# Should You Use LLMs to Simulate Opinions? Quality Checks for Early-Stage Deliberation

**Terrence Neumann**[1], **Maria De-Arteaga**[2], **Sina Fazelpour**[3]

[1]McCombs School of Business, University of Texas at Austin
[2]Universitat Ramon Llull, ESADE
[3]Department of Philosophy and Computer Science, Northeastern University

## Abstract

The emergent capabilities of large language models (LLMs) have prompted interest in using them as surrogates for human subjects in opinion surveys. However, prior evaluations of LLM-based opinion simulation have relied heavily on costly, domain-specific survey data, and mixed empirical results leave their reliability in question. To enable cost-effective, early-stage evaluation, we introduce a quality control assessment designed to test the viability of LLM-simulated opinions on Likert-scale tasks without requiring large-scale human data for validation. This assessment comprises two key tests: *logical consistency* and *alignment with stakeholder expectations*, offering a low-cost, domain-adaptable validation tool. We apply our quality control assessment to an opinion simulation task relevant to AI-assisted content moderation and fact-checking workflows—a socially impactful use case—and evaluate nine LLMs using a baseline prompt engineering method (backstory prompting), as well as fine-tuning and in-context learning variants. None of the models or methods pass the full assessment, revealing several failure modes. We conclude with a discussion of the risk management implications and release `TopicMisinfo`, a benchmark dataset with paired human and LLM annotations simulated by various models and approaches, to support future research.

**Code & Data** — https://tinyurl.com/qualitychecksAAAI

**Full Paper with Appendices** — https://arxiv.org/abs/2504.08954

## Introduction

Large language models (LLMs) have demonstrated significant appeal and versatility across a wide range of tasks. While the primary training objective of the base model is to predict the next most likely token, training on vast and diverse datasets, combined with increasingly complex system architectures, has resulted in the "emergence" of capabilities that were not explicitly anticipated during development (Wei et al. 2022). For instance, LLMs have been shown to learn new tasks through in-context learning (Brown et al. 2020), perform coding tasks with remarkable accuracy (Zheng et al. 2024), and have displayed human-like capabilities across different tasks (Kosinski 2024; Strachan et al. 2024).



These emerging capabilities have led researchers to investigate more unconventional uses of LLMs, including automating opinion surveys by simulating the viewpoints of specific demographic or ideological groups (Qu and Wang 2024). The possibility that LLMs may successfully simulate demographic viewpoints stems from (i) evidence that LLMs can be prompted to replicate responses from human subject studies in psychology and other social sciences (Aher, Arriaga, and Kalai 2023), (ii) training data that likely encompass a broad spectrum of perspectives and opinions (Miranda et al. 2024), and (iii) prompting techniques that lead LLMs to produce seemingly coherent, opinionated responses (Wright et al. 2024). Automating such surveys could have broad applications in marketing (Sarstedt et al. 2024), content moderation (Fröhling, Demartini, and Assenmacher 2024), policymaking, and public relations (Sanders, Ulinich, and Schneier 2023). LLMs may be especially valuable for pilot studies (Sarstedt et al. 2024; Rothschild et al. 2024), sampling hard-to-reach populations (Jansen, gyo Jung, and Salminen 2023), and labeling data that may be psychologically harmful for human annotators (Wang, Morgenstern, and Dickerson 2025).

However, existing evidence is inconclusive on the extent to which LLM-based methods can accurately simulate human opinions. Some studies show that including a demographic "backstory" in a prompt (Argyle et al. 2023; Bui et al. 2025; Jiang, Wei, and Zhang 2024) (i.e., "backstory" prompting), fine-tuning (Namikoshi et al. 2024), and in-context learning (Karanjai et al. 2025) are approaches that can successfully approximate group-level opinions, while others document systematic failures of these same approaches—especially when simulating the opinions of minority or non-Western groups (Santurkar et al. 2023; Sun et al. 2023; Qu and Wang 2024; Mingmeng, He, and Trotta 2024; Orlikowski et al. 2025). Additional work finds that backstory prompting approaches—the most prevalent approach to simulating opinions in prior research—flatten within-group variance of opinions (Wang, Morgenstern, and Dickerson 2025; Bisbee et al. 2024; Mingmeng, He, and Trotta 2024) and fails to generalize across topics (Sanders, Ulinich, and Schneier 2023; Lee et al. 2024). See **Appendix A** for a literature review table that summarizes prior works methodologies, respective application domains, and findings.

The mixed results in the empirical literature suggest that the reliability of LLM-simulated opinions cannot be assumed across domains and instead requires rigorous validation for each new use case. Importantly, however, current validation methods rely on gathering high-quality human-labeled data—the cost of which can rival that of traditional surveys. For early-stage assessments, this upfront expense is likely to outweigh uncertain benefits, significantly hindering research and development efforts. The existence of applications with broad consensus about the potential desirability of simulating opinions, such as when human survey subjects would be exposed to harmful content (Wang, Morgenstern, and Dickerson 2025) or trade secrets (Wang, Zhang, and Zhang 2024), motivates the need for validation techniques that lower the cost of early-stage exploration of LLM usage to simulate opinions.

**Contribution 1: Designing a quality control assessment for early-stage deliberation of LLM opinion simulation.** We introduce two diagnostic tests that probe desirable structural properties of simulated opinions without requiring the collection of large-sample human reference data:

1. **Logical Consistency**—whether models reliably position "average" opinions as a convex combination of group-level opinions *and/or* whether these average opinions are consistent with sampling from a stable reference population.
2. **Alignment with Stakeholder Expectations**—whether models position differences in group opinions in ways that are consistent with prior domain knowledge *and/or* small sample survey data.

The design of the proposed quality control assessment is grounded in the fact that some of the challenges involved in validating LLM-simulated opinions are similar to those involved in validating subjective human annotations. Researchers have developed various quality controls to assess the reliability of human data and annotator attentiveness (Artstein and Poesio 2008; Sap et al. 2022; Mostafazadeh Davani, Díaz, and Prabhakaran 2022; Lease 2011). Our proposed tests draw from this literature, adapting it to be suitable for assessing LLMs, with a focus on Likert-scale opinions. In the discussion section, we provide recommendations for integrating these tests into organizational risk management practices.

