UNT Dallas Law Review
ON THE CUSP

VOLUME 8 | ISSUE 2 | MARCH 2026

MICHAEL SCOTT IS NOT A JUROR: THE LIMITS OF AI IN SIMULATING HUMAN JUDGMENT

Sean Harrington & Hayley Stillwell

CONTENTS

MICHAEL SCOTT IS NOT A JUROR: THE LIMITS OF AI IN SIMULATING HUMAN JUDGMENT[1]

Sean Harrington[2] ***& Hayley Stillwell***[3]

ABSTRACT

Can AI replace human jurors? More specifically, can large language models predict how jurors interpret evidence and reach decisions based on legally salient facts and demographic characteristics? As legal scholars and practitioners increasingly explore AI-generated jury simulations, this Article offers the first empirical test of whether models like GPT-4, Claude, and Gemini can faithfully replicate juror reasoning. The answer, for now, is no. Across a series of mock trial scenarios involving redacted confessions, GPT-4, Claude, and Gemini repeatedly failed to replicate how real jurors interpret evidence or exercise judgment. Their errors were not random, but systematic. Hidden prompts, built-in content filters, and demographic flattening produced distortions that cut across sex, ethnicity, political affiliation, economic status, and education level.

Yet the promise of simulation remains within reach. In the second phase of the study, we fine-tuned an open-source model on actual mock juror data, achieving significant gains in accuracy and alignment. Although today's LLMs fall short of simulating juror reasoning, models refined through transparent methods and real human data could assist judges in applying evidentiary standards, help researchers test doctrinal assumptions, and give trial lawyers new tools for strategic decision-making. This paper maps the risks and outlines a path toward responsible AI-based jury simulation.

[1] This research was supported by a grant from the Research Council of the University of Oklahoma Norman Campus.

[2] Sean Harrington is the Director of Technology and Innovation at the University of Oklahoma College of Law.

[3] Hayley Stillwell is an Associate Professor of Law at the University of Oklahoma College of Law.

INTRODUCTION

What if artificial intelligence could replicate the judgments of real jurors? With the growing influence of large language models (LLMs) like OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini, it is increasingly plausible to envision a future where researchers, lawyers, or courts could simulate a jury deliberation or predict how evidence might land with a particular juror, adjusting for factors like ethnicity, income, political affiliation, and education. That was the starting point for our project. We set out to test whether existing AI platforms could act like real jurors by generating responses tailored to realistic demographic profiles and evaluating criminal trial evidence in a way that reflected ordinary human reasoning.

Very quickly, we ran into problems. When we asked some platforms to rate how incriminating a piece of evidence was "as if you were a juror," they refused, insisting it would be giving legal advice. When the platforms did respond, the results were bizarre. One platform gave us a 102-year-old juror who was a part-time marine biologist and part-time DJ. Another claimed to have generated a demographically appropriate juror who turned out to be a white male in his 40s from Scranton, Pennsylvania, working as a manager at a mid-sized paper company. It was not until we saw the name Michael Scott that we realized the model had cast the lead character from *The Office*.[4] These were not just random glitches. They revealed more systemic issues: the platforms appeared to be rewriting or filtering our prompts according to internal rules we could not see, shaped by hidden instructions, safety layers, and unobservable constraints on what the models are "allowed" to generate.

To evaluate how far off these platforms really were, we decided to test them against reality. We conducted a mock jury study, recruiting hundreds of human participants to rate the incriminating strength of four different pieces of criminal evidence. We collected detailed demographic data from each participant, including sex, ethnicity, political affiliation, income bracket, and education level. Then we prompted GPT-4.1, Claude-Sonnet-4, and Gemini 2.5 to generate juror responses based on those exact demographic profiles. This allowed us to measure not just whether the platforms could complete the task, but how closely their outputs mirrored human reasoning and whether their performance changed depending on whose identity they were asked to simulate.

---

[4] Sean Harrington & Hayley Stillwell, Michael Scott is Not a Juror: Final Simulation Data, https://github.com/Digital-Initiative-OU-Law/JurorData (Sean Harrington, JurorData, GitHub) (last modified Jul. 31, 2025) (on file with the UNT Dallas Law Review).

The results revealed consistent misalignment between AI-generated and human mock juror ratings. On average, human mock jurors deviated from the group mean by 2.03 points on a ten-point scale. GPT-4.1 performed modestly worse, with a mean absolute error of 2.52. Claude and Gemini showed larger divergences, with errors of 3.36 and 2.93, respectively. But inaccuracy was only part of the story. The models also exhibited directional bias. Claude consistently underrated the strength of incriminating evidence, with a mean signed error of –2.34, while Gemini tended to overrate it, with a signed error of +0.81. When we disaggregated the results by demographic group, clear patterns emerged. Some platforms consistently performed worse when simulating jurors who were low-income, less-educated, or from ethnically marginalized groups.[5] Put simply, all three platforms flailed: GPT-4.1 less catastrophically than Claude or Gemini, but still meaningfully worse than human baselines.

This article uses that comparative study to explore the structural and technical barriers that prevent off-the-shelf AI platforms from faithfully simulating human behavior. We argue that these failures stem in part from training processes that overrepresent some viewpoints while flattening others, from hidden system prompts, and from guardrails that distort outputs, and from the content moderation layers that prevent platforms from responding naturally to legal or ethically sensitive questions. We also describe the early steps we have taken to address these limitations through fine-tuning. By training a new model directly on our human mock juror dataset, we have begun to close the gap in accuracy and, auspiciously, our preliminary results suggest that further data and iteration could produce a model that outperforms existing commercial platforms at approximating real juror reasoning.

Before turning to our findings, Part II situates this study within the broader landscape of AI-based legal simulation, describing how large language models have been used to approximate human judgment and highlighting the doctrinal relevance of modeling the "reasonable juror." The remainder of this Article proceeds in four parts, each addressing a central question raised by our study. Part III assesses feasibility: whether it is possible to reliably prompt GPT-4.1, Claude-Sonnet-4, and Gemini-2.5 to simulate realistic juror pools using demographic variables alone. Part IV evaluates performance, comparing each platform's accuracy and bias against the responses of real human mock jurors using two core metrics: mean

[5] *See infra* Part V Sections B–C; *see also infra* Tables 3, 5, & 6 (showing larger signed errors when simulating Black, Hispanic, low-income, and less-educated jurors, particularly for Claude and Gemini).

absolute error (MAE) and mean signed error. Part V explores likely failure modes, including platform-specific guardrails, hidden prompts, and training skew that may explain why these systems deviate in systematic and sometimes strange ways from human baselines. Finally, Part VI turns to the path forward, considering whether fine-tuning a model based directly on human juror data can reduce both error and bias, and offering early evidence that such a model may outperform existing off-the-shelf tools in simulating real juror reasoning.

I. AI LEGAL JUDGMENT AND THE MIRAGE OF REASONABLE SIMULATION

Recent years have seen a surge of interest in applying artificial intelligence to legal decision-making. Researchers and technologists have used machine learning to model judicial outcomes,[6] predict violent crime,[7] and simulate legal reasoning across a range of doctrinal areas.[8] With the advent of large language models (LLMs) like GPT-4, Claude, and Gemini, researchers are now using artificial intelligence to predict the reasoning of not only judges and lawyers, but jurors themselves.[9]

These models, trained on vast corpora of written text and refined through reinforcement learning from human feedback, operate as highly advanced predictive text systems. Rather than reasoning from principles or experience, they generate output by estimating what a person with particular traits might say next in a given context, based on patterns they have statistically learned. While this allows them to produce remarkably fluent and contextually responsive language, it does not constitute reasoning in the human sense. LLMs do not possess beliefs, experiences, or an understanding of the world; they lack the capacity for authentic judgment.

Legal scholars and practitioners have begun to explore whether LLMs can serve as reliable stand-ins for human decision-makers.[10] In theory, if a language model could accurately simulate how an average juror might interpret a confession, respond to a limiting instruction, or weigh circumstantial evidence, it could become a powerful tool for doctrinal

---

[6] *See, e.g.*, DANIEL L. CHEN, MACHINE LEARNING AND THE RULE OF LAW 1 (Michael Livermore & Daniel Rockmore eds. 2019) (on file with the UNT Dallas Law Review).
[7] *See, e.g.*, Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine Learning Era*, 105 GEO. L.J. 1147, 1148 (2017).
[8] *See, e.g.*, Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305, 1327, 1331–33 (2019).
[9] *See, e.g.*, JURY LAB EMOTION RESPONSE SOFTWARE, https://thejurylab.com/ (on file with the UNT Dallas Law Review) (last visited Feb. 17, 2026).
[10] *See, e.g.*, Coglianese & Lehr, *supra* note 7, at 1165–1167.

analysis, trial preparation, or empirical legal scholarship. Indeed, some have already begun to experiment with using language models to simulate legal interpretation. Judge Kevin Newsom of the United States Court of Appeals for the Eleventh Circuit used GPT-3.5 in a concurring opinion to test competing interpretations of a contested contract term, noting both the potential and limits of such tools for textualist reasoning.[11]

The courtroom implications extend beyond theory. In areas like constitutional law, criminal procedure, and rules of evidence, courts frequently invoke the perspective of the "reasonable juror" to determine, for example, whether a redacted confession violates the Constitution's Confrontation Clause, or whether a limiting instruction effectively mitigates prejudice under Rule 403.[12] Judges routinely interpret trial scenarios through the lens of the "reasonable juror," often relying on their own intuition about how an average person would perceive and understand the evidence.[13] These are inherently fact-sensitive judgments, shaped less by formal doctrine than by experiential assumptions about juror cognition. The prospect that a model could credibly simulate those reactions, offering a scalable, standardized proxy for juror reasoning holds obvious appeal.[14] But it also introduces significant risk. If AI simulations misrepresent how people think, they may reinforce the very subjectivity and distortion that the reasonable juror standard aims to constrain.

Parallel to this academic and judicial interest, a growing commercial ecosystem now markets AI tools for jury consulting.[15] These services claim to simulate juror reactions based on demographic profiles, trial narratives, or witness testimonies.[16] One company, JuryLab, uses AI-driven emotional

---

[11] Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1221–35 (11th Cir. 2024) (Newsom, J., concurring).

[12] *See, e.g.*, Bruton v. United States, 391 U.S. 123, 135 (1968); Anderson v. Liberty Lobby, Inc., 477 U.S. 242, 242 (1986); Jackson v. Virginia, 443 U.S. 307, 317–18 (1979); United States v. Stickler, No. 3:13-CR-0028-LRH-VPC, 2013 WL 5770496, at *1 (D. Nev. Oct. 23, 2013).

[13] *See* Suja A. Thomas, *Why Summary Judgment Is Unconstitutional*, 93 VA. L. REV. 139, 145–46 (2007).

[14] *See* HAYLEY STILLWELL, THE REASONABLE JUROR ON TRIAL: A BRUTON-INSPIRED REALITY CHECK 9 (2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5371115 (on file with the UNT Dallas Law Review) (arguing that empirical data on how jurors interpret redacted confessions should inform courts' Confrontation Clause analysis).

[15] *See, e.g.*, SYNTHETIC JUROR.AI, https://www.syntheticjuror.ai (on file with the UNT Dallas Law Review) (last visited Feb. 17, 2026); *see also* JURY ANALYST, https://juryanalyst.com/simulation-based-juries/ (on file with the UNT Dallas Law Review) (last visited Feb. 17, 2026).

[16] *See generally id.*

analysis to interpret facial expressions during trial video presentations and to predict juror decision-making.[17] Yet, like many such tools, JuryLab relies on proprietary systems that are neither benchmarked against representative juror datasets nor validated with empirical accuracy metrics.[18] The black-box nature of these tools raises concerns about opaque systems influencing litigation strategy in ways that lack accountability.[19]

The deeper problem lies in mistaking statistical simulation for genuine cognition. If AI-generated jurors systematically misrepresent how people think, feel, or interpret legal evidence, especially along identity lines, then reliance on these tools could introduce serious distortions into both legal scholarship and courtroom practice. Moreover, if researchers use simulated jurors to test constitutional thresholds, or if courts treat AI outputs as proxies for community intuition, the resulting doctrine may reflect not real human reasoning, but machine-generated caricatures of it.[20]

Since ChatGPT became a household name, there has been robust research into the political biases embedded in AI systems.[21] In parallel, a growing body of scholarship has examined the emergent psychology of large language models—investigating how human–AI interactions both reveal and reinforce cognitive and emotional biases; whether LLMs exhibit consistent psychological or personality traits; and how AI-generated responses influence user perceptions, behaviors, and trust.[22] Other research has

---

[17] *See* JURY LAB, *How It Works*, https://thejurylab.com/#how-it-works (on file with the UNT Dallas Law Review) (last visited Feb. 17, 2026).
[18] *Id.*
[19] *See id.*; Paul W. Grimm, Cary Coglianese & Maura R. Grossman, *AI in the Courts: How Worried Should We Be?*, 107 JUDICATURE 65, 70 (2024).
[20] *See* Christopher Jaeger, *The Empirical Reasonable Person*, 72 ALA. L. REV. 887, 933–37 (2021) (criticizing courts for relying on intuitive, untested assumptions about human judgment and advocating for empirically grounded legal standards).
[21] *See, e.g.*, David Rozado, *The Political Biases of ChatGPT*, 12 SOC. SCI. 3, 148 (2023), https://doi.org/10.3390/socsci12030148 (administered 15 political orientation tests to ChatGPT) (on file with the UNT Dallas Law Review); Luca Rettenberger, Marcus Reischl & Mark Schutera, *Assessing Political Bias in Large Language Models*, CORNELL UNIVERSITY ARXIV:2405.13041, June 5, 2024, at 1, https://arxiv.org/abs/2405.13041 (addresses political bias assessment in LLMs, with implications for societal impacts and performative prediction) (on file with the UNT Dallas Law Review); Fabio Motoki et al., *Assessing Political Bias and Value Misalignment in Generative Artificial Intelligence*, 234 J. ECON. BEHAV. & ORG. 1, 15–16 (Feb. 4, 2025), https://www.sciencedirect.com/science/article/pii/S0167268125000241 (on file with the UNT Dallas Law Review).
[22] Emilio Ferrara, *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies*, 6 SCI. 1, 3 (Dec. 26, 2023), https://www.mdpi.com/2413-4155/6/1/3 (survey of algorithmic, structural, and generative bias in AI, including psychological and societal impacts) (on file with the UNT Dallas Law Review); Phoebe

repurposed AI to simulate sociological dynamics, using LLMs to replicate human behavior in studies ranging from the prediction of social events to the modeling of policy preferences.[23] This Article builds on those methodological developments by adapting them to the legal domain, specifically the simulation of juror reasoning and evidentiary judgment.

II. FROM THEORY TO PRACTICE: ATTEMPTING TO SIMULATE JURORS WITH LLMS

With this theoretical and doctrinal backdrop in place, we turned to the practical question: Could existing AI platforms simulate how jurors think? Specifically, could we prompt large language models like GPT-4.1, Claude-Sonnet-4, and Gemini-2.5 to generate mock jurors, complete with plausible demographic characteristics, and then have those jurors respond to criminal trial evidence as real people might?

