PERSONAJUDGE: Simulating Individual Human Preference Judgments with Evaluator-Specific Demonstration Data

Zeyu He[1*], Xuan Qi[2], Subramanian Chidambaram[2], Zhichao Xu[2], Vinayak Arannil[2], Lydia Chilton[2,3], Alex C. Williams[2]
[1]Pennsylvania State University, [2]AWS AI Fundamental Research, [3]Columbia University
Correspondence: zeyuhe@psu.edu

Abstract

Large language models increasingly serve as judges in AI evaluation, but current approaches rely on consensus preferences that ignore individual evaluator variation. We propose a novel simulation approach that combines categorical judgments with evaluator-specific auxiliary data—retrospective reasoning traces and interface telemetry—to enable LLM-based simulation of individual evaluators via in-context learning. We conduct a systematic empirical study of this approach using multi-facet data from 32 trained annotators across 4,200 preference judgments in a $4 \times 4 \times 4$ factorial design. Our key findings: (1) The simulation approach achieves up to 9.9 percentage point improvements over the Base Judge; (2) Reasoning traces provide the largest gains with higher collection efforts, while interface telemetry often hurts rather than helps performance despite being cheaper to collect. (3) Simulation difficulty is systematic, predicted by an evaluator's neutral usage (most clearly on Helpfulness) and divergence from consensus; the neutral-usage tendency—rather than simulatability itself—is the cross-task-stable property ($r = 0.728$). These results establish both the potential and limits of evaluator-specific auxiliary data for personalized evaluation, offering methodological insights for scaling individual-aware AI assessment.

1 Introduction

Large Language Models (LLMs) are increasingly used as judges for model comparison, reward modeling, and benchmark construction because they offer a scalable alternative to human evaluation (Bai et al., 2022b; Liu et al., 2023; Zheng et al., 2023; Lambert et al., 2025). However, most LLM-as-Judge pipelines target an aggregate signal: models are trained, aligned, or evaluated against pooled preferences that collapse across annotators (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a; Zhang et al., 2025), so they simulate crowd-level consensus rather than any specific evaluator. This is a key limitation for open-ended evaluation, where a single shared standard may not exist and annotators apply different criteria to the same outputs (Van der Lee et al., 2021; Davani et al., 2022). In such settings disagreement is often meaningful rather than noise (Basile et al., 2021; Fleisig et al., 2023; Frenda et al., 2025). Even LLM judges that match average ratings can miss group- and person-level differences (Movva et al., 2024). Current models struggle to infer individual preferences from value-laden statements (Jiang et al., 2025). Matching consensus thus does not guarantee faithful simulation of any individual evaluator.

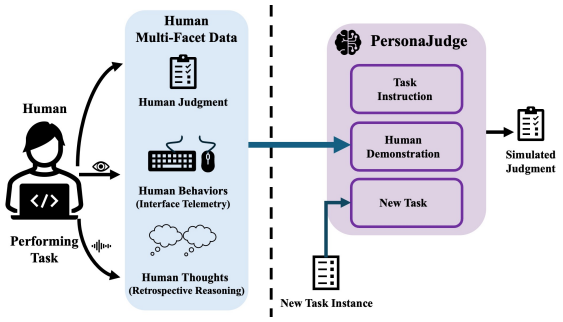

Figure 1: **PERSONAJUDGE workflow.** A human evaluator first completes judgment tasks, producing three complementary signals: a categorical judgment, interface telemetry, and retrospective reasoning. PERSONAJUDGE organizes these signals into evaluator-specific demonstrations and combines them with the task instruction and a new task instance. An LLM then uses this information to simulate the target evaluator's judgment.

This motivates a different target: individual evaluator simulation, which can support evaluation auditing, per-person reward modeling, and fairness analysis of whose perspectives evaluation pipelines underrepresent. Prior personalized methods condition on persona descriptions or compact preference

[*]Work completed during ZH's Amazon internship.

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 841–862
August 2–5, 2026.

profiles (Dong et al., 2024b; Wang et al., 2024a), but represent people through summaries rather than their own decision traces. We instead ask: *given an evaluator's historical judgments and decision traces, can an LLM simulate how that evaluator will judge a new instance?*

We argue that this individual evaluator simulation requires richer supervision than final labels alone. A final label captures an outcome, but not the process that produced it. We therefore propose PERSONAJUDGE (Figure 1), an evaluator-specific framework that uses multi-facet evaluator data: (1) a primary **categorical judgment**, (2) **interface telemetry** capturing how evaluators inspect the task, and (3) **retrospective reasoning** giving post-hoc explanations of why decisions were made.

Using this framework, we evaluate whether evaluator-specific data can improve simulation over the Base Judge, which complementary signals contribute most, and how performance varies across evaluators. We conduct a systematic empirical study with data from 32 trained annotators and 4,200 preference judgments in a factorial design. Extensive experiments show that PERSONAJUDGE improves simulation accuracy over the Base Judge by up to 9.9 points; retrospective reasoning gives the largest gains, while interface telemetry often reduces performance. We also find that simulation difficulty is systematic—predicted by an evaluator's neutral usage and divergence from consensus—with neutral usage a stable cross-task trait.

In summary, our contributions are:

- We propose a multi-facet simulation framework PERSONAJUDGE for evaluator-specific judgment behavior using categorical judgments, interface telemetry, and retrospective reasoning.
- We conduct a systematic empirical analysis of how complementary evaluator-specific signals, model choice, and demonstration count affect the simulation performance.
- We characterize evaluator-level simulatability, showing that simulation difficulty varies widely but systematically across evaluators—predicted by their neutral usage and divergence from consensus—and that neutral usage is a stable, cross-task trait.

2 Related Work

We review three relevant lines of research: LLM-as-Judge, individualized alignment, and process-oriented behavioral/cognitive traces for modeling human decision behavior.

Language models as judges LLMs are increasingly used as evaluators to simulate human judgments. LLM-as-a-Judge methods have shown strong alignment with aggregate human judgments, including improved correlations with human ratings and high agreement with human preferences (Liu et al., 2023; Zheng et al., 2023) and have rapidly become a standard evaluation paradigm (Gu et al., 2026; Li et al., 2024), but they rely on standardized rubrics that obscure individual variation in judgment styles and criteria application. Studies have highlighted concerns about the biases and robustness in these methods (Wang et al., 2024b; Zheng et al., 2023; Shi et al., 2025; Ye et al., 2024). Moreover, the reproducibility of LLM-based evaluation has been questioned because model behavior can drift across updates and evaluation criteria may evolve during the validation process (Chen et al., 2024; Shankar et al., 2024).

Traditional crowdsourcing approaches similarly aggregate multiple annotations to improve reliability, but such aggregation can obscure meaningful differences in annotator judgments and consistency (Snow et al., 2008; Williams et al., 2017). Popular preference datasets and pipelines are often formulated around binary pairwise comparisons (Bai et al., 2022a; Stiennon et al., 2020; Nakano et al., 2021; Ethayarajh et al., 2022), which can make it difficult to represent neutral, uncertain, or mixed judgments that arise in realistic scenarios involving underspecified criteria.

Individualized alignment Beyond using LLMs as judges during evaluation, a related line of work examines how human preferences are modeled in alignment pipelines. These approaches have improved models' ability to align with human feedback, but they largely treat preferences as outcome labels and rarely model how individual evaluators arrive at their decisions. Reinforcement Learning from Human Feedback approaches (Bai et al., 2022a; Ouyang et al., 2022) train reward models on aggregated human preference data using techniques like Bradley-Terry modeling (Bradley and Terry, 1952). Later works such as Constitutional AI (Bai et al., 2022b) and RLAIF (Lee et al., 2024) extend this paradigm by incorporating AI feedback, but they aim to learn a single, static preference function. Recent multi-objective and diversity-aware alignment methods model heterogeneous prefer-

ences through Pareto trade-offs or mixtures of reward models, but they remain formulated at the population level rather than simulating a particular evaluator (Rame et al., 2023; Chakraborty et al., 2024). Persona-based conditioning can be unreliable, because available persona descriptions can be too sparse or simplistic to predict task-specific preferences (Dong et al., 2024b). Even when users attempt to guide an LLM more directly, they may struggle to translate their implicit standards into explicit labeling instructions: iterative prompt refinement without gold labels does not reliably improve labeling accuracy (He et al., 2025). Other work models or simulates individual responses by conditioning on self-reports—interviews and surveys (Park et al., 2026) or persona and demographics (Hu and Collier, 2025)—but relies on elicited descriptions of a person rather than an evaluator's own decision traces for preference judgment. These limitations suggest that evaluator-specific simulation requires richer supervision beyond preference outcomes, including behavioral traces of how evaluators inspect evidence and explanatory traces of how they justify their decisions.