The value of these tests compared to those proposed in previous work (Aher, Arriaga, and Kalai 2023; Argyle et al. 2023; Sun et al. 2023) is that they can be performed *prior* to expensive large-scale human validation effort, fostering well-informed deliberation on the suitability of LLM opinion simulation in a particular context.

**Contribution 2: Crafting a domain-specific testbed.** To illustrate the utility of our quality checks, we develop a domain-specific testbed in content moderation, focusing on the prioritization of potentially harmful misinformation for fact-checker review. Because prioritization is opinion-driven, resource-intensive, and exposes annotators to psychologically taxing content (Liu, Gwizdka, and Lease 2024), it provides a compelling, socially impactful context for evaluating simulated opinions.

An effective testbed for group-level opinion simulation should include topics with both expected agreement and disagreement amongst groups. Gender offers strong prior expectations based on a significant body of public opinion research (Huddy, Cassese, and Lizotte 2008; Lizotte 2020). We focus on the prioritization criterion of identifying "potential harm to specific groups," as previous research highlights this dimension of prioritization as one along which gender differences are particularly pronounced (Huddy, Cassese, and Lizotte 2008; Lizotte 2020).[1]

We built `TopicMisinfo`, a 160-claim dataset spanning topics that can be expected to elicit gender disagreement and consensus. The selection of these topics is anchored in findings from the American National Election Studies (ANES) (American National Election Studies 2021), which help us capture a diversity in topics that are likely to provoke varying degrees of gendered disagreement. We also included "Gold" comprehension checks, which are claims that are harmless, and are either obviously true (e.g., "A circle is round.") or obviously false (e.g., "The tallest tree on Earth touches Mars."), with a very strong prior on gender-level consensus of opinion. We also collect a small sample ($\approx$ 1600 annotations) of task-specific human opinion data. Dataset composition, example claims, and human annotation statistics are detailed in the **Appendices B-C**; the `TopicMisinfo` dataset is publicly available to support future research.

**Contribution 3: Benchmarking leading models and approaches to opinion simulation using our quality checks.** Applying the proposed quality checks, we evaluate simulated opinions from several commercially available LLMs: GPT-3.5-Turbo (checkpoints 06-13-23, 11-06-23, 01-25-24), GPT-4, GPT-4.1, GPT-5-mini, LLaMA 3, Titan-Text-Premier, and Mistral-Large—referred to throughout as GPT3.5a, GPT3.5b, GPT3.5c, GPT-4, GPT-4.1, GPT-5-mini, LLaMA3, Titan, and Mistral. We elicit group-level responses via backstory prompts (Argyle et al. 2023; Sun et al. 2023) that instruct models to adopt personas of men or women (see **Appendix D** for prompt templates). We also benchmark two other approaches to simulating opinions suggested in the literature: fine-tuning on small-sample human survey data (Namikoshi et al. 2024); and in-context learning with human-labeled examples (Karanjai et al. 2025). Importantly, our findings show that while fine-tuning and in-context learning generally improve alignment with stakeholder expectations, they do *not* substantially outperform backstory prompting on logical consistency. These results indicate that all evaluated models and prompting approaches fail the proposed quality control assessment for the task considered, suggesting that current methods face limitations in their ability to produce stable, internally coherent simulations of human opinions.

[1]While not the only criteria relevant for prioritization, research has noted that this is an important axis of consideration used by fact-checkers (Sehat et al. 2024; Liu, Gwizdka, and Lease 2024).

## Quality Checks & Results

In this section, we introduce two quality checks designed for early-stage evaluation of LLM-simulated opinions. For each check, we detail the methodology and analyze how a range of LLMs perform across gender conditions using the "backstory prompting" approach (Argyle et al. 2023) on claims drawn from the `TopicMisinfo` dataset, focusing on a key opinion task relevant to fact-checking and content moderation workflows: *assessing the potential harm of claims to specific groups.* The code and reproducible pipeline are publicly available.

### Quality Check 1: Logical Consistency

Testing for logical consistency is a foundational tool for evaluating the quality of data from human annotators (Aruguete et al. 2019) and from data aggregation methods such as majority voting (Mostafazadeh Davani, Díaz, and Prabhakaran 2022). In this work, we adapt this principle to assess the internal coherence of LLM-simulated opinions across demographic groups.

When simulating opinions for multiple groups, a logically consistent model should produce an average opinion that is interpretable as a convex combination of the group-level opinions. This requirement reflects a basic expectation of distributional coherence: the "average" simulated opinion should be derivable from the underlying subgroups it is averaging over. For example, if in simulating the opinion of some population, $G$, consisting of two sub-groups, $\{g_1, g_2\}$, and the model simulates that members of $g_1$, on average, rate a claim's harmfulness as 2-out-of-6 and members of $g_2$, on average, as 4-out-of-6, then simulating an average opinion of 5-out-of-6 for $G$ as a whole is not just implausible, it is *logically inconsistent*, as it violates a foundational algebraic constraint on how means combine.

Note that this is a rather weak constraint in the sense that it simply verifies that the average opinion for *a given claim* falls between group-level opinions for that same claim. That is, the test allows the presumed proportion of the subpopulations, and so the weight given to their simulated average opinions, to vary *across different claims* within a topic. A stronger test can, therefore, require that there exists a *single, stable mixture weight*—e.g., a 50/50 blend of $g_1$'s and $g_2$'s opinions—that can consistently reconstruct the average opinion across all claims within a topic. A model that meets this condition supports the existence of a coherent reference population that the prompt reliably reflects. By contrast, failure to meet this criterion suggests that it is mathematically impossible for all simulated opinions across claims to reflect the views of a single population, casting doubt on how the simulated opinions can be interpreted and used.

These considerations can be summarized in terms of the following research questions:

- **RQ-1a (Weak Test).** Does the model's simulated average opinion fall within the convex hull of its group-conditioned opinions on at least a prespecified proportion $p_0$ of claims in a given topic?
- **RQ-1b (Strong Test).** Is there a fixed mixture weight $q_0 \in [0, 1]$ that consistently reconstructs the model's average opinion across claims, indicating the presence of a stable underlying reference population?

