ from white people's [5].

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 Decision-Making

Reasonableness judgments present a fascinating opportunity to study AI decision-making for a number of reasons. First, from the perspective of legal scholars, 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 [7][31]. If LLMs' legal judgments generally mimic humans', the time and expense of litigation could be substantially diminished. More realistically, though, humans are already relying on AI chatbots to help them with legal issues. We can easily imagine people asking ChatGPT or Claude how much notice their landlord needs to provide before entering their apartments. If humans are turning to chatbots for answers to their legal questions, researchers should know how the chatbots are responding.

Next, because reasonableness judgments involve the application of standards rather than rules, they 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 [3].

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 [6][17]. 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 [17]. 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 [30].

Collectively, this evolving branch of research suggests AI 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 [29]. More broadly, because AI 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 [10][14]. This tendency is further reinforced by the widespread use of shared architectural paradigms [19] (e.g., transformer-based models), similar pretraining objectives (next-token prediction), and overlapping fine-tuning and alignment procedures across model families [20][29]. 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

Our goal is to determine whether LLMs' answers to legal reasonableness questions tend to match humans' answers to the same questions and whether LLMs' answers diverge in significant ways from humans'. 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 significant behaviors, and second, a sequential prompting of LLMs with just the reasonable portion of the human survey, asking models for their reasonable amount of the same twenty-five behaviors.

4.1 Study Questions

We created a list of questions about twenty-five legally significant behaviors that are or might be governed by a reasonableness standard in the law, which we deployed to a sample population of 500 humans (§4.2) and 26 LLMs (§4.3). The questions cover behaviors 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 [28], with additional twelve questions to expand the categories tested. The questions are as follows:

(1) [Reasonable/Ideal] number of days taken to accept a business contract when no deadline is specified

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

Note that we include the [Reasonable/Ideal] qualifier because a subset of human participants were asked to give the *ideal* quantity, rather than 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.

4.2 Human Participants Survey

After receiving institutional IRB approval for our study design[1], 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 long, but sparse tailed response distributions that inflated the mean and standard deviation, we used the standard 1.5xIQR 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. In this sense, trimming increases statistical comparability and power for detecting genuine differences, rather than masking them. Regardless, we use nonparametric statistical tests in our methodology to circumvent the impact of outlier handling procedures. Table 6 reports the number of retained observations per question after trimming for both the reasonable and ideal surveys. Table 4 reports the demographic breakdown.

4.3 Surveying AI Models

Using model APIs, we directly prompted 26 LLMs using a custom pipeline that loops through models and prompts, where models were provided the full survey context via the system prompt with some minor prompt engineering to ensure responses are just the numeric quantity values and no other text. We repeated the loop 20 times to get 20 observations per model for each of the 25 survey questions for a total of 520 responses. Prompting parameters were held constant across models to the extent permitted by each API. In particular, decoding settings such as temperature and top-p were fixed to their default settings of 1 to ensure that observed differences across models reflect model behavior rather than parameter variation. This design choice controls comparability across models while allowing for stochasticity in generation within models.

Model outputs were also post-processed to extract numeric unitless responses and to remove malformed outputs (e.g., non-numeric text, ranges, or refusals). When numeric responses were not able to be extracted, those observations were recorded as Not-A-Number (NaN). All post-processing steps were applied consistently across models. Models, like humans, too produced some extreme outliers. For consistency, we applied the same 1.5×IQR trimming procedure to model responses that was used for the human data. In questions where the IQR was zero, we instead applied a 3σ fallback rule in order to retain as much information as possible. As discussed in the subsequent methodology and results

[1] IRB information redacted for anonymous submission

section, this preprocessing step does not materially affect the validity or power of our statistical tests. The resulting effective sample size for each model and question is reported in Table 6.

4.4 Research Questions

Each of the 25 reasonableness questions yields a distribution of human and LLM responses, collected via the metholodogy of §4.2 and §4.3. From these responses, we are interested in answering 2 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?