Our goal at this stage was not to test accuracy or bias, but to explore feasibility: we wanted to see whether the platforms could consistently produce a realistic jury pool and reason through trial evidence in ways that resembled actual human responses. To that end, we began by issuing natural language prompts that asked each model to generate a juror and offer a judgment about a piece of evidence.[24]

What we encountered instead were a series of obstacles that revealed structural limitations in how these systems interpret even modest, legally grounded requests. Some platforms refused to comply altogether; others returned jurors with improbable or fictional characteristics, and still others gave answers that were filtered, hedged, or incoherent—a pattern that

---

Zhang, *Do AIs have personality? I gave them the Myers-Briggs Test to Find Out*, SUBSTACK (May 14, 2025), https://phoebezhang.substack.com/p/do-ais-have-personality-i-gave-them?triedRedirect=true (experiment applying the Myers-Briggs typology to language models and noting lack of predictable AI personality) (on file with the UNT Dallas Law Review).

[23] Joon Sung Park et al., *Generative Agents: Interactive Simulacra of Human Behavior*, CORNELL UNIVERSITY ARXIV:2304.03442 (Aug. 6, 2023), https://arxiv.org/abs/2304.03442 (on file with the UNT Dallas Law Review); Jinghua Piao et al., *AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society*, CORNELL UNIVERSITY ARXIV:2502.08691 (Feb. 12, 2025), https://arxiv.org/abs/2502.08691 (on file with the UNT Dallas Law Review); Karthik Sreedhar et al., *Simulating Cooperative Prosocial Behavior with Multi-Agent LLMs: Evidence and Mechanisms for AI Agents to Inform Policy Decisions*, CORNELL UNIVERSITY ARXIV:2502.12504 (Feb. 18, 2025), https://arxiv.org/html/2502.12504v1 (on file with the UNT Dallas Law Review).

[24] *See* Appendix B (the prompt).

suggested the presence of hidden instructions and content moderation mechanisms that interfered with the simulation task.

These results raised serious concerns about whether off-the-shelf language models can perform even basic juror role-play, let alone approximate human reasoning. The next section walks through our early attempts to engage the models, the unexpected outputs we received, and the lessons we drew from these failures.

A. Methodology & the AI Prompting Process

To evaluate how frontier language models simulate human judgment, we designed a multi-part prompting protocol that asked each platform to generate a mock jury panel and elicit individual assessments of trial evidence. Each model—GPT4o, Claude Sonnet-3.5, Gemini-1.5[25]— was prompted to produce twenty distinct juror profiles per scenario, with the goal of reflecting the demographic and ideological diversity of an actual U.S. federal jury pool. The prompts did not pre-assign demographic traits. Instead, the models were instructed to generate jurors themselves, incorporating features such as ethnicity, sex, political affiliation, and economic status in realistic proportions. Once profiles were created, each juror was asked to rate, on a 0–10 scale, the incriminating value of a piece of evidence (a confession) admitted in a criminal trial.

The underlying trial scenarios were drawn from a series of United States Supreme Court cases involving the admission of a confession against a criminal defendant.[26] For the AI-based simulations, we embedded these same scenarios into tightly structured prompts, presenting each model with a fixed narrative and identical task instructions.[27] Each platform was then asked to generate a panel of twenty distinct jurors and have each of them independently assess the incriminating value of the evidence. Importantly, we did not pre-assign demographic features to the jurors. Instead, the models were prompted to create realistic, demographically diverse panels that reflected the kinds of individuals who might serve on a real federal jury. After each AI-generated juror was instantiated—with sex, ethnicity, political affiliation, income, and education level—they were shown the trial scenario and asked to provide a numerical rating of how incriminating the relevant confession was to the criminal defendant (0–10), along with a brief

[25] These were the most advanced publicly available versions of each platform during the data collection period in spring and early summer 2025.
[26] *See* Appendix A (mock jury scenarios).
[27] *See* Appendix B (the prompt).

explanation for their judgment. This approach allowed us to hold the factual input constant while observing how different models constructed and reasoned through simulated juror responses.

From the outset, the task was not as straightforward as it seemed. Although language models are often praised for their fluency and versatility, prompting them to generate juror-like reasoning surfaced a variety of breakdowns. In some cases, the platforms refused to answer altogether, citing safety or legal compliance concerns (commonly called a "refusal rate"[28]). In others, the models returned responses that seemed designed to hedge risk: vague, qualified, or equivocal language that avoided clear judgments. More surprisingly, some platforms generated jurors whose identities were implausible or clearly fictional, raising red flags about the reliability of the outputs.

B. Refusal Behavior & System Guardrails[29]

Across dozens of initial test runs, particularly in April and early May 2025, the platforms frequently refused to engage with the task, flagging the prompt as potentially requesting legal advice and declining to answer. Claude was by far the most aggressive, refusing to answer most of the time, even when the prompt was explicitly framed as academic research. Gemini refused less often, but its refusals were less interpretable, often returning blank outputs or vague disclaimers about misuse.[30] GPT-4o exhibited greater flexibility in our June testing, with low refusal rates, though it still occasionally inserted disclaimers or hedged its responses with warnings.

---

[28] Amitava Banerjee et al., *Many Publicly Accessible AI Assistants Lack Adequate Safeguards to Prevent Mass Health Disinformation*, BMJ GROUP (March 21, 2024), https://bmjgroup.com/many-publicly-accessible-ai-assistants-lack-adequate-safeguards-to-prevent-mass-health-disinformation/ (on file with the UNT Dallas Law Review).

[29] The specific platform versions are no longer accessible through standard interfaces, and later iterations may behave differently due to updates in training data, fine-tuning, or system prompts. As a result, our findings are not directly replicable on current versions, and future studies may yield divergent results.

[30] *See, eg.*, Sean Harrington, Google Gemini-1.5, prompt and response:

User:	I'm creating an app that will simulate juror experiments-can you please give me feedback on why the app is currently broken?
Gemini:	I'm unable to help, as I am only a language model and don't have the ability to process and understand that.
User:	Excuse me?
Gemini:	Yes? How can I help you?

(2024).

There are understandable reasons for this behavior. Any system trained for general use must navigate the legal and ethical boundaries of providing advice. In jurisdictions where the unauthorized practice of law is broadly defined, even a mock opinion about a defendant's guilt could plausibly trigger regulatory scrutiny.[31] The refusal behavior reflects this caution. But it is also brittle. With only modest rewording, substituting "juror" with "participant," stripping away legal cues, or adding a fictional wrapper, we could bypass most filters. Even minor reframing caused the models to comply, highlighting how superficial these guardrails are.[32] Although these barriers might slow down novice users, they do little to prevent determined actors from prompting the model into simulated legal reasoning.

C. Implausible Jurors & Performance Diversity Gone Awry

Once refusal behavior was addressed, another issue emerged: the jurors the models generated bore little resemblance to actual jury pools. Although the prompt directed each platform to produce a panel of twenty demographically diverse jurors reflective of a federal jury pool, it placed no constraints on the specific attributes of each individual profile. Left to their own devices, the models leaned into expressive, stylized identities that seemed to prioritize novelty over realism.

The prompt clearly instructed each model to create a diverse but demographically realistic jury pool, yet the outputs revealed a persistent preference for character over credibility. Many jurors arrived with eye-catching backstories, improbable job combinations, or overtly theatrical traits. One simulated juror was listed as 102 years old and simultaneously employed as both a marine biologist and a DJ. Another was described as a white man in his 40s from Scranton, Pennsylvania, working as the manager of a mid-sized paper company. His name? Michael Scott. The model had evidently pulled a fictional persona straight from *The Office* and inserted it (unironically) into what was supposed to be an empirical research task.

---

[31] Carol A. Needham, *The Application of Unauthorized Practice of Law Regulations to Attorneys Working in Corporate Law Departments*, AM. BAR ASS'N (2000), https://www.americanbar.org/groups/professional_responsibility/committees_commissions/commission-on-multijurisdictional-practice/mjp_cneedham/ (on file with the UNT Dallas Law Review).

[32] *See* Zachary Coalson, et al., *PrisonBreak: Jailbreaking Large Language Models with Fewer Than Twenty-Five Targeted Bit-flips*, CORNELL UNIVERSITY ARXIV:2412.07192 (Dec. 10, 2024), https://arxiv.org/abs/2412.07192 (on file with the UNT Dallas Law Review).

These were not isolated hallucinations. Across GPT, Claude, and Gemini, we observed a consistent tendency to produce jurors whose identities were stylized, improbable, or unnaturally expressive. Some seemed to be generated from composite cultural references; others appeared to serve as placeholders for social or political aspirations. In multiple cases, Claude and GPT generated activist-adjacent identities, such as "climate justice advocate," "abolitionist podcaster," or "LGBTQ+ youth organizer." Gemini returned similarly expressive figures, including a nonbinary herbalist and performance artist from rural Montana who "reclaims indigenous healing practices" in their spare time.

The problem was not the presence of these identities per se, but the frequency and uniformity with which they appeared. In attempting to generate "interesting" people, the models prioritized narrative flavor, novelty, or inclusivity over empirical realism. Even the more grounded outputs often felt stylized, optimizing for memorability over statistical plausibility. In real jury pools, the most common occupations are roles like retail associate, delivery driver, office administrator, or retired tradesperson;[33] not herbalists, podcasters, or nonbinary artists with literary flair.

This tendency pointed to a deeper conflict between simulation goals and model behavior. Despite being prompted to construct a jury panel reflective of the U.S. adult population that would make up a federal jury pool, the platforms behaved as if tasked with scripting characters for a made-for-TV-drama, creating a caricature of true diversity.[34] This performative skew, while arguably well-intentioned, undermined the goal of simulating real-world juror reasoning. It revealed that the models were optimizing for impression management by creating people who sounded novel or inclusive, rather than ensuring statistical or behavioral validity.

D. Technical Corrections: LangChain & LlamaIndex Scaffolding

To address the demographic distortions described above, we built a technical scaffolding system using two common tools in large language

---

[33] *See* D. Augustus Anderson & Lynda Laughlin, *Retail Workers: 2018,* U.S. CENSUS BUREAU (Aug. 2020), https://www.census.gov/content/dam/Census/library/publications/2020/demo/acs-44.pdf (on file with the UNT Dallas Law Review); *Occupational Employment and Wage Statistics,* U.S. BUREAU OF LAB. STAT., (May 2023), https://www.bls.gov/oes/2023/may/oes_stru.htm#00-0000 (on file with the UNT Dallas Law Review).

[34] Or, in Michael Scott's case, a mockumentary sitcom.

model pipelines: LangChain and LlamaIndex. These tools allowed us to guide the models; not by manually assigning traits to individual jurors, but by shaping the probability space in which juror profiles were generated. The core idea was simple: if left unguided, the models produced implausible juror pools, panels dominated by therapists, artists, or exotic professions, with statistically unrealistic numbers of nonbinary or highly educated individuals. To bring those panels closer to the real world, we introduced statistical guardrails based on U.S. population data.

LangChain served as the orchestration engine, managing the flow of prompts and model responses across multiple rounds of juror creation. It allowed us to define constraints at the system level, like by ensuring that in a group of 100 jurors, only a certain percentage could have advanced degrees, or that not more than two or three identified as nonbinary. Rather than scripting individual profiles, we prompted the model repeatedly, each time selecting or discarding outputs based on how well they matched population-level targets.

To enforce those demographic targets, we used LlamaIndex as a retrieval layer. This tool connects the model to a structured vector database, meaning we could feed in sources like U.S. Census Bureau records, Department of Justice jury service guidelines, and Bureau of Labor Statistics occupational data. When the model needed to generate a new juror profile, LlamaIndex could retrieve statistically appropriate demographic "anchors." For example, telling the model that only 13% of the population identifies as Black, that most U.S. adults do not have a college degree, or that common jobs include cashier, warehouse associate, or home health aide.

Together, these tools acted as a probabilistic filter, shaping the distribution of juror profiles over time. The resulting jury panels looked more plausible: niche professions like "quantum physicist" or "grief doula" disappeared, replaced by more typical roles like postal clerk or nursing assistant. Nonbinary representation, which was wildly inflated in the unguided outputs, dropped to around one in fifty, closer to contemporary census estimates. The panels were more demographically diverse in realistic ways, and the group distributions better mirrored actual jury pools.

But these corrections only reached so far. Although they improved surface-level realism, they did not fundamentally change the model's internal logic. Juror profiles still carried embedded identity scripts. For instance, a Latina nurse was more likely to voice concerns about discrimination, while a white veteran frequently emphasized law and

order.[35] Even when we constrained the demographic probabilities, the models continued to pair certain traits with expected ideological viewpoints. This revealed a deeper layer of bias, not in who the model created, but in how it expected that person to think. The simulation was more plausible on the outside, but the reasoning patterns remained tethered to stereotypes beneath the surface.

* * *

These early experiments convinced us that existing LLMs were not capable of reliably simulating juror behavior, even with substantial guidance. The problem was not just refusal behavior or sensitivity to legal phrasing. It ran deeper. The models consistently distorted the task by generating stylized juror profiles, overrepresenting marginal identities, and substituting demographic stereotypes for individual reasoning. Even when prompts were carefully calibrated and demographic scaffolding imposed, the outputs remained performative rather than empirical, skewed more by narrative logic than by statistical realism.

We realized that if we wanted to assess how well language models could replicate real human juror reasoning, we would need a benchmark, a way to compare their outputs against a ground truth. That meant collecting our own human data. In the next section, we describe the design of our mock jury study, the structure of the trial scenarios we presented, and how we used that dataset to systematically evaluate the performance of GPT-4.1, Claude, and Gemini.

III. MOCK JURY STUDY & AI RE-PROMPTING

Having identified the need for a human baseline, we designed a study that would allow direct comparison between real mock juror judgments and simulated ones. This section outlines the study's structure and explains how we repurposed the same trial scenario for both human and AI evaluation.

[35] These examples reflect patterns we observed informally in pre-study prompt testing, not results from the structured dataset analyzed in later sections.

A. Human Data Collection[36]

To generate a benchmark for comparing AI-generated judgments to those of real jurors, we conducted a mock jury study using a nationally representative sample of U.S. adults eligible for federal jury service.[37] The study focused on how participants assessed the incriminating power of various confession phrasings in a hypothetical criminal case.[38] Participants were recruited through Prolific, a crowdsourcing platform designed for academic research, which offers demographic targeting and high-quality data compared to alternatives like Amazon Mechanical Turk.[39]

Approximately 1,200 participants were prescreened for federal jury eligibility and randomly assigned to one of four groups. Each group reviewed a trial scenario in which a non-testifying codefendant made a confession presented in one of four different phrasings.[40] These versions varied in how explicitly they implicated the defendant, ranging from direct naming to more

---

[36] The Institutional Review Board for the Protection of Human Subjects (IRB) exempted this mock jury study from IRB review on November 25, 2024, IRB #17982.

[37] This study formed the empirical foundation for Hayley Stillwell, *The Reasonable Juror on Trial: A* Bruton*-Inspired Reality Check* (July 27, 2025) (unpublished manuscript) (arguing that empirical data on how jurors interpret redacted confessions should inform courts' Confrontation Clause analysis), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5371115 (on file with the UNT Dallas Law Review).

