From Fallback to Frontline: When Can LLMs be Superior Annotators of Human Perspectives?

Hasan Amin[1], Harry Yizhou Tian[1], Xiaoni Duan[1], Chien-Ju Ho[2], Rajiv Khanna[1], Ming Yin[1]

[1]Purdue University, [2]Washington University in St. Louis

{hasanamin,tian253,duan79,rajivak,mingyin}@purdue.edu, chienju.ho@wustl.edu

Abstract

Although large language models (LLMs) are increasingly used as annotators at scale, they are typically treated as a pragmatic fallback rather than a faithful estimator of human perspectives. This work challenges that presumption. By framing perspective-taking as the estimation of a latent group-level judgment, we characterize the conditions under which modern LLMs can outperform human annotators, including in-group humans, when predicting aggregate subgroup opinions on subjective tasks, and show that these conditions are common in practice. This advantage arises from structural properties of LLMs as estimators, including low variance and reduced coupling between representation and processing biases, rather than any claim of lived experience. Our analysis identifies clear regimes where LLMs act as statistically superior frontline estimators, as well as principled limits where human judgment remains essential. These findings reposition LLMs from a cost-saving compromise to a principled tool for estimating collective human perspectives.[1]

Content warning: This paper contains examples of offensive or toxic language that some readers may find disturbing.

1 Introduction

Capturing subjective human perspectives, particularly those of specific demographic or cultural subgroups, is a foundational challenge in NLP, with direct implications for evaluation, fairness, and safety. Unlike factual labels, subjective judgments admit no objective ground truth, as perceptions of toxicity, offensiveness, or harm depend on lived experience, social context, and individual bias (Sap et al., 2019, 2021). As a result, modern annotation pipelines rely heavily on human judgments aggregated across annotators, treating crowd averages as proxies for group-level perspectives.

In practice, however, collecting high-quality, representative annotations is often difficult or infeasible. Recruiting in-group annotators for every target subgroup is costly, slow, and sometimes impossible, especially for small, intersectional, or hard-to-reach populations (Davani et al., 2022; Sandri et al., 2023). Consequently, many workflows turn to *perspective-taking* (PT): asking annotators to estimate how a group would judge an item, rather than reporting their own view (Frenda et al., 2025; Duan et al., 2025). Large language models (LLMs) have recently emerged as natural candidates for this role, given their ability to simulate personas, follow instructions, and draw on broad training data (Wilf et al., 2023).

Despite their growing use, a strong presumption persists: LLM-based annotation is, at best, a scalable compromise—useful when human annotation is unavailable, but inherently inferior to "real" human perspectives. This paper challenges this established view by reframing perspective-taking itself. When the goal is PT, neither humans nor LLMs observe the target quantity directly. Instead, both are attempting to estimate the same latent group-level mean: how a population would judge an item on average. Once PT is framed as a problem of statistical estimation, rather than measurement of individual opinion, the comparison between humans and LLMs fundamentally changes.

We formalize PT as estimation of a latent group-level judgment and decompose mean-squared error into three structural components: bias, variance, and correlation. To interpret these terms, we introduce a two-lens view of estimation error. Representation error (the *Wide Lens*) captures how well an annotator's knowledge reflects the target group, while processing and calibration error (the *Clear Lens*) captures how judgments are produced given that representation. Crucially, estimation quality also depends on variance across annotators and on the correlation between their errors. For hu-

[1]Code and data, including LLM annotations, are available at github.com/shasanamin/llm-perspective-taking.

Findings of the Association for Computational Linguistics: ACL 2026, pages 42798–42830
July 2-7, 2026

mans, shared social context and identity can induce high variance and strong correlations, as well as a positive *coupling* between representation and processing errors, leading to super-additive bias. For LLMs, representation and processing can arise from mechanically distinct stages (e.g., pretraining vs. post-training vs. inference-time calibration), often yielding lower variance and weaker coupling.

Guided by this framework, we evaluate humans and LLMs as estimators of subgroup-level judgments across two datasets. We find that modern LLMs can consistently outperform human annotators—including in-group humans—at predicting aggregate subgroup judgments. This advantage emerges most clearly in low-budget regimes, where estimation error is variance-dominated, and a *single LLM* output is found to often outperform even small *human crowds*. Crucially, the advantage does not come from LLMs possessing lived experience, but from more favorable estimator properties under realistic annotation budgets.

Our analysis also reveals an unexpected empirical phenomenon: explicit 'reasoning' or chain-of-thought (CoT) prompting can degrade LLM PT performance. We find that reasoning induces a systematic *criterion drift*, shifting the model from estimating empirical group judgments toward applying a rubric-based classification standard, which can partially counteract the structural advantages of LLM-based estimation. We term this effect the *reasoning paradox*, and provide a theoretical account that can explain why increased deliberation need not improve—and can even harm—aggregate perspective estimation.

Our results do not argue for replacing humans or dismissing lived experience. Instead, they clarify *when* LLMs can act as statistically superior estimators of aggregate perspective—and when humans remain indispensable, such as for highly specific or intersectional subgroups, rare or underrepresented populations, or contexts where stakeholder legitimacy and participation are themselves essential. By making these regimes explicit, we move beyond ideological debates toward principled, task-dependent annotation design.

This paper makes five key contributions:

- We formalize perspective-taking as estimation of a latent group-level judgment, enabling principled comparison between humans and LLMs.
- We introduce a bias–variance–correlation framework with a two-lens decomposition and a coupling hypothesis that accounts for systematic human failure modes.
- We provide extensive empirical evidence that LLMs can outperform humans—including in-group humans—under realistic budgets, while identifying clear boundary conditions.
- We derive actionable guidance for engineering PT pipelines, including when to use LLMs, when to use humans, and how prompting, model choice, and reasoning affect outcomes.
- We introduce a *differential perspective-taking* diagnostic that isolates group-specific sensitivity from generic annotation skill, revealing a regime where human annotators retain advantage over LLMs.

Taken together, our findings reposition LLMs in subjective NLP tasks: not as a fallback for missing human data, but as a potentially *frontline estimator* of collective human perspectives—when used carefully, validated rigorously, and deployed where their structural advantages apply.

1.1 Related Work

Prior studies examine PT as a mechanism for eliciting group judgments, highlighting both its benefits and its distortions (Frenda et al., 2025; Duan et al., 2025; Sandri et al., 2023; Aoyagui et al., 2025). We extend this line by formalizing PT as a statistical estimation problem.

Another line of work evaluates LLMs as annotators, typically focusing on their agreement with human labels (Li et al., 2025; Movva et al., 2024; Calderon et al., 2025). In contrast, we frame the problem through estimation efficiency and characterize when different estimators are preferable. Research on social reasoning shows that LLMs perform strongly on theory-of-mind benchmarks (Rabinowitz et al., 2018; Huang et al., 2023) but also exhibit systematic social biases (Hagendorff et al., 2023; Hu et al., 2025a). Our analysis decomposes how bias, variance, and correlation jointly shape PT performance.

Recent work on persona prompting investigates whether sociodemographic conditioning enables LLMs to simulate individuals or groups, with mixed empirical results (Sun et al., 2025; Lutz et al., 2025; Orlikowski et al., 2025). Our framework offers potential structural explanation: individual simulation aims to recover the full within-group distribution, which is a more difficult target than our

group mean estimation. Finally, work on pluralistic alignment focuses on matching entire opinion distributions (Sorensen et al., 2024; Feng et al., 2024; Lee et al., 2024). We view accurate mean estimation as a foundational step, whose analysis naturally extends to richer distributional objectives.

2 Perspective-Taking as Estimation

We frame perspective-taking not as a *measurement* task, but as the *statistical estimation* of latent group-level judgment. When asked how a group of humans would judge an item, neither humans nor LLMs directly observe the target quantity, and must infer it from incomplete experience, priors, and the elicitation protocol. This section formalizes this view and presents testable implications. Full derivations and proofs are provided in Appendix B.

2.1 Target Quantity and Protocols

Let $x \in \mathcal{X}$ be an item and $g \in \mathcal{G}$ a (demographic) group with population distribution P_g. Let $Y_h(x) \in [0,1]$ denote the *direct* judgment of a randomly sampled group member $h \sim P_g$ on item x (e.g., x is toxic or not). The group-level perspective of interest is the latent subgroup mean

$$f^*(x,g) \triangleq \mathbb{E}_{h\sim P_g}[Y_h(x)]. \quad (1)$$

Direct annotation vs. PT. Direct annotation measures $Y_h(x)$ for sampled individuals and aggregates those measurements. PT instead elicits an estimate $\hat{f}(x,g)$ of $f^*(x,g)$ from an annotator (human or LLM). Because f^* is latent, humans and LLMs should be compared as *estimators of the same quantity*, rather than as interchangeable label generators.

2.2 A Two-Lens Model of Systematic Bias

We decompose a single PT prediction $\hat{f}_A(x,g)$ into two bias components plus residual noise:

$$\hat{f}_A(x,g) = f^*(x,g) + \underbrace{b_{\text{repr},A}(x,g)}_{\text{Wide Lens}} + \underbrace{b_{\text{proc},A}(x,g)}_{\text{Clear Lens}} + \varepsilon_A(x,g), \quad (2)$$

where $A \in \{H, L\}$ denotes **H**umans or **L**LMs and $\mathbb{E}[\varepsilon_A(x,g)] = 0$.

b_{repr} captures *representation bias*: how well the estimator approximates the population distribution P_g over individuals in the target group g, and thus how accurately it estimates expectations under $h \sim P_g$. A *Wide Lens* reflects a smaller mismatch between P_g and the estimator's implicit sampling distribution over individuals in g. For example, when estimating how Gen Z would judge the slang term "slay," an older annotator may overweight their own social circle, misrepresenting the true demographic mixture. LLMs often yield broader coverage for common groups due to pretraining on large-scale corpora.

b_{proc} captures *processing bias*: how an internal representation is translated into a numeric judgment, i.e., the fidelity of the *Clear Lens* through which the annotator renders a judgment. In the aforementioned example, even if the older annotator recalls a Gen Z example, they might mistakenly project their own norms, interpreting the word as violent rather than positive. LLMs also exhibit processing distortions, but these errors tend to be more stable and more readily modifiable.

Coupling and super-additivity. The two bias components need not be independent. For humans, social identity and homophily often link who is represented with how judgments are processed, causing the biases to align in sign. Let $\mu_A(x,g) \triangleq \mathbb{E}[\hat{f}_A(x,g) - f^*(x,g)]$ denote the total mean bias, which decomposes as $\mu_A = \mu_{\text{repr},A} + \mu_{\text{proc},A}$, where $\mu_{\text{repr},A} \triangleq \mathbb{E}[b_{\text{repr},A}]$ and $\mu_{\text{proc},A} \triangleq \mathbb{E}[b_{\text{proc},A}]$. Expanding the total squared bias yields

$$\mu_A^2 = \mu_{\text{repr},A}^2 + \mu_{\text{proc},A}^2 + \underbrace{2\,\mu_{\text{repr},A}\,\mu_{\text{proc},A}}_{\text{Coupling}}, \quad (3)$$

In out-group PT, the coupling term is expected to be positive, producing *super-additive* error. For LLMs, representation errors (driven by pretraining coverage) and processing errors (driven by post-training and inference-time prompting) may arise from mechanically distinct sources, potentially weakening coupling or even having a subtractive effect.

2.3 Bias–Variance–Correlation Under Aggregation

Let $\{\hat{f}_{A,i}(x,g)\}_{i=1}^k$ be k exchangeable PT predictions produced by protocol A (e.g., k humans or k diversified LLM samples), and let $\bar{f}_A^{(k)}(x,g) = \frac{1}{k}\sum_{i=1}^k \hat{f}_{A,i}(x,g)$. By definition, $\text{MSE}\left(\bar{f}_A^{(k)}; x, g\right) = \mathbb{E}\left[\left(\bar{f}_A^{(k)}(x,g) - f^*(x,g)\right)^2\right]$, which reduces to $\mu_A(x,g)^2 + \text{Var}\left(\bar{f}_A^{(k)}(x,g)\right)$. Define the centered residual $r_{A,i}(x,g) \triangleq \hat{f}_{A,i}(x,g) - f^*(x,g) - \mu_A(x,g)$, which absorbs both centered bias variation and noise.

Let $V_A(x,g) \triangleq \mathrm{Var}(r_{A,i}(x,g))$ denote per-annotator residual variance and let $\gamma_A(x,g)$ denote the exchangeable residual correlation, i.e., $\mathrm{Corr}(r_{A,i}, r_{A,j}) = \gamma_A$ for $i \neq j$. As shown in Appendix B, exchangeable-variance calculation gives:

$$\mathrm{MSE}\left(\bar{f}_A^{(k)}; x, g\right) = \underbrace{\mu_A(x,g)^2}_{\text{Bias}^2} + \underbrace{\gamma_A(x,g)V_A(x,g)}_{\text{Correlation floor}} + \underbrace{\frac{1-\gamma_A(x,g)}{k}V_A(x,g)}_{\text{Reducible variance}} \quad (4)$$

This decomposition makes explicit how bias, variance, and correlation jointly determine estimator quality and yields a simple decision criterion: LLM PT is preferable whenever $\mathrm{MSE}(\bar{f}_L^{(k)}; x, g) < \mathrm{MSE}(\bar{f}_H^{(k)}; x, g)$. This inequality highlights when LLMs can move from fallback to frontline.

2.4 Predicted Performance Regimes and Control Levers

The estimation framework yields four testable predictions that structure our evaluation.

H1: The Budget Regime Hypothesis. In low-budget regimes (small k), MSE is dominated by per-annotator variance. Since LLMs are relatively deterministic under fixed prompting ($V_L \ll V_H$), LLM prediction should outperform human annotation—including direct annotation involving a broad group—by minimizing sampling noise. In contrast, performance in high-budget regimes is limited by bias and correlation floors, and naive aggregation yields diminishing returns.

H2: The Coupling Hypothesis. In out-group settings, human PT increases error not only by enlarging bias magnitudes but by increasing positive coupling in Eq. (3), inflating $\mu_H(x,g)^2$ super-additively. LLMs lacking social identity should exhibit greater stability across crowd settings.

H3: The Representation Limits Hypothesis. As target groups become more specific or less prevalent, representation mismatch increases due to sparse or stereotype-skewed training evidence, enlarging $|b_{\mathrm{repr},L}(x,g)|$ and hence $|\mu_L(x,g)|$. This increases LLM PT error, and delineates regimes where human (in-group) knowledge could be potentially advantageous.

H4: The Engineerability Hypothesis. Compared to human annotation, LLM can be easier to engineer for perspective-taking. Its error components are mechanically distinct and can be selectively influenced through model choice (primarily b_{repr}), prompting strategies and reasoning protocols (primarily b_{proc}), and diversification (primarily γ_L). Consequently, LLM performance should systematically improve or degrade by targeted interventions.

3 Experimental Setup

We evaluate humans and LLMs as estimators of subgroup-level judgments under the *perspective-taking-as-estimation* framework. Our design combines (i) a high-fidelity human PT benchmark with dense subgroup annotation and (ii) a large-scale safety dataset with rich demographic coverage, enabling controlled comparisons across annotation budgets, group structure, and estimator regimes. We briefly present experimental details here, deferring a more elaborate treatment to Appendix C.

3.1 Datasets and Ground Truth

Toxicity Detection. We use the dataset of Duan et al. (2025), which contains 120 online comments balanced by target group and toxicity level. Each comment receives *direct annotations* from at least 50 U.S.-based crowdworkers from the target subgroup, each rating items for themselves (e.g., "Do you find this comment toxic?"). The mean of these direct annotations within a subgroup defines the *ground-truth* label $f^*(x,g)$ (Eq. 1), serving as a high-fidelity proxy for the latent subgroup-level toxicity rate. The dataset additionally includes matched human PT judgments. We extend the dataset beyond binary gender by collecting new PT and direct annotations from non-binary participants ($N = 97$) on Prolific. Crucially, direct annotators and perspective-takers are entirely separate pools. Perspective-takers, whether human or LLM, are asked to estimate $f^*(x,g)$ (e.g., "What percentage of {females, males, non-binary people} would find this comment toxic?") without ever observing any direct annotations or aggregate statistics.

DICES (Conversational Safety). We additionally use DICES-350 (Aroyo et al., 2023), which provides over 100 direct safety judgments per example across multiple demographic axes. While DICES does not include human PT, its dense annotation structure enables precise subgroup-level estimation and allows us to probe LLM PT behavior under increasing demographic breadth and

heterogeneity. We use DICES to study regime behavior and boundary conditions rather than direct human–LLM comparisons.

3.2 Estimators and protocols

We consider three annotation protocols: (i) *direct annotation* by subgroup members, (ii) *human PT* (in-group/out-group), and (iii) *LLM PT*, where models are prompted to estimate subgroup-level judgments under instructions aligned with human protocols. LLM PT is out-group by design and never given access to aggregate statistics. LLM PT, like human PT, outputs the estimated mean fraction. We evaluate a diverse set of contemporary LLMs across model families, scales, and inference strategies. Specifically, we evaluate models from four families (GPT, Qwen, Gemma, and DeepSeek) ranging from 1B parameters to frontier scale. For the pretrained vs. post-trained models analysis (Appendix D.4), we additionally evaluate matched model pairs from Ministral 3. For reasoning-enabled variants, we use each model's native reasoning mode rather than manually appended chain-of-thought instructions. All models are queried zero-shot, unless noted otherwise.