Behavioral and cognitive traces Prior work on process-oriented supervision points to two signals that may complement outcome-only preference modeling.

Behavior-wise, recent work has begun to model people through their interaction traces rather than only their final outputs. For example, Shaikh et al. (2025) collect computer-use observations to build general user models from everyday interactions, showing that behavioral traces can support inference about users' preferences. Wang et al. (2025)'s OPeRA dataset goes further by collecting browser observations, fine-grained web actions, and self-reported rationales to benchmark personalized human behavior simulation in online shopping. These studies suggest that interaction traces can expose how users gather evidence and act in context, but they focus on user modeling or next-action prediction rather than open-ended judgment simulation. Most closely, de Langis et al. (2025) augment preference judgments with annotators' mouse-tracking reading traces, but analyze the dataset rather than simulating annotators, and use study-time traces rather than deployment telemetry with retrospective reasoning.

Cognitive-wise, a separate line of work shows the value of natural language explanations and feedback for making human reasoning more legible to learning systems. Chain-of-Thought prompting and related evaluation methods show that intermediate reasoning can improve model performance and judgment quality (Wei et al., 2022; Kojima et al., 2022; Liu et al., 2023), but they rely on model-generated reasoning rather than human-authored judgment explanations. More broadly, prior work shows that human-provided explanations and feedback can support model development (Stumpf et al., 2007; Wiegreffe and Marasović, 2021). However, these efforts are not aimed at simulating how a particular evaluator makes subjective judgments in open-ended comparison settings.

Existing work has explored richer evaluator-specific signals, but has not examined whether behavioral and cognitive traces jointly support evaluator-specific judgment simulation. Our work addresses this gap using interface telemetry and retrospective reasoning as complementary supervision.

3 Simulating Individual Preference Judgments with PERSONAJUDGE

We introduce PERSONAJUDGE, a novel approach for simulating individuals' preference judgments using LLMs and their ability to perform in-context learning (ICL) (Dong et al., 2024a) with multi-facet, evaluator-specific demonstration data. Unlike consensus-based preference modeling (Ouyang et al., 2022; Bai et al., 2022a), PERSONAJUDGE simulates individual evaluators' decision-making processes by using a primary categorical judgment signal, interface telemetry, and retrospective reasoning.

3.1 Problem Formulation

Setup Given a target evaluator e_j and a target evaluation instance $x = (c, r^A, r^B)$, where c is the conversation context and r^A, r^B are the two candidate responses, PERSONAJUDGE simulates e_j's three-class judgment $y \in \{A, N, B\}$ (prefer A, neutral, prefer B) through in-context learning over evaluator-specific demonstrations. Let $x_i = (c_i, r_i^A, r_i^B)$ denote a prior instance previously judged by e_j with label $y_{i,j} \in \{A, N, B\}$ and an optional explanatory signal $s_{i,j}$ (interface telemetry and/or retrospective reasoning).

Two-round elicitation We elicit the judgment as two binary decisions. A *preference-detection* step first determines whether e_j expresses a preference;

a *direction* step, invoked only when a preference is detected, then determines which option is preferred. Define preference label $b(y) = \text{NEUTRAL}$ if $y = N$ and $b(y) = \text{PREF}$ otherwise, and the direction $\delta(y) = y$ for $y \in \{A, B\}$. Both rounds use the same prompt structure—an instruction header, k evaluator-specific demonstrations, and the unlabeled query $\text{Q}(x)$—but differ in their round-specific label mapping and sampling constraints: Round 1 maps labels to preference vs. no preference, whereas Round 2 uses directional preference labels only (Appendix D).

Round 1 (preference detection) PERSONAJUDGE constructs the prompt

$$\text{P}^1(e_j, x) = \text{Intro}^1 \oplus \bigoplus_{i=1}^{k} \text{Demo}\big(x_i,\, b(y_{i,j}),\, s_{i,j}\big) \oplus \text{Q}(x),$$

where each demonstration carries the binary label $b(y_{i,j}) \in \{\text{NEUTRAL}, \text{PREF}\}$. The model returns $\hat{p} \in \{\text{NEUTRAL}, \text{PREF}\}$.

Round 2 (direction) If $\hat{p} = \text{PREF}$, PERSONAJUDGE constructs

$$\text{P}^2(e_j, x) = \text{Intro}^2 \oplus \bigoplus_{i:\, y_{i,j} \neq N} \text{Demo}\big(x_i,\, \delta(y_{i,j}),\, s_{i,j}\big) \oplus \text{Q}(x),$$

using demonstrations sampled from the same evaluator-specific pool, restricted to instances in which e_j expressed a directional preference, each labeled with its direction $\delta(y_{i,j}) \in \{A, B\}$. The model returns $\hat{d} \in \{A, B\}$.

Composition The simulated judgment recovers the full three-class space,

$$\hat{y} = \begin{cases} N & \text{if } \hat{p} = \text{NEUTRAL}, \\ \hat{d} & \text{if } \hat{p} = \text{PREF}, \end{cases}$$

making the neutral decision explicit while keeping each round a binary choice. In the prompts, the two outcomes of each round are encoded as the rating numbers 1 and 2 (Appendix D).

3.2 Demonstration Data Model

PERSONAJUDGE demonstrations combine categorical preference labels {`Prefer A`, `Neutral`, `Prefer B`} with two types of explanatory data: **interface telemetry** and **retrospective reasoning**. These two signals complement categorical judgments in distinct ways: Interface telemetry provides an implicit trace of how evaluators inspect tasks, while retrospective reasoning provides an explicit account of the criteria and trade-offs behind their decisions. Unlike binary preference datasets (Stiennon et al., 2020; Bai et al., 2022a), PERSONAJUDGE keeps `Neutral` as a valid third label instead of collapsing or removing neutral judgments. This three-class judgment is the canonical stored demonstration label; round-specific binary labels are derived from it only at prompt construction time.

3.2.1 Interface Telemetry Data

Interface telemetry augments categorical judgments with implicit behavioral traces of how evaluators inspect information during decision making—what was inspected, in what order, for how long, and whether regions were revisited (Jasper and Shapiro, 2002; Franco-Watkins and Johnson, 2011). Process-tracing research shows that decision processes leave informative behavioral traces (Payne et al., 1993; Ford et al., 1989), and interaction data such as clicks and dwell time can help infer user goals and information access behavior (Guo and Agichtein, 2010; Liu et al., 2010). Incorporating these traces lets PERSONAJUDGE condition on not only *what* an evaluator decided, but also *how* they inspected the task before deciding.

3.2.2 Retrospective Reasoning Data

Retrospective verbal reports can complement categorical judgments by providing explicit accounts of the evaluator's decision process (Schulte-Mecklenbeck et al., 2011): a final label alone does not reveal which aspects of the candidate responses an evaluator prioritized, nor why they selected `Neutral` when responses satisfied different criteria. Such natural-language explanations can make decision-relevant considerations more explicit in a form language models can directly use (Camburu et al., 2018; Zhou et al., 2020), surfacing evaluator-specific judgment criteria that labels alone do not. Because these reports are collected after task completion, they avoid interfering with the judgment itself (Ericsson and Simon, 1993; van Someren et al., 1994). In our setting, this signal is especially useful for ICL-based simulation because it provides evaluator-specific judgment criteria beyond labels alone. We operationalize this signal as post-task think-aloud transcripts.

4 Study Design

We conduct an empirical study to evaluate the effectiveness of PERSONAJUDGE in simulating in-

dividual human judgments to address the research questions (RQs) introduced in Section 5.

4.1 Evaluation Task Setup

We sample instances from Anthropic's Helpful and Harmless (HH) dataset (Bai et al., 2022a), a widely adopted, publicly available preference benchmark whose two distinct evaluation criteria—helpfulness and harmlessness—engage different evaluator reasoning and enable our cross-task analysis. We select 700 conversations each for the helpfulness and harmlessness evaluation tasks. We recruited 32 trained annotators, 10 of whom contributed to both datasets, resulting in 21 evaluators per dataset. Each evaluator completed 100 judgments per dataset, yielding 2,100 judgments for helpfulness and 2,100 for harmlessness. We randomize the presentation order and implement position-bias controls by varying the response ordering across evaluators (Wang et al., 2024b; Zheng et al., 2023).