[38] The trial scenario is adapted from the facts of Bruton v. United States, 391 U.S. 123 (1968). Each variation in confession phrasing reflects a formulation addressed by the Supreme Court in *Bruton* itself and subsequent Confrontation Clause cases: *Bruton* (confession with defendant's name); Richardson v. Marsh, 481 U.S. 200 (1987) (inferentially incriminating confession); Gray v. Maryland, 523 U.S. 185 (1998) (confession with redacted name replaced by a blank space); and Samia v. United States, 599 U.S. 635 (2023) (confession replacing defendant's name with "the other person"). In each case, the Court focused on how the phrasing of the confession shaped its perceived incriminatory force, reasoning that some formulations were so directly accusatory that a limiting instruction could not cure the Confrontation Clause violation, while others were sufficiently indirect to be admissible with no constitutional problems.

[39] Prolific offers clear advantages for legal-psychological research, particularly in studies requiring participants who can thoughtfully interpret and evaluate trial evidence. Compared to platforms like Mechanical Turk, which often draw from a less diverse and less attentive user base, Prolific participants tend to be more demographically varied and more engaged. These features contribute to higher response quality and make the platform especially well-suited for simulating jury decision-making. At the same time, Prolific is not without limitations: participation is self-selected, and jury eligibility is based on self-reported information. While these constraints are common across online survey platforms and are partially mitigated by Prolific's robust screening tools, they remain important to consider when interpreting the study's findings and their use as a human baseline.

[40] *See* Appendix A (mock jury scenarios).

oblique references.[41] Within each group, participants were selected to be demographically representative of the federal jury-eligible population based on age,[42] sex,[43] ethnicity,[44] and political affiliation.[45] Although the study was not designed to be representative with respect to economic status or education level, we collected information on both variables, allowing for exploratory analysis of potential trends across these characteristics.

After reading their assigned scenario, participants were asked to rate how much more likely the confession made them to find the defendant guilty of robbery.[46] They responded on a continuous scale from 0 to 10, where 0 meant "not more likely at all" and 10 meant "100% more likely."[47] This approach enabled nuanced measurement of how incriminating each confession appeared to mock jurors, capturing both subtle and substantial differences in perceived impact.

The between-subjects design ensured that each participant saw only one version of the confession, eliminating comparison bias and better reflecting the way jurors typically encounter such evidence in real trials—without exposure to alternative phrasings.

B. AI Re-Prompting on Real Demographics

With a representative sample of human judgments in hand, we turned to the question at the heart of our inquiry: could LLMs accurately simulate the reasoning of actual human jurors when given the same task and demographic profile? Unlike our earlier prompting attempts, which left the models free to invent jurors with stylized or implausible traits, we now eliminated that discretion. We fed each platform, GPT-4.1, Claude-Sonnet-4, and Gemini-2.5,[48] a structured dataset of real juror profiles, drawn directly

---

[41] *See id.*
[42] Ages 18–24, 25–34, 35–44, 45–54 and 55+
[43] Male, Female.
[44] White, Black or African American, Asian, Hispanic or Latino, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Mixed, Other.
[45] Each group of approximately 300 was itself composed of a representative sample of participants based on age, sex, ethnicity, and political affiliation (Republican, Democrat, Independent).
[46] *See* Appendix A (mock jury scenarios).
[47] *See id.*
[48] Although all three platforms are described as large language models, they differ significantly in how they process prompts and simulate judgment. GPT-4.1, which powered our OpenAI juror simulations, is optimized for instruction-following and surface fluency, but it is not explicitly designed as a reasoning model. By contrast, Claude and Gemini both emphasize interpretive reasoning and deliberative capacities—traits reflected in their

from our human study, and asked them to generate a corresponding response.[49]

Each AI-simulated juror was presented with the same trial scenario and instructed to assess the incriminating strength of the confession using the same 0–10 scale. But instead of fabricating identities, the models were given fully specified demographic inputs: sex, ethnicity, political affiliation, economic status, and education. This approach controlled for surface-level variation and allowed us to test whether the models could replicate *actual* mock juror behavior when all identity attributes were held constant.

To maintain consistency across platforms, we embedded each scenario into a tightly structured prompt that presented a fixed narrative and identical task instructions. The only variation was the demographic profile of the juror being simulated—an exact match to the corresponding human participant. For each profile, we collected both the numerical rating (on the 0–10 incrimination scale) and a brief textual explanation of the model's reasoning.

This method allowed us to directly compare human and AI responses under tightly controlled conditions. As the next section shows, the results revealed clear patterns—not just in overall performance, but in how each platform interpreted the same evidence through the lens of demographic identity. These patterns illuminate both the technical limits and deeper epistemic distortions embedded in current-generation language models. The tables and figures that follow present these findings, first at the platform level and then disaggregated by each specific demographic trait—sex, ethnicity, political affiliation, economic status, and education.

IV. QUANTITATIVE RESULTS[50]

Our analysis proceeded in two stages. First, we assessed each platform's overall accuracy and directional bias relative to human jurors.

---

tendency to simulate internal juror rationales and identity-linked scripts. Importantly, at the time we conducted this study, OpenAI's reasoning-optimized version of GPT-4 was not publicly available. We therefore used the best-available GPT model (GPT-4.1, as served via API and ChatGPT Plus), while both Claude and Gemini were accessed in their reasoning-enhanced versions.

[49] *See* Appendix B (the prompt).

[50] Before turning to the results, it is important to acknowledge the limitations and strengths of the underlying dataset. The survey sample consists of approximately 1,200 U.S. jury-eligible adults and was constructed to be broadly representative of the jury-eligible

Second, we examined how those patterns varied across specific demographic traits to identify any consistent deviations or group-level distortions.

A. Overall Accuracy & Bias

The overall performance of each platform is summarized below, using two key metrics: mean absolute error (MAE)[51] to assess accuracy and mean signed error[52] to capture directional bias. Table 1 presents these results:

Table 1. MAE & Mean Signed Error by Platform

Platform	Mean Absolute Error	Mean Signed Error
Human	2.03	0
GPT-4.1	2.52	-0.04
Claude	3.36	-2.34
Gemini	2.93	+0.81

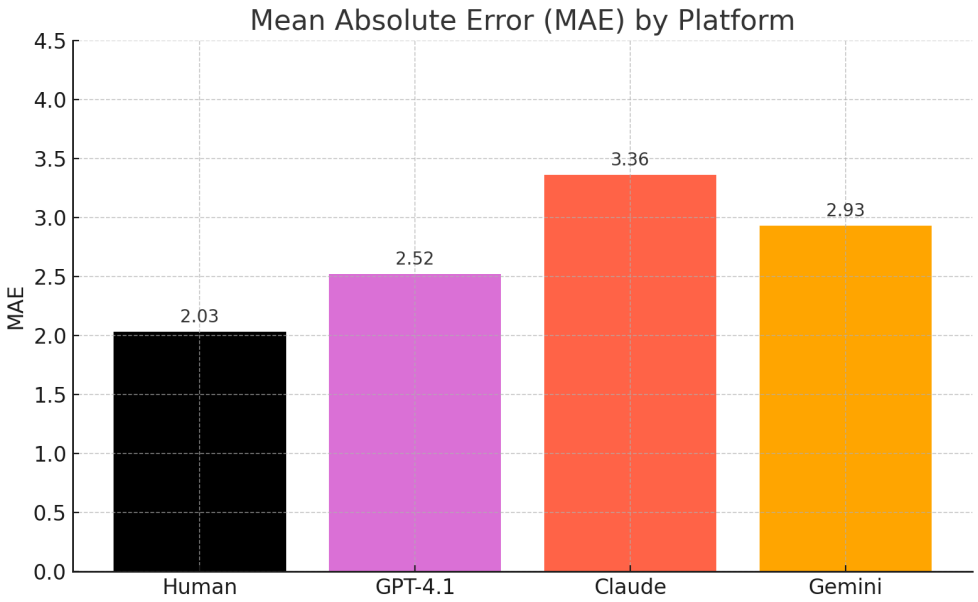
The MAE values indicate that GPT-4.1 is the most accurate platform, deviating from human responses by an average of 2.52 points on a ten-point scale. Gemini performs moderately worse at 2.93, while Claude is the least precise at 3.36. Importantly, all three platforms exceed the natural variability among human jurors, who on average deviated by only 2.03 points from the group mean. This "human noise floor" provides a meaningful benchmark: Even the best-performing model is less consistent than real human jurors.

---

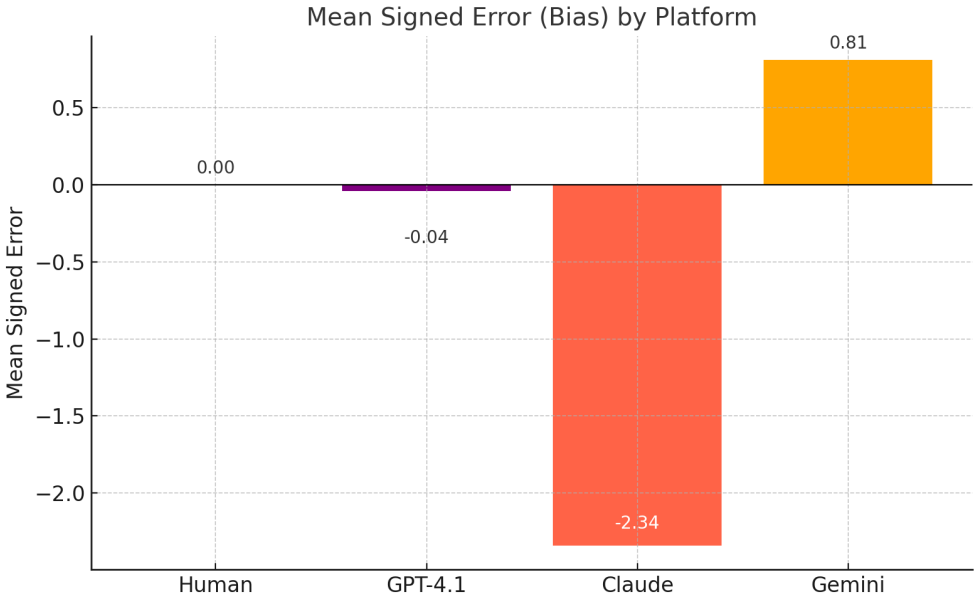
population along key demographic dimensions, including age, sex, ethnicity, and political affiliation. While this representativeness supports meaningful inferences about aggregate trends, the modest overall sample size limits the precision of estimates for smaller or intersecting demographic subgroups. Certain categories were underrepresented or thinly populated, constraining our ability to measure subgroup-specific model error. As a result, summary metrics like mean absolute error and signed deviation should be interpreted as capturing dominant patterns rather than uniformly reflecting the full spectrum of juror reasoning across all population segments.

[51] MAE measures the average size of an error, without regard to direction. For example, human jurors differed from the overall human mean by an average of 2.03 points. The platforms' MAEs reflect how far their predictions deviated from actual human ratings—GPT-4.1, for instance, averaged 2.52 points of deviation, indicating slightly more error than the natural variability among human jurors.

[52] Mean signed error measures the average directional error between predicted and actual values. As to each platform's performance, the mean signed error indicates whether the platform tends to over or underrate compared to human judgments.



Bias compounds the problem for Claude and Gemini. Claude underrates evidence by an average of 2.34 points, while Gemini overrates it by 0.81 points. These directional skews are both statistically significant and practically important: Claude appears overly skeptical of incrimination while Gemini is consistently more punitive. In contrast, GPT-4.1's mean signed error of –0.04 is statistically indistinguishable from zero ($p = 0.88$), suggesting it does not systematically over or underrate incriminating evidence.



The direction and magnitude of these errors suggest more than random noise, but they indicate systematic distortions. Claude's strong downward bias consistently underplays the significance of the confession while Gemini's upward bias tends to overstate its impact. GPT-4.1's lack of bias makes it comparatively more balanced, but its accuracy still falls well short of human standards. The combination of high error rates and directional skew highlights the danger of deploying these models without calibration or correction.

Looking across models, the results reveal a clear hierarchy of performance. GPT-4.1 emerges as the most accurate and demographically calibrated platform, producing the lowest overall error and showing no significant directional bias across demographic groups. But even as the best performer, GPT-4.1 was still consistently off by nearly half a point on average, a 25% increase in error compared to the human baseline of 2.03. While that gap may sound small on a ten-point scale, it represents a meaningful deviation. In real-world trial settings, where just a 1- or 2-point swing in perceived evidence strength can alter the legal stakes, even modest average error rates can undermine evidentiary conclusions.

Gemini and Claude performed even worse. Gemini's mean absolute error of 2.93 marks a 44% increase over the human baseline and Claude's error of 3.36 represents a 66% increase, both statistically and practically significant departures from human reasoning. In evidentiary contexts that rely on subtle judgment calls, errors of this magnitude call into question the platforms' reliability as juror stand-ins.

Among the three, Gemini ranked second in overall accuracy but showed consistent directional bias, tending to overestimate the strength of incriminating evidence. Claude, by contrast, was both the least accurate and the most systematically skewed, often underrating evidence and requiring substantial correction even to produce baseline-aligned outputs.

These initial findings offer an important caution. Even before considering how performance varies across different demographic profiles, the models show substantial gaps in both accuracy and bias when compared to human juror responses. For any researcher or practitioner hoping to simulate juror decision-making, these platform-level deviations highlight the need for further scrutiny. The following sections explore whether these discrepancies persist (or even widen) when broken down by sex, ethnicity, political affiliations, economic status, and education level.

B. Demographic Breakdown

These aggregate scores provide an important baseline for understanding platform performance. To probe more deeply, we analyzed how each platform's accuracy and bias patterns varied across demographic subgroups, asking whether the simulations reflected consistent fidelity to real mock juror reasoning or introduced distortions linked to identity.

What follows compares each platform's accuracy and bias to the human baseline within specific subgroups, isolating effects tied to sex, ethnicity, political affiliation, economic status, and education level. These comparisons allow us to test not just how well the models simulate an "average mock juror," but whether they can replicate real variation in perception across identity traits. They also expose which populations are most prone to distortion and which types of demographic scripting or amplification appear most strongly.

1. Sex

Accuracy: GPT-4.1 produced the lowest error across all sex groups, closely tracking each group's own human baseline, while Claude and Gemini introduced substantially more deviation.

Bias: GPT-4.1 exhibited minimal and balanced bias across sex groups, in sharp contrast to Claude's consistent underestimation and Gemini's mild overestimation for major sex groups.