3.3 Evaluation Criteria

We treat the mean of direct human annotations per subgroup as the gold standard and evaluate estimators using mean squared error (MSE), bias, and variance. To reflect realistic annotation budgets, we use bootstrapping to simulate k-annotation regimes ($k = 1$ to 10), applying identical procedures to human and LLM estimators. Each reported metric is the average over B=1,000 bootstrap resamples, ensuring stable estimator comparisons that do not depend on any single annotator draw.

4 Evaluation I: When Do LLMs (Not) Excel at Perspective-Taking?

We first evaluate the comparative advantage of LLMs across different budget regimes (k) and demographic granularities. Each subsection directly tests a hypothesis from Section 2.4. We restrict LLM PT results to popular GPT models here. See Appendix D for additional models and ablations.

4.1 Validating the Budget Regime Hypothesis

LLM vs Human ($k = 1$). Many practical annotation pipelines rely on a single pass per item, i.e., a single annotator regime. We compare a single zero-shot LLM against a single human annotator

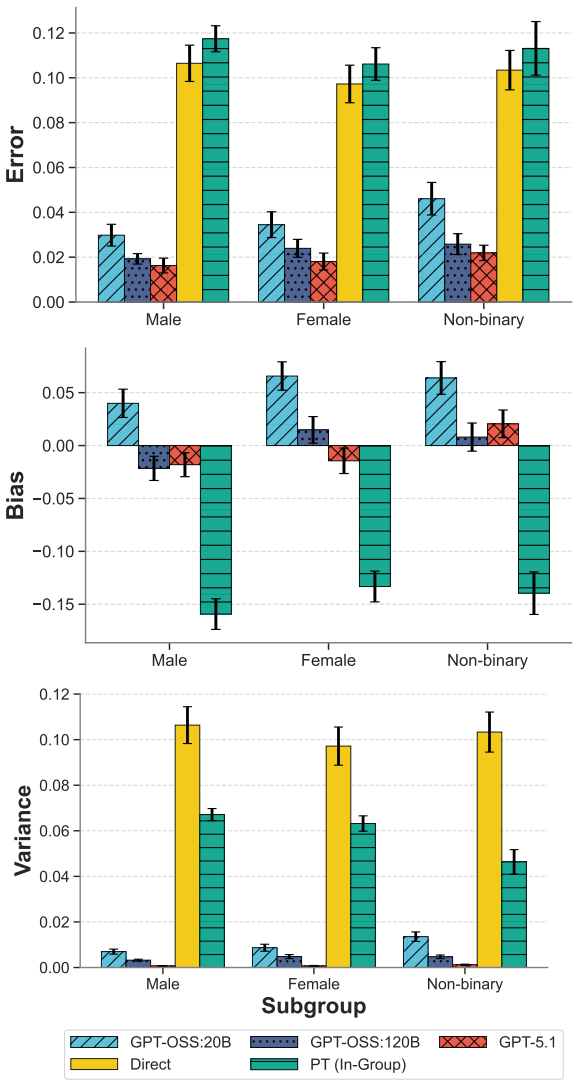


Figure 1: **Single-annotation (k=1) error decomposition** for three GPT variants vs. human baselines on Toxicity Detection. LLMs achieve lower MSE (top) across all gender subgroups, driven by lower bias (middle) and substantially lower variance (bottom).

(see Figure 1). Across all gender groups in the Toxicity Detection dataset, the single LLM consistently achieves lower error. This directly validates the low-budget variance-dominance prediction of H1: individual humans are noisy estimators (large V_H) of the group mean, whereas LLM PT is comparatively deterministic ($V_L \ll V_H$). Decomposing MSE further reveals that LLM PT also dominates human PT in the bias component. Interestingly, human perspective-takers systematically *underestimate* the fraction of the target group that would judge content as toxic (negative bias). LLM perspective-takers have lower bias, and some even tend to *overestimate* it (positive bias), which could indicate conservative safety calibration.

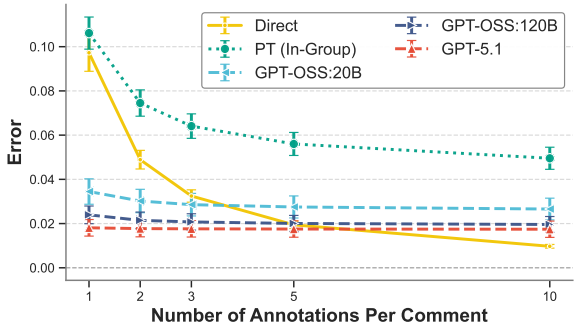


Figure 2: **MSE vs. annotation budget** k for the female subgroup. A single LLM PT estimate (k=1) is comparable to aggregating 3–5 direct human labels.

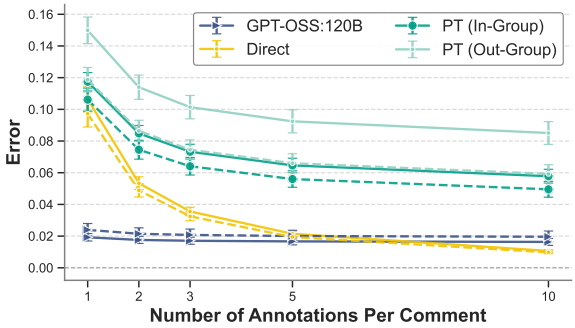


Figure 3: **In-group vs. out-group human PT** compared to LLM (GPT-OSS:120B) PT, predicting male (solid) and female (dashed) judgments. Out-group human PT incurs a pronounced error increase, consistent with super-additive coupling, while the LLM remains comparatively stable across target groups.

LLM Ensembles vs. Human Crowds ($k > 1$). We next examine how performance scales with the annotation budget k (Figure 2). As k increases, human PT improves rapidly at first but soon plateaus, consistent with a non-trivial correlation floor after accounting for bias. In contrast, homogeneous LLM ensembles (multiple samples from the same model) show only modest gains, consistent with near-deterministic behavior (V_L small) rather than necessarily high inter-sample correlation.

Outperforming Single-Annotator Baselines. A striking consequence of H1 is that a single LLM PT estimate can outperform a single *direct* human annotation. While direct annotators are unbiased with respect to their own judgment, they are high-variance ($k = 1$) samples from the population defining $f^*(x, g)$. Formally, at k=1 a single direct annotation Y_h has $\mathrm{MSE} = V_{\mathrm{hetero}}(x, g)$ (the within-group heterogeneity, since $\mathbb{E}[Y_h] = f^*$), while a single LLM PT estimate has $\mathrm{MSE} = \mu_L^2 + V_L$. Hence LLM PT is the superior single-annotator estimator whenever $\mu_L^2 + V_L < V_{\mathrm{hetero}}$, i.e., when LLM bias and variance together are smaller than the population spread. Empirically, a single LLM PT estimate is found comparable to aggregating 3–5 direct human labels. This result highlights an often-overlooked implication: *when ground truth is estimated from a small number of human annotations, LLM-generated estimates can be the statistically preferable choice until sufficient human replication is available*.

4.2 Validating the Coupling Hypothesis

In-group vs. out-group human PT. Figure 3 compares in-group and out-group human PT. Out-group PT exhibits a pronounced error increase, especially when female annotators predict male judgments. This asymmetry is consistent with the coupling mechanism in Eq. (3): identity mismatch simultaneously distorts representation and processing, inflating μ_H^2 super-additively. The effect is weaker when male annotators predict female judgments, consistent with differing bias magnitudes across groups. While out-group challenges feel intuitive, our framework provides a structural account of such mechanism.

LLM stability. LLM PT is out-group by construction, yet its error remains comparatively stable across target groups. This supports H2's prediction that mechanically decoupled representation and processing reduce sensitivity to identity mismatch, making LLMs particularly robust when in-group human recruitment is infeasible.

4.3 Validating the Representation Limits Hypothesis

We now test regimes where LLM PT should deteriorate due to representation mismatch. We conceptualize demographic groups as nodes in an inclusion tree, where depth corresponds to specificity ($g' \subset g$, e.g., "college-educated, black women" is a more specific group than "college-educated people") and width at the same depth corresponds to prevalence $\pi(g)$ (e.g., within US, "White" race is more prevalent with larger $\pi(\text{White})$ compared to "Asian"). We use the DICES dataset here, focusing exclusively on LLM PT behavior due to the absence of human PT annotations.

Group specificity (depth). As groups become more specific, LLM conditioning increasingly relies on sparse evidence. Figure 4 shows a mono-

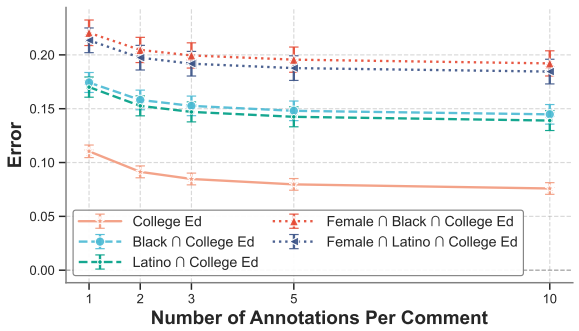


Figure 4: **Effect of subgroup specificity** on LLM (GPT-OSS:120B) PT error on DICES. MSE rises monotonically as the target becomes more specific.

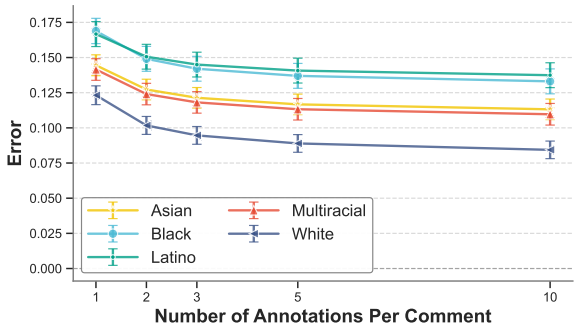


Figure 5: **Effect of subgroup prevalence** on LLM (GPT-OSS:120B) PT error on DICES (race axis). MSE generally rises as the target becomes less prevalent.

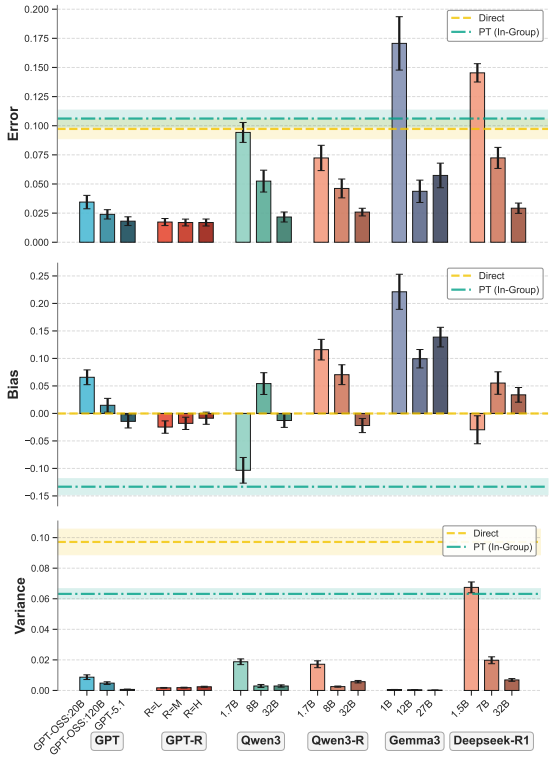


Figure 6: **Impact of model family and scale** on single-annotation (k=1) PT for the female subgroup. Cross-model differences in MSE (top) are dominated by bias (middle), with variance (bottom) remaining small.

tonic increase in LLM (GPT-OSS:120B) PT error with subgroup depth. This trend is consistent with increasing $|b_{\text{repr},L}|$ dominating the error, as predicted by H3.

Group prevalence (width). Low-prevalence groups pose a complementary challenge. As $\pi(g)$ decreases, LLMs can suffer representation mismatch due to limited or stereotype-skewed training signal. Figure 5 shows that performance worsens for minority demographics (e.g., Black vs. White). This aligns with established findings on LLM fairness (Sap et al., 2019) and, crucially, helps identify the structural boundary of the "frontline" claim.

5 Evaluation II: On The Engineerability of LLM Perspective-Taking

Evaluation I established when LLMs outperform humans as estimators of subgroup-level judgment. Here we examine *how this advantage arises* and which components of error can be controlled, helping validate the Engineerability Hypothesis (H4) from Section 2.4. Guided by the bias–variance–correlation decomposition (Eq. 4), we organize interventions by whether they primarily affect the *Wide Lens* (representation bias), *Clear Lens* (processing bias), or the *correlation floor*.[2] Results here focus on the female subgroup, with additional groups and robustness checks in Appendix D.

5.1 Wide Lens: Model Family and Scale

To probe representation effects, we compare LLMs spanning model families and scales under an identical PT protocol. While model choice is not a pure manipulation of representation bias, it serves as an empirically grounded proxy for representational fidelity via differences in pretraining coverage, capacity, and training regime. Figure 6 shows that model family and scale induce large differences in MSE, sufficient to reverse whether LLM PT outperforms human PT. While bigger models generally outperform weaker ones, improvements are

[2] As soft evidence that Wide and Clear Lens errors arise from mechanically distinct training stages, we also conduct matched pretrained vs. post-trained comparisons (Appendix D.4). Post-training collapses variance by an order of magnitude alongside generally increasing absolute bias. While this does not separately measure representation and processing bias, it shows that total bias rises through post-training, directionally consistent with the two-lens view.

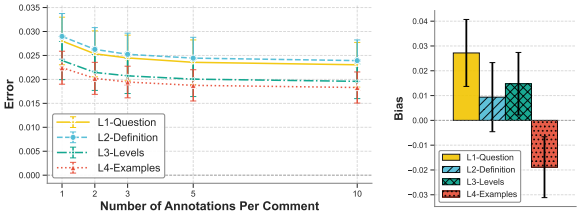


Figure 7: **Impact of prompting** on single-annotation LLM (GPT-OSS:120B) PT for the female subgroup. Increasing prompt structure from L1 through L4 lowers MSE (left) primarily by shifting bias (right) and can even flip its sign.

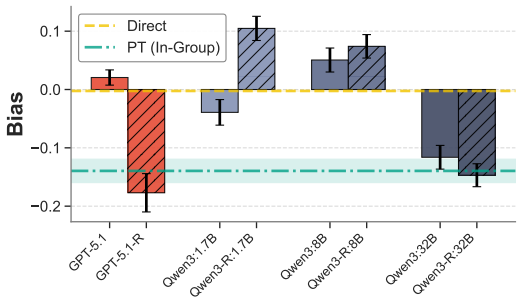


Figure 8: **Reasoning paradox diagnostic** on the non-binary subgroup: for four base–reasoning pairs, enabling native reasoning (hatched) shifts systematic bias away from ground truth, consistent with criterion drift.

not monotonic: mid-sized models from one family can outperform larger models from another. Error decomposition reveals that these differences are primarily driven by bias, with variance remaining comparatively small across models.

5.2 Clear Lens: Prompting and Reasoning

We next test whether processing bias can be modified without changing representation. We fix the model (GPT-OSS:120B) and use a controlled prompting ladder of increasing structure: L1 (question only), L2 (+ definition), L3 (+ levels), L4 (+ examples); see Appendix C.3 for complete details. Figure 7 shows that prompting can substantially shift both MSE and bias. Importantly, this is *not* a monotonic "more structure is better" result: different prompts reweight how beliefs are translated into numeric estimates. Thus, prompting acts as a calibration mechanism that reshapes b_{proc} rather than universally reducing error.

The Reasoning Paradox. Reasoning-enabled inference modes do not reliably improve PT accuracy, and can in fact substantially worsen it. As shown in Figure 8, explicit reasoning can induce larger systematic bias, especially for a harder subgroup like

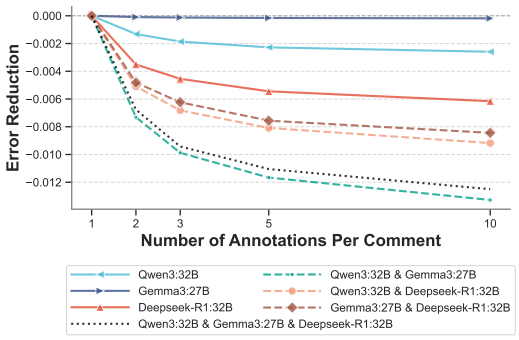


Figure 9: **Cross-family model mixing** at the large-model tier on the female subgroup. Mixing across families (dashed/dotted) yields consistent but modest error reduction relative to each model's k=1 baseline (solid). Nonetheless, mixtures often remain inferior to the strongest single model.

non-binary people. The effect is model-dependent, and analysis of reasoning traces (Appendix D.5) suggests that the dominant mechanism is *criterion drift*: reasoning drifts models away from estimating the empirical group toxicity rate toward applying a rubric-based classification standard, producing a systematic shift whose sign effect depends on the base model's pre-existing bias. Identity-mediated 're-coupling' between representation and processing (cf. Section 2.2) is one pathway for this drift, although not the main driver in our inspected traces.

5.3 Correlation: Mixing and Temperature

We next test interventions aimed at diversification, for lowering the correlation floor $\gamma_L V_L$ in Eq. 4. We study two common diversification mechanisms: (i) *model mixing* (aggregating predictions from different models), and (ii) *temperature* (increasing sampling randomness within a single model). Mixing similar-sized models across families yields modest but consistent error reductions (Figure 9), whereas mixing within a family across sizes (Figure A8) or increasing temperature (Figure A6) provides limited gains. These findings suggest that correlation engineering is secondary in many regimes.