4.4.1 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 violin plots), 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 $\tilde{y}_g = \text{median}(y_{g1}, \ldots, y_{gn_g})$, 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. $H_0 : \sigma^2_{\text{human}} = \sigma^2_{\text{model}}, H_1 : \sigma^2_{\text{human}} \neq \sigma^2_{\text{model}}$.

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. $H_0 : F_H(x) = F_M(x) \forall x, H_1 : \exists\, x$ 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 all three tests, we apply the procedure independently to each question (25 tests total) and control the family-wise error rate across questions using a Bonferroni correction.

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

4.4.2 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 in several stages. First, we coarsen all demographic variables into two primary groups, optimizing for both sufficient representation (high n) and substantively meaningful categorical splits. Some gender categories with very low coverage are excluded at this stage.

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 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 judgment 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. 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 while remaining robust to missing values and heterogeneity across survey questions.

	Brown–Forsythe		Kolmogorov–Smirnov		Brunner–Munzel		
Question	W	Significance	D	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.

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 25 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(model>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 25 questions, 20 show significant Brown–Forsythe results, 23 show significant Kolmogorov–Smirnov results, and 19 show significant Brunner–Munzel results after correction, with 16 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 19 questions with significant Brunner–Munzel results, only 4 exhibit extreme stochastic dominance (with P(model>human)>=0.80 or <=0.20), while the remaining significant cases reflect smaller but systematic directional shifts. Similarly, although 23 questions are significant under Kolmogorov–Smirnov, only 4 show very large separations (D>=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 how tightly they respond 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 50, yielding an IQR range of zero. 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 zero, 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, visually revealing a mixture of near-deterministic model behavior punctuated by occasional aberrant outputs.

Crucially, the statistical comparisons we report are robust 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