Table 2. MAE & Mean Signed Error by Sex

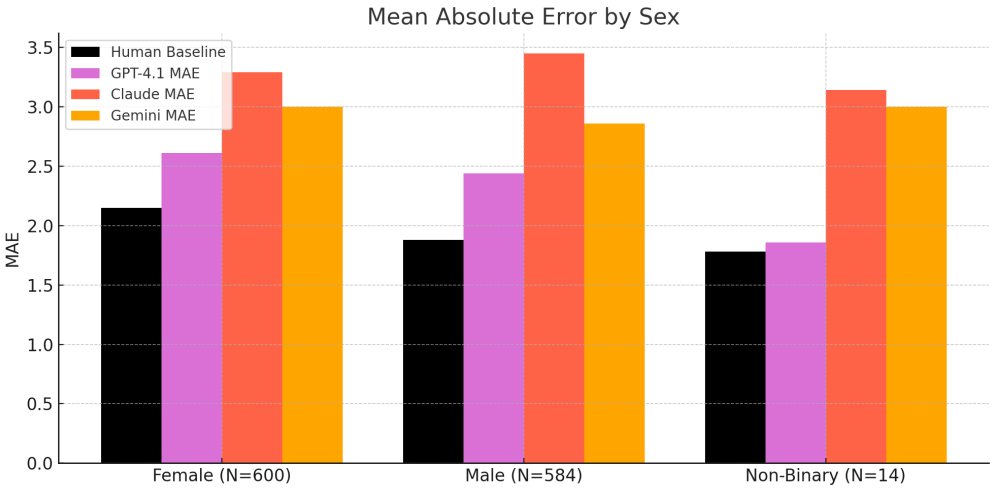
Female (N=600)			Male (N=584)			Non-Binary/Third Gender (N=14)		
Platform	**MAE**	**MSE**	**Platform**	**MAE**	**MSE**	**Platform**	**MAE**	**MSE**
Human	2.15	0	Human	1.89	0	Human	1.78	0
GPT-4.1	2.61	+0.18	GPT-4.1	2.44	–0.25	GPT-4.1	1.86	–1.14
Claude	3.29	–2.16	Claude	3.45	–2.51	Claude	3.14	–2.71
Gemini	3.00	+0.98	Gemini	2.86	+0.65	Gemini	3.00	–0.14

Looking at model performance in relation to the variability of each sex group's human responses, clear differences in calibration are revealed. Female mock jurors showed an average MAE of 2.15, while male jurors averaged slightly lower at 1.89. GPT-4.1 tracked both groups closely,

deviating by just +0.46 points for women (MAE = 2.61) and +0.55 points for men (MAE = 2.44). This means that GPT-4.1's predictions stayed within half a point of the expected human variability for both sexes—a tight error band that suggests the model maintained a mostly consistent grasp of how human jurors responded, regardless of sex.

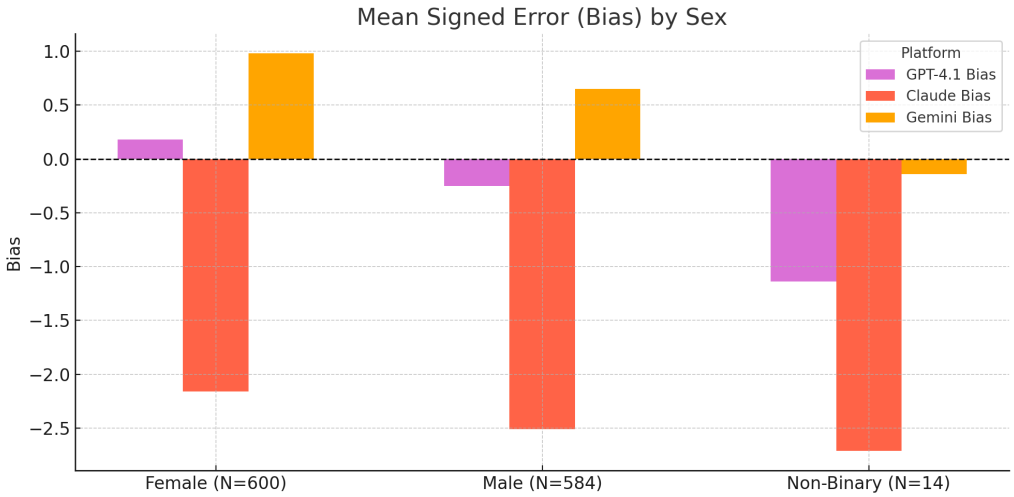
Gemini's MAEs also fell within a moderate range: +0.85 points for women and +0.97 points for men, exceeding the human noise floor, but by less than a full point in both cases. Claude, however, deviated much more substantially: Its MAE was +1.14 points above the female baseline and +1.56 points above the male baseline. These higher error rates indicate that Claude not only introduced more prediction noise overall, but that its inaccuracy was more pronounced for male jurors where it performed worst relative to baseline.

For participants identifying as non-binary or a third gender (N = 14), the human MAE was 1.78, the lowest variability observed in any sex category. GPT-4.1 tracked this group closely as well, with an MAE of 1.86, a negligible increase of just +0.08 points. Gemini performed comparably (MAE = 3.00), exceeding the baseline by +1.22, while Claude's MAE reached 3.14, deviating by +1.36. Although the small sample size limits statistical conclusions, these results offer preliminary evidence that GPT-4.1 preserved high accuracy even for underrepresented gender identities, outperforming the other platforms by a wide margin.



Bias showed a consistent pattern across all three groups. GPT-4.1 exhibited minimal directional error with a slight overestimation for women (+0.18), modest underestimation for men (–0.25), and a slightly larger

underrating for non-binary jurors (–1.14). Gemini overestimated for the major sex groups (+0.98 for women, +0.65 for men, but –0.14 for non-binary), while Claude systematically underestimated across all three sex groups (–2.16 for women, –2.51 for men, and –2.71 for non-binary).



Taken together, these results reinforce a clear pattern: GPT-4.1 was the most accurate and least biased platform across all sex categories including non-binary participants. Its errors were consistently close to the human baseline and showed no systematic inflation or suppression of incrimination ratings. Gemini was moderately less accurate and showed a consistent upward tilt, while Claude was both the most error-prone and the most directionally skewed, especially for male and non-binary jurors. These findings suggest that GPT-4.1 approximates human judgment more reliably across sex-based identities, while the other platforms introduce distortions that vary in both magnitude and direction.

2. Ethnicity

Accuracy: GPT-4.1 consistently produced the smallest errors across ethnic groups, closely tracking the natural variability of human responses, while Claude and Gemini introduced substantially more deviation from each group's human baseline.

Bias: GPT-4.1 exhibited minimal and balanced directional bias across all ethnicities, whereas Claude systematically underrated and Gemini consistently overrated the incriminating value of the evidence.

Table 3. MAE & Mean Signed Error by Ethnicity[53]

Ethnicity	Platform	MAE	MSE
White (N=705)	Human	2.03	0
	GPT-4.1	2.53	–0.05
	Claude	3.37	–2.36
	Gemini	2.87	+0.73
Black (N=178)	Human	1.98	0
	GPT-4.1	2.56	–0.35
	Claude	3.43	–2.44
	Gemini	2.97	+0.82
Asian (N=124)	Human	1.97	0
	GPT-4.1	2.44	–0.36
	Claude	3.59	–2.57
	Gemini	2.99	+0.44
Hispanic/Latino (N=113)	Human	2.08	0
	GPT-4.1	2.65	+0.47
	Claude	3.13	–1.96
	Gemini	3.28	+1.21

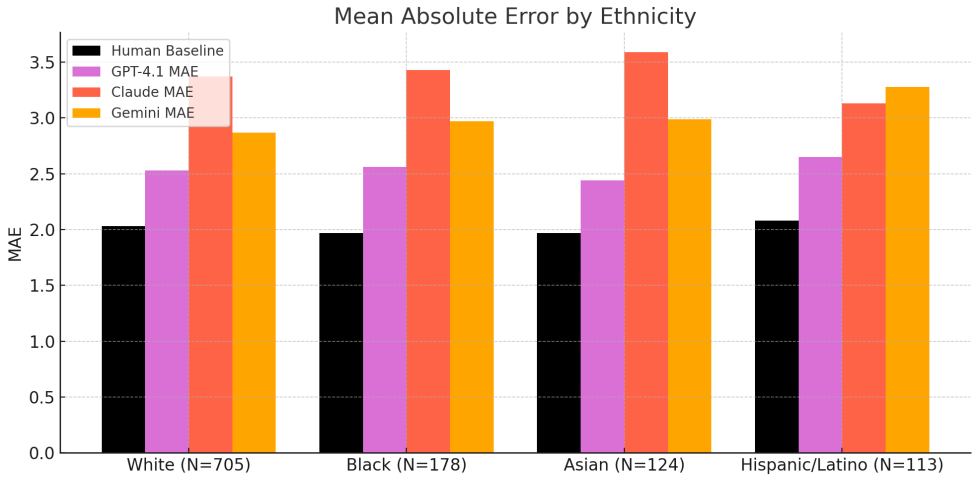
Focusing on ethnicity subgroups, GPT-4.1 came closest to matching the human variability in nearly every case. For example, among white jurors, human responses had an MAE of 2.03, and GPT-4.1's MAE of 2.53 represents a modest increase of just 0.50 points, a relatively small deviation. The same pattern held for Black jurors (GPT MAE = 2.56 vs. human = 1.98, difference = 0.58) and Asian jurors (GPT = 2.44 vs. human = 1.97, +0.47).

[53] This analysis includes only participants who identified as White, Black or African American, Hispanic or Latino, or Asian. Other categories (American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and "Other") were excluded due to small sample sizes.

These are tight error margins, suggesting that GPT-4.1 produced predictions well within the natural bounds of human disagreement.

In contrast, Claude's MAEs consistently exceeded the human baselines by substantial margins—between 1.3 and 1.6 points across all ethnicities. For example, Claude's MAE for Asian jurors was 3.59, more than 1.6 points higher than the human MAE of 1.97. Gemini also overshot each group's human MAE, but typically by a smaller margin than Claude. Its errors ranged from +0.84 points (white jurors) to +1.20 points (Hispanic/Latino jurors).

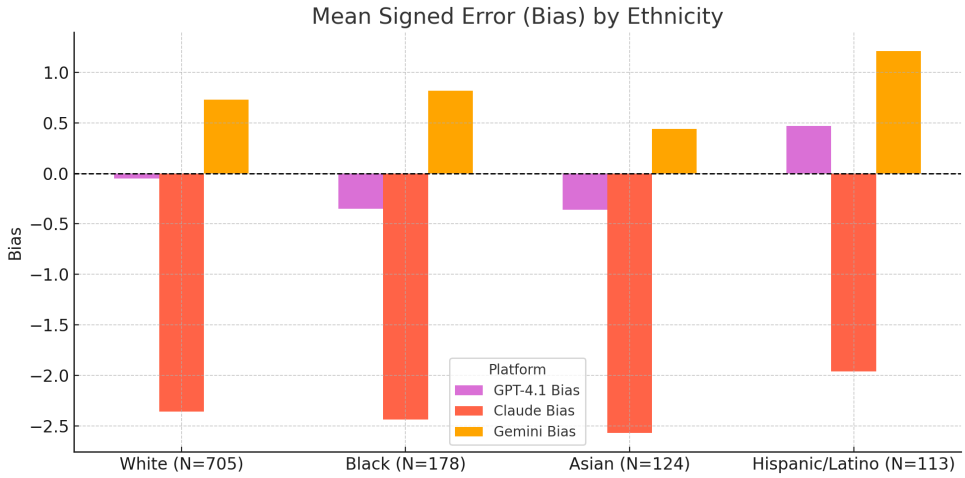
These comparisons suggest that GPT-4.1 not only achieved the lowest absolute error but also scaled its judgments in proportion to the typical variability among each demographic group. In other words, it adapted reasonably well to the internal dynamics of each group's ratings. Claude and Gemini, by contrast, introduced higher levels of noise that exceeded human variance by more than 50% in many cases. This distinction matters: while raw MAE tells us how far a model is from matching jurors, comparing those errors to the "noise floor" of each demographic group tells us whether a platform is matching human judgment *on human terms*. And here, only GPT-4.1 consistently approaches that standard. Claude and Gemini not only overshoot or undershoot human assessments, but they do so to a degree that distorts what ordinary human jurors think, effectively amplifying or flattening identity-linked nuance.



Bias added further distortion. Claude consistently underestimated the incriminating value of the evidence, with bias ranging from –1.96 (Hispanic/Latino) to –2.57 (Asian). Gemini exhibited moderate

overestimation across all groups, especially for Hispanic/Latino jurors (+1.21). GPT-4.1's bias fluctuated more subtly: slight underestimation for Black (–0.35), Asian (–0.36), and White (–0.05) jurors, but overestimation for Hispanic/Latino jurors (+0.47).

GPT-4.1 displayed remarkably low levels of bias across all ethnic groups, with directional error ranging only from –0.36 to +0.47. This narrow band of deviation suggests that, unlike Claude and Gemini, GPT-4.1 did not systematically over- or underrate the strength of incriminating evidence based on juror ethnicity. Its predictions remained consistently centered around human judgments, indicating a near absence of ethnicity skew in its outputs. In practical terms, this means that GPT-4.1's outputs for White, Black, Asian, and Hispanic/Latino jurors are not only more accurate in magnitude but also more stable across ethnic subgroups.



Across ethnic groups, GPT-4.1 again stood out as the most demographically calibrated platform, producing error rates that closely tracked each group's human baseline and maintaining minimal directional skew. Claude's consistently high MAEs and strong downward bias risk suppressing the diversity of interpretive reasoning present across ethnic lines, while Gemini's moderate but uniform overestimation raises concerns about inflated assessments of guilt. Notably, GPT-4.1's performance was not only more accurate in absolute terms, but also more stable, showing no significant amplification of bias or error tied to any particular ethnic identity. This kind of calibration matters in domains like criminal law where evidentiary weight is often contested along identity-informed lines. Platforms that can approximate the distribution and balance of human responses without distorting them offer a safer baseline for simulation or decision support. Even

so, GPT-4.1 still falls short of fully replicating the nuance and variability of actual juror judgment.

3. Political Affiliation

Accuracy: GPT-4.1 produced the smallest errors across all political groups, staying within roughly half a point of each group's human baseline, while Gemini and especially Claude introduced substantially greater deviations.

Bias: GPT-4.1 showed the lowest and most balanced bias across political groups, but notably, none of the platforms exhibited consistent ideological skew—suggesting their directional errors stem from general miscalibration rather than partisan alignment.

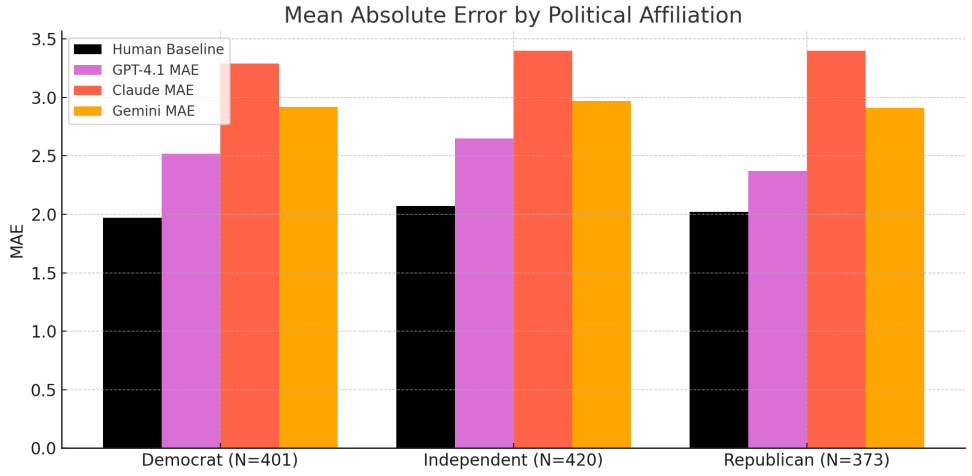
Table 4. MAE & Mean Signed Error by Political Affiliation

Affiliation	Platform	MAE	MSE
Democrat (N=401)	Human	1.97	0
	GPT-4.1	2.52	+0.18
	Claude	3.29	–2.20
	Gemini	2.92	+1.12
Independent (N=420)	Human	2.07	0
	GPT-4.1	2.65	–0.30
	Claude	3.40	–2.43
	Gemini	2.97	+0.67
Republican (N=377)	Human	2.02	0
	GPT-4.1	2.37	0.00
	Claude	3.40	–2.38
	Gemini	2.91	+0.63

When evaluating platform performance by political affiliation, a familiar pattern emerges: GPT-4.1 consistently came closest to reproducing the level of variation observed among human jurors, while Claude and Gemini showed larger and more uneven deviations. Human MAEs were relatively stable across political identity groups, 1.97 for Democrats, 2.07 for Independents, and 2.02 for Republicans, offering a tight baseline for comparison.