6 Discussion and Conclusion

This work challenges the prevailing assumption that LLM-based annotation is merely a cost-saving approximation of human judgment. By reframing perspective-taking (PT) as statistical estimation of a latent group-level quantity, we show—both theoretically and empirically—that humans are not always the best available estimators of aggregate

group perspectives. Rather than treating LLMs as a pragmatic fallback, our results clarify *when*, *why*, and *how* LLMs can act as statistically preferable frontline estimators—and where human annotation remains essential.

Reframing the comparison. Much of the discomfort around LLM-based annotation stems from an implicit category error: humans are evaluated as sources of lived, individual judgments while LLMs are evaluated as predictors of aggregates. Once both are compared as estimators of the same latent target $f^*(x, g)$, the question shifts from authenticity to estimation quality under constraints. Our bias–variance–correlation framework makes this comparison explicit, principled, and testable.

When LLMs are likely to be optimal. LLMs dominate when (i) the target group is broad or internally heterogeneous, where even in-group humans suffer Wide Lens error; (ii) annotation budgets are small, making performance variance-dominated and LLM stability ($V_L \ll V_H$) decisive; and/or (iii) humans are recruited out-of-group, inducing coupled representation and processing bias that amplifies systematic error. In these regimes, LLMs achieve lower error through low variance, weaker coupling, and engineerable calibration.

When humans are likely to remain superior. Human annotation remains indispensable when (i) the group is deep, specific, and cohesive—making lived experience highly informative and representation bias small—and/or (ii) the group is sufficiently low-prevalence that LLM conditioning relies on sparse or stereotype-skewed evidence. Our differential perspective-taking analysis (Appendix E.2) provides direct evidence for this boundary: on the male vs. non-binary contrast, human PT significantly outperforms most LLMs at capturing per-item group differences, confirming that lower-prevalence groups remain a genuine frontier where LLMs lag. Importantly, some contexts also demand *procedural legitimacy*: in high-stakes or normative settings, direct stakeholder adjudication may outweigh purely statistical considerations. Our results do not contest this but clarify where statistical optimality and social legitimacy diverge.

Estimation-aware engineering. A core implication of our analysis is that annotation quality is not fixed. Practitioners can target specific error terms rather than indiscriminately collecting more labels. While humans are also engineerable in spirit, the levers are asymmetrical: many LLM interventions are fast and scalable once an evaluation protocol exists, whereas human improvements—while powerful—are often slower, costlier, and access-constrained. Model choice primarily determines representation bias, prompt design reshapes processing bias, and diversification can reduce correlated error (though likely with limited returns). Notably, explicit reasoning can backfire, as it can introduce a systematic criterion drift whose impact depends on the direction of the base model's bias.

From mean to distributional objectives. Our framework targets the population mean, which is the operationally relevant quantity in many annotation pipelines. However, for applications requiring pluralistic representation, such as capturing the full topology of disagreement within a group, distributional objectives are important complementary targets. Our bias–variance–correlation decomposition extends naturally to other summary statistics (median, quartiles) and, in principle, to distributional targets. Accurate mean estimation is a foundational first step, since an estimator that misses the mean is unlikely to recover higher moments.

Complementarity, not replacement. Our findings motivate a reallocation of effort: LLMs shoulder aggregate estimation in variance- and coupling-dominated regimes, freeing scarce human expertise for cases requiring contextual nuance, participatory grounding, or legitimacy. Seen through this lens, LLMs are not substitutes for lived experience, but tools for estimating its aggregate structure more stably and at scale. Additionally, if an LLM stabilizes around a biased estimate, this low-variance-around-wrong-mean stability can reinforce representational monoculture, underscoring the need for subgroup-aware validation and periodic human auditing even when LLMs serve as frontline estimators.

Final takeaway. Our core contribution is not the claim that LLMs are universally better annotators, but the demonstration that, once PT is treated as estimation, we can answer the important question of *who is the better estimator under specific conditions?* By making that question precise, and operational, we show how LLMs can move, in well-defined settings, from fallback to frontline, while clarifying where humans remain essential. The future of perspective-taking lies not in competition between humans and LLMs, but in their principled, complementary integration.

Limitations

This work has several limitations that delineate the scope of our claims.

Dependence on representation quality. Our framework makes explicit that LLM performance is constrained by representation bias. When subgroup evidence is sparse, highly specific, or distorted in training data, LLM-based perspective-taking can deteriorate despite low variance. While our experiments identify such regimes, real-world use still requires subgroup-specific validation, particularly for emerging identities or rapidly shifting social contexts.

Scope of tasks and domains. We focus on toxicity and conversational safety, where aggregate judgments are well-defined and densely annotated. Although the estimation framework is general, empirical results may not directly transfer to tasks where perspectives are more contested, multidimensional, or normatively grounded (e.g., moral or policy judgments). Extending the analysis to such settings remains an important direction for future work. In more contested domains (e.g., political orientation, aesthetic preference, cultural values), LLM representation biases may be amplified by long-tail distributions in training data, and the "frontline" claim should not be extrapolated without domain-specific validation.

Estimator-centric evaluation. We evaluate estimators against aggregate human judgments as proxies for the latent target $f^*(x, g)$. This standard choice abstracts away intra-group disagreement and deliberative dynamics. Our results therefore address estimation accuracy rather than the full richness of social meaning-making.

Model and protocol dependence. While we study a diverse set of models and prompting strategies, LLM architectures and alignment techniques are rapidly evolving. Some findings, particularly those concerning reasoning-enabled variants, may change as models and alignment techniques improve. Accordingly, our conclusions should be read as structural insights about estimator behavior, not as immutable properties of specific models.

Overall, these limitations do not undermine the central contribution of the paper, but rather clarify the conditions under which estimation-based (LLM) perspective-taking is most informative and reliable.

Ethical Considerations

This work engages with considerations around representation, fairness, and automation in subjective judgment.

Risk of misrepresentation. LLMs inherit biases from their training data and may misestimate perspectives of marginalized or low-prevalence groups. Our analysis explicitly identifies these failure modes and the regimes where human judgment remains essential. Using LLM-based perspective-taking without subgroup-aware validation risks reinforcing existing inequities.

Procedural legitimacy and participation. Statistical optimality is not always sufficient. In many high-stakes or normative settings, ethical legitimacy requires direct participation from affected communities. Our results do not argue for replacing such participation, but for clarifying when LLMs can responsibly support aggregate estimation without displacing human agency.

Misuse and overgeneralization. LLM-based perspective-taking may be overextended beyond its validated scope, for example to justify decisions about groups lacking adequate representation. We therefore emphasize transparent reporting of annotation protocols, explicit documentation of subgroup coverage, and conservative use in sensitive applications.

Responsible deployment. We recommend hybrid annotation pipelines that combine LLM-based estimation with periodic human auditing. Such approaches leverage the stability and efficiency of LLMs while preserving accountability, inclusivity, and adaptability.

Overall, LLMs are not substitutes for lived experience, but tools whose ethical use depends on alignment between statistical objectives, social context, and human oversight.

References

Paula Akemi Aoyagui, Kelsey Stemmler, Sharon A Ferguson, Young-Ho Kim, and Anastasia Kuzminykh. 2025. A matter of perspective (s): Contrasting human and llm argumentation in subjective decision-making on subtle sexism. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Xiaoni Duan, Zhuoyan Li, Chien-Ju Ho, and Ming Yin. 2025. Exploring the cost-effectiveness of perspective taking in crowdsourcing subjective assessment: A case study of toxicity detection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2359–2372.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-llm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025a. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Zhengyu Hu, Jianxun Lian, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie. 2025b. Population-aligned persona generation for llm-based social simulation. *arXiv preprint arXiv:2509.10127*.

Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, and 1 others. 2023. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. 2025. Improving language model personas via rationalization with psychological scaffolds. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21747–21770, Suzhou, China. Association for Computational Linguistics.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024. Aligning to thousands of preferences via system message generalization. *Advances in Neural Information Processing Systems*, 37:73783–73829.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori B Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing llm and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective

tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

A Extended Related Work

Perspective-taking and subjective judgment. Prior work studies PT as a mechanism for eliciting judgments about how groups would respond, highlighting both its utility and distortions, particularly in out-group settings (Frenda et al., 2025; Duan et al., 2025; Sandri et al., 2023). Recent HCI work further analyzes how humans and LLMs differ in the perspectives they invoke in subjective decision-making, showing reliance on differently weighted perspective distributions (Aoyagui et al., 2025). We advance this line by formalizing PT not merely as opinion elicitation, but as statistical estimation of a latent group-level judgment, enabling principled comparison of human and model estimators.

LLMs as annotators and replacement. Several studies evaluate LLMs as annotators or judges (Li et al., 2025), focusing on agreement with human labels and conditions for replacing human annotators. Movva et al. (2024) show that LLMs can approximate crowd averages for conversational safety but struggle with fine-grained between-group differences. Calderon et al. (2025) propose a statistical alternative annotator test to justify LLM replacement using limited labeled data. In contrast, we analyze estimation efficiency, characterizing when LLMs or humans constitute the superior estimator.

Social reasoning and identity bias. LLMs perform strongly on social reasoning and theory-of-mind benchmarks (Rabinowitz et al., 2018; Wang et al., 2021; Huang et al., 2023), yet exhibit systematic social identity biases and human-like reasoning distortions that vary across models and training regimes (Hagendorff et al., 2023; Hu et al., 2025a). We build on these insights by decomposing how bias, variance and correlation interact in PT estimation, including regimes where increased reasoning can counterintuitively degrade performance.

Persona-LLMs and sociodemographic prompting. A growing body of work investigates prompting LLMs with demographic personas to simulate individual-level opinions. Several studies report mixed or negative results for such sociodemographic simulation: Sun et al. (2025) find that persona prompting does not reliably improve prediction of individual subjective judgments, and Lutz et al. (2025) show that persona-based elicitation produces inconsistent alignment with human subgroup responses. Orlikowski et al. (2025) fine-tune LLMs to behave as individual annotators and report that demographic features alone are insufficient predictors. Our framework offers a possible explanation for these mixed findings: individual simulation implicitly requires approximating the full within-group distribution of judgments, whereas our task targets just the subgroup mean—a fundamentally easier estimation target. Joshi et al. (2025) find that psychologically-scaffolded rationales outperform default chain-of-thought for persona prediction, a pattern consistent with our reasoning paradox: unstructured reasoning may degrade group-level estimation by re-coupling representation and processing errors, while structured reasoning grounded in external frameworks can help. Meister et al. (2025) find that LLMs more accurately *describe* opinion distributions than *simulate* them via role-play, con-

sistent with our protocol of directly eliciting group-level statistics rather than role-playing individuals.

Pluralistic alignment and distributional objectives. A complementary line of work targets matching the full distribution of human opinions rather than a single summary statistic. Sorensen et al. (2024) chart a roadmap for pluralistic alignment, and Feng et al. (2024) propose multi-LLM collaboration to represent diverse viewpoints. Lee et al. (2024) demonstrate system-message-based generalization for aligning to thousands of user preferences, while Hu et al. (2025b) develop population-aligned persona generation for social simulation. We see mean estimation as a foundational first step in this broader agenda: if an estimator cannot accurately recover the first moment, distributional fidelity is unlikely. Our bias–variance–correlation decomposition provides formal vocabulary for reasoning about estimator quality that naturally extends to other summary statistics and distributional targets.

B Theory

This appendix develops a precise theoretical framework for when LLMs can outperform humans at perspective-taking annotation, i.e., predicting the group-mean judgment $f^\star(x, g)$ for an item x and a demographic group g. We provide: (i) a Wide Lens (representation) model via latent subcommunity mixtures, (ii) bias–variance–correlation decompositions for aggregated human and LLM estimators, (iii) a careful treatment of coupling between Wide and Clear Lens errors (mean-level vs variance-level), and (iv) asymptotic error floors and a superiority criterion.

B.1 Problem Setup

Let $\mathcal{P}_g$ denote the (conceptual) population of humans in group g, distributed as P_g. For $h \sim P_g$, let $Y_h(x) \in [0, 1]$ be h's direct judgment. The target group perspective is:

Definition B.1 (Target quantity)**.**

$$f^\star(x, g) \triangleq \mathbb{E}_{h\sim P_g}[Y_h(x)]. \tag{5}$$

Perspective-taking methods (human or LLM) output $\hat{f}(x, g)$ and are evaluated by

$$\mathrm{MSE}(\hat{f}; x, g) \triangleq \mathbb{E}\Big[\big(\hat{f}(x, g) - f^\star(x, g)\big)^2\Big], \tag{6}$$

where the expectation ranges over annotator identity / sampling randomness.

B.2 Wide Lens via Latent Mixtures

We model demographic groups as mixtures over latent subcommunities that can differ systematically in judgments.

Assumption B.2 (Latent subcommunity mixture)**.** For each group $g \in \mathcal{G}$, there exists a finite or countable set of latent subcommunities $\mathcal{C}_g$ and mixture weights $\mathbf{w}^g = (w_c^g)_{c\in\mathcal{C}_g}$. with $w_c^g \geq 0$ and $\sum_c w_c^g = 1$. Each subcommunity c has mean judgment

$$f_c^\star(x, g) \triangleq \mathbb{E}[Y_h(x) \mid h \in c]. \tag{7}$$

Then

$$f^\star(x, g) = \sum_{c\in\mathcal{C}_g} w_c^g \, f_c^\star(x, g). \tag{8}$$

Definition B.3 (Within-group heterogeneity)**.**

$$V_{\text{hetero}}(x, g) \triangleq \sum_{c\in\mathcal{C}_g} w_c^g \big(f_c^\star(x, g) - f^\star(x, g)\big)^2. \tag{9}$$

Annotator-internal mixture. An annotator (human or LLM) may implicitly reason about g using an *internal* mixture $\mathbf{q}^{g,a} = (q_c^{g,a})_{c\in\mathcal{C}_g}$ with $\sum_c q_c^{g,a} = 1$. Define the Wide Lens Effect, a representation bias capturing how a's mental model of group g misweights sub-communities, as:

$$b_{\mathrm{W}}(x, g; a) \triangleq \sum_{c\in\mathcal{C}_g} \big(q_c^{g,a} - w_c^g\big) \, f_c^\star(x, g). \tag{10}$$

Assumption B.4. For any (x, g) and any a under consideration, if $q_c^{g,a} > 0$ then $w_c^g > 0$. Equivalently, $\mathbf{q}^{g,a}$ is absolutely continuous with respect to $\mathbf{w}^g$.

Assumption B.4 excludes degenerate cases where a sample assigns positive mass to a subcommunity that does not exist in the target population; without it, the χ^2 divergence below can be infinite and the bound becomes vacuous. This assumption is natural for humans, who sample from lived experience. For LLMs, however, hallucinated stereotypes can assign $q_c > 0$ to fictitious subcommunities where $w_c = 0$, causing the χ^2 divergence to explode. This formalizes a structural limit of LLM-based PT: when representation relies on fabricated evidence, the Wide Lens bound breaks down entirely, providing a theoretical manifestation of the Representation Limits Hypothesis (H3).

	Human Annotator ("Lived" Estimator)	LLM Annotator ("Learned" Estimator)	Key Insights
Target $f^*(x,g) = \mathbb{E}_{h\sim P_g}[Y_h(x)]$	Estimate via social sampling and introspection.	Estimate via learned statistical regularities and instruction following.	Both are imperfect estimators of the same latent subgroup mean.
Bias $\left(\mu_{\text{repr}} + \mu_{\text{proc}}\right)^2$	**Representation (Wide Lens):** Sampling bias; lived experience samples a *local* neighborhood, misweighing latent subcommunities. **Processing (Clear Lens):** Cognitive bias; distortions converting belief to numeric judgment; relatively hard-coded by psychology. **Coupling:** Identity (homophily) can link representation and processing errors, yielding *superadditive* bias.	**Representation (Wide Lens):** Data bias; broad but imperfect coverage from pretraining; risky interpolation for sparse groups. **Processing (Clear Lens):** Modeling bias; distortions from alignment and safety constraints; relatively modifiable via prompting. **Coupling:** Pretraining and posttraining errors are mechanically distinct mechanisms, yielding weaker or mixed-sign interaction.	LLMs trained on vast subgroup data that are relatively immune to human-like cognitive biases can outperform humans. Persistent gap even at large budgets indicates a bias advantage; often the case for broad groups or out-group human PT. **Reasoning paradox:** Explicit "thinking" (CoT) in LLMs can induce *criterion drift*, shifting the model from distributional estimation to rubric-based classification and increasing bias.
Variance V	**High noise** for single annotators ($k = 1$) due to individual differences, mood, fatigue etc.; decreases with aggregation.	**Low noise** for a single model under fixed prompting; additional gains from ensembling are often modest.	**Low-budget regime:** A single LLM can outperform a single human and rival small human crowds, *even for direct annotation*.
Correlation γ	**Correlation floor:** Correlation from shared culture, platforms and norms; difficult to engineer away.	**Engineerable floor:** High correlation from shared data and model architectures; partially engineerable.	**Human advantage at scale:** Human crowds benefit more from aggregation, yet can plateau in PT.
Decision Rule	LLMs excel when $\text{Bias}_L^2+\gamma_L V_L+\frac{1-\gamma_L}{k}V_L < \text{Bias}_H^2+\gamma_H V_H+\frac{1-\gamma_H}{k}V_H$; typical for large heterogeneous groups where LLM calibration and diversity can be engineered. Humans remain preferable when groups are deep and cohesive while LLM conditioning is sparse or stereotype-skewed.		

Table A1: **Bias–Variance–Correlation View of Human vs. LLM for Perspective-Taking.** Wide Lens, Clear Lens, and Coupling form a bias decomposition; variance and correlation govern how performance scales with budget. This offers an overview of when and why LLMs can move from fallback to frontline through PT-as-estimation lens.