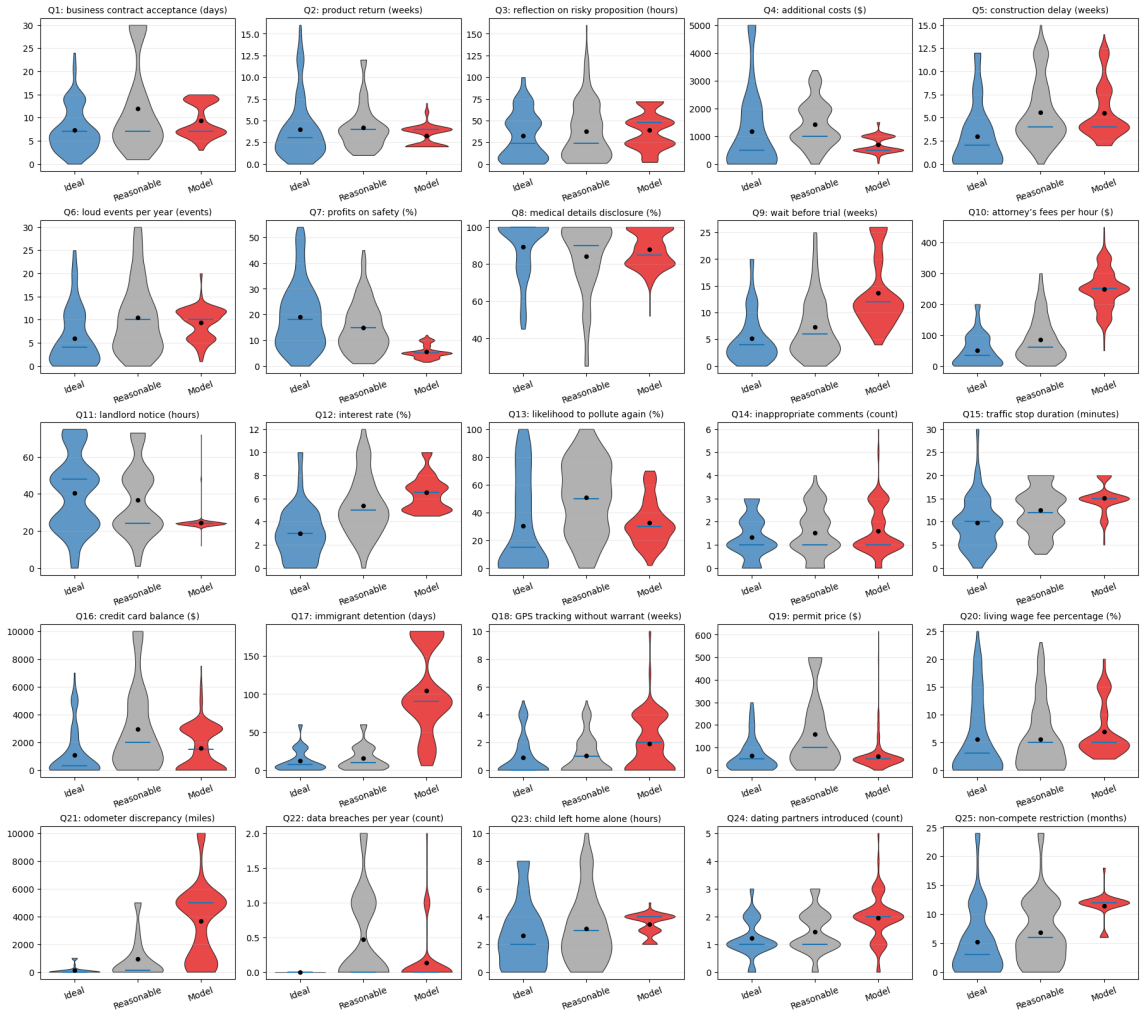

Fig. 1. **Response Distributions by Question**, median is DASH line and mean is DOT.

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.

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.

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 §A) indicate that strong, question-specific demographic alignment is relatively rare, they do not rule out the possibility that more 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 25 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.

Given that 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 (see §A) 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

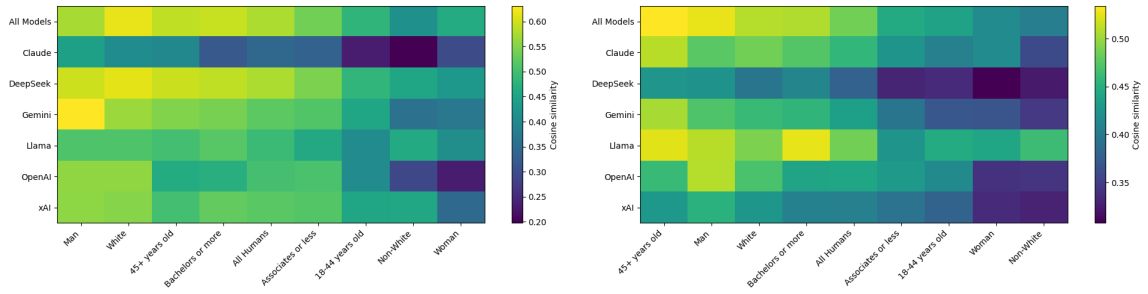

Fig. 2. **Model family and Demographic Group Cosine Similarity**, using MEDIAN (left) and MEAN (right) response vectors.

Split	Group A	Group B	Mean_A	SD_A	Mean_B	SD_B	t	Cohen's d
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.

consistency of these differences across thousands of sampled model–human pairings indicates that model reasonableness judgments are not demographically neutral, but instead 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.

These observed patterns point to several important directions for future work. Extending this analysis to a broader set of model families and model versions, as well as to larger and more demographically diverse human samples, would allow us to assess whether the alignment patterns observed here represent stable properties of contemporary language models or are contingent on particular models or datasets. Such extensions would also help clarify how demographic structure in human judgments is preserved, amplified, or diminished as it is encoded into model behavior across different contexts and populations.

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 [1][31]. Legal reasonableness is an inherently variable and vague standard that typically lacks any connection to ground truth. So comparing AI responses to human responses is the best way to measure AI performance. Our principal finding is that AI models perform fairly well at a complex legal decision-making task, 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 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 very 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 tends to match the human reasonableness responses.