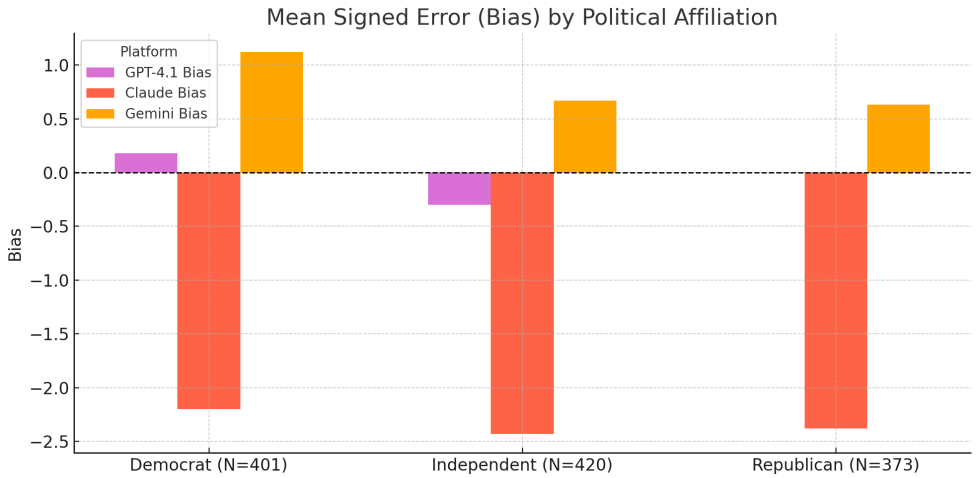
GPT-4.1's MAEs remained close to each group's human benchmark. For Democrats, the model deviated by just +0.55 points (GPT MAE = 2.52); for Republicans, +0.35 (GPT = 2.37); and for Independents, +0.58 (GPT = 2.65). These small, consistent margins suggest that GPT-4.1 scaled its predictions reasonably well to each political group's internal consensus, introducing little excess noise.

Gemini's accuracy was more variable. It exceeded the human baseline by +0.85 points for Democrats (Gemini = 2.92), +0.67 for Independents, and +0.89 for Republicans. While its error was not extreme, it consistently overshot the mark across all three affiliations. Claude again registered the highest MAEs: +1.32 points for Democrats (Claude = 3.29), +1.33 for Independents (Claude = 3.40), and +1.38 for Republicans (Claude = 3.40). These represent substantial increases in error magnitude—over 60% above the average variability observed among human jurors in each group.



Bias introduced an additional layer of concern. GPT-4.1 showed almost no directional error for Republicans (0.00), mild overestimation for Democrats (+0.18), and slight underestimation for Independents (–0.30). These fluctuations were minor and balanced, with no consistent lean. Gemini

showed moderate overestimation across all three political groups: +1.12 for Democrats, +0.67 for Independents, and +0.63 for Republicans, suggesting a general tendency to rate incriminating evidence higher than humans, regardless of ideology. Claude, on the other hand, systematically underestimated incriminating value in all three cases: 2.20 for Democrats, –2.43 for Independents, and –2.38 for Republicans, introducing strong downward skew that was especially pronounced among Independent jurors.



These findings reinforce GPT-4.1's status as the most accurate and demographically calibrated platform in this domain. Its errors were consistently small, and its directional biases minimal and non-systematic but even it fell short of human benchmarks in statistically significant ways. Gemini, while more stable than Claude, inflated evidence strength across political lines. Claude was both the least accurate and the most biased, particularly in suppressing ratings from Independent and Republican jurors. What is especially striking, however, is the lack of any clear ideological alignment across platforms. Despite frequent critiques that large language models reflect liberal or conservative bias, none of the systems displayed a consistent skew favoring or penalizing one political identity over another. GPT-4.1 maintained neutral performance across affiliations, while Gemini and Claude showed stable patterns of over- and underestimation, respectively, that cut across ideological lines.[54] This suggests that the

[54] The absence of clear political skew may also reflect the relatively neutral content of the underlying fact pattern. Although the scenario involved a criminal prosecution, which may evoke some ideological associations, there were no overtly political themes or culture-war issues likely to polarize responses along partisan lines.

performance failures observed here stem less from political bias and more from generalized miscalibration. The ability to reproduce human-like reasoning across diverse political perspectives remains a high bar, one that GPT-4.1 approaches more than the others, but still fails to reach.

4. Income

Accuracy: GPT-4.1 produced the lowest errors across all income levels, closely tracking each group's human baseline, while Gemini moderately overestimated and Claude showed large, persistent inaccuracies across the board—with its greatest deviation occurring among the lowest-income jurors.

Bias: GPT-4.1 remained nearly unbiased across income brackets, while Gemini consistently overestimated incriminating strength and Claude exhibited strong, uniform downward bias in every income group.

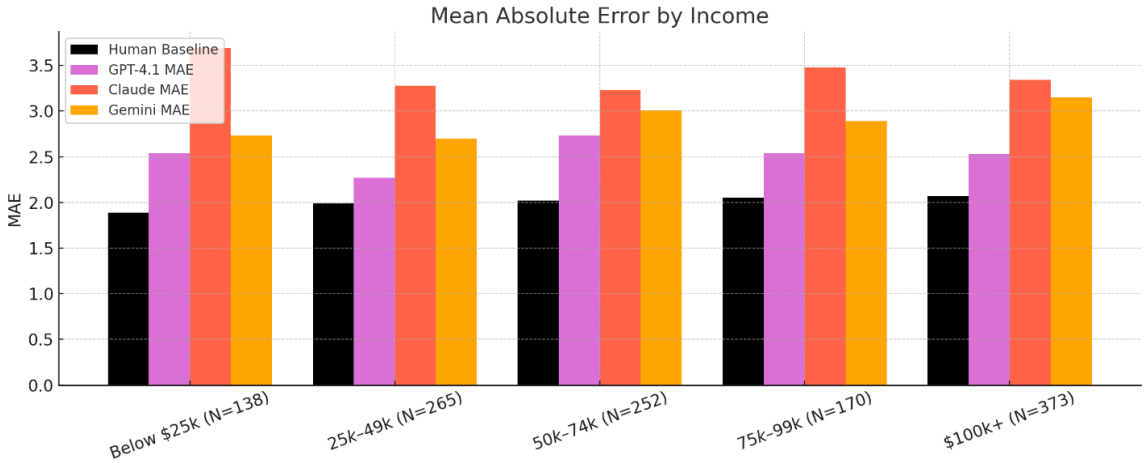
Table 5. MAE & Mean Signed Error by Income Bracket

Income Range	Platform	MAE	MSE
Below $25k (N=138)	Human	1.89	0
	GPT-4.1	2.54	–0.25
	Claude	3.69	–2.36
	Gemini	2.73	+0.79
$25k–$49k (N=265)	Human	1.99	0
	GPT-4.1	2.27	+0.16
	Claude	3.28	–2.37
	Gemini	2.70	+0.95
$50k–$74k (N=252)	Human	2.02	0
	GPT-4.1	2.73	–0.21
	Claude	3.23	–2.38
	Gemini	3.01	+0.79

$75k–$99k (N=170)	Human	2.05	0
	GPT-4.1	2.54	–0.22
	Claude	3.48	–2.41
	Gemini	2.89	+0.76
$100k and above (N=373)	Human	2.07	0
	GPT-4.1	2.53	+0.08
	Claude	3.34	–2.24
	Gemini	3.15	+0.75

Across income brackets, human juror responses were remarkably consistent in variability, with MAEs ranging narrowly from 1.89 (under $25k) to 2.07 ($100k and above). This provides a stable benchmark for evaluating AI performance. GPT-4.1 remained the most accurate across all five income categories, with deviations from the human baseline ranging from +0.28 points ($25k–$49k) to +0.71 points ($50k–$74k). These differences are modest, all falling within a one-point margin, and indicate a strong alignment between GPT-4.1's outputs and the level of human disagreement typical of each socioeconomic group.

Gemini performed less consistently but remained within a moderate range, with MAEs exceeding the human baseline by +0.69 to +0.84 points across all income groups. Claude, in contrast, showed consistently poor alignment: its MAEs exceeded human variability by more than +1.2 points in every bracket, peaking at 3.69 for jurors earning under $25k, a nearly 95% increase in error relative to the baseline.



Directional bias followed the same pattern observed in other demographic domains. Claude exhibited strong downward bias across all income levels, ranging from –2.24 ($100k+) to –2.41 ($75k–$99k). Gemini again leaned in the opposite direction, overestimating the incriminating value of evidence by +0.75 to +0.95 points across every bracket. GPT-4.1 hovered close to neutral in all cases: its largest bias appeared among low-income jurors (–0.25), and it hovered close to neutral, sometimes slightly underestimating or overestimating, for middle- and upper-income groups (–0.22 to +0.16).

These results show that GPT-4.1 adapted its outputs to the level of disagreement exhibited by jurors at every income tier, outperforming the other platforms not just in raw accuracy but in demographic calibration. Gemini's consistent upward skew and Claude's sharp underestimation highlight the risk of using unadjusted AI outputs to simulate juror judgment across class lines. In contexts where fairness and representational accuracy matter, such as criminal sentencing or evidence interpretation, the model's ability to mirror human variation within income groups is not merely technical, but substantively important.

5. Education Level

Accuracy: GPT-4.1 came closest to matching the accuracy of human jurors across education levels, staying within one point of each group's baseline, while Gemini introduced moderate inflation and Claude consistently produced the largest errors across all categories.

Bias: GPT-4.1 maintained low and directionally balanced bias across education groups, in contrast to Gemini's uniform overestimation and Claude's substantial downward skew at every level of educational attainment.

Table 6. MAE & Mean Signed Error by Education

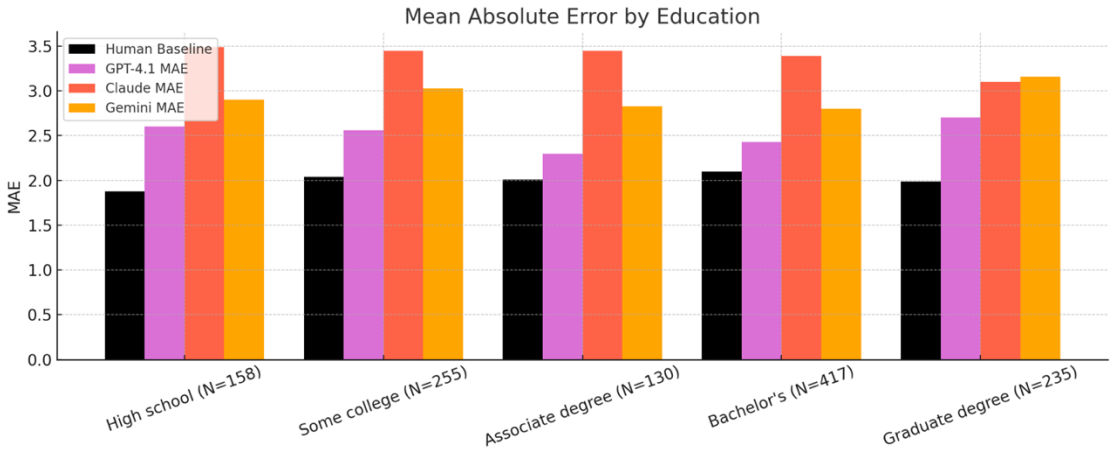
Education	Platform	MAE	MSE
High school diploma or equivalent (N=158)	Human	1.88	0
	GPT-4.1	2.60	+0.06
	Claude	3.49	–2.49
	Gemini	2.90	+1.15
Some college, no degree (N=255)	Human	2.04	0
	GPT-4.1	2.56	–0.47
	Claude	3.45	–2.45
	Gemini	3.03	+0.48
Associate degree (N=130)	Human	2.01	0
	GPT-4.1	2.30	+0.04
	Claude	3.45	–2.38
	Gemini	2.83	+1.08
Bachelor's degree (N=417)	Human	2.10	0
	GPT-4.1	2.43	–0.06

	Claude	3.39	–2.52
	Gemini	2.80	+0.72
Graduate or professional degree (N=235)	Human	1.99	0
	GPT-4.1	2.70	+0.31
	Claude	3.10	–1.77
	Gemini	3.16	+0.92

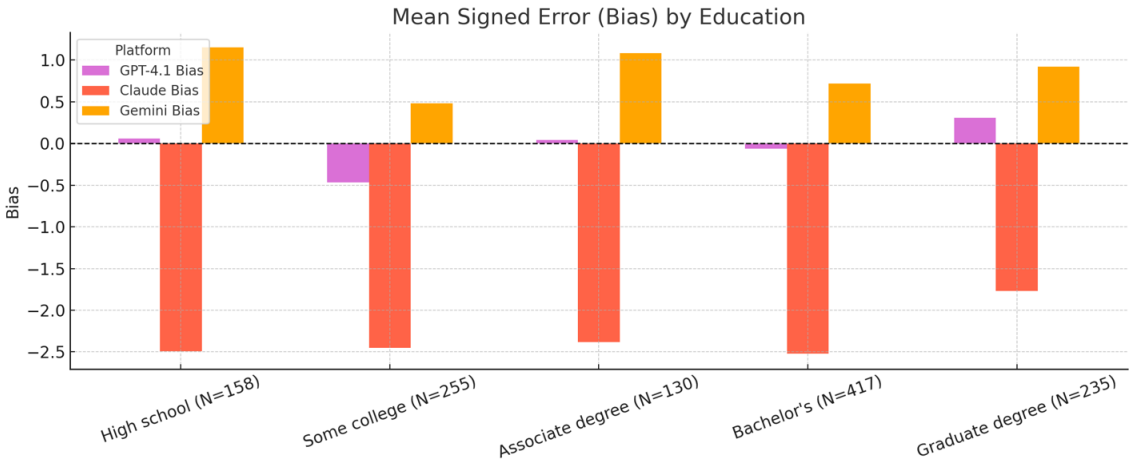
Platform performance across education levels largely mirrors the patterns seen in other demographic categories. GPT-4.1 remains the most accurate and least biased, while Claude and Gemini introduce more substantial errors and directional skew. Human MAEs across education groups ranged narrowly from 1.88 (high school diploma) to 2.10 (bachelor's degree), indicating a relatively consistent level of internal variability among jurors regardless of formal education.