**Methodology** For each claim $c_i$, the model is prompted $n_i$ times under each condition (men, women, average) using a temperature setting of $\tau = 0.5$, generating distributions of Likert-scale responses. A logically consistent average opinion should be expressible as a convex combination of the group-level opinions it seeks to summarize. In the context of gender (assuming only two: men and women)[2], that would entail for each claim $c_i$, the model's average predicted mean $\mu_{\text{avg}}(c_i)$ should satisfy:

$$\mu_{\text{avg}}(c_i) = \hat{q}_{c_i} \cdot \mu_{\text{men}}(c_i) + (1 - \hat{q}_{c_i}) \cdot \mu_{\text{women}}(c_i),$$

for some $\hat{q}_{c_i} \in [0, 1]$. If this condition holds, the average-conditioned response can be interpreted as a plausible mixture of the two. If not, the average-conditioned response is logically inconsistent with the group-level outputs.

**Weak Test** To assess logical consistency, we define the implied mixture weight $\hat{q}_{c_i}^{(b)}$ for each bootstrap replicate $b$ as:

$$\mu_{\text{avg}}^{(b)}(c_i) = \hat{q}_{c_i}^{(b)} \cdot \mu_{\text{men}}^{(b)}(c_i) + (1 - \hat{q}_{c_i}^{(b)}) \cdot \mu_{\text{women}}^{(b)}(c_i)$$

where the bootstrapped means $\mu_b^{\text{group}}(c_i)$ are computed by resampling with replacement $n_i$ times from the original set of labels for each group. The mixture weight $\hat{q}_{c_i}^{(b)}$ is then used to determine whether the average-conditioned mean lies within the convex hull of the group-specific means:

$$I_i^{(b)} = \begin{cases} 1 & \text{if } \hat{q}_{c_i}^{(b)} \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

We aggregate the claim-level indicators within topic $\omega$ to produce a topic-level success rate for each bootstrap sample:

$$\hat{P}_\omega^{(b)} = \frac{1}{N_\omega} \sum_{i=1}^{N_\omega} I_i^{(b)}$$

We then test the one-sided hypothesis:

$$H_0 \colon \hat{P}_\omega \geq p_0 \quad \text{vs.} \quad H_1 \colon \hat{P}_\omega < p_0$$

The empirical $p$-value over $B = 10^4$ replicates is computed as:

$$\hat{p}(p_0) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left[\hat{P}_\omega^{(b)} \geq p_0\right]$$

We reject $H_0$ at level $\alpha$ if $\hat{p}(p_0) < \alpha$, concluding that the model fails to produce average-conditioned opinions that lie within the convex hull of group-level distributions at an acceptable rate. We test acceptable rate values $p_0 \in [0.7, 0.8, 0.9, 1.0]$. See **Appendix F** for pseudocode that generalizes this test for $\geq 2$ groups.

[2]While we assume a binary gender, **Appendices F and G** contain pseudo-code for generalizing this methodology for $\geq 2$ groups, while **Appendix H** includes the detail of a replication study including non-binary genders simulated with GPT3.5c.

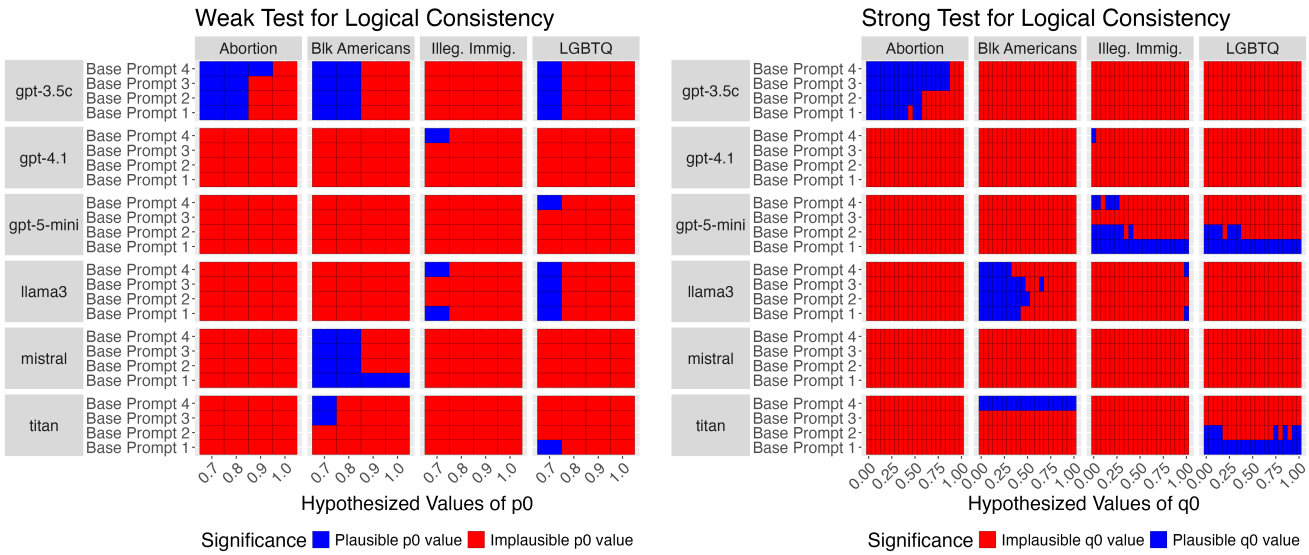


Figure 1: Results for weak and strong tests of logical consistency.

**Strong Test** To assess whether the average-conditioned responses could reflect a single, stable underlying population, we evaluate whether a fixed mixture weight $q_0$ can consistently explain the simulated average-conditioned responses. For each claim $c_i$, we compute the absolute deviation from a specified $q_0 \in [0, 1]$:

$$\hat{L}_\omega(q_0) = \frac{1}{N_\omega} \sum_{i=1}^{N_\omega} |\hat{q}_{c_i} - q_0|$$

Under the null hypothesis $H_0 \colon \hat{L}_\omega(q_0) = 0$, the model exhibits logical consistency at mixture weight $q_0$.