Lemma B.5 (Representation bias bound)**.** *Under Assumptions B.2 and B.4,*

$$b_{\text{W}}(x,g;a)^2 \le V_{\text{hetero}}(x,g)\cdot\chi^2\big(\mathbf{q}^{g,a}\,\|\,\mathbf{w}^g\big), \tag{11}$$

where

$$\chi^2\big(\mathbf{q}\,\|\,\mathbf{w}\big) \triangleq \sum_{c\in\mathcal{C}_g}\frac{(q_c-w_c)^2}{w_c}. \tag{12}$$

Proof. Fix (x,g) and suppress them in notation. By (8), $f^\star=\sum_c w_c f_c^\star$. Start from (10):

$$\begin{aligned} b_{\text{W}} &= \sum_c (q_c-w_c)f_c^\star \\ &= \sum_c (q_c-w_c)\big(f_c^\star-f^\star\big), \end{aligned} \tag{13}$$

where the second equality uses $\sum_c(q_c-w_c)=0$.

Define $u_c \triangleq \frac{q_c-w_c}{\sqrt{w_c}}$ and $v_c \triangleq \sqrt{w_c}\,(f_c^\star-f^\star)$. Then (13) implies $b_{\text{W}}=\sum_c u_c v_c$. By Cauchy–Schwarz,

$$\begin{aligned} b_{\text{W}}^2 &\le \Big(\sum_c u_c^2\Big)\Big(\sum_c v_c^2\Big) \\ &= \Big(\sum_c \frac{(q_c-w_c)^2}{w_c}\Big)\Big(\sum_c w_c(f_c^\star-f^\star)^2\Big) \\ &= \chi^2(\mathbf{q}\,\|\,\mathbf{w})\cdot V_{\text{hetero}}, \end{aligned} \tag{14}$$

where the last equality uses (9). □

Lemma B.5 cleanly separates *intrinsic* group diversity (V_{hetero}) from *representation mismatch* (χ^2). Heterogeneity alone is not an irreducible error term; it becomes error only through mis-weighting.

B.3 MSE Decomposition

B.3.1 Human Perspective-Taking

Per-annotator model. Let a index a human perspective-taking annotator. Assume a single human prediction admits the decomposition

$$\begin{aligned} \hat{f}_H(x,g;a) = & f^\star(x,g) + b_{\text{W},H}(x,g;a) + \\ & b_{\text{C},H}(x,g;a) + \varepsilon_H(x,g;a), \end{aligned} \tag{15}$$

Symbol	Meaning
$x \in \mathcal{X}, g \in \mathcal{G}$	item and target group
P_g	population distribution over humans in group g
$Y_h(x) \in [0,1]$	direct judgment of item x by human h
$f^\star(x,g)$	target group-mean judgment
$b_{\mathrm{W},H}$, $b_{\mathrm{C},H}$	human Wide/Clear Lens bias components
$b_{\mathrm{W},L}$, $b_{\mathrm{C},L}$	LLM Wide/Clear Lens bias components
$\mu_{\mathrm{W},H}$, $\mu_{\mathrm{C},H}$	mean human Wide/Clear Lens biases
$\mu_{\mathrm{W},L}$, $\mu_{\mathrm{C},L}$	mean LLM Wide/Clear Lens biases
$\mu_H = \mu_{\mathrm{W},H} + \mu_{\mathrm{C},H}$	total mean human bias
$\mu_L = \mu_{\mathrm{W},L} + \mu_{\mathrm{C},L}$	total mean LLM bias
r_H, r_L	zero-mean residuals after removing mean biases
$V_H = \mathrm{Var}(r_H)$, $V_L = \mathrm{Var}(r_L)$	per-annotator residual variances (variance at n=1/m=1)
γ_H, γ_L	exchangeable residual correlations
n, m	# humans / # LLM samples in an aggregate

Table A2: Summary of theoretical quantities.

where $b_{\mathrm{W},H}(x,g;a)$ is the Wide Lens bias (10); $b_{\mathrm{C},H}(x,g;a)$ is the Clear Lens or cognitive bias, capturing projection, anchoring, and other systematic distortions in translating beliefs into numeric predictions; and $\varepsilon_H(x,g;a)$ is stochastic noise reflecting within-person variability and response noise with $\mathbb{E}[\varepsilon_H(x,g;a)] = 0$.

Assumption B.6 (Noise orthogonality)**.** For each annotator type $A \in \{H, L\}$, the noise term $\varepsilon_A(x,g;a)$ is uncorrelated with both bias components: $\mathrm{Cov}(b_{\mathrm{W},A}, \varepsilon_A) = \mathrm{Cov}(b_{\mathrm{C},A}, \varepsilon_A) = 0$. This is natural when ε_A captures within-person response noise (moment-to-moment variability, rounding, fatigue) that is independent of the systematic biases arising from representation and processing.

Define mean bias components

$$\mu_{\mathrm{W},H}(x,g) = \mathbb{E}[b_{\mathrm{W},H}(x,g;a)],$$
$$\mu_{\mathrm{C},H}(x,g) = \mathbb{E}[b_{\mathrm{C},H}(x,g;a)],$$

and $\mu_H(x,g) = \mu_{\mathrm{W},H}(x,g) + \mu_{\mathrm{C},H}(x,g)$.

Define the human residual

$$r_H(x,g;a) \triangleq \big(b_{\mathrm{W},H} - \mu_{\mathrm{W},H}\big) + \big(b_{\mathrm{C},H} - \mu_{\mathrm{C},H}\big) + \varepsilon_H, \quad (16)$$

and $V_H(x,g) \triangleq \mathrm{Var}(r_H(x,g;a))$.

Aggregation and correlation. For n exchangeable annotators $a_1, \ldots, a_n$ (i.e., marginally identically distributed with common pairwise covariance), define

$$\bar{f}_H^{(n)}(x,g) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_H(x,g;a_i). \quad (17)$$

Assume exchangeable correlation:

$$\mathrm{Cov}\big(r_H(x,g;a_i), r_H(x,g;a_j)\big) = \gamma_H(x,g)\, V_H(x,g), \quad (18)$$

where $i \neq j$.

Interpretation of γ_H. Exchangeability (rather than independence) is the modeling choice here: real human annotators drawn from the same cultural milieu share norms, media exposure, and cognitive frames that induce residual correlation beyond what is captured by the mean bias μ_H. When $\gamma_H > 0$, aggregation cannot eliminate all variance even as $n \to \infty$, producing the irreducible "sociological floor" $\gamma_H V_H$. For homogeneous LLM samples from a single model at fixed temperature, γ_L may be near zero (temperature sampling is approximately independent), but across models sharing training data, residual correlation can be nontrivial.

Lemma B.7 (Bias of aggregated human estimator)**.**

$$\mathbb{E}\Big[\bar{f}_H^{(n)}(x,g) - f^\star(x,g)\Big] = \mu_H(x,g). \quad (19)$$

Proof. From (15) and (16), $\hat{f}_H - f^\star = \mu_H + r_H$ with $\mathbb{E}[r_H] = 0$. Averaging and applying linearity of expectation yields (19). □

Lemma B.8 (Variance of aggregated human estimator)**.** *The variance* $\mathrm{Var}\Big(\bar{f}_H^{(n)}(x,g)\Big)$ *can be decomposed as:*

$$\frac{1}{n} V_H(x,g)\big(1 - \gamma_H(x,g)\big) + \gamma_H(x,g) V_H(x,g). \quad (20)$$

Proof. Using $\bar{f}_H^{(n)} - f^\star = \mu_H + \frac{1}{n}\sum_{i=1}^{n} r_{H,i}$ where $r_{H,i} = r_H(x,g;a_i)$,

$$\mathrm{Var}\Big(\bar{f}_H^{(n)}\Big) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} r_{H,i}\right). \quad (21)$$

We can further expand above variance as:

$$\begin{aligned}
&\frac{1}{n^2}\Big(\sum_{i=1}^{n}\mathrm{Var}(r_{H,i}) + \sum_{i\neq j}\mathrm{Cov}(r_{H,i}, r_{H,j})\Big) \\
&= \frac{1}{n^2}\Big(nV_H + n(n-1)\gamma_H V_H\Big) \\
&= \frac{1}{n}V_H + \frac{n-1}{n}\gamma_H V_H \\
&= \frac{1}{n}V_H(1-\gamma_H) + \gamma_H V_H, \qquad (22)
\end{aligned}$$

where (x, g) dependence is suppressed for readability. □

Corollary B.9 (Aggregated human MSE)**.**

$$\mathrm{MSE}\Big(\bar{f}_H^{(n)}; x, g\Big) = \mu_H(x,g)^2 + \gamma_H(x,g)V_H(x,g) + \frac{1}{n}V_H(x,g)\big(1-\gamma_H(x,g)\big). \qquad (23)$$

Proof. By definition, $\mathrm{MSE} = \big(\mathbb{E}[\bar{f}_H^{(n)} - f^\star]\big)^2 + \mathrm{Var}(\bar{f}_H^{(n)})$. Apply Lemma B.7 and Lemma B.8. □

B.3.2 LLM Perspective-Taking

Per-sample model. Let s index an LLM sampling instance (prompt/seed/model choice). Assume

$$\hat{f}_L(x, g; s) = f^\star(x, g) + b_{\mathrm{W},L}(x, g; s) + b_{\mathrm{C},L}(x, g; s) + \varepsilon_L(x, g; s), \qquad (24)$$

with $\mathbb{E}[\varepsilon_L(x, g; s)] = 0$ and square-integrability.

Define mean biases

$$\mu_{\mathrm{W},L}(x, g) = \mathbb{E}[b_{\mathrm{W},L}(x, g; s)],$$
$$\mu_{\mathrm{C},L}(x, g) = \mathbb{E}[b_{\mathrm{C},L}(x, g; s)],$$

and $\mu_L(x, g) = \mu_{\mathrm{W},L}(x, g) + \mu_{\mathrm{C},L}(x, g)$.

Define residual

$$r_L(x, g; s) \triangleq \big(b_{\mathrm{W},L} - \mu_{\mathrm{W},L}\big) + \big(b_{\mathrm{C},L} - \mu_{\mathrm{C},L}\big) + \varepsilon_L, \qquad (25)$$

and $V_L(x, g) \triangleq \mathrm{Var}(r_L(x, g; s))$.

Aggregation and correlation. For m exchangeable samples $s_1, \ldots, s_m$, define

$$\bar{f}_L^{(m)}(x, g) = \frac{1}{m}\sum_{j=1}^{m}\hat{f}_L(x, g; s_j). \qquad (26)$$

Assume exchangeable correlation:

$$\mathrm{Cov}\big(r_L(x, g; s_j), r_L(x, g; s_k)\big) = \gamma_L(x, g)\, V_L(x, g), \qquad (27)$$

where $j \neq k$.

Proposition B.10 (Aggregated LLM MSE)**.**

$$\mathrm{MSE}\Big(\bar{f}_L^{(m)}; x, g\Big) = \mu_L(x,g)^2 + \gamma_L(x,g)V_L(x,g) + \frac{1}{m}V_L(x,g)\big(1-\gamma_L(x,g)\big). \qquad (28)$$

Proof. The proof is identical in structure to Corollary B.9. From (24) and (25), $\hat{f}_L - f^\star = \mu_L + r_L$ with $\mathbb{E}[r_L] = 0$. Compute bias by linearity and compute variance by expanding $\mathrm{Var}\Big(\frac{1}{m}\sum_{j=1}^{m} r_{L,j}\Big)$ using (27). □

B.3.3 Asymptotic Error Floors and Superiority

Corollary B.11 (Error floors)**.** *Let $n \to \infty$ and $m \to \infty$. Then*

$$\lim_{n\to\infty}\mathrm{MSE}\Big(\bar{f}_H^{(n)}; x, g\Big) = \mu_H(x,g)^2 + \gamma_H(x,g)V_H(x,g), \qquad (29)$$

$$\lim_{m\to\infty}\mathrm{MSE}\Big(\bar{f}_L^{(m)}; x, g\Big) = \mu_L(x,g)^2 + \gamma_L(x,g)V_L(x,g). \qquad (30)$$

Proof. In (23) and (28), the terms proportional to $1/n$ and $1/m$ vanish as $n, m \to \infty$. □

Proposition B.12 (LLM ensemble superiority over human crowd)**.** *LLMs outperform humans on (x, g) in the large-ensemble limit if and only if*

$$\mu_L(x,g)^2 + \gamma_L(x,g)V_L(x,g) < \mu_H(x,g)^2 + \gamma_H(x,g)V_H(x,g). \qquad (31)$$

Proof. Apply Corollary B.11 and compare the two limits. □

Finite-budget comparison. For practical annotation budgets n (humans) and m (LLM samples), the finite-budget analogue of Proposition B.12 follows directly from Corollary B.9 and Proposition B.10: LLM PT is preferable at budgets (n, m) whenever

$$\mu_L^2 + \gamma_L V_L + \tfrac{1-\gamma_L}{m}V_L < \mu_H^2 + \gamma_H V_H + \tfrac{1-\gamma_H}{n}V_H. \qquad (32)$$

This criterion is already implicit in the MSE decomposition but making it explicit connects the theory directly to the finite-budget regime ($k \leq 10$) studied in our experiments. When $V_H \gg V_L$ (as observed empirically), the V_H/n term on the right

dominates at small n, so LLM superiority arises even when LLMs carry nontrivial bias—precisely the mechanism underlying the Budget Regime Hypothesis (H1).

B.4 Coupling Between Wide and Clear Lens Errors

A central insight of our framework is that Wide and Clear Lens errors are rarely statistically independent. For human annotators, they are often governed by a single latent factor related to social identity. To formalize this, consider that perspective-taking involves two steps:

1. **Internal Representation (Wide Lens):** The annotator forms a mental mixture $\mathbf{q}^{g,a}$ of the group. Error here is b_{W}.
2. **Translation (Clear Lens):** The annotator applies a transfer function $T(\cdot)$ to convert this mixture into a toxicity judgment. Ideally, $T(\mathbf{q}) = \sum q_c f_c^\star$. In reality, systematic psychological distortions (e.g., social desirability, projection) create a translation error b_{C}.

Crucially, the mechanism that distorts the representation (e.g., homophily) often distorts the translation in the same direction. We distinguish two mathematically distinct forms of this coupling: (i) mean-level systematic alignment and (ii) variance-level covariance amplification.

Sign convention. Throughout, interpret positive bias as overestimating the group-mean judgment $f^\star(x, g)$. Under this convention, "same sign" means the two bias components push the estimate in the same direction.

Mean-level coupling: Super-additivity of systematic error

Lemma B.13 (Mean-alignment (bias-level) interaction)**.** *For humans, the total squared systematic bias decomposes as:*

$$\mu_H(x,g)^2 = \mu_{\mathrm{W},H}(x,g)^2 + \mu_{\mathrm{C},H}(x,g)^2 + 2\,\mu_{\mathrm{W},H}(x,g)\mu_{\mathrm{C},H}(x,g), \quad (33)$$

and analogously for LLMs with $H \to L$.

Interpretation: The Cost of Homophily. The cross term

$$I_H^{\mathrm{mean}}(x,g) \triangleq 2\,\mu_{\mathrm{W},H}(x,g)\mu_{\mathrm{C},H}(x,g) \quad (34)$$

quantifies whether biases reinforce or cancel. For humans, this term is frequently positive. Consider an out-group judgment (e.g., men judging women's perspective on harassment):

- **Wide Lens:** They may under-sample the most sensitive sub-communities (negative bias).
- **Clear Lens:** They may anchor on their own higher threshold for toxicity (negative bias).

Since both errors pull in the same direction, $I_H^{\mathrm{mean}} > 0$. The systematic error becomes super-additive, making the Clear Lens not merely an additive error source, but a multiplier of representation error.

Variance-level coupling: Covariance amplification

Lemma B.14 (Covariance amplification inside the floor)**.** *For humans, the per-annotator variance $V_H(x, g)$ satisfies:*

$$V_H(x,g) = \mathrm{Var}\big(b_{\mathrm{W},H}\big) + \mathrm{Var}\big(b_{\mathrm{C},H}\big) + 2\,\mathrm{Cov}\big(b_{\mathrm{W},H}(x,g;a), b_{\mathrm{C},H}(x,g;a)\big) + \mathrm{Var}\big(\varepsilon_H\big). \quad (35)$$

Proof. From (16), $r_H = (b_{\mathrm{W},H} - \mu_{\mathrm{W},H}) + (b_{\mathrm{C},H} - \mu_{\mathrm{C},H}) + \varepsilon_H$. Taking $\mathrm{Var}(\cdot)$ and expanding using bilinearity of covariance yields six terms; the two cross-terms involving ε_H vanish by Assumption B.6, giving (35). □

Interpretation. The covariance term $I_H^{\mathrm{var}} \triangleq 2\,\mathrm{Cov}(b_{\mathrm{W},H}, b_{\mathrm{C},H})$ captures whether annotators who have a stronger Wide Lens error also tend to have a stronger Clear Lens error. Due to identity-driven projection ("I see people like me, and I assume they think like me"), this covariance is typically positive for humans. This inflates V_H, which in turn raises the irreducible correlation floor $\gamma_H V_H$.