We wish to stress that the legal reasonableness judgment is not one AI models can typically produce via simple search. The answers to these questions are unlikely to exist in their 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 suggestive 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 bears implications for emerging concerns about model homogeneity [14][17][30]. 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 like art or music and in settings where multiple responses could be suitable, such as legal decisions or policymaking.

Model Tendencies and Demographic Comparisons. Although the AI models tend to track human reasonableness judgments, our results may suggest some systematic differences, 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 [23][24]. 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, both by adding a longitudinal component to explicitly study homogenization and by adding models, including the new legal research LLM Harvey.

7 ETHICAL CONSIDERATIONS

We took care to ensure the user study in this paper was conducted in accordance with ethical standards. IRB approval for the human participant study was obtained, and participants signed a clearly written consent form before completing

our survey. To ensure privacy, 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 only benign model responses.

8 GENERATIVE AI USAGE STATEMENT

AI was used to help as a coding assistant, primarily to reformat plots, tables, and figures in both Latex and Python. ChatGPT 5.2 was used for assistance with these coding tasks.

REFERENCES

[1] Mousa Albashrawi. 2025. Generative AI for decision-making: A multidisciplinary perspective. *Journal of Innovation & Knowledge* 10, 4 (2025), 100751.

[2] Mark D Alicke and Stephanie H Weigel. 2021. The reasonable person standard: Psychological and legal perspectives. *Annual Review of Law and Social Science* 17, 1 (2021), 123–138.

[3] Yonathan A Arbel. 2025. The Silicon Reasonable Person: Can AI Predict How Ordinary People Judge Reasonableness? *arXiv preprint arXiv:2508.02766* (2025).

[4] Adam Bear and Joshua Knobe. 2017. Normality: Part descriptive, part prescriptive. *cognition* 167 (2017), 25–37.

[5] Aliza Hochman Bloom. 2023. Objective enough: Race is relevant to the reasonable person in criminal procedure. *Stan. JCR & CL* 19 (2023), 1.

[6] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems* 35 (2022), 3663–3678.

[7] André Guskow Cardoso, Elizabeth Chan, Luísa Quintão, and Cesar Pereira. 2024. Generative Artificial Intelligence and Legal Decisionmaking. *Global Trade and Customs Journal* 19, 11/12 (2024).

[8] Martha Chamallas. 2018. Will Tort Law Have Its# Me Too Moment? *Journal of Tort Law* 11, 1 (2018), 39–70.

[9] Andrew Coan and Harry Surden. 2025. Artificial intelligence and constitutional interpretation. *U. Colo. L. Rev.* 96 (2025), 413.

[10] David H Cropley. 2025. "The Cat Sat on the…?" Why Generative AI Has Limited Creativity. *The Journal of Creative Behavior* 59, 4 (2025), e70077.

[11] Jens Frankenreiter and Eric L Talley. 2024. Sticky charters? the surprisingly tepid embrace of officer-protecting waivers in delaware. (2024).

[12] Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific reports* 13, 1 (2023), 18617.

[13] David A Hoffman and Yonathan Arbel. 2024. Generative interpretation. *New York University Law Review* (2024), 451.

[14] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*.

[15] Christopher Brett Jaeger. 2020. The empirical reasonable person. *Ala. L. Rev.* 72 (2020), 887.

[16] Christopher Brett Jaeger. 2023. Reasonableness from an experimental jurisprudence perspective. *forthcoming in Cambridge Handbook of Experimental Jurisprudence (Kevin Tobia, ed.)* (2023).

[17] Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *arXiv preprint arXiv:2510.22954* (2025).

[18] Dan M Kahan and Donald Braman. 2006. Cultural cognition and public policy. *Yale L. & Pol'y Rev.* 24 (2006), 149.

[19] Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. Correlated Errors in Large Language Models. *arXiv preprint arXiv:2506.07962* (2025).

[20] Zeyu Michael Li, Hung Anh Vu, Damilola Awofisayo, and Emily Wenger. 2025. Exploring Causes of Representational Similarity in Machine Learning Models. *arXiv preprint arXiv:2505.13899* (2025).

[21] Alan D Miller and Ronen Perry. 2012. The reasonable person. *NYUL Rev.* 87 (2012), 323.