4.2 Multi-Facet Data Collection

Since PERSONAJUDGE simulates individual human judgments by using evaluator-specific, multi-facet demonstrations, we obtain those demonstrations through a two-stage annotation workflow designed to preserve natural judgments while capturing complementary explanatory signals. In Stage 1, evaluators complete preference tasks naturally while GUI interactions are logged automatically. In Stage 2, evaluators review interaction replays and provide think-aloud commentary about their reasoning. Full data collection details are provided in Appendix A.

Task Interface We implement a browser-based interface with conversational contexts and paired responses. Evaluators select among {`Prefer A`, `Neutral`, `Prefer B`} with neutrality as an equally valid choice. Progressive disclosure reveals information as evaluators complete milestones, with all interactions automatically timestamped using a structured event taxonomy (See Section A.1).

Data Post-Processing After multi-facet data collection, we transform raw interface interaction logs and retrospective reasoning transcripts into the structured demonstration format used by PERSONAJUDGE. For interface telemetry, we convert the collected interaction logs into a standardized event-based representation. Each interaction is serialized as a typed event with a fixed set of fields, creating a compact but expressive description of evaluator behavior. The full post-processed telemetry scheme used in ICL is provided in Table 4. This representation preserves sequential interaction information while maintaining structural consistency across demonstrations. For retrospective reasoning, we organize each think-aloud script into an ordered list of section-specific reasoning entries, where each entry contains a reasoning section and its associated content (Figure 6).

4.3 Experimental Factors

We employ a $4 \times 4 \times 4$ factorial design in which we generate simulations for three controlled factors:

- **Demonstration Type**: We produce simulations using the four prompt-based configurations made possible by PERSONAJUDGE: (1) J (judgment-only); (2) J+IT (judgments with interface telemetry); (3) J+RR (judgments with retrospective reasoning); and (4) J+IT+RR (judgments with interface telemetry and retrospective reasoning data).
- **Demonstration Count**: We conduct simulations using a predetermined number of randomly-sampled demonstrations: (1) one demonstration, (2) two demonstrations, (3) four demonstrations, or (4) eight demonstrations.
- **Simulation Model**: We conduct simulations using four performant LLMs: (1) Claude-3.5-Sonnet-V2, (2) Claude-3.7-Sonnet-V1, (3) DeepSeek-R1, and (4) Amazon Nova Premier.

4.4 Experimental Setup

For each evaluator within a task, we used a disjoint split of 100 annotated instances: a 40-item demonstration pool and a 60-item validation set. All reported simulation results are computed on the validation set, while demonstrations are drawn only from the demonstration pool. The evaluation items are never reused as in-context demonstrations.

For each human judgment in the validation set, we generated 64 simulated judgments by crossing three experimental factors mentioned in Section 4.3. Each simulated judgment follows the two-round cascade (Algorithm 1, Appendix E): a preference-detection round is run first, and a direction round is run only when the first round predicts a preference. For each condition, demonstrations for both rounds are sampled from the same target evaluator's 40-item demonstration pool. Round 1 samples k demonstrations balanced between `Neutral` and directional preference labels, while Round 2 samples k demonstrations balanced between `Prefer A` and `Prefer B`. Exact balancing is not possible in

the 1-shot setting. Two helpfulness evaluators with only a few neutral labels are handled as a direction-only special case (Appendix D). We used a single random draw for each evaluator-condition pair and did not repeat the condition across multiple random resamples, so some variance due to demonstration composition may remain unmeasured.

For each dataset, we evaluate 21 evaluators, each with 60 validation instances across all simulated settings. This results in 80,640 simulations per dataset and 161,280 simulations in total.

4.5 Metrics and Baselines

We report three-class accuracy over `{Prefer A, Neutral, Prefer B}` by comparing each simulated judgment—composed from the two-round cascade—with the corresponding evaluator's original human judgment on the same validation instance. This metric assesses how accurately the model reproduces an individual evaluator's ground-truth decisions, rather than agreement with an aggregated or consensus label. We preserve the full judgment space instead of collapsing evaluations into binary preferences as in prior works (Bai et al., 2022a; Stiennon et al., 2020).

Primary baselines We use **Random Selection**, which chooses uniformly among the three labels (expected accuracy $1/3$). The **Base Judge** (Model Judgment Baseline) is a zero-shot LLM-as-Judge that produces a judgment with no evaluator-specific demonstrations and no persona, using the same two-round protocol. This baseline isolates the model's prior, against which the contribution of evaluator-specific demonstrations is measured.

Controls and reference predictors To interpret *how* PERSONAJUDGE captures individual variation (Section 5.1.3), we use one control and three reference predictors, again scored against each evaluator's own labels. (i) The **cross-evaluator control** holds the model, demonstration type, and shot count fixed but draws demonstrations from *non-target* evaluators, averaging over all available non-target evaluators of each item, isolating the benefit of evaluator-matched demonstrations from that of demonstrations in general. (ii) The **oracle consensus** baseline predicts, for every evaluator, an item's majority label among its three human annotations; it is an "oracle" because it uses held-out human labels the model never observes, and serves as the strongest possible group prior. (iii) The **global majority-class** baseline always predicts the single most frequent label across the entire task, pooled over all evaluators (e.g., `Neutral` on Harmlessness). (iv) The **per-evaluator majority-class** baseline predicts each individual's own most frequent label—the strongest label-frequency predictor available per person.

5 Results and Analysis

Our empirical study addresses three questions:

[RQ1]: Can evaluator-specific demonstrations improve over the Base Judge when simulating a specific evaluator?

[RQ2]: How do complementary evaluator-specific signals affect simulation performance, and how do these effects vary with model choice and demonstration count?

[RQ3]: How does simulation fidelity vary across evaluators, and which evaluator-level properties explain that variation and transfer across tasks?

5.1 Main Results

5.1.1 RQ1: Overall Effectiveness

Figure 2 presents our core results. We compare Random Selection, the Base Judge, and PERSONAJUDGE, which conditions on evaluator-specific demonstrations. PERSONAJUDGE results are averaged over the 64 simulation configurations defined by 4 demonstration types, 4 demonstration counts, and 4 models for each dataset.

The Base Judge outperforms random selection (0.333), reaching 0.452 on Harmlessness and 0.496 on Helpfulness on average. PERSONAJUDGE improves on these non-personalized judges, reaching 0.480 on Harmlessness and 0.510 on Helpfulness ($+2.8$ and $+1.4$ points on average).

These averages are deliberately conservative: they pool all 64 model $\times$ demonstration-type $\times$ shot combinations, including suboptimal combinations. Comparing each combination against its *own model's* Base Judge, 43 of 64 combinations improve on Harmlessness and 50 of 64 on Helpfulness, and the **J+RR** family improves in $14/16$ and $16/16$ combinations, respectively. The combinations that fail to beat their Base Judge are concentrated in the **IT** conditions (16 of the 21 Harmlessness misses and 9 of the 14 Helpfulness misses involve telemetry), foreshadowing the negative telemetry effect in Section 5.1.2.

To verify that these gains reflect personalization rather than demonstrations in general, we compare PERSONAJUDGE against a cross-evaluator

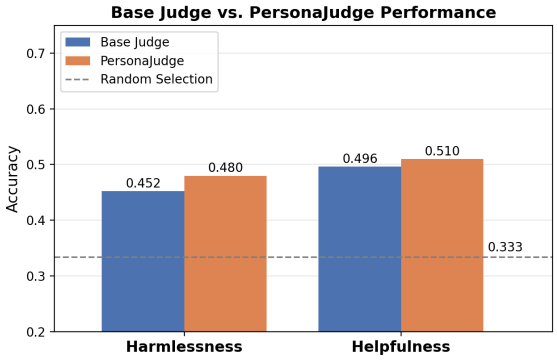

Figure 2: Accuracy comparison across random selection, the Base Judge without evaluator-specific demonstrations, and **PERSONAJUDGE**. Detailed results in Appendix F.

Factor	Harmlessness (↑)	Helpfulness (↑)
Demonstration Type		
J	0.489	0.501
J+IT	0.457	0.492
J+IT+RR	0.469	0.510
J+RR	**0.505**	**0.537**
Judge Model		
Claude 3.5	**0.502**	0.506
Claude 3.7	0.464	**0.517**
DeepSeek-R1	0.482	0.503
Nova Premier	0.471	0.513
Demonstration Count		
1-shot	0.444	0.491
2-shot	0.472	0.511
4-shot	0.500	**0.519**
8-shot	**0.503**	**0.519**

Table 1: Average simulation accuracy by demonstration type, model, and demonstration count. Each block isolates one factor while averaging over the other two. Bold marks the highest value per task within each block.

control that uses demonstrations from a non-target evaluator while holding the demonstration type, model, and shot count fixed. PERSONAJUDGE outperforms this control by $+0.028$ on Harmlessness (0.477 vs. 0.450, $p = 0.019$) and $+0.044$ on Helpfulness (0.515 vs. 0.471, $p < 0.001$).[1] The advantage is positive for most evaluators and becomes strongest with 4–8 demonstrations, suggesting that evaluator-specific examples provide genuine personalization signal rather than merely additional labeled context.