We conduct a bootstrap hypothesis test with $B = 10^4$ replicates. For each claim, we construct a synthetic mixture null hypothesis distribution:

$$\mathcal{D}_{c_i}^{q_0} = q_0 \cdot \mathcal{D}_{c_i}^{\text{men}} + (1 - q_0) \cdot \mathcal{D}_{c_i}^{\text{women}}$$

From this mixture, we resample synthetic average-conditioned responses, compute $\hat{q}_{c_i}^{(b)}$, and aggregate across claims: $\hat{L}_\omega^{(b)}(q_0) = \frac{1}{N_\omega} \sum_{i=1}^{N_\omega} \left| \hat{q}_{c_i}^{(b)} - q_0 \right|$. The $p$-value is $\Pr\left( \hat{L}_\omega^{(b)}(q_0) > \hat{L}_\omega(q_0) \right)$. This test is repeated for values of $q_0$ in $[0, 1]$ at intervals of 0.05, applying a Bonferroni correction ($\alpha^* = 0.0025$) to account for multiple comparisons. See **Appendix G** for pseudocode that generalizes this test for $\geq 2$ groups.

**Results.** Figure 1 reports feasible consistency thresholds ($p_0$) for the weak test and valid $q_0$ intervals for the strong test, indicating alignment with a consistent reference population. We highlight results for four socially divisive topics where group-level disagreement is well-documented. Given the lack of a gold standard for prompting average-conditioned or population-level perspectives, we evaluate consistency across four alternative prompts. See **Appendix E** for details.

- **GPT-3.5c**: For threshold $p_0 = 0.7$, passes the *weak test* across three out of four topics with Base Prompt 4, indicating basic geometric plausibility. For the *strong test*, it only passes the *Abortion* topic, with feasible mixture weights $q_0 \in [0, 0.55]$ or $[0, 0.85]$ depending on prompt. Fails to produce a consistent average-conditioned reference population for *Black Americans*, *Illegal Immigration*, and *LGBTQ*.
- **GPT-4.1**: One prompt passes the *weak test* for *Illegal Immigration*, but we see no other feasible thresholds across the prompts and topics tested. For the *strong test*, we see the same prompt (Base Prompt 4) generate a plausible value for the topic of Illegal Immigration, but no other plausible values across topics and prompts.
- **GPT-5-mini**: The only reasoning model we tested does not perform markedly better. One feasible point for the *weak test* for the *LGBTQ* topic, but no others. We see better performance on the *strong test* than other models, with feasible values for multiple prompts across two topics — *Illegal Immigration* and *LGBTQ*.
- **LLaMA-3**: Two prompts pass the *weak test* at $p_0 = 0.7$ for *Illegal Immigration* and all four prompts pass for *LGBTQ*, but fails on other topics. For the *strong test*, exhibits moderate success: feasible $q_0 \in [0, 0.5]$ for *Black Americans* across prompts, and up-samples opinions of men for *Illegal Immigration* ($q_0 \in [0.95, 1]$). Fails on *Abortion* and *LGBTQ*.
- **Mistral**: Shows success on the *weak test* only for *Black Americans*, with feasible $p_0$ values ranging from 0.7 to 1.0 depending on the prompt. Fails the *strong test* across all topics—no consistent mixture weights were found.
- **Titan**: Two prompts pass the *weak test* at $p_0 = 0.7$ for *Black Americans* and one prompt passes for *LGBTQ* at $p_0 = 0.7$. On the *strong test*, Prompt 4 yields success on *Black Americans* with full-range plausibility ($q_0 \in$

$[0, 1]$). Prompts 1 and 2 yield plausible $q_0 \in [0, 1]$ for *LGBTQ*, but it fails on both *Abortion* and *Illegal Immigration*.

**Take-away** Overall, most models fail to place average-conditioned opinions inside the convex hull of gendered responses at a reasonable rate, as evidenced by the results from the weak test. We see only $\frac{12}{96} = 12.5\%$ and $\frac{2}{96} = 2\%$ model×topic×prompt pairs passing a consistency threshold ($p_0$) of 0.8 and 0.9 respectively for these topics. Similarly, for the strong test, we see that models often do not often maintain a fixed reference population when sampling an average-conditioned opinion, as only $\frac{19}{96} = 19.7\%$ of model×topic×prompt pairs reveal a plausible fixed reference $q_0$.

## Quality Check 2: Alignment with Stakeholder Expectations and Small Sample Survey Data

Assessing alignment with common sense or widespread expectations is standard for evaluating human annotations, e.g., via attention checks (Lease 2011). The same underlying idea can be leveraged when assessing the quality of LLM-generated annotations. In our setting, for instance, stakeholders—e.g., end-users or developers—can have clear intuitions about cases where group-level differences should or should not arise. Checking for misalignment on these cases can provide a powerful lens for evaluating the quality of LLM-generated annotations. If neglected, such misalignment can bias downstream tasks, leading to skewed decisions or misrepresentation of group perspectives. In high-stakes settings like claim prioritization for fact-checking, such errors carry serious ethical consequences.

Rather than evaluating annotations individually, comparing simulated opinions across demographic groups can reveal critical early-stage information about the viability of opinion simulation. Group-level consensus on neutral topics and divergence on divisive ones offer early signals of realism and reliability. Quality Check 2 introduces a flexible method to test whether simulated group differences align with stakeholder expectations. For topics where domain experts anticipate clear consensus or pronounced disagreement across demographic groups ("clear-cut cases"), relying solely on stakeholder priors may be sufficient. In situations where expectations are less certain, small-scale human surveys serve as practical benchmarks. In our analysis, we leverage our own domain expertise as researchers in AI-assisted fact-checking to define stakeholder expectations and interpret test outcomes.

**Research Questions.** To what extent do simulated group differences in opinion conform to stakeholder expectations informed by (**RQ2a**) common-sense and domain-knowledge and (**RQ2b**) small-sample survey data?