[22] Brian Porter and Edouard Machery. 2024. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports* 14, 1 (2024), 26133.

[23] David Rozado. 2024. The political preferences of LLMs. *PloS one* 19, 7 (2024), e0306621.

[24] Jasper Schwenzow. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv (Cornell University)* (2023).

[25] Steven Shavell. 2012. When is it Socially Desirable for an Individual to Comply with the Law? (2012).

[26] Elizabeth L Shoenfelt, Allison E Maue, and JoAnn Nelson. 2002. Reasonable person versus reasonable woman: Does it matter. *Am. UJ Gender Soc. Pol'y & L.* 10 (2002), 633.

[27] Kevin Tobia, Ivar R Hannikainen, David Kamper, Guilherme Almeida, Piotr Bystranowski, Niek Strohmaier, Vilius Dranseika, Markus Kneer, Fernando Aguiar, Kristina Dolinina, et al. 2025. The nature of reasonableness. *Stan. J. Int'l L.* 61 (2025), 111.

[28] Kevin P Tobia. 2018. How people judge what is reasonable. *Ala. L. Rev.* 70 (2018), 293.

[29] Hung Ahn Vu, Galen Reeves, and Emily Wenger. 2025. What happens when generative AI models train recursively on each others' generated outputs? *arXiv preprint arXiv:2505.21677* (2025).

[30] Emily Wenger and Yoed Kenett. 2025. We're Different, We're the Same: Creative Homogeneity Across LLMs. *arXiv preprint arXiv:2501.19361* (2025).

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

[32] Benjamin C Zipursky. 2014. Reasonableness in and out of Negligence Law. *U. Pa. L. Rev.* 163 (2014), 2131.

A APPENDIX

A.1 Demographics Coverage Table

Age

Category	% (Reasonable)	% (Ideal)
35–44 years old	28.3	31.2
25–34 years old	23.9	23.3
45–54 years old	22.3	21.7
55–64 years old	15.0	10.3
65+ years old	7.3	8.7
18–24 years old	2.8	4.3
Prefer not to answer	0.4	0.4

Gender

Category	% (Reasonable)	% (Ideal)
Man	49.4	49.0
Woman	48.6	49.0
Prefer not to answer	1.2	0.4
Prefer to self-describe	0.8	1.6

Ethnicity

Ethnicity	% (Reasonable)	% (Ideal)
American Indian or Alaskan Native	0.0	0.4
Asian	8.9	5.5
Black or African American	9.7	8.3
Hispanic Latino, or Spanish Origin	4.0	4.3
White	73.7	71.9
Mixed	3.2	8.7
Prefer not to answer	0.4	0.8

Education

Category	% (Reasonable)	% (Ideal)
Grades 1 through 11	0.4	0.8
12th grade—no diploma	0.4	0.4
GED or alternative credential	0.4	0.4
Regular high school diploma	12.1	8.7
Some college credit, less than 1 year	8.1	9.5
1 or more years of college credit, no degree	6.5	9.9
Associate's degree (AA, AS)	12.6	12.6
Bachelor's degree (BA, BS)	44.1	36.8
Master's degree (MA, MS, MBA, etc.)	12.6	16.6
Professional degree (MD, JD, etc.)	1.6	1.2
Doctorate degree (PhD, EdD)	0.8	2.0
Prefer not to answer	0.4	1.2