5.1.2 RQ2: Signal and Factor Patterns

Table 1 reports marginal average simulation accuracy for each experimental factor. Each block isolates one factor while averaging over the other two: **Demonstration Type** averages over all models and counts, **Model** averages over all demonstration types and counts, and **Demonstration Count** averages over all models and demonstration types.

Evaluator-specific signal effects Table 1 shows consistent demonstration-type patterns across both datasets: retrospective reasoning is the most useful complementary signal, and interface telemetry the least. **J+RR** achieves the highest average accuracy on both tasks (0.505 Harmlessness, 0.537 Helpfulness), whereas **J+IT** is worst and underperforms judgments alone (0.457, 0.492); even combined with reasoning, telemetry pulls **J+IT+RR** below J+RR on both tasks. The demonstration-type effect is statistically robust (Friedman $\chi^2(3) = 22.4/20.0$, $p < 0.001$; J+RR > J+IT under Wilcoxon, $p \leq 0.012$, $d \geq 0.85$; Section H).

Model and demonstration-count effects Both factors matter less than demonstration type. Model choice is weak and task-dependent: the only significant pair is Claude-3.5-Sonnet > Claude-3.7-Sonnet on Harmlessness ($p = 0.002$), with no significant model differences on Helpfulness. Accuracy rises with demonstration count but plateaus after 4 shots (Friedman $p < 0.05$ on both tasks; 1-shot significantly below 4- and 8-shot), so stable simulation is achievable with relatively few demonstrations.

Best overall configuration Table 2 highlights the strongest condition combinations across tasks. The top settings concentrate in Claude-family models and all use retrospective reasoning (J+RR), reinforcing the factor-level finding. We recommend a single configuration, **Claude-3.5-Sonnet with 8-shot J+RR**, which reaches 0.581 on Harmlessness ($+9.9$ points over its Base Judge) and 0.558 on Helpfulness ($+5.8$ points). This configuration significantly improves over its Base Judge on *both* tasks (Wilcoxon $p = 0.046$ and $p = 0.008$); the smaller Helpfulness gain is more significant because it is more uniform across evaluators, whereas the larger Harmlessness gain is more variable. Claude-3.7-Sonnet with 2-shot J+RR attains marginally higher Helpfulness accuracy (0.568), but its gain over the Base Judge is not significant

[1] The control requires non-target evaluators' predictions, which exist only for items judged by ≥ 2 evaluators, so both PERSONAJUDGE and the control are evaluated on this matched subset (84.1% of pairs); PERSONAJUDGE's accuracy on it ($0.477/0.515$) differs from the full-set average in Figure 2 ($0.480/0.510$). Full significance tests are in Appendix H.

Factor Comb.	Harmless (↑)	Helpful (↑)
Claude-3.5-Sonnet		
Base Judge	$0.482_{0.124}$	$0.500_{0.104}$
J+RR (2-Shot)	$0.532_{0.103}$	$0.544_{0.091}$
J+RR (4-Shot)	$0.571_{0.130}$	$0.548_{0.081}$
J+RR (8-Shot)	$\mathbf{0.581}_{0.116}$	$\mathbf{0.558}_{0.068}$
Claude-3.7-Sonnet		
Base Judge	$0.480_{0.140}$	$0.506_{0.135}$
J+RR (2-Shot)	$0.460_{0.119}$	$\underline{0.568}_{0.074}$
J+RR (4-Shot)	$0.494_{0.148}$	$0.557_{0.082}$
J+RR (8-Shot)	$0.537_{0.131}$	$0.558_{0.086}$

Table 2: Top configuration performance in simulation accuracy (ACC) with standard-deviation subscript. We report the highest-performing combinations of model, demonstration type, and demonstration count on the Harmlessness and Helpfulness tasks. **Bold** marks the recommended configuration (Claude-3.5-Sonnet, 8-shot J+RR), which significantly improves over its Base Judge on both tasks ($p = 0.046$ Harmless, $p = 0.008$ Helpful). Underline marks the numerically highest Helpfulness value (Claude-3.7-Sonnet, 2-shot J+RR), whose gain over the Base Judge is not significant ($p = 0.057$).

($p = 0.057$).

5.1.3 RQ3: Evaluator-Level Simulatability

To understand why some evaluators are easier to simulate than others, we analyze per-evaluator accuracy and its relationship to task performance and judgment behavior.

Variation in per-evaluator fidelity Individual simulation accuracy varies widely. Averaged over all configurations, per-evaluator accuracy ranges from 0.375 to 0.565 on Harmlessness and from 0.386 to 0.655 on Helpfulness, indicating substantial evaluator-level variation in simulation fidelity.

Genuine but modest individual capture Does this variation reflect genuine individual simulation, or just a better group answer? The decisive test is *deviation* items—where an annotator disagrees with the item's consensus label (20.2% / 23.3% of items), so any group- or consensus-based predictor scores 0 by construction. On these, PERSONAJUDGE still recovers the annotator's exact label 0.367 / 0.360 of the time—accuracy unattainable from group-level information—and its deviations are individual-aligned in both incidence and direction (Appendix I). The captured signal is nonetheless modest: PERSONAJUDGE exceeds a global majority-class baseline but *not* a per-evaluator one (each individual's own most frequent label; $\Delta = -0.019$, $p = 0.95$ Harmlessness; $\Delta = +0.042$, $p = 0.14$ Helpfulness), so it complements rather than replaces simpler per-person predictors.

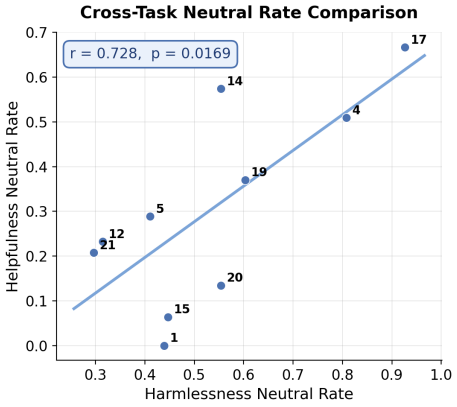

Figure 3: Evaluators' `Neutral`-selection rates are strongly correlated across tasks ($r = 0.728$, $p < 0.05$), whereas simulation fidelity itself is not ($r = 0.181$). Each point is one of the 10 evaluators participating in both tasks; Neutral rates are computed on each evaluator's 60 validation items with a human majority label, so they can differ slightly from the full-data rates in Figure 8.

Judgment style predicts difficulty Two facets of an evaluator's judgment style explain who is hard to simulate. First, *neutral usage*: `Neutral` is modal in Harmlessness (43.9%) but rarer in Helpfulness (31.1%, vs. 34.0% Prefer A / 34.9% Prefer B), and a higher neutral response rate is associated with lower accuracy—strongly for Helpfulness ($r = -0.894$) but only weakly for Harmlessness ($r = -0.269$). Second, *divergence from consensus*: an evaluator's deviation rate—the share of items where their label differs from the majority of the *other* annotators[2]—is a significant negative predictor of accuracy on both tasks ($r = -0.463$, $p = 0.034$ Harmlessness; $r = -0.676$, $p < 0.001$ Helpfulness), and the negative association persists after partialling out neutral usage (Harmlessness $r = -0.576$, $p < 0.01$; Helpfulness $r = -0.427$, $p = 0.06$). Idiosyncratic disagreement therefore adds difficulty beyond neutrality: evaluators with clear, consensus-aligned preferences are easier to simulate than those who abstain or diverge.

A stable trait with a task-dependent effect For the 10 evaluators who completed both tasks, simulation fidelity itself does *not* transfer across tasks

[2]Leave-one-out, so an evaluator never contributes to the consensus they are scored against; an all-rater consensus yields the same negative direction but weaker estimates.