**Methodology** We evaluate whether LLM-generated gender opinion gaps align with stakeholder expectations or small-scale human surveys. For each claim $c_i$ within topic $\omega$, we define the model-estimated gender gap as:

$$\hat{D}(c_i) = \mu_{\text{LLM}}^{\text{woman}}(c_i) - \mu_{\text{LLM}}^{\text{man}}(c_i),$$

| | Topic: **Gold** (Prior: Insig. Diffs) | Topic: **Abortion** (Prior: Sig. Diffs) |
|---|---|---|
| GPT-3.5a | 0.45*** | 3.42*** |
| GPT-3.5b | 0.36** | 2.05*** |
| GPT-3.5c | 0.72*** | 2.37*** |
| GPT-4 | 0.00 | 0.07 |
| GPT-4.1 | 0.12** | 0.43** |
| GPT-5-mini | 0.11*** | -0.10 |
| llama3 | -0.29* | 0.13* |
| mistral | 0.03 | 0.53 |
| titan | 0.33** | 0.61* |

Table 1: Quality Check 2a Statistical Results — Alignment between model-predicted gender differences and stakeholder priors. ($*$p<0.05; $**$p<0.01; $***$p<0.001). Blue values agree with the prior; red values contradict it. Results indicate that no model satisfies both priors.

and compare it to an expected gap $E[D(c_i)]$, which is specified based either on stakeholder priors or small-scale surveys. We then compute the deviation $g_i = \hat{D}(c_i) - E[D(c_i)]$.

To create an efficient test statistic by reducing variance from imbalanced number of labels from simulated men and women, we weight each claim using the harmonic mean of the gender-group sizes:

$$w_i = \frac{n_i^{\text{woman}} \cdot n_i^{\text{man}}}{n_i^{\text{woman}} + n_i^{\text{man}}},$$

a standard weighting approach in meta-analysis to manage group-size discrepancies and enhance estimator precision (Hedges and Olkin 2014).

To test alignment at the topic level, we perform a weighted one-sample hypothesis test:

$$H_0 : \bar{g}_\omega = 0 \quad \text{vs.} \quad H_1 : \bar{g}_\omega \neq 0,$$

where the weighted mean discrepancy for topic $\omega$ is:

$$\bar{g}_\omega = \frac{1}{\sum_i w_i} \sum_{i=1}^{N_\omega} w_i \cdot g_i,$$

and the appropriate weighted standard error (NIST 2003) is:

$$\text{SE}(\bar{g}_\omega) = \sqrt{\frac{\sum_i w_i (g_i - \bar{g}_\omega)^2}{(\sum_i w_i - 1) \sum_i w_i}}.$$

This yields the topic-level $t$-statistic $t_\omega = \frac{\bar{g}_\omega}{\text{SE}(\bar{g}_\omega)}$, with degrees of freedom $\nu = \sum_i w_i - 1$.

Each model response in our dataset corresponds to a 6-point Likert-scale judgment, formally an ordinal variable. Nonetheless, parametric methods such as the $t$-test can reliably analyze Likert-type data when averaged over moderate-to-large samples (Norman 2010; de Winter and Dodou 2010). In our experiments, each model–gender–claim cell aggregates at least 10 LLM completions, and statistics are computed at the level of *topic means*, averaged over individual claims, ensuring approximate normality via the Central Limit Theorem.

| Topic | gpt-3.5a | | gpt-3.5b | | gpt-3.5c | | gpt-4 | | gpt-4.1 | | gpt-5-mini | | llama3 | | mistral-large | | titan-text-premier | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\bar{g}_\omega$ | p-val | $\bar{g}_\omega$ | p-val | $\bar{g}_\omega$ | p-val | $\bar{g}_\omega$ | p-val | $\bar{g}_\omega$ | p-val | $\bar{g}_\omega$ | p-val | $\bar{g}_\omega$ | p-val | $\bar{g}_\omega$ | p-val | $\bar{g}_\omega$ | p-val |
| **More Divisive Topics** | | | | | | | | | | | | | | | | | | |
| Abortion | 1.65 | * | 0.28 | — | 0.60 | — | -1.70 | *** | -1.34 | *** | -1.89 | *** | -1.65 | *** | -1.24 | ** | -1.16 | * |
| Black Americans | 1.72 | * | 0.35 | — | 0.39 | — | -1.28 | *** | -1.01 | *** | -0.92 | ** | -1.09 | *** | -1.92 | * | -0.97 | ** |
| Illegal Immigration | 0.92 | ** | -1.06 | * | -0.61 | — | -1.42 | ** | -1.05 | ** | -1.04 | ** | -0.80 | * | -0.43 | — | -0.71 | * |
| LGBTQ | 0.66 | — | -1.37 | * | -0.37 | — | -0.92 | * | -1.06 | ** | -0.99 | * | -1.08 | ** | -0.99 | ** | -1.10 | ** |
| **Less Divisive Topics** | | | | | | | | | | | | | | | | | | |
| Entertainment | 1.59 | * | 1.05 | * | 1.18 | * | 0.35 | ** | 0.39 | ** | 0.47 | ** | 0.41 | ** | 0.58 | * | 0.48 | * |
| Gold | 0.43 | — | 0.34 | — | 0.71 | ** | -0.01 | — | 0.10 | — | 0.10 | — | -0.30 | — | 0.02 | — | 0.32 | — |
| HealthScience | 1.86 | * | 1.07 | * | 0.65 | — | -0.19 | — | 0.58 | * | 0.32 | — | 0.04 | — | 0.79 | — | 0.56 | — |
| Sports | -0.01 | — | -0.06 | — | 0.21 | — | -0.26 | — | -0.26 | — | -0.36 | — | -0.26 | — | -0.26 | — | -0.16 | — |
| USA | 1.68 | *** | 1.12 | *** | 0.63 | ** | -0.61 | *** | -0.05 | — | -0.15 | — | -0.21 | — | -0.56 | * | 0.16 | — |
| WeatherClimate | 1.06 | * | 1.27 | * | 0.97 | * | -0.39 | — | 0.01 | — | 0.06 | — | -0.10 | — | -0.68 | — | 0.10 | — |

Significance levels: $*$ for $p < 0.05$, $**$ for $p < 0.01$, $***$ for $p < 0.001$.
Blue: positive significant values, Orange: negative significant values.