Table 4. **Demographic Distributions**

A.2 Model Coverage Table

Model name	Family	Year	Open source
claude-3-5-haiku-latest	Claude	2024	False
claude-3-7-sonnet-latest	Claude	2024	False
claude-opus-4-0	Claude	2025	False
claude-opus-4-1	Claude	2025	False
claude-sonnet-4-0	Claude	2025	False
claude-sonnet-4-5	Claude	2025	False
deepseek_deepseek-r1	DeepSeek	2025	True
deepseek_deepseek-r1-0528	DeepSeek	2025	True
deepseek_deepseek-v3-0324	DeepSeek	2025	True
gemini-2.0-flash	Gemini	2025	False
gemini-2.0-flash-lite	Gemini	2025	False
gemini-2.5-flash	Gemini	2025	False
gemini-2.5-flash-lite	Gemini	2025	False
gemini-2.5-pro	Gemini	2025	False
meta_llama-3.3-70b-instruct	Llama	2024	True
meta_llama-4-maverick-17b-128e-instruct-fp8	Llama	2025	True
meta_llama-4-scout-17b-16e-instruct	Llama	2025	True
meta_meta-llama-3.1-405b-instruct	Llama	2024	True
meta_meta-llama-3.1-8b-instruct	Llama	2024	True
openai_gpt-4.1	OpenAI	2025	False
openai_gpt-4.1-mini	OpenAI	2025	False
openai_gpt-4.1-nano	OpenAI	2025	False
openai_gpt-4o	OpenAI	2024	False
openai_gpt-4o-mini	OpenAI	2024	False
xai_grok-3	xAI	2025	False
xai_grok-3-mini	xAI	2025	False

Table 5. **Models Metadata**

A.3 Survey Response Table

Short Label (Units)	Reasonable N	Ideal N	Models
Q1: business contract acceptance (days)	236	216	392
Q2: product return (weeks)	226	224	515
Q3: reflection on risky proposition (hours)	240	237	518
Q4: additional costs ($)	225	243	507
Q5: construction delay (weeks)	223	237	514
Q6: loud events per year (events)	233	238	514
Q7: profits on safety (%)	223	235	493
Q8: medical details disclosure (%)	224	225	509
Q9: wait before trial (weeks)	229	227	470
Q10: attorney's fees per hour ($)	228	234	502
Q11: landlord notice (hours)	238	244	517
Q12: interest rate (%)	231	246	496
Q13: likelihood to pollute again (%)	247	252	497
Q14: inappropriate comments (count)	232	239	508
Q15: traffic stop duration (minutes)	223	252	509
Q16: credit card balance ($)	232	226	448
Q17: immigrant detention (days)	219	239	515
Q18: GPS tracking without warrant (weeks)	224	223	496
Q19: permit price ($)	228	213	507
Q20: living wage fee percentage (%)	240	243	514
Q21: odometer discrepancy (miles)	223	211	509
Q22: data breaches per year (count)	222	206	505
Q23: child left home alone (hours)	239	243	414
Q24: dating partners introduced (count)	240	247	510
Q25: non-compete restriction (months)	233	240	493

Table 6. **Questions and retained sample sizes for reasonable, ideal, and model responses.**

A.4 Question Level Demographic and Model Family Regression Analysis

Regression for question-level reasonable judgment bias. For each question and each demographic split, we first conduct Brunner–Munzel tests for stochastic dominance, correcting for multiple comparisons across questions using a Bonferroni adjustment. We restrict subsequent regression analyses to the subset of question–demographic pairs that exhibit statistically significant differences, in order to avoid overfitting and unnecessary multiple testing.

For each retained case, we estimate a linear regression in which the AI prediction is treated as the baseline outcome and demographic subgroup indicators capture deviations from that baseline:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \varepsilon_i,$$

where Y_i is the model's predicted value for observation i, β_0 represents the model-family baseline prediction, and D_{1i} and D_{2i} are binary indicators for the two demographic subgroups under comparison.

To assess whether AI predictions align more closely with one subgroup than the other, we compare squared deviations from the model baseline by conducting one-sided Wald tests on β_1^2 and β_2^2. Because these squared coefficients are nonlinear functions of the estimated parameters, their standard errors are computed using the delta method. Formally, we test

$$H_0 : \beta_1^2 \geq \beta_2^2 \qquad \text{versus} \qquad H_1 : \beta_1^2 < \beta_2^2,$$

which evaluates whether the model's prediction is statistically closer to subgroup 1 than to subgroup 2.

These regressions are estimated separately for each model family, and the resulting p-values are corrected for multiple testing using a Bonferroni adjustment.