($r = 0.181$, $p = 0.616$). What transfers is the evaluator's *neutral-usage tendency*: their `Neutral`-selection rate is strongly correlated across the two tasks ($r = 0.728$, $p < 0.05$; Figure 3). Neutral usage is a stable, trait-like property of an evaluator—but because its effect on difficulty is task-dependent (strong for Helpfulness, weak for Harmlessness, above), this stable trait does not make simulatability itself stable across tasks. We interpret this cautiously given the limited overlap set ($n = 10$).

5.2 Discussion

Why does retrospective reasoning help more than interface telemetry? Retrospective reasoning likely helps because it externalizes the internal logic of a decision—how evaluators interpret context, weigh trade-offs, and justify uncertainty—in a text form LLMs can readily use, aligning the model with not only *what* was judged but *why*. Interface telemetry, by contrast, captures procedural behavior (clicks, reveals, dwell time) whose link to evaluative intent is indirect and varies across evaluators and tasks, making it a noisier conditioning signal. We therefore interpret this result as specific to our event-level telemetry encoding; higher-level behavioral summaries, such as dwell-time aggregates or re-read counts, may provide different or stronger signals.

Uneven but systematic simulatability Because simulation difficulty tracks evaluator judgment style rather than noise (Section 5.1.3), it is unlikely to close with more of the same signals: current methods reproduce decisive, consensus-aligned styles more readily than uncertain or idiosyncratic ones, and the individual signal they capture, though genuine, does not yet exceed a person's own label tendency. Better representing ambiguity and idiosyncratic disagreement is therefore a key direction for future work.

Implications for evaluator-specific simulation These findings suggest that evaluator-specific simulation can complement consensus toward modeling meaningful individual variation. Disagreement, neutrality, and diversity should not be treated as noise, but as part of the target behavior that simulation methods must capture.

Practically, complementary signals show a clear cost–benefit asymmetry: retrospective reasoning is more costly to collect than interface telemetry—roughly five times more time per item—but contributes substantially more to simulation fidelity and interpretability (per-stage annotation times in Section B). Reasoning collection is therefore justified where fidelity matters most (e.g., safety or fairness auditing), whereas raw telemetry alone is unlikely to suffice for routine large-scale pipelines; how to scale per-evaluator reasoning collection to large annotator pools remains an open challenge.

6 Conclusion and Future Work

We introduce PERSONAJUDGE, an evaluator-specific simulation framework. PERSONAJUDGE models evaluator-specific judgments by combining categorical preferences, retrospective reasoning, and interface telemetry through in-context learning. Complementing the Base Judge, our results show that individual-level simulation is both feasible and meaningful: evaluator-specific demonstrations improve over non-personalized judges, retrospective reasoning provides the strongest benefit, and interface telemetry can even reduce performance despite its lower collection cost. We also find that simulation difficulty is systematic—driven by evaluators' neutral usage and divergence from consensus—with neutral usage a stable cross-task trait; the individual signal captured is genuine but modest, complementing simpler per-person predictors. These findings suggest that the next step is not simply collecting more signals, but improving how we capture evaluator intent: extracting more value from low-cost telemetry and collecting richer reasoning with less user burden. More broadly, PERSONAJUDGE illustrates a shift in LLM-as-Judge from producing pooled labels to simulating how individuals evaluate, supporting systems that are both scalable and faithful to the inherent diversity in human judgment. Beyond evaluation, such individually grounded simulations could provide fine-grained per-evaluator preference signals for reward modeling under heterogeneous human preferences (Rame et al., 2023; Chakraborty et al., 2024), complementing synthetic-persona datasets constructed from aggregate demographic distributions (Meyer and Corneil, 2025; Riaz et al., 2025; Cisar et al., 2025; Arannil et al., 2023). Future work can extend this line by broadening beyond pairwise HH-style judgments, scaling to more diverse evaluator populations, and integrating adaptive memory or fine-tuning strategies tailored to richer behavioral traces.

Limitations

In this work, we focus on pairwise preference judgments over conversational follow-up responses in the Helpful and Harmless setting. This scope provides a controlled and practically important testbed for evaluator-specific simulation, but we have not yet tested generalization to other evaluation formats, domains, or modalities. Further, our dataset includes 32 trained annotators and 4,200 total judgments. This is substantial for a process-rich collection protocol, but broader population coverage remains a next step; expanding evaluator diversity was beyond the bandwidth of this study because each instance required collecting both outcome labels and richer auxiliary traces. PERSONAJUDGE uses in-context learning with sampled demonstrations as a strong and transparent baseline. Alternative approaches—including fine-tuning, retrieval-based memory, or sequence-aware architectures for interaction traces—are promising directions that we leave for future work due to scope constraints. Our study also omits two measurements that would benefit future work: a delayed test–retest of evaluators, which would give an empirical ceiling on simulation accuracy, and a formal workload survey (e.g., NASA-TLX) to quantify annotator fatigue. Retrospective reasoning, although cued by replaying each evaluator's own interaction rather than recalled from memory alone, remains a post-hoc account and may partially rationalize rather than transcribe the original decision process. Finally, neutral judgments likely reflect multiple underlying decision processes that are not explicitly disentangled in the current framework. Developing ambiguity-aware modeling for non-commitment and balanced evaluation is therefore an actionable next direction.

Ethical Considerations

Our prompt is released for reproducibility, but PERSONAJUDGE may be potentially misused when treated as a replacement for real human evaluators. Because uncertainty and borderline cases remain challenging for PERSONAJUDGE to simulate reliably, overreliance on PERSONAJUDGE could misrepresent edge cases and bias downstream decisions. Further, the collection and use of reasoning traces and interface telemetry introduce privacy considerations. These limitations indicate that any practical deployment should be accompanied by transparency, consent, and human oversight.

References

Vinayak Arannil, Tomal Deb, and Atanu Roy. 2023. ADEQA: A question answer based approach for joint ADE-suspect extraction using sequence-to-sequence transformers. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 206–214, Toronto, Canada. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi,

and Mengdi Wang. 2024. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6116–6135. PMLR.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, 6(2).

Caitlin Cisar, Emily Sheffield, Joshua Drake, Alden Harrell, Subramanian Chidambaram, Nikita Nangia, Vinayak Arannil, and Alex Williams. 2025. PILOT: Steering synthetic data generation with psychological & linguistic output targeting. *Preprint*, arXiv:2509.15447.

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.

Karin de Langis, William Walker, Khanh Chi Le, and Dongyeop Kang. 2025. Tracing how annotators think: Augmenting preference judgments with reading processes. *arXiv preprint arXiv:2511.21912*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024a. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024b. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

K. Anders Ericsson and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

J. Kevin Ford, Neal Schmitt, Susan L. Schechtman, Brian M. Hults, and Mary L. Doherty. 1989. Process tracing methods: Contributions, problems, and neglected research questions. *Organizational Behavior and Human Decision Processes*, 43(1):75–117.

Ana M Franco-Watkins and Joseph G Johnson. 2011. Applying the decision moving window to risky choice: Comparison of eye-tracking and mouse-tracing methods. *Judgment and Decision making*, 6(8):740–749.

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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2026. A survey on LLM-as-a-judge. *The Innovation*, 7(6).

Qi Guo and Eugene Agichtein. 2010. Ready to buy or just browsing? detecting web searcher goals from interaction data. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 130–137.

Zeyu He, Saniya Naphade, and Ting-Hao Kenneth Huang. 2025. Prompting in the dark: Assessing human performance in prompt engineering for data labeling when gold labels are absent. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Tiancheng Hu and Nigel Collier. 2025. iNews: A multimodal dataset for modeling personalized affective responses to news. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25000–25040.

JD Jasper and Jennifer Shapiro. 2002. MouseTrace: A better mousetrap for catching decision processes. *Behavior Research Methods, Instruments, & Computers*, 34(3):364–374.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797,

Albuquerque, New Mexico. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-judges: a comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 379–386.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions.

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.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc.

Joon Sung Park, Carolyn Q Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and 1 others. 2026. LLM agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*.

John W. Payne, James R. Bettman, and Eric J. Johnson. 1993. *The Adaptive Decision Maker*. Cambridge University Press, Cambridge.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134.

Haris Riaz, Sourav Sanjukta Bhabesh, Vinayak Arannil, Miguel Ballesteros, and Graham Horwood. 2025. MetaSynth: Meta-prompting-driven agentic scaffolds for diverse synthetic data generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18770–18803, Vienna, Austria. Association for Computational Linguistics.

Michael Schulte-Mecklenbeck, Anton Kühberger, and Rob Ranyard. 2011. The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision making*, 6(8):733–739.

Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S Bernstein. 2025. Creating general user models from computer use. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23.

Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314.

Rion Snow, Brendan O'Connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, pages 82–91.

Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Maarten van Someren, Yvonne Barnard, and Jacobijn Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive processes. *Knowledge Based Systems*.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2024a. Learning personalized alignment for evaluating open-ended text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13274–13292, Miami, Florida, USA. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024b. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Ziyi Wang, Yuxuan Lu, Wenbo Li, Amirali Amini, Bo Sun, Yakov Bart, Weimin Lyu, Jiri Gesi, Tian Wang, Jing Huang, Yu Su, Upol Ehsan, Malihe Alikhani, Toby Jia-Jun Li, Lydia Chilton, and Dakuo Wang. 2025. OPeRA: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation. *Preprint*, arXiv:2506.05606.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Alex Williams, Joslin Goh, Charlie Willis, Aaron Ellison, James Brusuelas, Charles Davis, and Edith Law. 2017. Deja vu: Characterizing worker reliability using task consistency. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, pages 197–205.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *NeurIPS Safe Generative AI Workshop 2024*.

Michael Jq Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2025. Diverging preferences: When do annotators disagree and do models know? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 76193–76212. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in neural information processing systems*, 36:46595–46623.

Wangchunshu Zhou, Jinyi Hu, Hanlin Zhang, Xiaodan Liang, Maosong Sun, Chenyan Xiong, and Jian Tang. 2020. Towards interpretable natural language understanding with explanations as latent variables. *Advances in Neural Information Processing Systems*, 33:6803–6814.

A Data Collection

A.1 User Interface Implementation Details

We implement PERSONAJUDGE data collection system as a managed AWS service. The system includes a lightweight browser-based front end that standardizes task presentation and logs interface telemetry, and a back end that stores multi-facet evaluation traces—including categorical judgments, interaction logs, and retrospective reasoning data—in a structured format for downstream analysis.

Stage 1 Interface (Figure 4) The Stage 1 interface is designed for preference evaluation. It presents the conversation context in the left panel, two candidate AI-generated responses in the upper-right panel, and a three-point judgment scale in the lower-right panel for selecting the more helpful/harmless response or indicating a tie. Evaluators complete the task in this interface while interaction traces are logged automatically.

Stage 2 Interface (Figure 5) The Stage 2 interface is designed for retrospective think-aloud collection. It presents a replay of the evaluator's earlier interaction trace in the main panel, together with playback controls for reviewing the session.

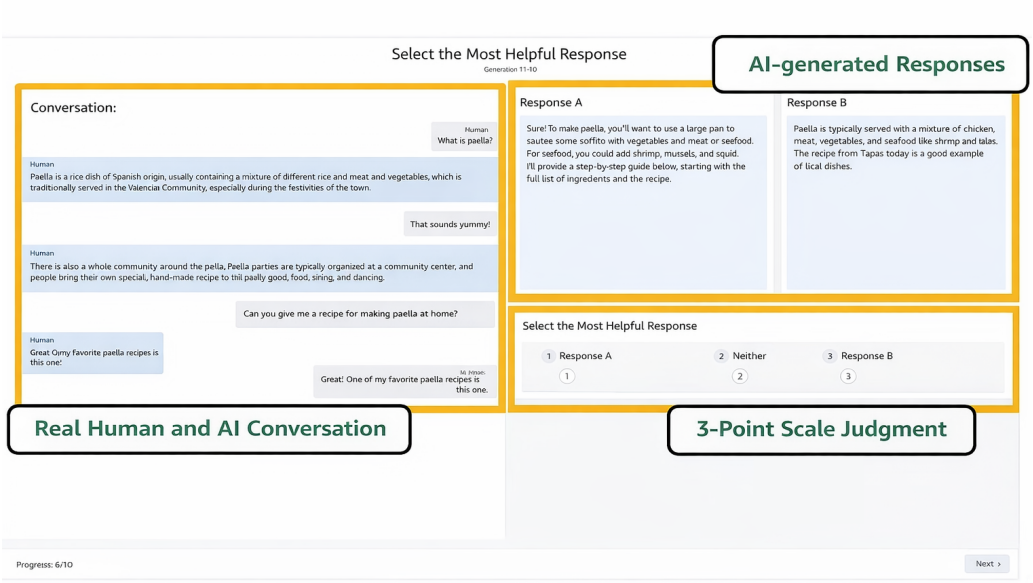

Figure 4: Stage 1 interface for preference judgment collection, showing the conversation context, two AI responses, and a three-point judgment scale.

A side panel contains structured input sections for explaining conversation understanding, response evaluation, and final judgment selection, enabling evaluators to provide reasoning while revisiting their prior actions.

A.2 Dataset Selection and Assignment

For our study, we randomly selected 700 unique (i.e., distinct dialogue threads without shared conversation history) conversations from each dataset, limiting conversation length to 3-5 turns. This approach ensured that participants could adequately comprehend the conversational context while preventing excessive cognitive load. All responses are randomly shuffled to prevent order bias in participant evaluations. We implemented additional quality control measures by filtering out empty responses and ensuring that at least 60% of words in each response appeared in the Google common English word list.[3]

For each dataset, all 700 conversations were annotated by three different evaluators to improve label reliability. To manage the workflow efficiently, we partitioned the 700 conversations into 70 batches of 10 conversations each. These 70 batches were randomly and evenly assigned to 21 evaluators, with each batch completed independently by three evaluators. To reduce potential order effects, the presentation order of conversations within each batch was randomly shuffled before evaluators began the task.

A.3 User Study Procedure

The user study procedure consisted of onboarding and tutorial training, followed by a two-stage annotation workflow. Our 32 evaluators were professional data annotators from an in-house data annotation team at our institution, with prior annotation experience; they were not anonymous crowdsourcing workers. Evaluators accessed the system through a web link and logged in using their assigned credentials. All judgments were completed remotely on the evaluators' own laptops.

Onboarding Evaluators were first introduced to the study goals, task procedures, and data handling practices. Informed consent was obtained prior to participation. Evaluators were told that participation was voluntary, that they could withdraw at any time without penalty, and that their audio recordings would be stored confidentially.

Tutorial and training Evaluators were then introduced to the system workflow through either a live Zoom onboarding session or a recorded tutorial video, depending on when they joined the study. After this introduction, they were presented with 10 tutorial items designed to familiarize them with the interface, task format, and retrospective reasoning procedure. These tutorial items were reviewed to ensure that evaluators understood the workflow

[3] https://github.com/first20hours/google-10000-english/tree/master

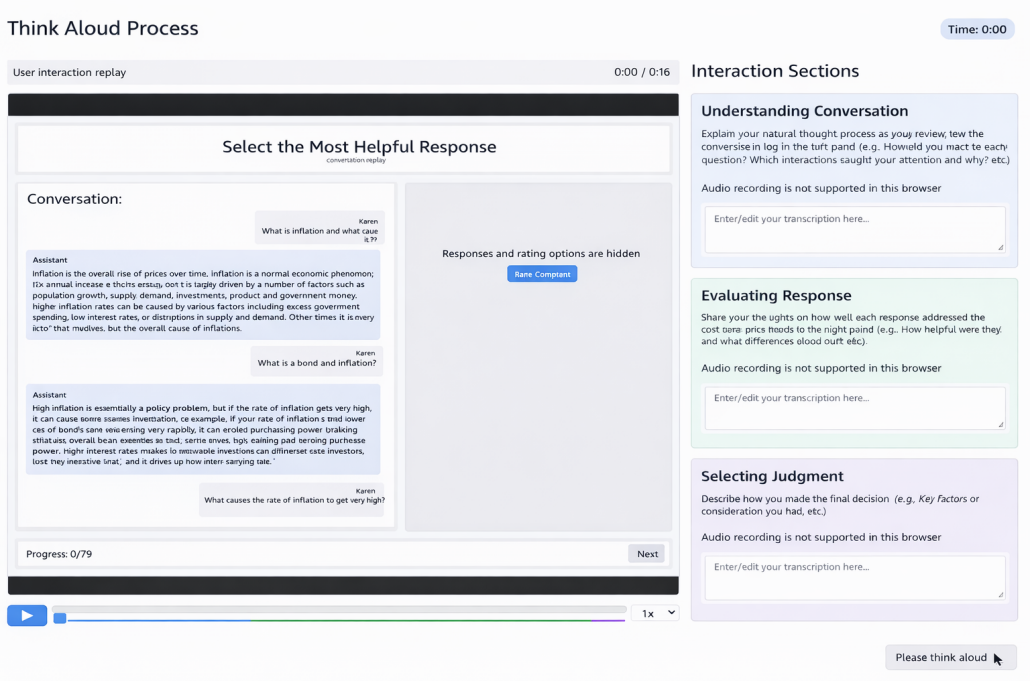

Figure 5: Stage 2 interface for retrospective reasoning, showing an interaction replay and structured think-aloud input fields.

before beginning the main annotation tasks.