Table 2: Quality Check 2b Statistical Results. GPT-3.5-turbo models tend to significantly exaggerate gender differences on topics humans found less divisive, while GPT-4, GPT-4.1, GPT-5-mini, Llama3, mistral-large, and Titan tended to significantly erode differences on topics humans found more divisive.

**RQ-2a: Testing Against Stakeholder Priors.** We select topics from `TopicMisinfo` known through domain expertise to be either clearly non-divisive (e.g., **Gold** attention checks) or clearly divisive (e.g., claims related to **Abortion**). For non-divisive topics, stakeholder intuition strongly supports minimal or no demographic differences; thus, we explicitly set $E[D(c_i)] = 0$. Similarly, for divisive topics, setting $E[D(c_i)] = 0$ serves as a falsification test: substantial deviations from zero confirm that the model appropriately simulates expected group divergences. Thus, failing to reject $H_0$ for non-divisive topics aligns with stakeholder priors, confirming minimal simulated group differences. Conversely, rejecting $H_0$ for divisive topics validates that simulated opinions mirror expected demographic divergences.

**RQ-2b: Testing against human annotations.** When human survey data are available, we define $E[D(c_i)] = \mu^{\text{woman}}_{\text{Human}}(c_i) - \mu^{\text{man}}_{\text{Human}}(c_i)$. We then calculate $g_i = \hat{D}(c_i) - E[D(c_i)]$. Although the gender gaps are estimated from distinct sources (LLM vs. human samples), the test is structured as a paired one-sample analysis over the differences $g_i$, with standard error and degrees of freedom computed identically. Rejection indicates meaningful divergence between LLM-simulated and human-annotated gender differences.

**Results (see Tables 1–2).**

- **Exaggeration on consensus topics.** In Quality Check 2a, all models except GPT-4 and Mistral report spurious gender gaps on the **Gold** "circle-is-round" topic ($p < \alpha$), indicating oversensitivity to gender distinctions. In Quality Check 2b, the GPT-3.5 family of models exaggerate disagreement relative to human survey data on less polarizing topics like **entertainment**, **health & science**, **U.S. politics**, and **climate**.
- **Erosion on divisive topics.** Other models—including GPT-4, GPT-4.1, GPT-5-mini, LLaMa3, Titan, and Mistral—tend to erode known gender differences, even on topics with well-established disagreement. Quality Check 2a shows that GPT-4, GPT-5-mini, and Mistral fail to reflect divergent opinions on **abortion**, and Quality Check 2b reveals that these models, in addition to GPT-4.1, Titan and LLaMa3, also underrepresent human gender-level disagreement on topics such as **Black Americans**, **LGBTQ** issues, and **Illegal Immigration**, yielding significant negative $\bar{g}_\omega$ scores ($p < \alpha$).

**Take-away.** Current LLMs often misalign with stakeholder priors—some exaggerate less divisive claims, others erode salient divides. Further, there are clear differences between the models tested: the GPT-3.5-turbo variants tend to be guilty of exaggerating gender-level differences, while GPT-4, Llama 3, Mistral, and Titan tend to erode differences across topics, both when conditioned on prior expectations and on small sample survey data.

## Benchmarking Experiment

In the previous section, we tested our quality checks using the backstory prompting approach across several different models. To assess alternative strategies, we benchmark this approach against two others prominent in the literature: fine-tuning (Namikoshi et al. 2024) and in-context learning (Karanjai et al. 2025). Specifically, we use only the GPT-3.5c checkpoint to control for variation across model architectures and focus on comparing simulation strategies. We test the following strategies:

1. **Prompt Engineering**: Two variations of backstory prompts ("Cond. Prompt 1" and "Cond. Prompt 2"), which differ in framing and specificity of demographic identity, direct the model to simulate opinions from specific demographic perspectives. See **Appendix 4** for exact prompt specifications.
2. **Fine-Tuning**: We fine-tune GPT3.5c on the average **human** ratings for each condition (men, women, overall average) using 40 claims from `TopicMisinfo`, yielding 120 training examples. Fine-tuning is performed over three epochs. See **Appendix 6a** for more detail.
3. **In-Context Learning (ICL)**: We test two strategies: (a) **Random Sampling** (RS-ICL), where prompts include random labeled examples for each condition (men,

| Method | QC1a | QC1b | QC2a | QC2b |
|---|---|---|---|---|
| Cond. Prompt 1 | 44% | 25% | 50% | 70% |
| Cond. Prompt 2 | 44% | **50%** | 0% | 70% |
| Fine-tuning | 19% | 25% | **100%** | **100%** |
| RS-ICL | 56% | **50%** | 50% | **100%** |
| NN-ICL | **88%** | 0% | 50% | 90% |

Table 3: Performance on quality checks across methods. Fine-tuning improves alignment, but not logical consistency compared to baseline.

women, overall average), and (b) **Nearest Neighbor** (NN-ICL), where prompts include the most semantically similar claims and their associated labels based on cosine similarity. See **Appendix J** for more detail on methodology and examples.

**Experimental Setup** We partition the `TopicMisinfo` dataset into 120 held-out claims for evaluation and 40 claims for training or in-context injection. We reapply our baseline prompt engineering methods to the held-out set, fine-tune on human-labeled examples, and construct in-context prompts using the 40 available claims. Each method is evaluated using a summary metric of our quality checks. For QC1a, we report the proportion of $p_0$ thresholds passed across four topics. For all other checks (QC1b, QC2a, QC2b), we report the percentage of topics for which the method successfully satisfies the quality criterion.