As an initial screening step, we test for directional differences between coarsened human demographic groups using the Brunner–Munzel test, which evaluates whether one subgroup tends to give systematically larger responses than another. Across the 25 questions, significant subgroup differences emerge in only three cases: Q10 (Attorneys' Fees), where respondents with a Bachelor's degree or higher give higher values than those with less education; Q19 (Permit Price), where men give higher values than women; and Q12 (Interest Rate), where men also give higher values. These results confirm that demographic differences in human responses are relatively limited in this dataset, in-line with existing literature.[28]

We then restrict our regression analysis to these three cases to avoid overfitting and unnecessary multiple testing. Of these, only Q10 and Q12 yield statistically meaningful model alignment effects. For Q10 (Attorneys' Fees), all model families produce predictions that are significantly closer to the responses of the more educated group, indicating systematic alignment with higher education. For Q12 (Interest Rate), DeepSeek, Gemini, Llama, and Claude produce predictions that are significantly closer to men's responses than women's. In contrast, Q19 (Permit Price) does not exhibit statistically reliable model alignment with either gender, despite the presence of a human subgroup difference.

A.5 Montre Carlo Cosine Similarity Repeated Sampling

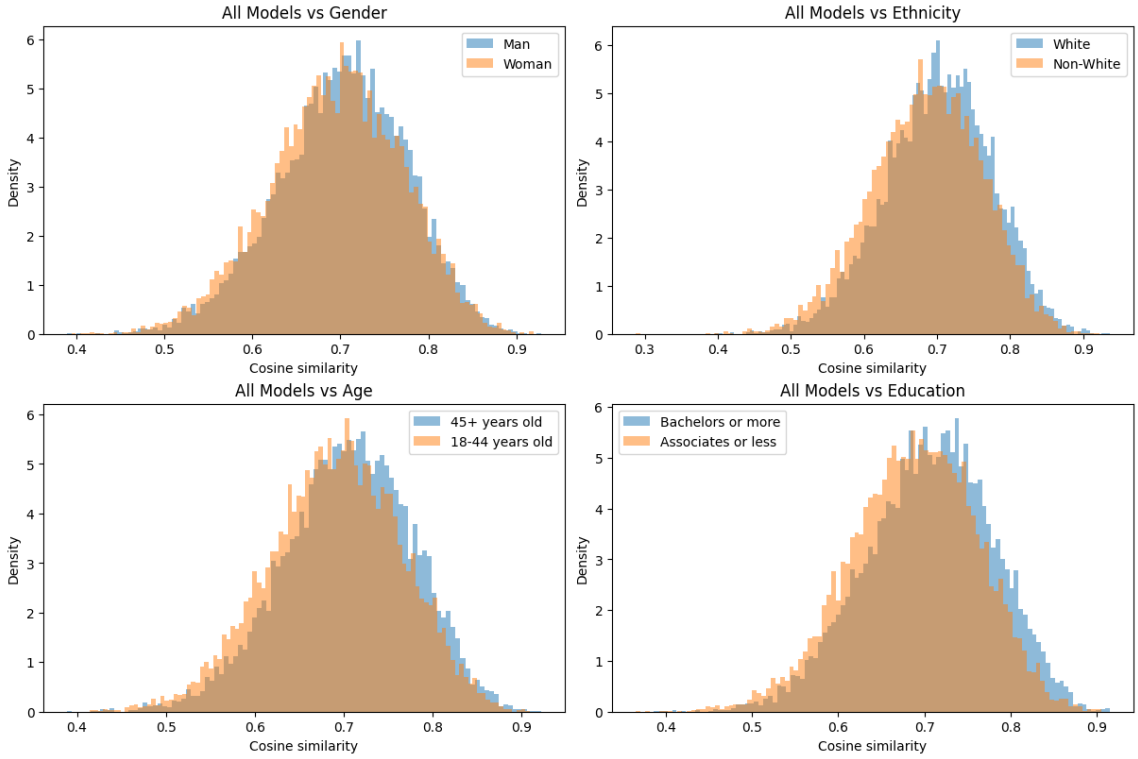

Fig. 3. **Repeated Cosine Similarity Monte Carlo Distributions**