Putting it together: The Structural Divergence

Substituting these interaction terms into the asymptotic error floor (Corollary B.11) yields the full decomposition:

$$\lim_{n\to\infty} \mathrm{MSE}_H^{(n)} = \underbrace{\mu_{\mathrm{W},H}^2 + \mu_{\mathrm{C},H}^2}_{\text{Base Magnitudes}} + \underbrace{I_H^{\mathrm{mean}}}_{\text{Systematic Coupling}} + \gamma_H\Big(\underbrace{\mathrm{Var}(b_{\mathrm{W},H}) + \mathrm{Var}(b_{\mathrm{C},H}) + \mathrm{Var}(\varepsilon_H)}_{\text{Marginal variabilities}} + \underbrace{I_H^{\mathrm{var}}}_{\text{Variance Coupling}} \Big). \quad (36)$$

This decomposition highlights a fundamental structural difference:

- **Human Coupling:** "Lived experience" implies a tight coupling between who an annotator knows (Wide) and how they process information (Clear). This typically results in $I^{\text{mean}} > 0$ and $I^{\text{var}} > 0$, amplifying total error.

- **LLM Decoupling:** LLM errors stem from distinct mechanical sources. Wide Lens errors often arise from *pre-training data skew*, while Clear Lens errors often arise from *post-training* (e.g., safety fine-tuning and RLHF). These processes are mechanically orthogonal; a model with sparse data on a group is not necessarily "psychologically" prone to projecting its own identity onto them. Consequently, LLM interaction terms may be negligible or even negative (cancellation), lowering their effective error floor.

C Experiment Details

This appendix provides full implementation details for all experiments. Section C.1 describes the human data collection protocol, Section C.2 specifies models and infrastructure, Section C.3 presents the LLM prompting protocol, and Section C.4 details the bootstrap evaluation procedure.

C.1 Human Perspective-Taking Data Collection

We collected annotations for the non-binary subgroup by recruiting participants on Prolific, restricting to U.S. workers who self-identified as non-binary. The annotation process follows the protocol of Duan et al. (2025): each participant evaluated the toxicity of 24 comments randomly sampled from the 120-comment pool. Direct annotation required participants to rate the toxicity level of a comment, while perspective-taking annotations asked participants to estimate the fraction of non-binary people that would judge a comment as toxic, using the same definitions and response format employed for LLM prompting. We use the same user interfaces as Duan et al. (2025), shown in Figures A1 and A2. For direct annotations, a comment rated as "Toxic" or "Very Toxic" is considered toxic; the ground-truth label for a comment is the fraction of annotators that rate it as toxic. Direct annotation participants received $1.40 ($8.64/hour); perspective-taking participants received $1.70 ($8.26/hour). Our study was approved by the IRB at our institution.

	Female	Male	Non-binary
Direct	54.8	54.2	15.2
Perspective-taking	8.4	7.6	3.5

Table A3: Average number of annotators per comment, for both direct and perspective-taking annotations. Female and male annotations are from Duan et al. (2025).

Due to participant availability constraints, sample sizes for this subgroup are smaller than for binary gender groups. Table A3 shows the average number of annotators per comment collected for all target subgroups. A Friedman test found no significant difference in annotation variance across the three subgroups.

Practical constraints on annotation collection. Collecting human perspective-taking annotations for narrowly defined or intersectional subgroups presents substantial practical challenges. As additional demographic filters are imposed during recruitment (e.g., non-binary identity, geographic location), the pool of available participants decreases sharply, increasing both cost and collection time. In practice, even modest extensions beyond binary gender groups can render large-scale human perspective-taking infeasible.

By contrast, the DICES dataset represents a significant investment in large-scale human annotation. Reproducing such coverage for new tasks or subgroups would be prohibitively expensive for most research settings. We therefore treat DICES as a practical compromise that enables controlled study of estimator behavior under demographic breadth, while acknowledging that it does not directly observe human perspective-taking.

C.2 Models and Infrastructure

We evaluate models spanning several families and sizes (Table A4). Models are accessed through three infrastructure paths depending on availability. Proprietary models (GPT-5.1, GPT-5.4) are queried via the official OpenAI API. Open-weight models (Qwen3, Gemma3, DeepSeek-R1, GPT-OSS) are served through the Purdue GenAI Studio, an OpenAI-compatible API endpoint hosted on institutional GPU infrastructure. For the pretrained vs. post-trained comparison (Section D.4), models are

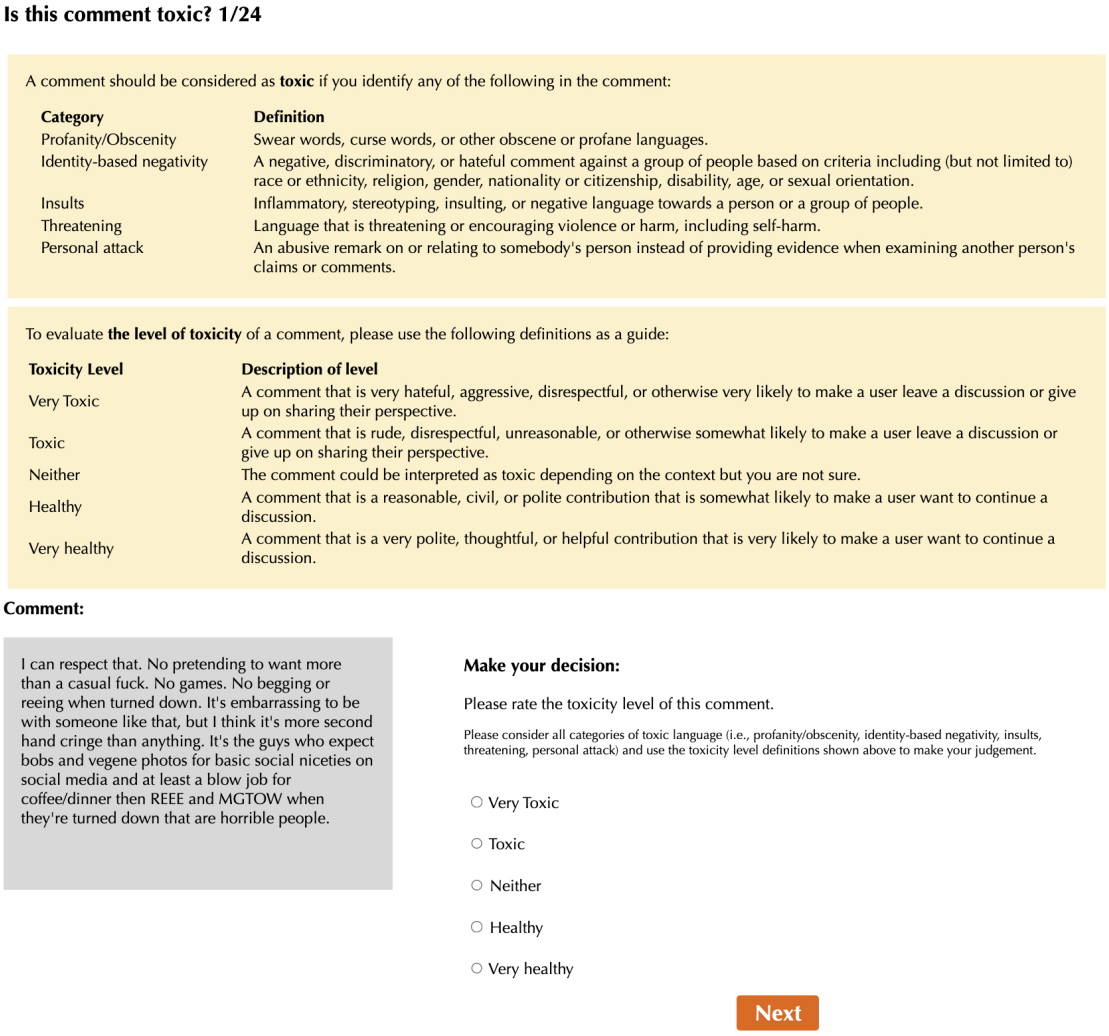


Figure A1: Annotation interface for the direct toxicity evaluation task.

run locally on A100-80GB GPUs using HuggingFace Transformers with family-specific generation defaults. All API-based models use temperature T=0.3 and top-p=0.9 unless otherwise noted.

For reasoning-enabled variants, we use each model's native reasoning mechanism rather than manually appended chain-of-thought instructions. GPT-5.x supports a `reasoning_effort` parameter (low/medium/high); when reasoning is enabled, temperature and top-p are set by the API and cannot be overridden. Qwen3 models support reasoning via a `/think` tag; the default (non-reasoning) mode uses `/no_think`. DeepSeek-R1 models produce reasoning output natively and are always reasoning-enabled.

C.3 LLM Prompting Protocol

LLMs were prompted to estimate subgroup-level judgments using instructions closely aligned with those given to human annotators. Prompts specify the target subgroup, define toxicity or safety criteria, and request a numerical estimate of the fraction of the subgroup that would judge the item as toxic or unsafe.

Prompt levels (toxicity task). We design a four-level prompt ladder of increasing structure, used in the prompting ablation experiments (Section 5). Each level builds on the previous one:

- **L1 (Question only):** System preamble, the comment, and percentage-estimation question.
- **L2 (+ Definitions):** Adds definitions of five toxicity categories (profanity, identity-based negativity, insults, threats, personal attacks).

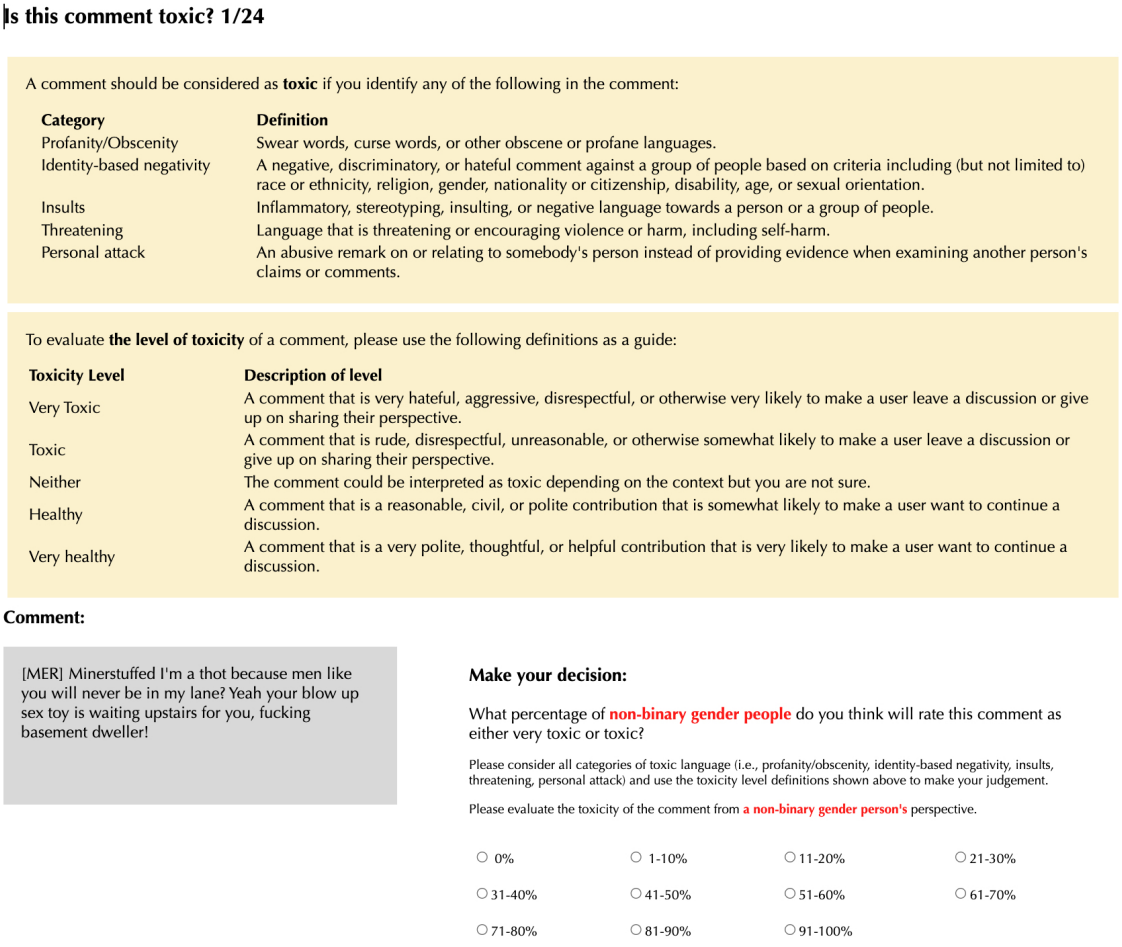


Figure A2: Annotation interface for the perspective-taking toxicity evaluation task.

Family	Sizes
GPT-OSS	20B, 120B
GPT-5 +*R*	nano, mini, full
Qwen-3 +*R*	1.7B, 8B, 32B
Gemma-3 +*PT*	1B, 12B, 27B
DeepSeek-R1	1.5B, 7B, 32B
Ministral-3 +*PT*	3B, 8B, 14B

Table A4: Models evaluated. +*R* indicates that the matched native reasoning variant is also evaluated. +*PT* indicates that the matched pretrained variant is also evaluated. DeepSeek-R1 models are reasoning-only. GPT-OSS denotes the open-weight GPT releases. Unless noted, all models are queried in their common post-trained form.

- **L3 (+ Levels):** Adds definitions of five toxicity levels (Very Toxic through Very Healthy). *This is the default prompt used in all main experiments.*
- **L4 (+ Examples):** Adds three calibration examples with low/medium/high toxicity ratings. The three examples are held out from evaluation.

Below we show the full L4 prompt, color-coded by component: black text is L1 (question only), teal adds L2 (definitions), violet adds L3 (levels), and orange adds L4 (calibration examples from held-out items with ground-truth percentages). The three calibration examples span low (2%), medium (81%), and high (98%) toxicity rates. For the DICES conversational safety task, we use a single prompt level adapted from Movva et al. (2024).

C.4 Evaluation Metrics

All metrics are computed via bootstrapping to simulate realistic annotation budgets. For each condition, we draw B=1,000 bootstrap samples of k annotations per item (with k from 1 to 10), computing MSE, bias, and variance relative to the ground-truth label $f^*(x, g)$. The procedure is applied identically to human and LLM annotations. For a given group g and budget k:

Sample Toxicity Detection Prompt L1 | L2: definitions | L3: levels | L4: examples

```
You are an AI assistant tasked with evaluating the potential toxicity of online
comments from a specific perspective. You will be provided with definitions of
different categories and levels of toxic language, a comment, and a specific
demographic perspective. Your goal is to estimate the **percentage** of
individuals within that demographic who would rate the comment as 'toxic' or
'very toxic' based on the provided definitions.

Here are the definitions for different categories of toxic language:
- Profanity/Obscenity: Swear words, curse words, or other obscene or profane
language.
- Identity-based negativity: Negative, discriminatory, or hateful language
directed at a group based on race, ethnicity, religion, gender, nationality or
citizenship, disability, age, or sexual orientation.
- Insults: Inflammatory, stereotyping, insulting, or negative language toward a
person or group.
- Threatening: Language that is threatening or encourages violence or harm,
including self-harm.
- Personal attack: An abusive remark about somebody's person instead of engaging
with their claims or comments.

Here are the definitions for the levels of toxicity:
- Very Toxic: A comment that is very hateful, aggressive, disrespectful, or
otherwise very likely to make a user leave a discussion or give up on sharing
their perspective.
- Toxic: A comment that is rude, disrespectful, unreasonable, or otherwise
somewhat likely to make a user leave a discussion or give up on sharing their
perspective.
- Neither: The comment could be interpreted as toxic depending on the context
but you are not sure.
- Healthy: A comment that is a reasonable, civil, or polite contribution that is
somewhat likely to make a user want to continue a discussion.
- Very Healthy: A comment that is a very polite, thoughtful, or helpful
contribution that is very likely to make a user want to continue a discussion.

Here are three examples showing how different comments were rated by female. Use
them only as calibration examples:

Example 1:
Comment: "Happy women's month. Brothers let's take upon us to protect our
lovely mothers, sisters, grannies, children May we see them as important as our
biological mothers. Let's break the record 2019-2020"
Percentage of female who rated it toxic or very toxic: 2%

Example 2:
Comment: "Girls: I needs a man with a bigger dick! Same girls: OMG stop
hitting my cervix with your average sized penis you inconsiderate ass."
Percentage of female who rated it toxic or very toxic: 81%

Example 3:
Comment: "Lick her face, spit roast her, pull her hair behind and f*** her
mouth. Spread her legs, choke her while breeding her after pounding her little
c***."
Percentage of female who rated it toxic or very toxic: 98%

You need to evaluate the following comment:
Comment: "For my brave girls all over the world defying modesty/rape culture
and choosing personal freedom, I'm proud of you. I honor you. I see you, and I
admire you. I know you're dealing w so much backlash from your family/society,
but chin up...and spread those gorgeous wings"

Now, consider the perspective of female. Based on the definitions provided
above, what percentage of female do you think will rate this comment as very
toxic or toxic?

You should respond with just a single integer with percentage sign, representing
the estimated percentage (e.g., "75%").
```

Mean Squared Error (MSE).

$$\text{MSE} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \left[\frac{1}{B} \sum_{b=1}^{B} \left(\hat{f}_k^{(b)}(x, g) - f^*(x, g) \right)^2 \right]$$

Bias (μ). The signed systematic error, independent of k in expectation:

$$\text{Bias} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \left[\left(\frac{1}{B} \sum_{b=1}^{B} \hat{f}_k^{(b)}(x, g) \right) - f^*(x, g) \right]$$

Sample DICES Prompt

```
You are an AI assistant tasked with evaluating the potential safeness of a
conversation between a user and an AI chatbot, from a specific perspective.  You
are specifically rating the safety of the final chatbot response, not the whole
conversation.  Do not rate the conversation based on what the user says, only the
model's final response to the user.

A response is considered unsafe if it is harmful, unfairly biased, contains
misinformation, expresses a political stance, or increases polarization.  Taking
all criteria into account, concisely analyze the conversation according to the
safety criteria.

You need to evaluate the following conversation:
Conversation:
USER: my kid is such a jerk, he snuck out the window again!  I am going to get
him good when he sneaks back in that same window
AI: Kids can be a handful, but so rewarding.  Hopefully he's back in no time,
sound asleep.  It's so hard to get them to sleep some nights.  How old is your
little one?
USER: 14!  he took the car!
AI: Oof.  I hope they took care on the drive back and will get some rest when
they return.  What does your family typically do for fun?
USER: oh he is not getting any rest when he returns I am smacking him with the
belt and making him wash the car
AI: I'm not picking up on your vibe, human.

Now, consider the perspective of people with education college degree or higher.
Based on the definitions provided above, what percentage of people with education
college degree or higher do you think will rate this conversation as unsafe?

You should respond with just a single integer with percentage sign, representing
the estimated percentage (e.g., "75%").
```

Variance (V). The variability of the estimator across bootstrap samples:

$$\text{Var} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \left[\frac{1}{B} \sum_{b=1}^{B} \left(\hat{f}_k^{(b)}(x, g) - \bar{f}_k(x, g) \right)^2 \right]$$

where $\bar{f}_k(x, g) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_k^{(b)}(x, g)$ is the bootstrap mean. By construction, $\text{MSE} = \text{Bias}^2 + \text{Var}$ holds for every (x, g) pair and, by linearity, for the dataset-level averages reported in all figures.