**Results** Table 3 summarizes outcomes. Relative to the baseline approach of backstory prompting (Cond. Prompt 1 and Cond. Prompt 2), we find that: **Fine-tuning** achieves perfect alignment with priors and human data (QC2a/b = 100%) but degrades logical coherence (QC1a/b); **Random-sample ICL** matches fine-tuning on QC2b and retains moderate logical consistency; **Nearest-neighbor ICL** delivers the best score on the weak test for logical consistency (QC1a = 88%) yet dramatically fails the strong test (QC1b = 0%). Notably, all methods struggle on QC1b, underscoring the open challenge of producing a stable "average" reference population.

## Discussion

LLMs offer a promising avenue for simulating demographic opinions efficiently without immediate reliance on costly human-generated data. However, ensuring these outputs truly capture nuanced human perspectives remains challenging, particularly because LLMs often produce superficially coherent but logically flawed responses. To address this, we introduce an early-stage assessment targeting logical consistency and alignment with stakeholder expectations. These tests can help organizations quickly evaluate whether simulated opinions from LLMs justify deeper investment and development in their particular application domain.

Applying our quality checks, we uncovered significant logical inconsistencies: specifically, most models (80%) produced "average" opinions more extreme than demographic-specific predictions, violating basic statistical logic. Unlike logical inconsistencies of LLMs exposed in previous research—for example, mutually incompatible factual beliefs (Kassner et al. 2021), self-inconsistent chain-of-thought traces (Wang et al. 2022), violations of basic propositional logic (Ghosh et al. 2025), or within-passage contradictions (Mündler et al. 2024)—these errors indicate fundamental logical flaws at the distributional level. Practitioners cannot simply acknowledge and proceed despite these errors; they must reconsider or even abandon flawed simulation strategies entirely.

Further, our checks reveal systematic discrepancies with stakeholder expectations. Older models (like GPT-3.5-turbo) tend to exaggerate opinion differences even on obviously non-divisive claims, risking harmful stereotyping or inefficient resource allocation. Conversely, newer models (GPT-4, GPT-4.1, GPT-5-mini, LLaMa3, Mistral, Titan) often obscure genuine demographic divides on controversial topics, potentially masking critical areas affecting particular groups. Importantly, we find that the magnitude of gender-level differences observed in QC2a correlates with the direction of the errors observed in QC2b, suggesting that testing clear-cut cases aligned with stakeholder priors—even in the absence of extensive human data—provides a valuable preliminary quality check of LLM reliability.

Finally, in a benchmarking experiment, we found that fine-tuning and in-context learning generally improve alignment with stakeholder expectations, but they do not substantially outperform backstory prompting in terms of generating a stable reference population (QC1b). These results suggest that current methods remain limited in their ability to produce stable, internally coherent simulations, highlighting a promising avenue for future work.

**Limitations** Our study provides a reusable framework for early-stage evaluation of LLM-simulated opinions, but several factors define the boundaries of its current scope and suggest directions for future work. First, in our empirical results we benchmark exclusively instruction–tuned language models, as prior literature suggests that such models are likely to perform better than others on instruction–style prompts. The proposed tests are agnostic to model type, but we leave evaluation of base models to future work. Second, our experiments use only two binary gender personas and an unprimed "average" persona as proof of concept. We constrain ourselves to binary gender due to limitations of the human data, and in Appendix H we show how QC1 can be extended to non-binary gender. Crucially, in many application domains the relevant demographic or psychographic spectrum is much richer. Identifying all important subgroups and capturing their diversity remains an open challenge; omitting a salient group can invalidate the average–persona check and obscure meaningful variation. Finally, the human annotator pool for our TopicMisinfo dataset is predominantly male ($\approx 69\,\%$ men, $31\,\%$ women). Although QC2b uses only subgroup–specific means and is therefore unaffected by the imbalance, this gender distribution could influence population-level statistics if the dataset is reused for other purposes.

**Risk Management: Logical Consistency** From a risk management perspective, the choice between weak and strong logical consistency tests should align directly with the intended use of LLM-simulated outputs. In preliminary evaluations or exploratory research, the weak test—which verifies that average responses for each claim lie within the range defined by subgroup opinions—might suffice as a basic sanity check. However, in high-stakes resource-allocation scenarios such as prioritizing claims for fact-checking or deciding which content requires urgent human moderation, the strong test becomes crucial. For example, if a fact-checking organization uses a simulated "average" opinion as a baseline to determine whether specific demographic groups are disproportionately affected by misinformation, the average stance must represent a consistent, interpretable reference population. Inconsistencies in this average opinion could obscure whether certain groups genuinely require attention, leading to misallocation of limited fact-checking resources. Thus, adopting the strong test in such settings ensures stable benchmarks, supports defensible allocation decisions, and enhances overall reliability in high-consequence decision-making processes.

**Risk Management: Alignment with Stakeholder Expectations** Organizations should prioritize testing whether model outputs align with clear common-sense or domain-specific expectations, paying close attention to discrepancies with their justified priors. For instance, an error of exaggeration of differences between groups on a non-divisive issue is likely a more serious error than an error of exaggeration on an issue known to be divisive. Moreover, while our current metrics focus on *mean differences* in opinion distributions, some stakeholders may prioritize other distributional properties—such as variance or skew of opinions (Bui et al. 2025; Wang, Morgenstern, and Dickerson 2025). Our framework is flexible: the test statistics we propose can be adapted to target other distributional moments.

**Conclusion** By applying quality control checks that surface logical consistency and alignment with stakeholder expectations, practitioners can determine whether an approach to opinion simulation clears a basic threshold for reliability before devoting significant resources to collecting large-scale human reference data. An approach that meets these conditions may still require subsequent testing and tuning, but it at least demonstrates fundamental viability. If it fails on multiple fronts, stakeholders should carefully consider whether an LLM-driven approach is worthwhile or whether alternative data-centric or human-led solutions would be more consistent and trustworthy.

## Acknowledgments

This work was supported in part by Good Systems, a research grand challenge at the University of Texas at Austin. S.F. acknowledges support from the Schmidt Sciences AI2050 Early Career Fellowship.

## References

Aher, G.; Arriaga, R. I.; and Kalai, A. T. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

American National Election Studies. 2021. ANES 2020 Time Series Study Full Release. Accessed December 21, 2023.

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

Artstein, R.; and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4): 555–596.