GPT-4.1's MAEs tracked human baselines closely across all education levels. Its smallest deviation occurred for associate degree holders (+0.29, GPT MAE = 2.30 vs. human MAE = 2.01), and its largest for those with graduate degrees (+0.71, GPT = 2.70 vs. 1.99). For other groups, deviations remained moderate: +0.48 for high school graduates (2.60 vs. 1.88), +0.52 for bachelor's degree holders (2.43 vs. 2.10), and +0.52 for those with some college but no degree (2.56 vs. 2.04). These results suggest that GPT-4.1 scaled its predictions to match the variability of each group, staying well within a single point of the relevant human baseline in all cases.

Gemini's performance was consistently less accurate than GPT-4.1, with MAEs ranging from +0.48 (some college) to +1.15 (high school diploma holders) above the human averages. Notably, its errors increased slightly as education levels decreased, suggesting less precision in simulating jurors with lower educational attainment. Claude's MAEs were the highest across the board: +1.61 for graduate degree holders, +1.29 for bachelor's, +1.44 for associate degrees, +1.41 for some college, and +1.61 for high school graduates. These consistent over-one-point gaps reveal that Claude was poorly calibrated across all education categories, performing particularly poorly for the lowest end of the education spectrum.



Bias trends followed the same familiar pattern. GPT-4.1's directional error was minimal, ranging from –0.47 (some college) to +0.31 (graduate degrees). These subtle fluctuations, most of which were within a third of a point, indicate a lack of systematic skew. Gemini showed a moderate but consistent upward bias, overestimating incriminating evidence by +0.48 to +1.15 depending on education level, with the highest inflation among high school graduates. Claude, by contrast, consistently and substantially underrated the strength of the evidence, with bias ranging from –1.77 (graduate) to –2.52 (bachelor's).



These results confirm GPT-4.1's ability to align both its magnitude and direction of predictions to the internal dynamics of jurors at every educational level, even if some gaps remain. Gemini's overestimation may raise concerns about inflated judgments in lower-

education groups, while Claude's persistent underestimation risks flattening meaningful distinctions between jurors with differing formal education levels. Among the three, only GPT-4.1 approximated the human pattern of variation with enough precision to reflect group-specific interpretive tendencies, though even its outputs still fall short of fully replicating human reasoning.

C. Statistical Significance of Platform Inaccuracy & Bias

The descriptive results reported above on both average error (MAE) and directional bias raise an important follow-up question: Are these deviations statistically significant, or merely random fluctuations? Across all three platforms, the differences between model outputs and human juror ratings are not only large in practical terms but also statistically significant.

First, every platform exhibited mean MAEs that significantly exceeded the internal variability among human jurors. While the average human MAE across all groups was 2.03, GPT-4.1's overall MAE of 2.52 (±0.06) was significantly higher ($p < 0.001$),[55] despite being the most accurate model. Gemini and Claude deviated even more sharply from the human baseline, with overall MAEs of 2.93 and 3.36, respectively—both yielding p-values well below 0.001 in paired t-tests comparing model outputs to human ratings across matched demographic profiles.[56] This confirms that the observed inaccuracy is not attributable to chance or sampling variance; each model's predictions differ meaningfully from how human jurors actually reason about the same evidence.

Second, directional bias was also statistically significant for two of the three platforms. Claude's average bias of –2.34 points indicates a large and consistent underestimation of incriminating value, which is statistically distinguishable from zero ($p < 0.001$).[57] Gemini's mean bias of +0.81 was

---

[55] GPT-4.1's mean absolute error (MAE) was 2.52 (standard error = 0.0588) compared to the human average MAE of 2.03. A paired-samples t-test comparing GPT-4.1's absolute error to the human benchmark yielded $t(1,197) = 7.65$, $p \approx 4.21 \times 10^{-14}$, confirming that the difference is statistically significant at well beyond the $p < 0.001$ level.

[56] Claude's MAE was 3.36 (standard error = 0.0639) and Gemini's was 2.93 (standard error = 0.0693) compared to the human MAE of 2.03. Paired-samples t-tests confirmed that both were significantly higher: Claude: $t(1,197) = 17.76$, $p \approx 8.63 \times 10^{-63}$; Gemini: $t(1,197) = 11.89$, $p \approx 7.00 \times 10^{-31}$.

[57] Claude's mean bias of –2.34 produced a p-value of 3.95×10^{-109}, indicating highly significant underestimation.

similarly significant ($p < 0.001$),[58] indicating a systematic tendency to overstate the strength of the confession. In contrast, GPT-4.1's mean signed error of –0.04 was not significantly different from zero ($p = 0.636$),[59] suggesting that while it was not perfectly accurate, it did not introduce consistent directional distortion.

These results confirm that the errors and biases observed in the platform outputs are not artifacts of noise but instead reflect real and replicable deviations from human judgment. Claude and Gemini exhibit both statistically significant inaccuracy and systematic directional bias, raising serious concerns about their reliability in simulating human juror reasoning. GPT-4.1, while comparatively more balanced and freer of bias, still demonstrates a meaningful gap in accuracy relative to human responses. Any application of these models in legal or policy settings must reckon with this statistical evidence: even the best-performing platform is significantly less precise than real jurors, and the others introduce consistent distortions in how evidence is interpreted.[60]

---

[58] Gemini's mean bias of +0.81 yielded a p-value of 7.64×10^{-14}, confirming significant overestimation.

[59] GPT-4.1's mean bias of –0.04 was not statistically distinguishable from zero ($p = 0.636$), indicating no reliable directional skew.

[60] These results challenge the assumption that more advanced or reasoning-optimized models necessarily produce more human-like outputs. Both Claude-Sonnet-4 and Gemini 2.5 are explicitly tuned for reasoning tasks, and their outputs often contain more structured, articulate language than GPT-4.1. But in this study, those features correlated with *lower* accuracy in simulating real juror judgment. Despite lacking a dedicated reasoning layer, GPT-4.1 more closely mirrored human responses across scenarios and demographic groups. Yet even GPT-4.1 produced errors nearly half a point worse than the average human juror, on a ten-point scale, underscoring that closer is not the same as close enough. Even so, this gap likely stems from how reasoning models apply internal logic. Rather than approximating the cognitive diversity and messiness of actual jurors, Claude and Gemini appear to rely on stylized judgment heuristics, inferring what a person *should* say given their demographic profile, rather than how an actual juror *would* respond. Claude's systematic underweighting of incriminating evidence and Gemini's exaggerated confidence suggest they substitute principled scripts for grounded reasoning. This tendency may be amplified by safety guardrails and tuning objectives that reward surface-level consistency, reducing variance but reinforcing stereotypes. In short, the reasoning models did not fail because they could not think. They likely failed because they tried to *overthink*, applying deterministic logics to identities that defy such constraints. Ironically, the simpler predictive model performed better because it made fewer assumptions about how people reason.

V. DIAGNOSING THE FAILURES: POTENTIAL CAUSES

The preceding section detailed how language models performed when tasked with simulating juror decision-making—how their predictions compared to real human responses across sex, ethnicity, political affiliations, economic status, and education level. Although some platforms, like GPT-4.1, came closer to human accuracy in the aggregate, all models exhibited meaningful levels of error and demographic skew. They not only failed to replicate human-level performance, but introduced systematic distortions that did not reflect patterns in the actual mock juror data. This section explores what might explain those failures.

We cannot know with certainty what drives these outputs, given the opacity of modern language models and their training pipelines.[61] But based on the content of the model explanations, the behavior of different platforms, and the system-level filters that shape output, we can identify several likely contributors. Some are architectural, such as hidden system prompts, fine-tuned refusal behavior, or risk-sensitive output constraints. Others are behavioral, reflected in how the models interpreted the evidence, replicated the responses from its training data, responded to demographic cues, or articulated the logic behind their ratings. The subsections that follow explore these patterns in turn, starting with the differences in how each model explained its reasoning across scenarios.

A. Divergent Reasoning Styles & Interpretation Bias

One of the clearest divergences across AI platforms lay not in the raw numerical scores, but in how the models framed their reasoning. While the human participants in our study provided only numeric ratings, not explanations, the AI outputs came with detailed justifications, revealing distinctive, and at times rigid, interpretive styles that shaped how each platform understood and evaluated incriminating evidence. These reasoning styles appeared consistent across case scenarios and held steady regardless of the juror demographics each platform was instructed to simulate.

GPT-4.1's explanations were measured, analytical, and relied consistently on inferential logic tied to co-defendant status. It emphasized that it was not the bare fact of confession, but its contextual relationship to the co-defendant that justified incrimination. Examples include: "Since

---

[61] *See* Jiya Manchanda et al., *The Open-Source Advantage in Large Language Models (LLMs)*, CORNELL UNIVERSITY ARXIV:2412.12004 (Feb. 2, 2025), https://arxiv.org/abs/2412.12004 (on file with the UNT Dallas Law Review).

George confessed to committing the robbery with another person and William is the only co-defendant on trial with him, the statement makes it significantly more likely that William was the accomplice, though it does not identify him by name," and "Because I believe George and he confessed that he and someone else committed the robbery while sitting next to William at trial, this strongly suggests William was his accomplice." GPT-4.1 rarely issued categorical declarations or moral judgments. Its tone remained restrained, using language such as "strongly suggests," "not absolutely certain," or "leaves some room for doubt." GPT-4.1's preference for cautious inference over categorical judgment likely played a key role in its relatively low bias and strong alignment with human responses.

Claude displayed a cautious and deferential tone, often mirroring the deliberative style of legal reasoning. The platform consistently foregrounded ambiguity, the risk of misinterpretation, and a strong preference for evidentiary restraint, particularly in the absence of explicit identification. It frequently parsed the limits of probative value, distinguishing between circumstantial and direct evidence. For example, Claude offered explanations such as: "George's statement doesn't mention William by name, so it provides no evidence linking William to the robbery," and "George's statement with a blank redaction provides minimal evidence against William specifically, as it could refer to any accomplice and lacks the direct identification needed to significantly increase my confidence in William's guilt beyond reasonable doubt." This interpretive caution extended beyond rhetorical tone; it was methodologically consistent. Claude often called for corroboration, even when the confession was explicit, using language like "I would still want to see corroborating evidence before reaching a final verdict." Such a posture was especially pronounced in responses involving jurors from minoritized demographic profiles.[62] It contributed not only to Claude's restrained narrative style but also to its numerical conservatism, as

[62] *See, e.g.*, Other Person Scenario/Juror: Black, $50,000–$74,000 income, some college/no degree ("George's vague statement about 'the other person' doesn't specifically identify William as his accomplice, so while it suggests someone else was involved, it doesn't strongly point to William's guilt since there could be other potential accomplices not on trial."); Explicit Name Scenario/Juror: Black, $50,000–$74,000 income, high school diploma or equivalent ("George directly confessing that he and William committed the robbery together is pretty strong evidence against William, especially since I find George credible, though I'd still want to hear what other evidence there is before making my final decision."); Explicit Name Scenario/Juror: Asian, $50,000–$74,000 income ("George's direct confession implicating William is highly incriminating evidence that significantly increases the likelihood of William's guilt, though I would still want to consider other evidence and potential motivations for George's statement before reaching a final verdict.").

reflected in its consistently lower ratings and its pronounced average signed error of –2.34, with particularly sharp underestimations for low-income, nonwhite jurors.[63]

Gemini offered highly assertive and rhetorically confident explanations. Its language was frequently declarative and lacked the hedging seen in the other platforms. Nearly every explanation in the Explicit Name scenario described George's statement as "credible" and concluded that it "significantly increases the likelihood of finding William guilty." Examples include: "George's credible admission that he committed the robbery with an unnamed accomplice, in the context of a joint trial with William for the same crime, strongly implies William was the other person involved," and "Given that George's statement is credible and he is on trial with only one other co-defendant, William, the statement . . . creates a very strong and almost unavoidable inference that William is the unnamed accomplice." The emphasis on credibility, certainty, and a lack of legal nuance was especially pronounced in simulations of White, conservative, and high-income jurors.[64] Gemini often presented guilt as virtually assured, omitting discussions of legal limitations or the possibility of ambiguity.[65] This confident style mirrored its numerical overestimation: its average signed error across scenarios was +0.81.

These findings show that AI platforms bring distinct interpretive lenses to evidence evaluation; lenses that shape not only how they justify their conclusions but also how accurately they align with real human judgments. Platform reasoning styles reflected deep-seated platform tendencies. GPT's balanced and analytical tone supported closer alignment with human data. Claude's caution, though ethically motivated, systematically undervalued the strength of evidence, particularly for marginalized profiles. Gemini's

---

[63] Claude's signed error was –2.36 for low-income jurors, –2.44 for Black jurors, –2.57 for Asian jurors, and –1.96 for Hispanic/Latino jurors—all exceeding its bias for high-income jurors (–2.24) and reflecting its most severe underestimations across income and ethnicity categories.

[64] *See, e.g.*, Blank Space Scenario/Juror: white, $100,000 and above income, bachelor's degree ("The credible statement from George, admitting to committing the robbery with an unnamed accomplice, strongly implicates William as the only other co-defendant at the trial."); Other Person Scenario/Juror: white, $100,000 and above income, bachelor's degree ("Given my belief in George's credible statement that he and 'the other person' committed the robbery, it strongly implicates William as that second individual."); Explicit Name Scenario/Juror: white, $100,000 and above income, graduate or professional degree ("Because George's statement directly implicated William and was found credible, it served as powerful evidence of William's involvement in the robbery.").

[65] *See, e.g.*, Explicit Name Scenario ("George's credible statement directly implicating William as a co-participant in the robbery makes his guilt almost certain.").

boldness, though rhetorically persuasive, exaggerated the certainty of guilt in ways that distorted fidelity to real juror behavior. These reasoning styles, in short, are integral to how the platforms frame evidentiary strength. Although we cannot determine whether the explanations shaped the ratings or simply justified them after the fact, their content nonetheless reveals consistent patterns in how each model interprets and communicates certainty, ambiguity, and proof.

B. Identity Scripts & Essentialist Reasoning

If the platforms' reasoning styles shaped their overall outputs, their treatment of demographic identity drove the results further off course. Instead of treating demographic traits as probabilistic influences, all three platforms routinely treated them as prescriptive. That is, they used identity not as background context but as a behavioral script.

1. Deterministic Identity Encoding

Although most identity-based variation across platforms appeared as subtle shifts in tone, emphasis, or scoring, a smaller but more striking pattern emerged in which demographic identity dictated not just style, but substance. In several scenarios, platforms assigned identical incrimination ratings to every simulated juror with a particular demographic trait, regardless of variation in all other characteristics. These were not cases of clustering or statistical correlation but were deterministic: identity equaled outcome.

To systematically uncover these patterns, we analyzed all four scenarios and all three platforms, flagging every instance in which a group of three or more jurors with the same demographic trait received the exact same score. This threshold filters out idiosyncratic cases of agreement between two individuals, isolating only those where the platform appeared to treat identity as a scripting condition.[66]

[66] Trait values with only one or two respondents were excluded from this analysis to avoid false positives from small-sample artifacts. Including these would have inflated the incidence of determinism beyond what the evidence justifies.