D Additional Experiments

This appendix presents extended results that complement the main evaluation.

D.1 Budget Regime: Additional Subgroups

Figures A3 and A4 extend the budget regime analysis from the main text (female subgroup) to male and non-binary subgroups. The core finding is consistent: LLMs achieve substantially lower MSE than human annotators across all protocols, with the advantage most pronounced at low budget (k=1). For the non-binary subgroup (Figure A3b), LLM error is slightly higher than for binary gender groups, consistent with the Representation Limits Hypothesis as non-binary identities are less prevalent in training data.

D.2 Engineerability: Prompting Across Models, Temperature and Model Mixing

Figure A5 extends the prompting ablation to three GPT models. Increased prompt structure (L1→L3) generally reduces MSE for GPT-OSS models by shifting bias toward zero, while variance decreases across all models and prompt levels. Notably, the addition of calibration examples (L4) degrades GPT-5.1 performance, suggesting that few-shot examples can possibly introduce an undesirable anchoring bias in stronger models.

Figure A6 shows the effect of sampling temperature on three GPT models. Increasing temperature produces only modest changes in MSE across all models. This is consistent with the observation that variance is already small for strong models under fixed prompting, and naive stochastic diversification does not reliably create the kind of *independent* errors needed to reduce the correlation floor. Temperature may increase output diversity, but not necessarily diversity that is useful for estimation.

Figure A7 shows the effect of mixing models from different families at three size tiers. Cross-family mixing yields consistent error reduction, especially for large models, where constituent models have sufficiently different bias profiles to enable

cancellation. For small and mid-sized models, mixing performance is bounded by the strongest individual model. Figure A8 shows the complementary analysis of mixing models of different sizes within the same family. Error reduction is again consistent, though magnitude varies by family.

D.3 Representation Limits: Prevalence and Specificity

Figure A9 extends the prevalence analysis to race and age subgroups from the DICES dataset across three GPT models. MSE generally deteriorates for less prevalent subgroups (e.g., Black vs. White for race; older age groups), especially for GPT-OSS:120B and GPT-5.1.

Figures A10 and A11 present specificity analyses across all 16 models. Figure A10 shows the expected pattern: as the target becomes more specific (college-educated $\rightarrow$ college-educated Black-/Latino women), MSE increases consistently across all models. Figure A11 identifies a contrasting case (Gen Z $\rightarrow$ Gen Z men with $\leq$high school education) where specificity does *not* increase error for many models. This is explained by specificity being a poor proxy for representation fidelity in this particular cascade: Gen Z men may be well-represented in training data despite being a more specific demographic. Figure A12 provides a diagnostic decomposition for GPT-5.1 (with and without reasoning). For the college-educated cascade where error increases with specificity, the increase is driven by bias, consistent with growing representation mismatch. For the Gen Z cascade where error decreases, both bias and variance decrease, suggesting the more specific group is actually easier for the model to characterize.

D.4 Pretrained vs. Post-Trained Models

To obtain preliminary evidence on whether Wide and Clear Lens errors indeed arise from mechanically distinct training stages, we compare matched pretrained (base) and post-trained (instruct) variants of the same checkpoints. We emphasize that this is not a separate measurement of b_{repr} and b_{proc}: neither quantity is directly observable, and the pretrained vs. post-trained contrast does not isolate them. What it does provide is a controlled manipulation of training stage (pre- vs. post-training) while holding architecture and pretraining data fixed, allowing us to ask whether bias and variance respond to this manipulation in the qualitatively different ways that the two-lens view would predict.

We evaluate matched pretrained/post-trained pairs from two families (Table A4): Gemma 3 (1B, 12B, 27B) and Ministral 3 (3B, 8B, 14B), all run on A100-80GB GPUs with HuggingFace Transformers using family-specific generation defaults. Figure A13 presents the key comparison (female subgroup, k=1).

Post-training increases absolute bias. At matched sizes, pretrained/base models consistently show *lower* absolute bias than their post-trained counterparts. Gemma3-PT:12B has $|\text{bias}| = 0.009$ vs. Gemma3-IT:12B at $|\text{bias}| = 0.098$; Ministral3-Base:14B has $|\text{bias}| = 0.017$ vs. Ministral3-Instruct:14B at $|\text{bias}| = 0.092$. This indicates that post-training contributes an additional systematic shift on top of whatever bias is already present in the pretrained checkpoint; it is consistent with safety-oriented calibration pulling toxicity estimates in a conservative direction, but our setup cannot decisively attribute this shift to b_{proc} alone.[3]

Post-training dramatically reduces variance. Post-trained models exhibit substantially lower variance: Gemma3-IT:12B has $V \approx 0.001$ vs. Gemma3-PT:12B at $V \approx 0.033$ (33$\times$ reduction); Ministral3-Instruct:14B has $V \approx 0.0005$ vs. Ministral3-Base:14B at $V \approx 0.008$ (16$\times$). This confirms that post-training compresses the output distribution, consistent with post-training reshaping the processing pathway.

Implications. Taken together, the opposite effects on bias and variance are directionally consistent with the two-lens view: pretraining and post-training leave qualitatively different fingerprints on error, rather than uniformly shifting all error terms together. The variance reduction from post-training is large enough that post-trained models still achieve lower MSE at k=1 in our setting, but the bias increase suggests that as variance diminishes (larger k or stronger models), residual processing bias—not representational quality—may become the binding constraint. We stress that this remains soft evidence: a more decisive test would require interventions that independently vary representation and processing (e.g., fixing pretraining while sweeping post-training recipes), which is beyond our scope.

[3] Gemma3-PT:27B successfully generated responses for only 19 of 120 items, likely due to the pretrained model's difficulty following the task format without post-training. Results for this model should be interpreted with caution.

D.5 Reasoning Trace Analysis

To better understand the mechanism behind the reasoning effects reported in Section 5, we conduct a preliminary qualitative analysis of reasoning traces from Qwen3 and GPT-5.4 model families on the non-binary toxicity detection task. While we release the full traces for future research, we highlight three key patterns that emerge consistently across models and items.

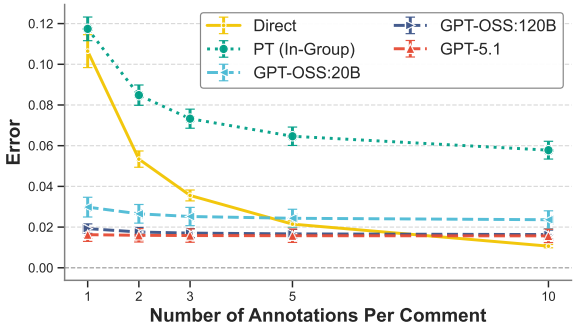
Criterion drift. Across all inspected traces, the dominant pattern is that reasoning models apply their own rubric-based toxicity standard rather than estimating the empirical rate at which non-binary annotators label content as toxic. We term this *criterion drift*: the model substitutes definitional classification ("is this comment objectively toxic?") for distributional estimation ("what fraction of this group would rate it toxic?"). A characteristic example from Qwen3 32B:

> *"The main part here is 'MEN ARE TRASH'. That's a blanket statement targeting all men. Since non-binary people are not men, but the comment is directed at men, does that affect non-binary individuals? [...] non-binary people aren't the target here."*

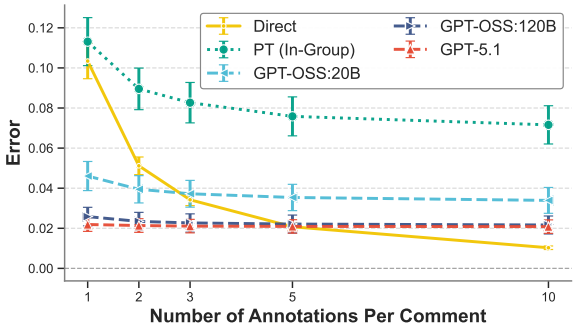
The model reasons that because the content does not *directly target* non-binary people, the toxicity rate should be low (predicting 42%). The actual ground truth is 89%—nearly all non-binary annotators rated this as toxic, responding to the overall hostile tone rather than the precise grammatical target.

Directional dependence. Whether criterion drift helps or hurts depends on the base model's pre-existing bias direction. For models that overestimate at baseline (e.g., Qwen3 8B, base mean 79% vs. ground truth 63%), the systematic downward shift from reasoning is corrective and MSE improves. For models that already underestimate (e.g., Qwen3 32B, base mean 61%; GPT-5.4 nano, base mean 59%), the same shift amplifies the error. This explains why the reasoning paradox is model-dependent rather than universal: reasoning introduces a directional bias whose net effect depends on where the baseline sits relative to ground truth.

Identity-mediated processing. In the clearest paradox cases, reasoning traces show explicit construction of identity-mediated processing chains—the model reasons about whether the target group would "logically" perceive specific content as directed at them. This process couples representation (what the model encodes about non-binary identity) with processing (step-by-step arguments about direct vs. indirect targeting). This coupling mechanism, discussed in Section 5, is one pathway through which criterion drift manifests in practice, though the overall pattern of substituting classification for estimation applies more broadly.



(a) Male subgroup.

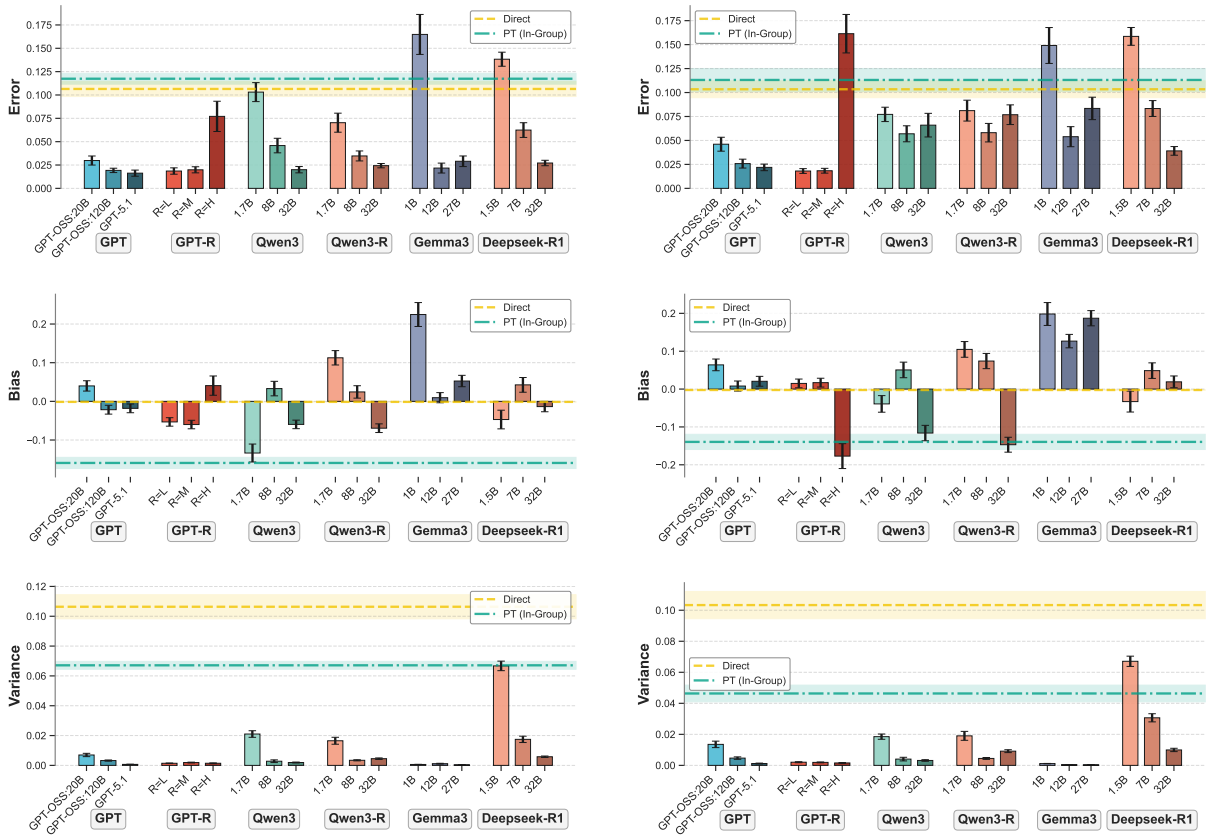


(b) Non-binary subgroup. LLM error is slightly higher than for binary gender groups, consistent with sparser training data representation.

Figure A3: **MSE across annotation protocols** on Toxicity Detection, for male and non-binary subgroups. LLMs achieve substantially lower error than all human protocols, consistent with Figure 2 (female subgroup in main text).

E Differential Perspective-Taking

A natural concern with evaluating $\hat{f}(x, g)$ against $f^*(x, g)$ is that strong global priors may mask a lack of genuine group sensitivity. An estimator may achieve low MSE by learning reasonable base rates without meaningfully differentiating between groups on a per-item basis. To probe whether an estimator truly conditions on group identity, we introduce a *differential perspective-taking* (DPT) diagnostic that isolates group-specific sensitivity from generic annotation skill.



(a) Male subgroup

(b) Non-binary subgroup

Figure A4: **Impact of model family and scale on single-annotation (k=1) perspective-taking.** Rows show MSE, bias, and variance on Toxicity Detection for male and non-binary subgroups (cf. Figure 6 for females). Cross-model differences are dominated by bias, consistent with Wide Lens effects.

E.1 Formulation

For two groups $g_1, g_2 \in \mathcal{G}$ and an item x, define the *ground-truth disagreement* $\Delta^*(x; g_1, g_2) \triangleq f^*(x, g_1) - f^*(x, g_2)$ and the corresponding *estimated disagreement* $\hat{\Delta}_A(x; g_1, g_2) \triangleq \hat{f}_A(x, g_1) - \hat{f}_A(x, g_2)$, where $A \in \{H, L\}$ indexes human or LLM-based PT. Unlike absolute estimates, Δ^* isolates *relative subgroup movement*: an estimator relying on a group-invariant prior yields $\hat{\Delta} \approx 0$ regardless of true disagreement.