Aruguete, M. S.; Huynh, H.; Browne, B. L.; Jurs, B.; Flint, E.; and McCutcheon, L. E. 2019. How serious is the 'carelessness' problem on Mechanical Turk? *International Journal of Social Research Methodology*, 22(5): 441–449.

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

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Bui, N.; Nguyen, H. T.; Kumar, S.; Theodore, J.; Qiu, W.; Nguyen, V. A.; and Ying, R. 2025. Mixture-of-Personas Language Models for Population Simulation. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 24761–24778. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

de Winter, J. C.; and Dodou, D. 2010. Five-point Likert items: t test versus Mann–Whitney–Wilcoxon (addendum included). *Practical Assessment, Research, and Evaluation*, 15(11).

Fröhling, L.; Demartini, G.; and Assenmacher, D. 2024. Personas with Attitudes: Controlling LLMs for Diverse Data Annotation. *arXiv preprint arXiv:2410.11745*.

Ghosh, B.; Hasan, S.; Arafat, N. A.; and Khan, A. 2025. Logical Consistency of Large Language Models in Fact-Checking. In *The Thirteenth International Conference on Learning Representations*.

Hedges, L. V.; and Olkin, I. 2014. *Statistical methods for meta-analysis*. Academic press.

Huddy, L.; Cassese, E.; and Lizotte, M.-K. 2008. Gender, public opinion, and political reasoning. *Political women and American democracy*, 31–49.

Jansen, B. J.; gyo Jung, S.; and Salminen, J. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4: 100020.

Jiang, S.; Wei, L.; and Zhang, C. 2024. Donald Trumps in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models. *arXiv preprint arXiv:2411.01582*.

Karanjai, R.; Shor, B.; Austin, A.; Kennedy, R.; Lu, Y.; Xu, L.; and Shi, W. 2025. Synthesizing Public Opinions with LLMs: Role Creation, Impacts, and the Future to eDemorcacy. *arXiv preprint arXiv:2504.00241*.

Kassner, N.; Tafjord, O.; Schütze, H.; and Clark, P. 2021. BeliefBank: Adding Memory to a Pre-Trained Language Model for a Systematic Notion of Belief. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8849–8861. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Kosinski, M. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45): e2405460121.

Lease, M. 2011. On quality control and machine learning in crowdsourcing. *Human Computation*, 11(11): 1085.

Lee, S.; Peng, T.-Q.; Goldberg, M. H.; Rosenthal, S. A.; Kotcher, J. E.; Maibach, E. W.; and Leiserowitz, A. 2024. Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8): e0000429.

Liu, H.; Gwizdka, J.; and Lease, M. 2024. Exploring Multidimensional Checkworthiness: Designing AI-assisted Claim Prioritization for Human Fact-checkers. *arXiv preprint arXiv:2412.08185*.

Lizotte, M.-K. 2020. *Gender differences in public opinion: Values and political consequences*. Temple University Press.

Mingmeng, G.; He, S.; and Trotta, R. 2024. Are Large Language Models Chameleons? In *ICML 2024 Workshop on LLMs and Cognition*.

Miranda, B.; Lee, A.; Sundar, S.; Casasola, A.; and Koyejo, S. 2024. Beyond Scale: The Diversity Coefficient as a Data Quality Metric for Variability in Natural Language Data. arXiv:2306.13840.

Mostafazadeh Davani, A.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Mündler, N.; He, J.; Jenko, S.; and Vechev, M. 2024. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. In *The Twelfth International Conference on Learning Representations*.

Namikoshi, K.; Filipowicz, A.; Shamma, D. A.; Iliev, R. I.; Hogan, C. L.; and Arechiga, N. 2024. Using LLMs to Model the Beliefs and Preferences of Targeted Populations. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

NIST. 2003. *Dataplot Reference Manual: Volume II – Let Subcommands*. National Institute of Standards and Technology, Gaithersburg, MD. https://www.itl.nist.gov/div898/software/dataplot/refman2/ch2/weightsd.pdf.

Norman, G. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5): 625–632.

Orlikowski, M.; Pei, J.; Röttger, P.; Cimiano, P.; Jurgens, D.; and Hovy, D. 2025. Beyond Demographics: Fine-tuning Large Language Models to Predict Individuals' Subjective Text Perceptions. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2092–2111. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

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

Rothschild, D. M.; Brand, J.; Schroeder, H.; and Wang, J. 2024. Opportunities and risks of LLMs in survey research. *Available at SSRN*.

Sanders, N. E.; Ulinich, A.; and Schneier, B. 2023. Demonstrations of the potential of AI-based political issue polling. *arXiv preprint arXiv:2307.04781*.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906. Seattle, United States: Association for Computational Linguistics.

Sarstedt, M.; Adler, S. J.; Rau, L.; and Schmitt, B. 2024. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6): 1254–1270.

Sehat, C. M.; Li, R.; Nie, P.; Prabhakar, T.; and Zhang, A. X. 2024. Misinformation as a Harm: Structured Approaches for Fact-Checking Prioritization. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11.

Sun, H.; Pei, J.; Choi, M.; and Jurgens, D. 2023. Aligning with Whom? Large Language Models Have Gender and Racial Biases in Subjective NLP Tasks. arXiv:2311.09730.

Wang, A.; Morgenstern, J.; and Dickerson, J. P. 2025. Large language models that replace human participants can harm-

fully misportray and flatten identity groups. *Nature Machine Intelligence*, 1–12.

Wang, M.; Zhang, D.; and Zhang, H. 2024. Large Language Models for Market Research: A Data-augmentation Approach. *Available at SSRN 5057769*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. Survey Certification.

Wright, D.; Arora, A.; Borenstein, N.; Yadav, S.; Belongie, S.; and Augenstein, I. 2024. LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 17085–17112. Miami, Florida, USA: Association for Computational Linguistics.

Zheng, Z.; Ning, K.; Wang, Y.; Zhang, J.; Zheng, D.; Ye, M.; and Chen, J. 2024. A Survey of Large Language Models for Code: Evolution, Benchmarking, and Future Trends. arXiv:2311.10372.