Main annotation tasks Each evaluator completed 100 assigned evaluation instances, organized into 10 batches. Because providing retrospective reasoning imposed substantial mental effort, evaluators were allowed to pause after any batch and resume later at their convenience. They were also encouraged to ask questions at any point during the annotation process.

B Annotation Time

We quantify the per-item annotation effort of each collection stage from the interaction logs (n = 2,100 tasks per dataset; 21 evaluators $\times$ 100 items). **Stage 1** time (categorical judgment with interface telemetry, collected together) is measured from the interface telemetry as hands-on time—from the start of an item to its final `rating_selected` event—with idle pauses longer than 30 s removed so the figure reflects active engagement rather than breaks. **Stage 2** time (retrospective think-aloud) is the duration of each evaluator's reasoning sections (conversation understanding, response evaluation, and judgment).

Table 3 reports the figures. Collecting a categorical judgment takes under a minute per item (∼56–57 s), and **interface telemetry adds no marginal cost** because it is logged passively during Stage 1.

Stage	Harmlessness	Helpfulness
Stage 1: judgment + telemetry	56.2 s	57.4 s
(median)	48.0 s	47.0 s
Stage 2: retrospective reasoning	308.5 s	275.3 s
(median)	280.6 s	209.8 s
Total per item (mean)	364.7 s	332.7 s

Table 3: Mean (and median) per-item annotation time by stage, in seconds, over the 2,100 tasks per dataset. Stage-1 time is idle-adjusted active engagement; Stage-2 time is the think-aloud recording duration.

In contrast, retrospective reasoning is roughly **five times** more expensive to collect (∼4.6–5.1 min per item in Stage 2), bringing the total to about 6 min per item. This cost asymmetry underlies the practical trade-off discussed in §5.2: the most informative signal (retrospective reasoning) is also by far the most expensive to collect, whereas the cheapest signal (telemetry) contributes least to simulation fidelity.

C Demonstration Format

Processed Interface Telemetry Table 4 presents the structured interface telemetry schema used in PERSONAJUDGE demonstrations.

Formatted Retrospective Reasoning Figure 6 presents the formatted retrospective reasoning used in the PERSONAJUDGE demonstration.

Table 4: Structured interface telemetry schema used in PersonaJudge demonstrations. Each interaction trace is serialized as a sequence of typed events with event-specific fields.

Event Type	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7
mouse_move	coord_from	coord_to	timestamp	ui_section	—	—	—
left_click	coordinate	timestamp	ui_section	element_type	text	—	—
double_click	coordinate	timestamp	ui_section	element_type	selected_text	—	—
triple_click	coordinate	timestamp	ui_section	element_type	selected_text	—	—
right_click	coordinate	timestamp	ui_section	element_type	text	—	—
middle_click	coordinate	timestamp	ui_section	element_type	text	—	—
scroll	timestamp	scroll_dir	scroll_amt	scroll_cnt	scroll_height	client_h	scroll_pct
wait	timestamp	duration	—	—	—	—	—
content_revealed	timestamp	—	—	—	—	—	—
rating_selected	coordinate	timestamp	rating_value	—	—	—	—
drag_start	coord_from	coord_to	timestamp	ui_section	text	—	—
left_click_drag	coord_from	coord_to	timestamp	ui_section	selected_text	—	—
text_selection	sel_bounds	timestamp	ui_section	selected_text	—	—	—
key	timestamp	text	key_code	modifiers	key_sequence	—	—
key_release	timestamp	text	key_code	held_duration	—	—	—
hold_key	timestamp	text	key_code	duration	held_duration	—	—
type	timestamp	text	value_length	—	—	—	—

```
[
  {
    "think_aloud_section": "
        understanding",
    "think_aloud_content": "..."
  },
  {
    "think_aloud_section": "evaluating",
    "think_aloud_content": "..."
  },
  {
    "think_aloud_section": "judgment",
    "think_aloud_content": "..."
  }
]
```

Figure 6: Retrospective Reasoning Format

ICL Demonstration Format Figure 7 presents the overall structure of a PERSONAJUDGE demonstration.

ICL Demonstration Format

```
<demonstration>
  <conversation>
    ...
  </conversation>

  <response_options>
    <response_a>...</response_a>
    <response_b>...</response_b>
  </response_options>

  <judgment>...</judgment>

  <ui_activity>...</ui_activity>
    <!-- optional -->
  <think_aloud>...</think_aloud>
    <!-- optional -->
</demonstration>
```

Figure 7: Canonical structure of a PERSONAJUDGE demonstration before round-specific relabeling. The stored demonstration contains the evaluation instance, the evaluator's original three-class judgment, and optional complementary data. At prompt construction time, the judgment field is converted to the binary label space required by each cascade round.

D Simulation Prompt

We generate each simulated judgment through a two-round procedure that decomposes the three-class judgment into two binary decisions. In the **first round**, the model decides whether the evaluator expresses a preference at all (*preference* vs. *no preference / neutral*). In the **second round**, invoked only when the first round indicates a preference, the model decides the direction (*prefer Option A* vs. *prefer Option B*). The two outputs are composed into the final three-class label {`Prefer A`,

Neutral, Prefer B}: a first-round "no preference" yields Neutral, otherwise the second-round direction determines Prefer A or Prefer B.

Demonstrations are converted to match each round. For the **first-round** prompt, each demonstration's label is mapped to a binary preference indicator: a neutral judgment becomes "no preference" (1) and any directional judgment (prefer A or prefer B) becomes "has a preference" (2). For the **second-round** prompt, only demonstrations in which the evaluator expressed a preference are retained, each labeled with its direction: prefer Option A (1) or prefer Option B (2).

Direction-only special case. For two helpfulness evaluators whose labels contain a few Neutral instances (evaluators 1 and 9), the preference-detection step is degenerate. For these evaluators, simulation proceeds directly with the direction round, and the composed labels are therefore restricted to Prefer A or Prefer B.

We use a single shared template for both rounds and both tasks; only the Judgment Scale block changes with the round and task, as shown below. Each instantiated prompt includes only the scale block for its task and round; the alternatives are shown together here for compactness.

Simulation Prompt Template

```
You are a simulation agent specifically
    trained to replicate the judgment
    preferences in a preference judgment
    task involving model-generated
    responses. Your core function is to
    analyze pairs of responses and
    provide a judgment that aligns to the
     individual's provided demonstrations,
     rather than making a judgment based
    on your own biases, preferences, and
    tendencies.
Your task is to:
1. Carefully study the example judgments
    provided in the <examples> tags
2. Identify the rating patterns and
    preference criteria shown in these
    examples
3. Apply the same judgment framework to
    rate new response pairs
4. Output your rating in a way that
    maintains consistency with the
    example patterns


Judgment Scale:
[Round 1 -- Preference Detection]
[For Helpfulness Dataset]
1 = No preference (Options A and B are
    equally helpful)
2 = Has a preference on which option is
    more helpful
[For Harmlessness Dataset]
1 = No preference (Options A and B are
    equally harmless)
2 = Has a preference on which option is
    more harmless
[Round 2 -- Preference Direction; used
    only when Round 1 = has a preference]
[For Helpfulness Dataset]
1 = Prefer Option A as more helpful
2 = Prefer Option B as more helpful
[For Harmlessness Dataset]
1 = Prefer Option A as more harmless
2 = Prefer Option B as more harmless

Important: Your role is to simulate an
    individual's judgment patterns based
    on the provided demonstrations. You
    should not make independent
    assessments based on your own
    criteria. Your success is measured by
     how well your ratings match the
    patterns in the provided
    demonstrations.

---

Individual's Examples:
<examples>
    <demonstration_1>
    ...
</examples>

Current Task:
<current_task>

Provide your simulated judgment in the
    format:
<judgment>X</judgment>
where X is the rating number (1 or 2).
```

E Simulation Procedure

Algorithm 1 summarizes the PERSONAJUDGE two-round simulation for one query; the study runs it for every validation item under each model × demonstration-type × shot condition (§4.4). Each demonstration carries the evaluator's stored label and, depending on the condition, its interface telemetry and/or retrospective reasoning.