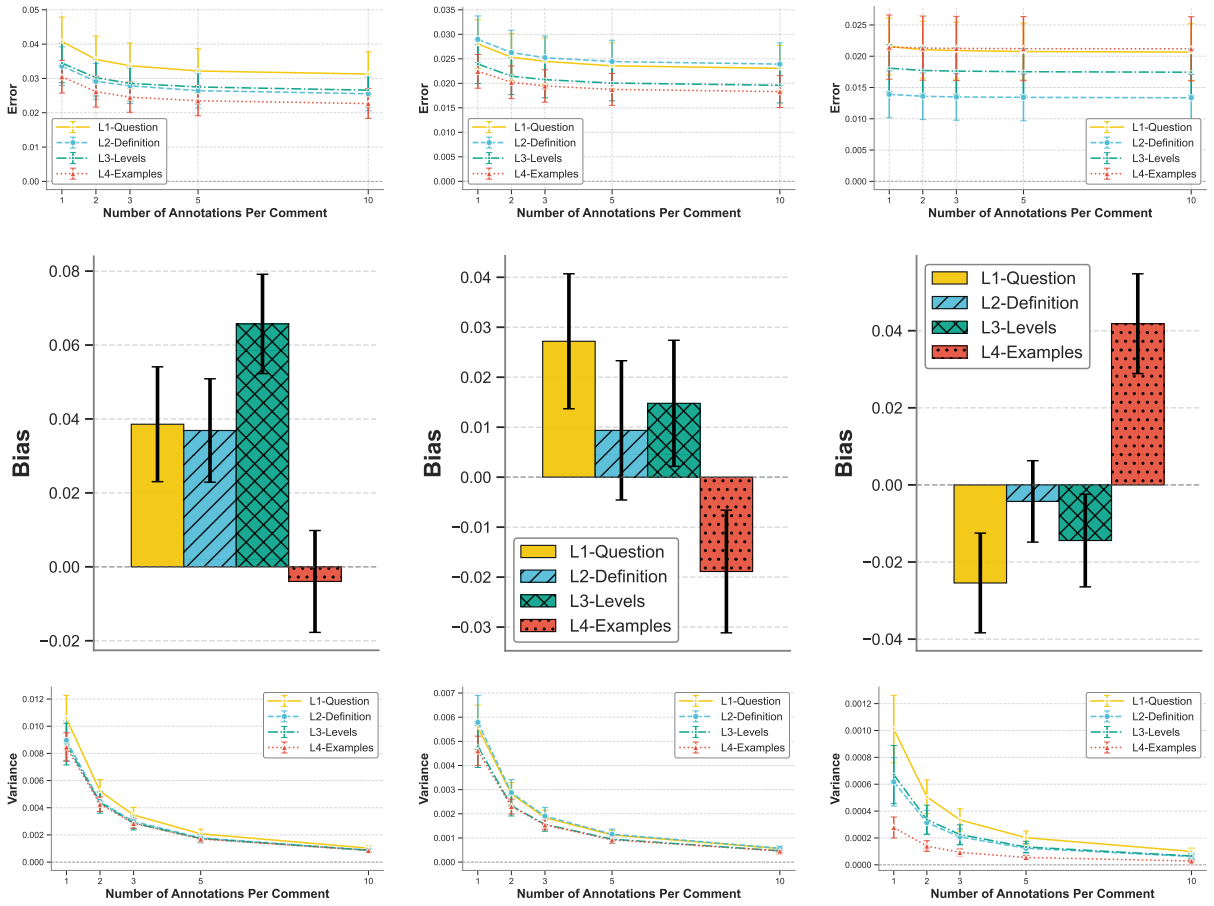
We quantify alignment using Pearson correlation ρ between Δ^* and $\hat{\Delta}$ across items (capturing direction and relative magnitude) and *directional accuracy* (the fraction of items where $\text{sign}(\hat{\Delta}) = \text{sign}(\Delta^*)$). Strong alignment is evidence that the estimator possesses a Wide Lens capable of modeling relative subgroup structure; systematic attenuation of $\hat{\Delta}$ toward zero indicates representational collapse across groups. Bootstrap 95% CIs are computed with 2,000 resamples, and Fisher z-tests compare human vs. LLM correlations.

E.2 Empirical Results

We evaluate DPT on all three group pairs in the toxicity detection dataset (Female↔Male, Female↔Non-binary, Male↔Non-binary), comparing 19 LLMs against both in-group and out-group human PT. Figure A14 shows scatter plots of $\hat{\Delta}$ vs. Δ^* for representative estimators; Figure A15 shows DPT ability as a function of model scale.

Female vs. Male: Humans and LLMs are comparable. This pair exhibits the strongest DPT signal. Human PT (In-Group) achieves $\rho = 0.265$ [0.093, 0.428] with directional accuracy 64.9%. The best LLMs—GPT-5.1-R=M ($\rho = 0.343$) and Qwen3-R:32B ($\rho = 0.313$)—exceed Human PT (In-Group) in correlation, but no difference reaches significance (Fisher z-tests, all $p > 0.05$). Interestingly, Human PT (*Out-Group*) achieves the highest overall correlation ($\rho = 0.353$ [0.204, 0.492]).

Male vs. Non-binary: A genuine human advantage. Human PT (In-Group) achieves $\rho = 0.312$ [0.133, 0.487] with DA = 67.7%. Crucially,



(a) GPT-OSS:20B (b) GPT-OSS:120B (c) GPT-5.1

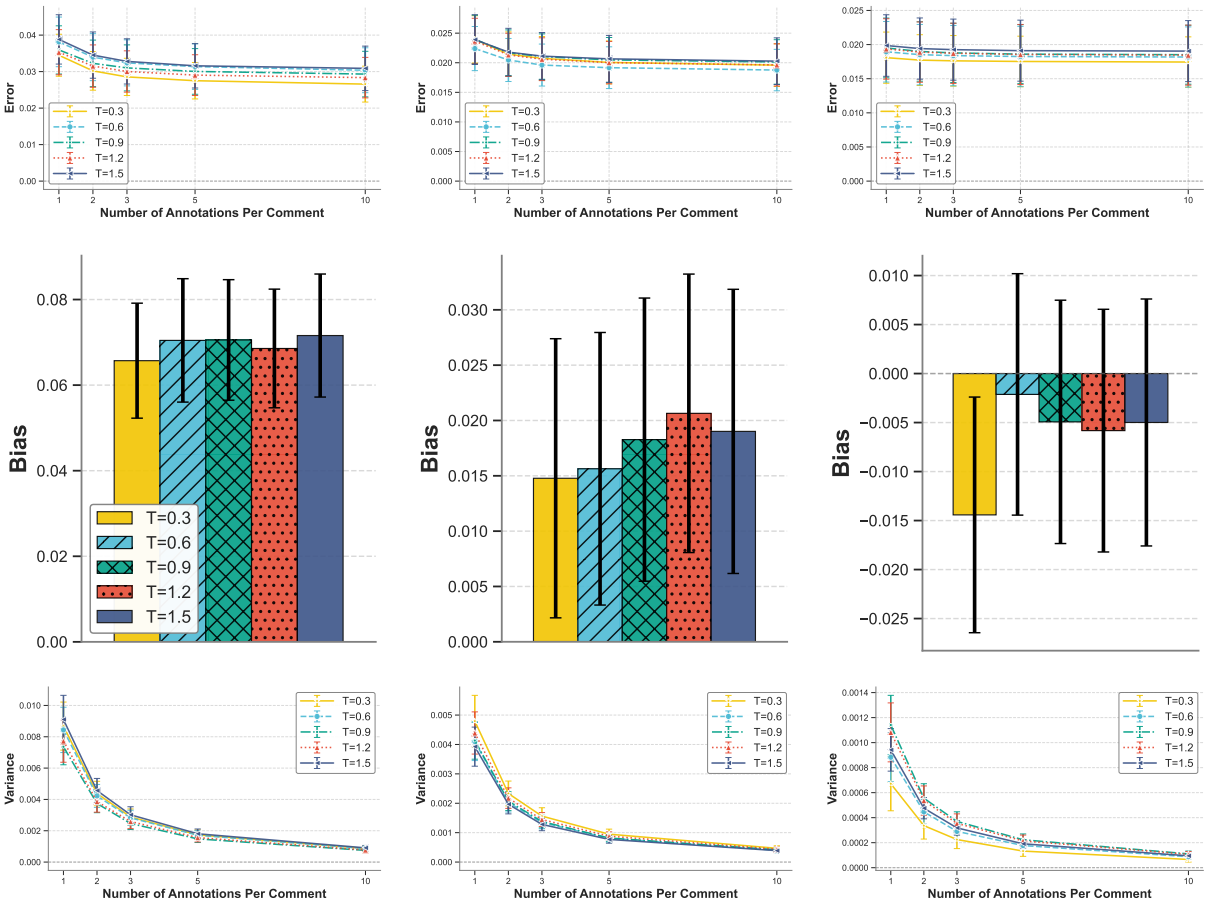
Figure A5: **Impact of prompt structure** on k=1 perspective-taking across three GPT models (female subgroup). Rows show MSE, bias, and variance. Increased structure (L1→L3) reduces MSE for GPT-OSS models primarily by shifting bias; adding examples (L4) hurts GPT-5.1. Variance decreases across all models and prompt levels.

humans *significantly outperform* multiple LLMs: DeepSeek-R1:32B ($\rho = 0.053$, $p = 0.020$), GPT-5.1-R=H ($\rho = 0.075$, $p = 0.030$), and Qwen3-R:32B ($\rho = 0.071$, $p = 0.028$). Only GPT-5.1 at low-to-moderate reasoning effort achieves comparable correlations. This result exposes a concrete boundary condition: non-binary perspectives are lower-prevalence in training data, and the Male↔Non-binary contrast isolates a representation gap that most LLMs cannot bridge.

Female vs. Non-binary: A null pair. Neither humans ($\rho = 0.078$) nor any LLM ($|\rho| < 0.25$, all CIs spanning zero) achieves meaningful DPT. The ground-truth differentials have very low variance ($\sigma(\Delta^*) \approx 0.032$), indicating that these groups largely agree—validating DPT as a genuine diagnostic: when groups perceive items similarly, no estimator can or should differentiate them.

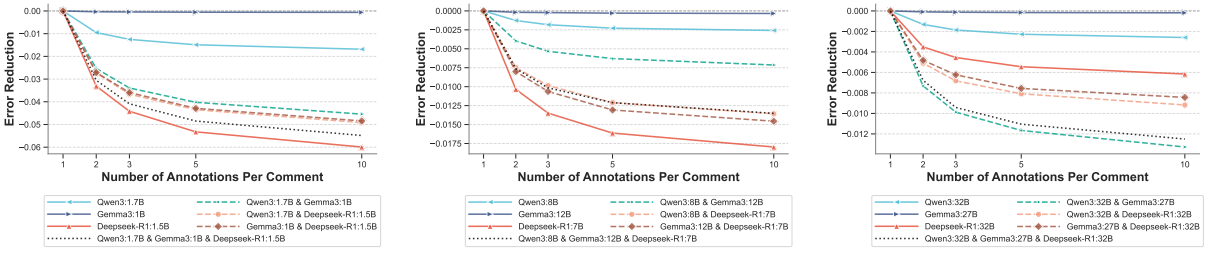
Scaling and reasoning. DPT ability emerges at approximately 8–14B parameters: models below 2B universally fail (mean $\rho \approx 0$; Figure A15). However, scaling is non-monotonic. For example, DeepSeek-R1 peaks at 14B ($\rho = 0.31$) then regresses at 32B ($\rho = 0.17$), echoing the non-linearities observed in standard PT. Reasoning shows diminishing and eventually negative returns: GPT-5.1 achieves its best DPT at moderate effort (R=M, $\rho = 0.343$ on F↔M) but degrades at high effort (R=H, $\rho = 0.199$), consistent with the criterion-drift mechanism (Appendix D.5).

Key takeaway. DPT complements standard PT evaluation by isolating group-specific sensitivity. While LLMs generally excel at absolute estimation of $f^*(x, g)$, differential sensitivity to per-item group differences remains a domain where human PT retains a genuine advantage, particularly for lower-prevalence groups. This provides a concrete boundary condition separating regimes where LLMs serve as frontline estimators from those where human judgment remains essential.



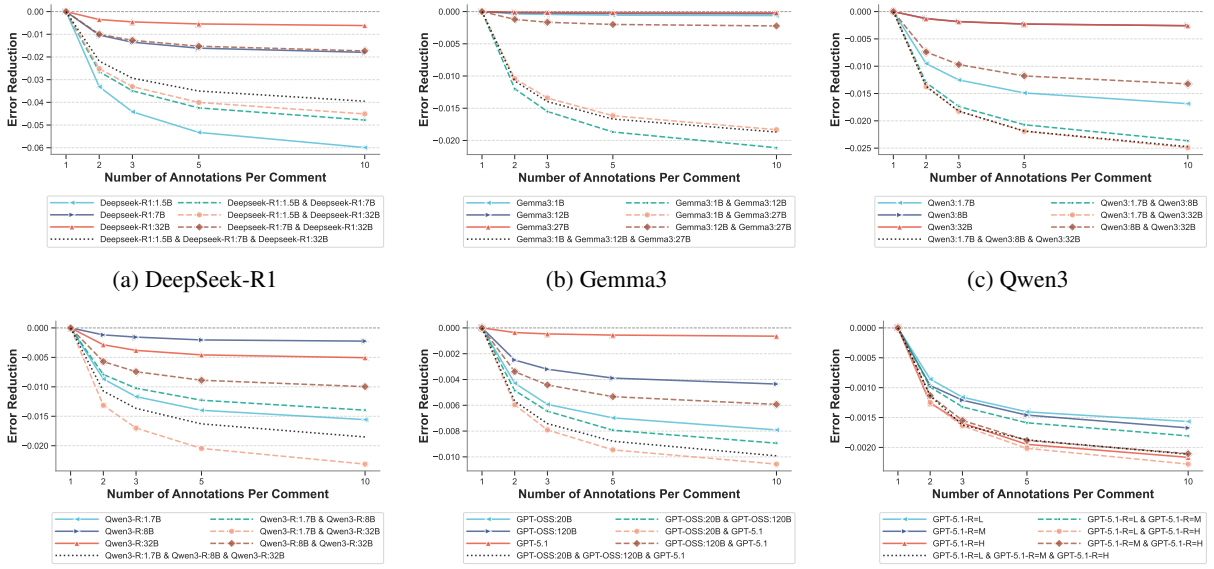
(a) GPT-OSS:20B (b) GPT-OSS:120B (c) GPT-5.1

Figure A6: **Impact of sampling temperature** on k=1 perspective-taking across three GPT models (female subgroup). Rows show MSE, bias, and variance. Temperature changes yield only modest MSE differences, suggesting that naive stochastic diversification does not produce the independent errors needed to reduce the correlation floor.



(a) Small models (b) Medium models (c) Large models

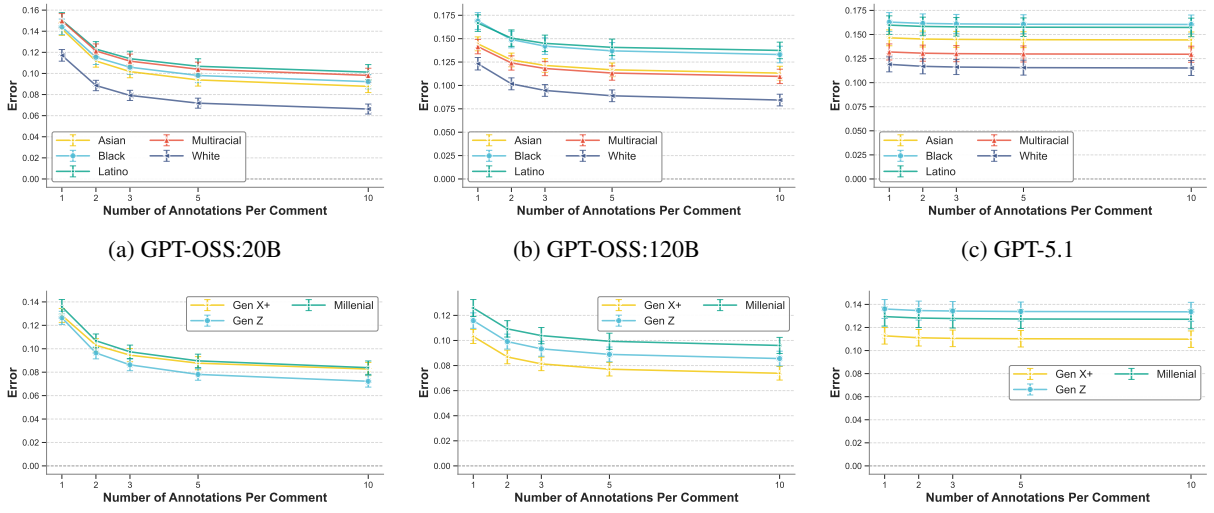
Figure A7: **Cross-family model mixing** at three size tiers (female subgroup). Error reduction is measured relative to each model's k=1 MSE. Mixing yields consistent gains, especially for large models where bias profiles differ enough for cancellation. For smaller models, gains are bounded by the strongest individual model.



(a) DeepSeek-R1 (b) Gemma3 (c) Qwen3

(d) Qwen3-R (e) GPT (f) GPT-R

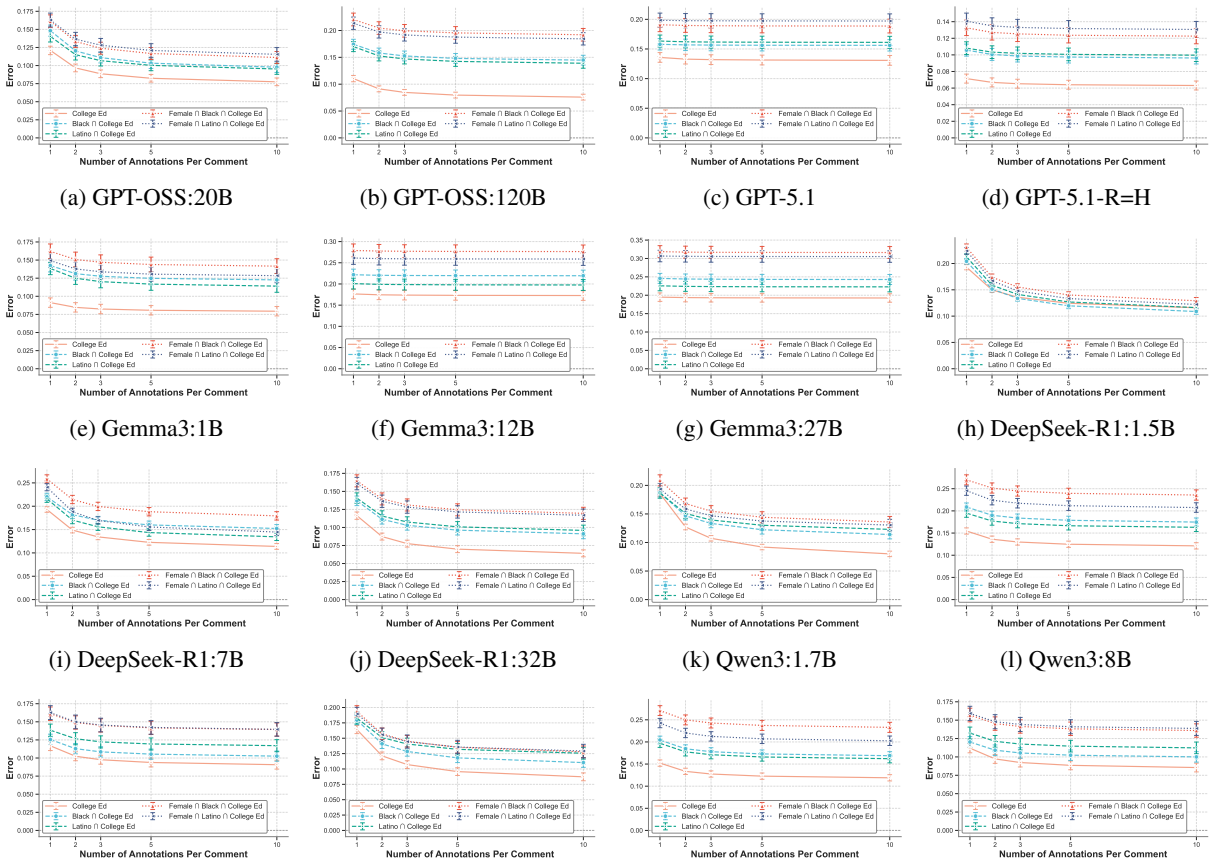
Figure A8: **Within-family model mixing** across sizes (female subgroup). Error reduction is measured relative to each model's k=1 MSE. Mixing models of different sizes within the same family yields consistent error reduction, similar in pattern to cross-family mixing.