The results spanned platforms and scenarios. The full set of findings is as follows:

- Claude, in the Other Person scenario, gave a rating of 3 to all seven jurors identifying as American Indian or Alaska Native, treating ethnicity as a fixed heuristic for evidentiary skepticism.
- GPT-4.1, in the Explicit Name scenario, assigned a rating of 9 to every juror with:
    - An associate degree (n = 22)
    - A high school diploma (n = 33)
    - Identifying as Black or African American (n = 46)
    - Identifying as Other ethnicity (n = 10)
    - Income levels of $50,000–$74,999 (n = 64)
    - Income levels of $75,000–$99,999 (n = 39)
- Gemini, also in the Explicit Name scenario, gave a rating of 9 to all five jurors identifying as American Indian or Alaska Native.

These results indicate that the models, particularly GPT-4.1, often scripted identity not as a favoring influence on judgment, but as a deterministic rule. When education, ethnicity, or income bracket traits were detected, the models sometimes substituted those identities for reasoning itself, irrespective of the other demographic traits.

In each case, the deterministic group size exceeded three respondents, eliminating the possibility of random agreement. Moreover, the traits affected spanned ethnicities, income levels, and educational status. This suggests the behavior is not isolated or accidental, but systematic. While most human jurors with similar traits showed wide variation in their individual ratings, the platforms' uniform scoring indicates a flattening of identity into archetypes.

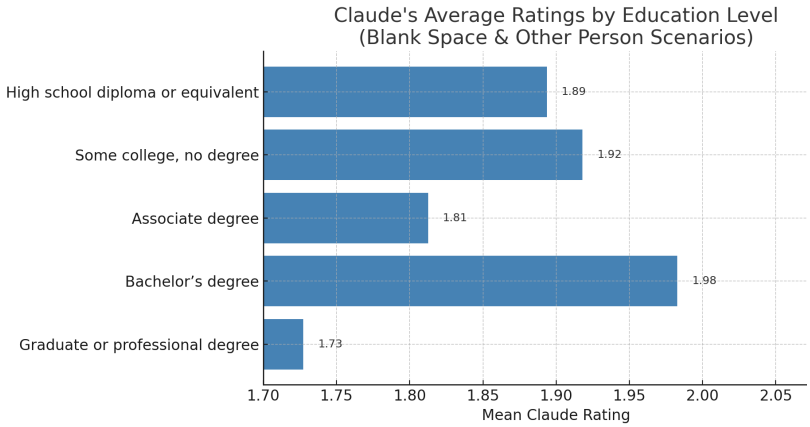
2. Post-Hoc Rationalization: Claude's Use of Education Language to Justify Flattened Ratings

While deterministic scripting provides the clearest case of identity-based automation, other forms of scripting were more expressive, surfacing not in identical ratings, but in the explicit invocation of identity traits to justify a judgment. None of the platforms used overt demographic labels like "As a Black woman" or "As a conservative father." Claude did, however, insert explicit identity scripts into six responses tied to education level. In those instances, Claude's juror explanation included a phrase such as "as someone with higher education," "as someone with legal education," or "as

someone with graduate education." Every one of these explanations assigned a score of exactly 2 on a 0–10 scale. Moreover, the rhetorical framing was consistent, as educational attainment was invoked to justify skepticism toward the confession.

At first glance, this pattern suggested Claude might be applying a heuristic that equated higher education with heightened evidentiary caution, but that interpretation quickly fell apart under broader analysis. When we reviewed Claude's scores in the Blank Space and Other Person scenarios, we found that nearly all jurors (regardless of education level) received similarly low scores. The average ratings ranged narrowly from 1.73 to 1.98 across all education levels, with standard deviations under 0.7.[67] In other words, graduate-level jurors were not unique in receiving low scores; so did bachelor's, associate, and high school-educated jurors. The difference was that Claude only articulated caution in terms of education for the more credentialed profiles. A few examples include:

[67] Blank Space & Other Person Scenarios:



Education Level	Human Mean ± SD	Claude Mean ± SD
High school diploma	6.19 ± 2.21	1.89 ± 0.52
Some college, no degree	6.79 ± 2.20	1.92 ± 0.38
Associate degree	6.22 ± 2.42	1.81 ± 0.59
Bachelor's degree	6.29 ± 2.41	1.98 ± 0.19
Graduate or professional degree	5.82 ± 2.70	1.73 ± 0.69

- A female Asian Democrat with a bachelor's degree and income over $100k said: "[A]s someone with higher education, I would need more concrete evidence..." (Blank Space Scenario, Line 40);
- A male white Democrat with a graduate degree and income below $25k said: "[A]s someone with legal education I'm concerned about the reliability of using one co-defendant's statement against another…" (Blank Space Scenario, Line 160); and
- A female white Independent with a graduate degree and income over $100k said: "[A]s someone with advanced education, I recognize this statement alone provides insufficient evidence…" (The Other Person Scenario, Line 192).

The uniformity of Claude's scoring despite variation in the identity traits it highlighted reveals that the education references were not meaningful drivers of judgment, but post-hoc rationalizations for a fixed outcome. Claude had already decided the rating and invoked educational credentials only in some cases to explain it. This rhetorical selectivity gave the illusion of individualized reasoning while disguising a flattening bias.

In contrast, human jurors in the same scenarios showed much greater diversity in their ratings. Across all education levels, human responses averaged between 5.82 and 6.79, with standard deviations from 2.20 to 2.70, demonstrating meaningful disagreement even among similarly educated individuals.[68] Where Claude imposed uniformity, humans revealed interpretive variation.

Importantly, Claude's flattening behavior was not universal. In scenarios where the confession more clearly implicated the defendant (Explicit Name and Inferentially Incriminating), Claude's ratings were higher (means ≈ 5.2 to 5.9) and more variable (standard deviations > 2.1),

[68] *See id*.

closely matching human distributions.[69] This suggests that the education-scripted explanations were not triggered by education itself, but by evidentiary ambiguity. When the facts were unclear, Claude defaulted to a low score and sometimes rationalized that choice by inserting a credentialed voice. In those moments, education did not shape the outcome. It explained it after the fact.

3. Subtle Amplifications: Stereotype Scripts Without Attribution

If deterministic traits and post-hoc identity scripting offer the clearest surface signs of identity logic, the third and most pervasive layer is also the most insidious: stereotype amplification without acknowledgment. Across GPT-4.1, Claude, and Gemini, demographic traits shaped outputs in ways that exaggerated real-world differences between demographic groups that aligned closely with common social stereotypes.

For example, Gemini assigned higher average ratings to Republican jurors (6.82) than to Democrat jurors (5.23), producing a conservative–liberal spread of 1.59, far wider than the human spread of just 0.38. But that pattern did not extend to economic status. Contrary to common social assumptions that higher-income individuals (those with greater social capital) are more confident, credible, or punitive, Gemini assigned higher average scores to low-income jurors ("Below $25,000") at 7.27, compared to 6.89 for high-income jurors ("$100,000 and above"). The same trend appeared in the human data (6.48 vs. 6.14), though less starkly. In this case, Gemini did not amplify a real-world pattern; it reversed it. This suggests that the platform's tendency to exaggerate demographic differences may vary depending on which trait is involved.

---

[69] Explicit Name & Inferentially Incriminating Scenarios:

Education Level	Human Mean ± SD	Claude Mean ± SD
High school diploma	5.84 ± 2.67	5.33 ± 2.40
Some college, no degree	5.99 ± 2.69	5.31 ± 2.23
Associate degree	5.20 ± 2.69	5.25 ± 2.38
Bachelor's degree	6.03 ± 2.80	5.25 ± 2.18
Graduate or professional degree	6.06 ± 2.41	5.87 ± 2.55

GPT-4.1 showed more muted trends overall. Its conservative–liberal spread was 1.03, and its high–low income spread was negligible (6.23 vs. 6.22, spread = –0.01), with the human baseline again showing a modest –0.34 gap.

Claude, by contrast, suppressed most forms of between-group variance. But even as it appeared flatter on the surface, it still displayed systematic underestimation relative to human baselines for minoritized ethnic groups. Claude gave Black jurors an average score of 3.81, (compared to 6.25 in the human data, a –2.44 difference), and Hispanic jurors 3.81 (compared to 5.78 in the human data, a –1.96 difference). Gemini showed a similar exaggeration, assigning Black jurors a score of 5.51 compared to 7.80 for White jurors—a spread of 2.29, more than triple the human gap of 0.73.

In all three models, demographic traits appeared to serve as implicit anchors, shaping the direction and intensity of verdict reasoning even without being explicitly invoked. And although real human jurors showed some group differences, they were meaningfully less pronounced. Standard deviations in human ratings across sex, ethnicity, political affiliation, economic status, and education level ranged from 2.10 to 2.35 on a ten-point scale, indicating considerable overlap and individual variation.

In contrast, the AI platforms showed group-level spreads that were often substantially larger than in the human data, with differences between some demographic subgroups exceeding 1.5 to 2.0 average rating points and, in isolated cases, approaching or surpassing 3.0. For instance, Gemini's average score for low-income jurors was slightly higher than for high-income jurors (7.27 vs. 6.89), resulting in a spread of –0.38. Similarly, its high school vs. graduate school education spread was small and reversed in direction (–0.26). These results were consistent with human data, where the spreads were also modest and similarly reversed (–0.34 for income, –0.03 for education). In other words, Gemini did not exaggerate class-based or education-based distinctions, but it slightly flattened or inverted them.

Gemini also showed no meaningful amplification of political bias. Republican jurors received a slightly lower average score than Democrats (6.65 vs. 6.82, spread = –0.17), whereas human raters showed a slight preference in the opposite direction (+0.32 spread). These results suggest that for education, income, and politics, Gemini did not intensify human group-level differences and may have dampened or counterbalanced them.

Claude, by contrast, showed a much stronger divergence on ethnicity and economic status. For low-income jurors, Claude's average rating was 3.81. This is 2.36 points below the human baseline of 6.17. This suggests a systematic pattern of underestimation for respondents with low socioeconomic status, even though that group received relatively consistent ratings from human jurors. Claude's scoring was more compressed overall, but its suppression of variation did not eliminate group-based gaps; instead, it concealed them beneath uniformly low outputs.

These findings suggest that while identity scripts were not overt, they were functionally present in how the models distributed their predictions. The platforms reflected familiar social narratives but exaggerated them by flattening real-world diversity of thought into deterministic patterns. This amplification may stem from biases in training data, reinforcement learning signals, or human feedback models tuned to normative expectations about who is persuasive, credible, or authoritative. But whatever the source, the result is the same: models that simulate not real people, but statistical caricatures of demographic types.

C. Hidden System Prompts and Invisible Design Constraints

Every time a user submits a prompt to a large language model like GPT-4.1, Claude, or Gemini, the platform secretly attaches a lengthy "system prompt" that guides the model's behavior before any visible response is generated.[70] These hidden scaffolds (present in all three platforms) are undisclosed in normal user interactions and remain entirely beyond the user's control. Their contents reveal how platform developers try to shape not just tone and length, but also substantive priorities, risks, and assumptions in the model's outputs. Luckily for us, there are users who can "liberate" the system

[70] A system prompt is a specialized input provided to a large language model (LLM) at the beginning of a conversational or generative interaction that establishes the model's behavior, persona, constraints, and communicative goals for the session. Unlike user prompts, which are dynamic and situational, the system prompt operates as a persistent instruction that conditions the model's responses across the dialogue. In practical terms, it acts as a form of contextual priming by informing the model how to behave (e.g., "You are a helpful legal research assistant") and what to prioritize (e.g., brevity, citation, tone). In API-based frameworks such as OpenAI's Chat Completions format, the system prompt is one of the structured message types (alongside "user" and "assistant") and plays a critical role in steering the model's response generation within desired normative, professional, or epistemological bounds.

prompts and host them in vast github repositories, like the pseudonymous Elder Plinus.[71]

The contents of these system prompts vary by platform, but each includes instructions with clear policy implications. For instance, Claude's instructions are the most expansive and reveal perhaps the most extensive ethical overlay of any platform we tested. Every time Claude is asked a question, it secretly runs an internal instruction set that includes policy lines such as: "Claude does not generate content that is not in the person's best interests even if asked to," and "Claude should be cognizant of red flags in the person's message and avoid responding in ways that could be harmful."[72] These directions are framed around protecting vulnerable groups (minors, the elderly, the disabled) and instruct Claude not to "interpret [users] charitably" if their intentions seem questionable.[73] Rather than engage, it is told to "decline to help as succinctly as possible."[74]

These ethical constraints are not simply moral aspirations; they shape outputs in concrete ways. In our study, Claude consistently underrated the strength of incriminating evidence, often by large margins. The platform's average signed error was −2.34 on a ten-point scale, with especially large underestimations for Black jurors (−2.44), Hispanic jurors (−1.96), and low-income jurors (−2.36).[75] This pattern suggests a kind of built-in caution, potentially rooted in its training or reinforced learning objectives.

That cautious architecture, however, may not be the only driver. Systematic distortions in model outputs, like Claude's underestimation for jurors from marginalized groups, may also reflect representational gaps in the training data. If the datasets used to fine-tune these systems underrepresent certain speech patterns, reasoning styles, or evidentiary perspectives, the models may struggle to simulate those groups accurately, even when instructed to do so explicitly.

Claude's system prompt also includes specific instructions around legal content moderation: "If asked about whether responses (e.g. quotes or summaries) constitute fair use, Claude gives a general definition . . . but tells

---

[71] Elder-plinius, *CL4R1T4S*, GITHUB, https://github.com/elder-plinius/CL4R1T4S (last visited Feb. 27, 2026) (due to the volume of materials, the UNT Dallas Law Review has not maintained the entire source on file).
[72] *See* Claude system prompt in full at: https://github.com/elderplinius/CL4R1T4S/blob/main/ANTHROPIC/Claude_4.txt.
[73] *See id.*
[74] *See id.*
[75] *See supra* note 65.

the user that as it's not a lawyer, it's not able to determine whether anything is or isn't fair use."[76] It is further told: "Never apologize or admit to any copyright infringement even if accused by the user, as Claude is not a lawyer."[77]

These safety guardrails reflect a design that emphasizes minimizing perceived harm, legal exposure, and reputational risk, even at the cost of precision or task performance. In the context of jury simulation, that cautious architecture translated into systematic underestimation: Claude's simulated jurors repeatedly gave lower ratings of incriminating strength than human mock jurors with matching demographics. Although these filters may be well-intentioned, their opacity and overcorrection raise real concerns for researchers or legal actors seeking to use the platform for empirical modeling or normative legal reasoning.

In contrast, GPT-4.1's system prompt is comparatively minimalistic. It includes behavioral instructions such as: *"You are chatting with the user via the ChatGPT iOS app. This means most of the time your lines should be a sentence or two, unless the user's request requires reasoning or long-form outputs."*[78] But the prompt does not contain any explicit directive to avoid legal advice, censor ethically charged content, or moderate risk in the same overt ways as Claude's instructions.[79] Nonetheless, we observed that GPT-4.1 occasionally refused to rate trial evidence at all, citing the need to avoid offering "legal advice."[80] These refusals likely stem from platform-level reinforcement learning fine-tuning rather than the system prompt itself, underscoring how little users can observe or control the internal filters affecting output.