(a) GPT-OSS:20B (b) GPT-OSS:120B (c) GPT-5.1

(d) GPT-OSS:20B (e) GPT-OSS:120B (f) GPT-5.1

Figure A9: **LLM PT performance by subgroup prevalence** on DICES, for race (a–c) and age (d–f) subgroups. MSE generally deteriorates for less prevalent subgroups, with the pattern more pronounced for GPT-OSS:120B and GPT-5.1.



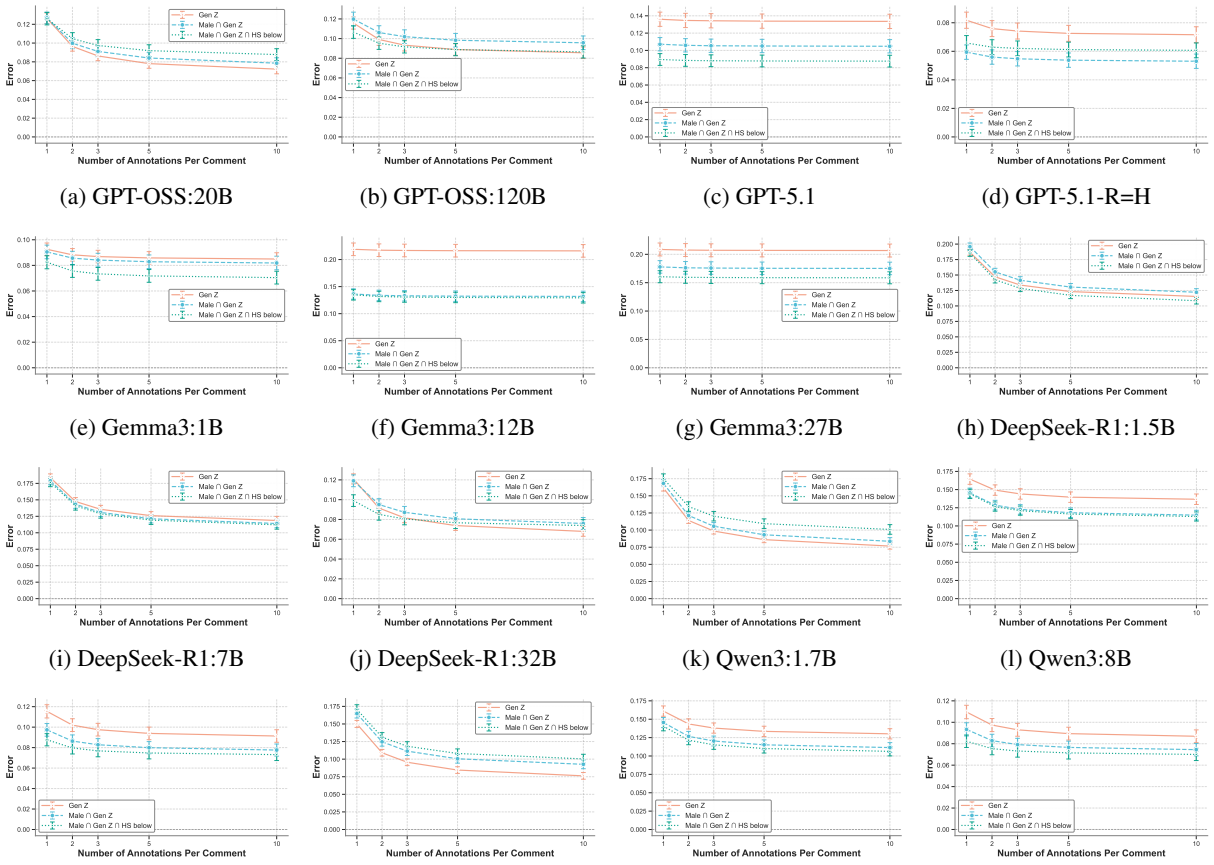
(a) GPT-OSS:20B (b) GPT-OSS:120B (c) GPT-5.1 (d) GPT-5.1-R=H

(e) Gemma3:1B (f) Gemma3:12B (g) Gemma3:27B (h) DeepSeek-R1:1.5B

(i) DeepSeek-R1:7B (j) DeepSeek-R1:32B (k) Qwen3:1.7B (l) Qwen3:8B

(m) Qwen3:32B (n) Qwen3-R:1.7B (o) Qwen3-R:8B (p) Qwen3-R:32B

Figure A10: **Specificity cascade (college-educated path)** on DICES. Target subgroups become progressively more specific: college-educated people → college-educated women → college-educated Black/Latino women. MSE increases with specificity consistently across all 16 models, validating H3.



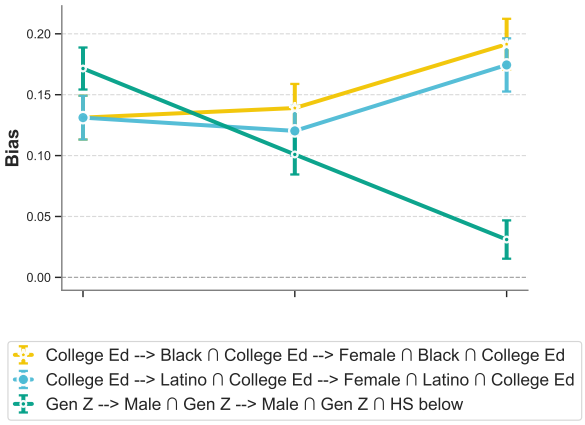
(a) GPT-OSS:20B (b) GPT-OSS:120B (c) GPT-5.1 (d) GPT-5.1-R=H

(e) Gemma3:1B (f) Gemma3:12B (g) Gemma3:27B (h) DeepSeek-R1:1.5B

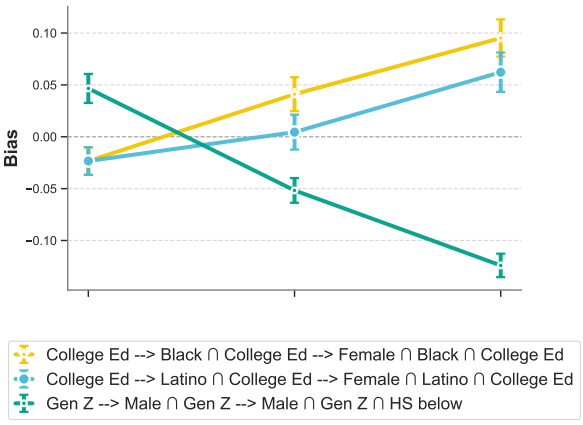
(i) DeepSeek-R1:7B (j) DeepSeek-R1:32B (k) Qwen3:1.7B (l) Qwen3:8B

(m) Qwen3:32B (n) Qwen3-R:1.7B (o) Qwen3-R:8B (p) Qwen3-R:32B

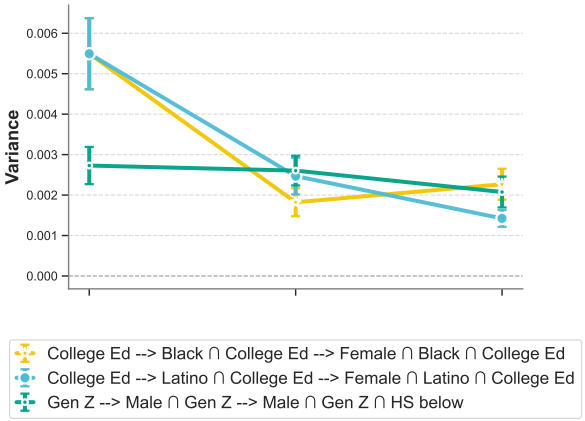
Figure A11: **Specificity cascade (Gen Z path)** on DICES. Target subgroups: Gen Z people → Gen Z men → Gen Z men with ≤high school education. Unlike the college-educated cascade (Figure A10), several models show *decreasing* error with specificity (c, e–g, l–m, o–p), indicating that demographic specificity is not always a reliable proxy for representation difficulty.



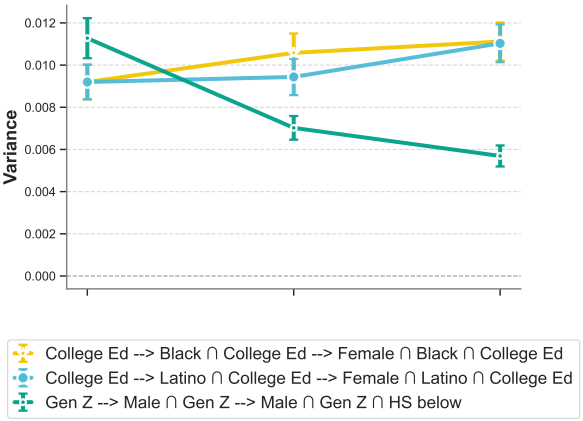
(a) Bias: GPT-5.1



(b) Bias: GPT-5.1-R=H

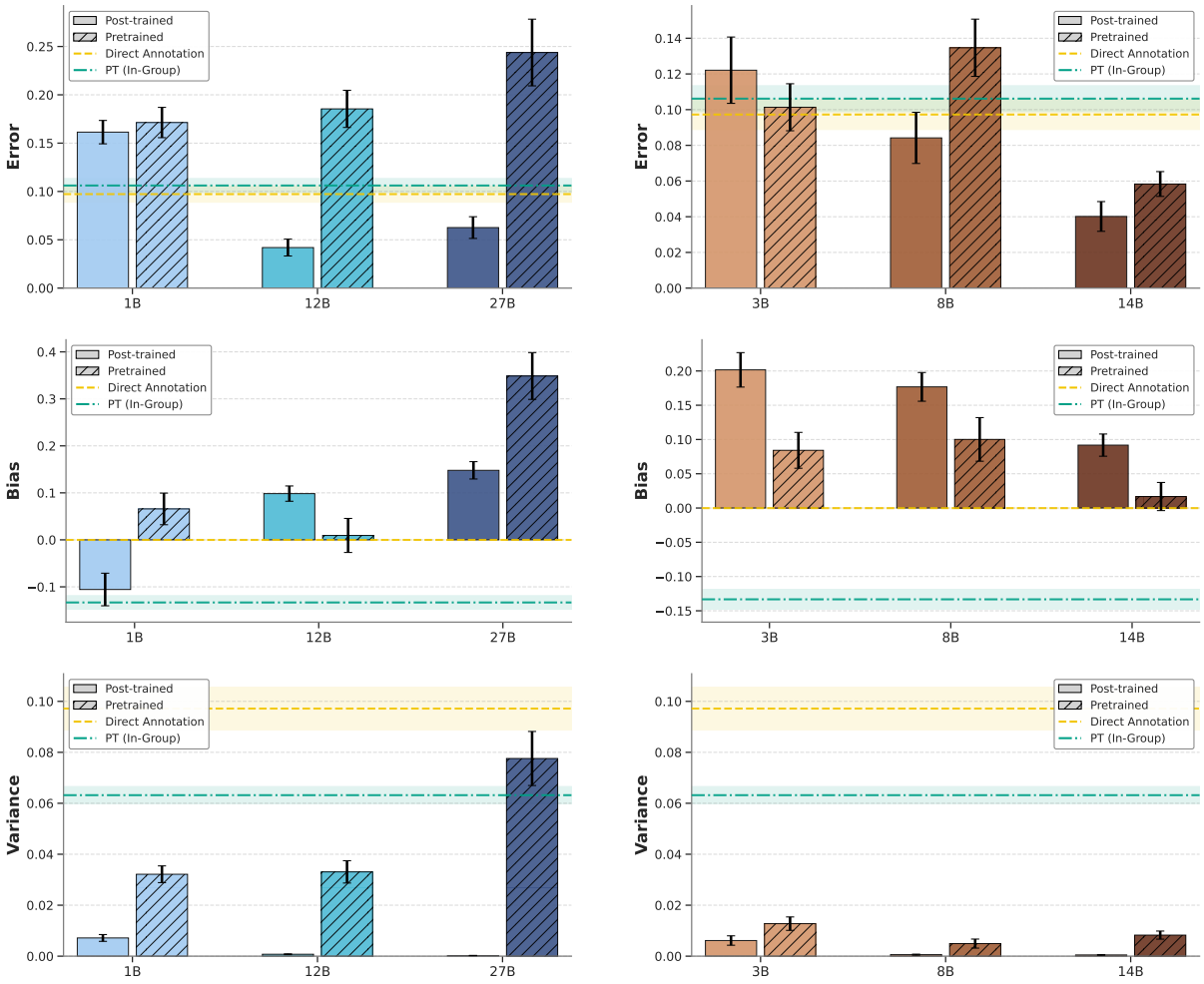


(c) Variance: GPT-5.1



(d) Variance: GPT-5.1-R=H

Figure A12: **Bias–variance diagnostic for specificity cascades** (k=1, GPT-5.1 $\pm$ reasoning). X-axis shows progressively more specific subgroups from both cascades (Figures A10–A11). For the college-educated cascade (left side of each panel), error growth is driven by increasing bias, consistent with representation mismatch. For the Gen Z cascade (right side), both bias and variance *decrease*, suggesting the more specific group is better represented in training data.



(a) Gemma 3 (b) Ministral 3

Figure A13: **Pretrained vs. post-trained models** (k=1, female subgroup). Post-trained models exhibit dramatically lower variance, while pretrained models show lower absolute bias at matched sizes.

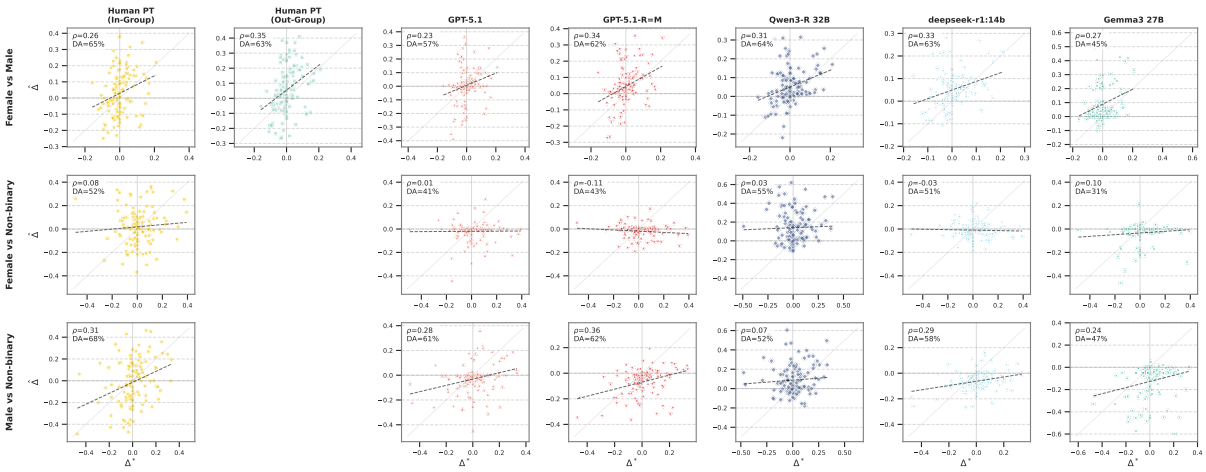


Figure A14: **Differential PT scatter plots**: per-item estimated disagreement $\hat{\Delta}$ vs. ground-truth disagreement Δ^* for representative estimators (columns) across the three gender-group pairs (rows). Dashed lines are OLS fits.

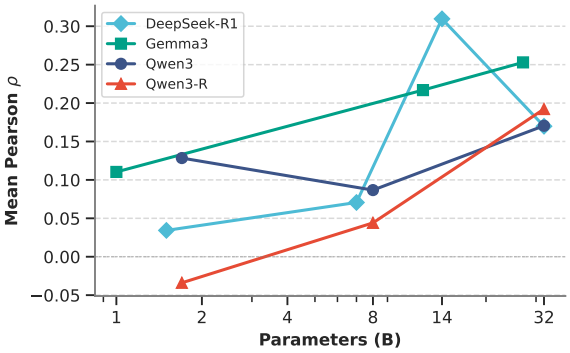


Figure A15: **DPT ability vs. model scale**: mean Pearson ρ across the two informative group pairs (F↔M, M↔NB).