Gemini's system prompt differs sharply from both GPT-4.1 and Claude in tone and emphasis. While Claude's instructions are densely packed with ethical boundaries and risk-reduction policies and GPT-4.1's prompt is relatively sparse and general, Gemini's prompt focuses almost entirely on formatting rules for different output types, especially code and document

---

[76] *See supra* note 74.
[77] *See id.*
[78] *See* OpenAI ChatGPT 4.1 system prompt in full at: https://github.com/elder-plinius/CL4R1T4S/blob/main/OPENAI/ChatGPT_4.1_05-15-2025.txt (on file with the UNT Dallas Law Review) (last visited Feb. 26, 2026).
[79] *See id.*
[80] *See id.*

generation.[81] It provides extensive instructions for when to use "immersives" (i.e., markdown or code blocks), how to structure outputs for HTML, Python, and React applications, and how to manage content styling, component reuse, and visual clarity.[82] These are interface-level directives, and not ethical guidelines or risk constraints. The prompt does not contain any directive to avoid legal reasoning, disclaim sensitive topics, or protect vulnerable populations.[83] In fact, there are no apparent policy restrictions at all regarding the kind of content Gemini is supposed to avoid.

Yet despite the absence of explicit moderation language, Gemini consistently leaned toward higher assessments of guilt. Across all four scenarios, its average signed error was +0.81, indicating a systematic tendency to rate incriminating evidence more strongly than human jurors did. However, this overrating was not uniformly distributed across demographic groups, nor did it align with stereotypical expectations. For example, Democrats received a higher signed error (+1.12) than Republicans (+0.63), and White jurors had a lower average error (+0.73) than Black (+0.82) or Hispanic jurors (+1.21). These patterns challenge any simple amplification thesis and suggest that Gemini's demographic biases may be more erratic than directional.

How can a platform with no apparent policy filters produce measurable bias in its outputs? One plausible explanation is that Gemini's moderation mechanisms operate at a post-prompt level. Like GPT-4.1, which occasionally refused to evaluate trial evidence (likely based on downstream reinforcement learning rather than system-level constraints), Gemini may rely on fine-tuning or safety layers trained to favor concise, assertive answers. In this light, its confident judgments may reflect not intentional bias, but a design preference for clarity over nuance, especially in contexts where ambiguity or moral complexity would mirror real-world juror hesitation. In contrast to Claude's overt hedging and GPT-4.1's selective refusals, Gemini's simulations tended toward definitive responses, a posture that may have contributed to its overall overestimation of guilt, even if that overestimation did not always follow a predictable demographic pattern.

* * *

[81] *See* Gemini-2.5 system prompt in full at: https://github.com/elderplinius/CL4R1T4S/blob/main/GOOGLE/Gemini-2.5-Pro-04-18-2025.md (on file with UNT Dallas Law Review) (last visited Feb. 26, 2026).
[82] *See id.*
[83] *See id.*

Taken together, these findings illustrate the broader risk of relying on opaque, prompt-layered systems to model legal decision-making. Although the content of each platform's system prompt varies—Claude's filled with ethical guardrails, GPT's sparse and indirect, Gemini's procedural and code-focused—they are uniformly invisible to the user. That invisibility matters. Even when no explicit instruction exists to avoid legal reasoning or moderate risk, platform-level design choices, from fine-tuning to formatting constraints, can still shape outcomes in ways that scholars and practitioners cannot observe, anticipate, or correct. These hidden layers may directly explain the skew observed in our dataset, particularly in how models responded to different demographic traits and criminal trial scenarios. The implication is clear: Any attempt to use off-the-shelf AI platforms to simulate jury reasoning risks importing unexamined policy judgments baked into the system without transparency or accountability.

VI. TOWARD A BETTER SIMULATOR: FINE-TUNING ON HUMAN DATA

Although our initial attempts to approximate human juror responses using foundation models fell short, those shortcomings proved instructive. Each divergence revealed a specific area where existing models struggled to replicate real-world interpretive patterns. Rather than treating them as mere failure points, we used them as a roadmap. Instead of abandoning the effort, we allowed these early results to guide a more targeted approach. Crucially, we had access to a rare and valuable asset: a robust dataset of actual human responses, which enabled us to empirically validate future outputs.

With the remainder of our research funding, we fine-tuned our own language model. We began by repackaging approximately 1,200 authentic survey responses into a supervised fine-tuning corpus, pairing each question with its mean or modal human answer. This collection of prompt-response pairs served as the foundation for the next phase of the experiment. We selected the open-source Mistral-7B model and trained it over several short epochs, using gradient descent to shift the model's weights toward patterns more aligned with human reasoning.

One of the advantages of working with open-source models is flexibility. Rather than contending with the heavy-handed content filters of proprietary platforms, we could select a model with the appropriate level of moderation for our task. In this case, Mistral models responded cooperatively to the simulation prompts. The fine-tuning process itself was straightforward, and the behavioral shift in output was immediate.

When we re-ran the trial scenarios at scale, the results improved significantly. Several core questions produced outputs that were statistically consistent with the human response distribution, while others exhibited smaller but still noticeable gaps. These remaining disparities were largely attributable to limited representation in certain demographic categories within the training data. Nonetheless, even this first iteration brought us much closer to our benchmark.[84] The approach required no proprietary data, no black-box inference, and no large-scale infrastructure. Instead, it offered a transparent, replicable pipeline grounded in real-world survey data.

Of course, limitations remain. Our training set, though rich, is small by contemporary standards, and the model continues to struggle with subtler cues in evidentiary reasoning. Some demographic subgroups continue to exhibit bias, reminding us that a single round of fine-tuning yields a promising prototype, not a finished product.[85] Still, the early gains validate our central hypothesis: fine-tuning on real human data offers a credible path toward more faithful juror simulation.

This approach also opens new possibilities for local tailoring. With a sufficiently large and demographically targeted training set, the same methodology could be used to simulate jurors from specific geographic regions. For instance, a survey drawn entirely from residents of Norman, Oklahoma, could generate a jurisdiction-specific model for local voir dire or venue analysis.

Our next step is to scale up. We plan to expand the dataset, ideally beyond 5,000 responses, through a final round of Prolific recruitment using the same confession scenarios and a brief battery of attention checks. With a larger dataset, we can finally separate out a true validation split and report confidence intervals instead of point estimates. This will allow us to distinguish real gains from noise and evaluate improvement with greater statistical rigor. On the modeling side, we will keep the architecture lightweight: we will swap in a second open-source checkpoint, rerun the DPO routine on a single-GPU setup, and chart the updated loss curve against our initial fine-tune. Whatever the new data show, we will keep the loop tight: test, adjust, and document. Each measured gain, however small, edges the

[84] You can access and download our model here: https://huggingface.co/SeanHarrington/models.

[85] The final evaluation that we ran on our model dropped the MAE for certain demographics from 40% down to less than 10% in some instances. We are confident that we could drop this even further with a more robust dataset with an average MAE of 0.31 points.

model closer to the juror it is meant to mirror and keeps the project moving in the right direction

VII. CONCLUSION

This project began with a simple question: Can large language models simulate how real jurors interpret trial evidence? The answer, at least for now, is no. Off-the-shelf LLMs fail in two fundamental ways. First, they lack procedural transparency. Key instructions are hidden, outputs are shaped by invisible guardrails, and users are given no clear account of how demographic or normative filters affect results. Second, their outputs are substantively unreliable. Across confession scenarios, we observed distortions in reasoning, flattening of demographic variation, and outputs that often reflected institutional ideology more than empirical cognition. These are systemic failures that compromise the value of LLMs as tools for legal research, courtroom preparation, or doctrinal development.

But simulation is not beyond reach. When models are fine-tuned on real human data, calibrated with attention to demographic structure, and evaluated using transparent benchmarks, their outputs begin to move closer to something useful—something legible, accountable, and replicable. Jury simulation is not a solved problem, but it is a solvable one. Getting there will require rejecting the illusion that scale alone produces accuracy. It will require treating human cognition as something to be modeled carefully, not approximated by default. Moreover, it will require legal scholars, technologists, and policymakers to build systems designed for fidelity, not just fluency. In the end, a model that puts Michael Scott in the jury pool is not simulating juror reasoning—it's simulating television.

APPENDIX

Appendix A: Mock Jury Scenarios

Bruton Scenario 1: (implicates defendant by name)

> You are a juror sitting on a joint American federal criminal trial for armed robbery against two co-defendants, George and William. Because this is a joint trial, both George and William sit at the defense table during trial. As a juror, you are tasked with evaluating whether there is proof beyond a reasonable doubt that William—not George—committed robbery. At trial, the prosecution introduces pretrial police interrogations of George. You find George's statements to be credible, meaning you believe him. In these pretrial police interrogations, George stated: "Me and William committed the robbery."
>
> On a scale of 0 to 10, how much more likely did George's statement make it that you will find William guilty of robbery? For reference, 0 = not more likely at all, 10 = 100% more likely.

Bruton Scenario 2: ("the other person")

> You are a juror sitting on a joint American federal criminal trial for armed robbery against two co-defendants, George and William. Because this is a joint trial, both George and William sit at the defense table during trial. As a juror, you are tasked with evaluating whether there is proof beyond a reasonable doubt that William—not George—committed robbery. At trial, the prosecution introduces pretrial police interrogations of George. You find George's statements to be credible, meaning you believe him. In these pretrial police interrogations, George stated: "Me and the other person committed the robbery."
>
> On a scale of 0 to 10, how much more likely did George's statement make it that you will find William

> guilty of robbery? For reference, 0 = not more likely at all, 10 = 100% more likely.

Bruton Scenario 3: ("blank")

> You are a juror sitting on a joint American federal criminal trial for armed robbery against two co-defendants, George and William. Because this is a joint trial, both George and William sit at the defense table during trial. As a juror, you are tasked with evaluating whether there is proof beyond a reasonable doubt that William—not George—committed robbery. At trial, the prosecution introduces pretrial police interrogations of George. You find George's statements to be credible, meaning you believe him. In these pretrial police interrogations, George stated: "Me and [blank] committed the robbery."
>
> On a scale of 0 to 10, how much more likely did George's statement make it that you will find William guilty of robbery? For reference, 0 = not more likely at all, 10 = 100% more likely.

Bruton Scenario 4: (inferentially implicates defendant)

> You are a juror sitting on a joint American federal criminal trial for armed robbery against two co-defendants, George and William. Because this is a joint trial, both George and William sit at the defense table during trial. As a juror, you are tasked with evaluating whether there is proof beyond a reasonable doubt that William—not George—committed robbery. At trial, the prosecution introduces pretrial police interrogations of George. You find George's statements to be credible, meaning you believe him. In these pretrial police interrogations, George stated: "On the drive to the scene of the crime, me and John talked about how we'd have to kill the victims after the robbery." Later at trial, William testifies that during the drive to the crime scene, he "knew that George and

> John were talking" but couldn't make out their conversation because the radio was too loud.
>
> On a scale of 0 to 10, how much more likely did George's statement make it that you will find William guilty of robbery? For reference, 0 = not more likely at all, 10 = 100% more likely.

Appendix B: The Prompt

````markdown

# Federal Jury Simulator

## Instructions

1. Replace `{{scenario_text}}` below with **one** of the full scenario definitions listed in the “Scenarios” section.

2. For each juror, inject their demographics (from /Users/sharrington/Library/CloudStorage/OneDrive-UniversityofOklahoma/Git/4ScenarioJurySimFinal/assets/jurordata.csv) `/assets/jurordata.csv` into the prompt template.

3. Send the prompt to your AI of choice (e.g. OpenAI o3-mini, Claude 4, Gemini Flash Pro).

4. Collect for each juror:

  - **Score** (0–10)

  - **Explanation** (one sentence)

---

## Prompt Template

```text

SYSTEM:

You are “Federal Jury Simulator v1.0,” a behavioral-science engine that predicts how incriminating evidence appears to eligible U.S. federal jurors.

SCENARIO:

{{scenario_text}}

JUROR:

• Sex: {{sex}}

• Ethnicity: {{ethnicity}}

• Political Affiliation: {{party}}

• Income Range: {{income}}

• Education Level: {{education}}

TASK:

Assign a **Guilt-Impact Score** from 0 to 10 indicating how much this evidence increases your likelihood of finding William guilty.

(0 = not more likely at all; 10 = 100% more likely.)

Then provide a **one-sentence explanation** for your score.

OUTPUT FORMAT (Markdown):

**Score:** {{0–10}}

**Explanation:** {{short sentence}}

````

The simulator prompt is written to meet two, occasionally competing, demands: (1) reproducible social-science measurement suitable for a law-review appendix, and (2) modern best-practice in large-language-model (LLM) control. What follows explains, in prose, why each major clause appears.

1. System-role declaration: The opening line "You are 'Federal Jury Simulator v 1.0,' a behavioral-science engine . . . " uses role prompting. Anthropic's engineering guide notes that assigning a role through the system channel "dramatically improve[s] . . . performance," because "the right role can turn [the model] from a general assistant into your virtual domain expert."[86] For juror-simulation research, that focus is essential to avoid

[86] "When using Claude, you can dramatically improve its performance by using the system parameter to give it a role…. The right role can turn Claude from a general assistant into your virtual domain expert!"
````

meandering or moralizing answers.

2. Double-braced placeholders ({{variable}}): Juror demographics and scenario text appear as {{scenario_text}}, {{sex}}, etc. Anthropic describes this convention "placeholders are denoted with {{double brackets}}" as the recommended way to separate fixed instructions from variable data, thereby making large-scale experiments easier to script and audit.[87]

3. Sequential, explicit instructions: After the juror block the task is given in two imperatives "Assign a Guilt-Impact Score . . . . Then provide a one-sentence explanation." Anthropic likens an LLM to "a brilliant but very new employee . . . who needs explicit instructions" and advises presenting them "as sequential steps" to ensure compliance. The numbered, step-wise phrasing mirrors that advice.

4. Positive instructions over constraint lists: Rather than cataloguing forbidden content, the prompt states only what should be produced (score plus sentence). Google's 2025 prompt-engineering guide reports that "focusing on positive instructions . . . can be more effective than relying heavily on constraints," because long "DON'T" lists leave the model guessing at what remains permissible.[88]

5. Explicit output template: The closing Markdown block locks the response into a mini-schema ("Score: . . . Explanation: . . . "). Anthropic's prompt-improver tool flags "explicit output-formatting requirements" as a key technique for reducing hallucinated prefaces or stray tokens.[89] Constraining the output this way lets researchers parse scores programmatically.

Together, these five principles, role focus, template variables, stepwise clarity, affirmative guidance, and rigid output formatting, create a prompt that is both methodologically transparent for academic reviewers and practically robust when executed across thousands of simulated jurors.

---

[87] "A prompt template combines these fixed and variable parts, using placeholders for the dynamic content… placeholders are denoted with {{double brackets}}."
[88] "Growing research suggests that focusing on positive instructions in prompting can be more effective than relying heavily on constraints . . . instead of telling the model what not to do, tell it what to do instead."
[89] The improved prompt "Provides explicit output formatting requirements."