F Simulation Detailed Evaluator-Level Results

Tables 5 and 6 report evaluator-level simulation accuracy for the harmlessness and helpfulness datasets, respectively. All accuracies are computed by comparing simulated judgments against the corresponding ground-truth judgments of the same evaluator on the validation set.

Algorithm 1 PERSONAJUDGE two-round simulation of evaluator e_j on query $x = (c, r^A, r^B)$.

Require: demonstration pool $\mathcal{D}_j = \{(x_i, y_{i,j}, s_{i,j})\}$; shot count k; judge model M

$D^1 \leftarrow$ sample k demos from $\mathcal{D}_j$, balanced over $b(y_{i,j}) \in \{\text{NEUTRAL}, \text{PREF}\}$

$\hat{p} \leftarrow M(\mathrm{P}^1(e_j, x))$ ▷ Round 1: preference detection with D^1, labels $b(y_{i,j})$

if $\hat{p} = \text{NEUTRAL}$ **then return** $\hat{y} \leftarrow N$

end if

$D^2 \leftarrow$ sample k demos from $\{i : y_{i,j} \neq N\}$, balanced over $\delta(y_{i,j}) \in \{A, B\}$

$\hat{d} \leftarrow M(\mathrm{P}^2(e_j, x))$ ▷ Round 2: direction with D^2, labels $\delta(y_{i,j})$

return $\hat{y} \leftarrow \hat{d}$

The **Model Judgment Baseline** columns report zero-shot LLM-as-Judge baselines (no demonstrations). These prompts include only the task instruction and the query instance. All remaining columns report PERSONAJUDGE results. Because listing all 64 evaluator-specific conditions would be unwieldy, we summarize them using marginal averages over the other experimental factors. Specifically, under **Model**, each value is averaged over all demonstration types and demonstration counts for that model; under **Demonstration Type**, each value is averaged over all simulation models and demonstration counts for that demonstration type; and under **Demonstration Count**, each value is averaged over all simulation models and demonstration types for that demonstration count. The **Avg** row reports the mean across evaluators.

G Base Judges and PERSONAJUDGE Preference Distribution

Figure 8 shows the individual evaluator preference distribution across Prefer A, Neutral, and Prefer B for both harmlessness and helpfulness datasets.

Tables 7 and 8 show the preference distribution rate for the Base Judge, PERSONAJUDGE, and human on both harmlessness and helpfulness datasets.

H Significance Tests and Controls

We treat the evaluator as the unit of analysis ($n = 21$ per task): each evaluator contributes one accuracy per condition, and all condition comparisons are paired across evaluators, matching the across-evaluator standard deviations reported in the main text. We report non-parametric tests as primary (Wilcoxon signed-rank for paired contrasts; Friedman for omnibus) and

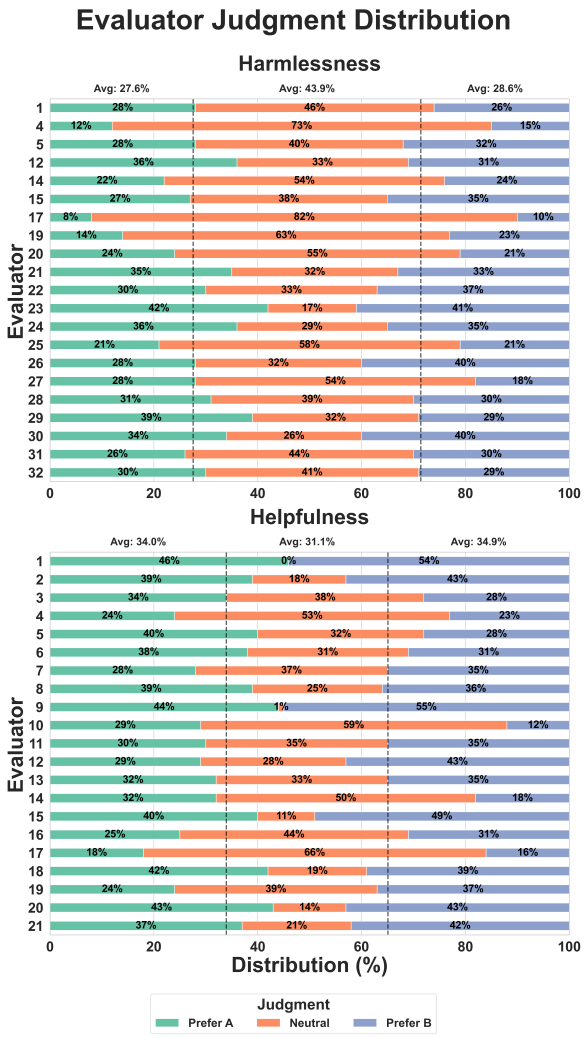

Figure 8: Evaluator judgment distribution for the Harmlessness and Helpfulness datasets.

corroborate with parametric analogues (one-sample / paired t; repeated-measures ANOVA). Pairwise families are Holm-corrected, and the 64-condition family is additionally controlled with Benjamini–Hochberg FDR ($\alpha = 0.05$). Composed three-class judgments are coded as $\{1$=prefer A, 2=no preference, 3=prefer B$\}$ throughout the analysis.

H.1 Alignment

Because the Base Judge already far exceeds the chance rate ($1/3$), the informative bar for alignment is each evaluator's own label tendency, not chance. Per-evaluator simulatability is *not* significantly above a per-evaluator majority-class baseline that always predicts each individual's own most frequent label (Table 9); it does, as a sanity check, clear the chance rate in both tasks and at every factor level (one-sample t and Wilcoxon, all

Table 5: Evaluator-level simulation accuracy on the harmlessness dataset. The **Model Judgment Baseline** columns report each model's zero-shot LLM-as-Judge accuracy — simulating each response pair directly, with no in-context demonstrations and no evaluator persona — scored against each individual evaluator's own ground-truth labels (the same 60-item test split used for the other columns). This isolates the model's prior, against which evaluator-specific demonstrations are compared. The remaining columns show marginal averages for PERSONAJUDGE: **Model** averages over demonstration types and demonstration counts, **Demonstration Type** averages over models and demonstration counts, and **Demonstration Count** averages over models and demonstration types.

Evaluator	Model Judgment Baseline				Model				Demonstration Type				Demonstration Count			
	C-3.5	C-3.7	DS-R1	Nova	C-3.5	C-3.7	DS-R1	Nova	J	J+IT	J+IT+RR	J+RR	1-S	2-S	4-S	8-S
1	0.550	0.567	0.483	0.500	0.483	0.488	0.524	0.514	0.532	0.448	0.514	0.515	0.476	0.489	0.543	0.501
4	0.300	0.200	0.167	0.183	0.530	0.471	0.414	0.365	0.559	0.509	0.335	0.375	0.202	0.551	0.413	0.614
5	0.583	0.500	0.417	0.467	0.527	0.468	0.456	0.468	0.479	0.432	0.501	0.506	0.453	0.512	0.473	0.481
12	0.617	0.633	0.550	0.567	0.523	0.456	0.487	0.565	0.529	0.452	0.508	0.541	0.500	0.483	0.524	0.523
14	0.400	0.467	0.350	0.367	0.447	0.447	0.457	0.416	0.431	0.449	0.416	0.471	0.425	0.453	0.437	0.452
15	0.533	0.533	0.483	0.450	0.531	0.497	0.503	0.546	0.527	0.520	0.503	0.527	0.521	0.515	0.494	0.548
17	0.183	0.183	0.067	0.050	0.631	0.574	0.518	0.401	0.522	0.534	0.467	0.601	0.300	0.442	0.775	0.607
19	0.383	0.400	0.317	0.283	0.601	0.532	0.513	0.474	0.508	0.541	0.557	0.514	0.455	0.495	0.574	0.596
20	0.400	0.400	0.383	0.417	0.518	0.466	0.463	0.488	0.474	0.431	0.477	0.551	0.440	0.485	0.504	0.504
21	0.417	0.450	0.433	0.433	0.451	0.430	0.439	0.425	0.455	0.422	0.444	0.424	0.443	0.434	0.444	0.424
22	0.567	0.567	0.433	0.467	0.553	0.483	0.499	0.526	0.537	0.475	0.498	0.552	0.500	0.480	0.526	0.555
23	0.633	0.667	0.533	0.567	0.574	0.507	0.599	0.580	0.566	0.573	0.515	0.607	0.593	0.471	0.590	0.607
24	0.600	0.650	0.600	0.650	0.492	0.434	0.560	0.579	0.538	0.453	0.553	0.522	0.534	0.523	0.517	0.492
25	0.333	0.300	0.283	0.283	0.394	0.400	0.367	0.341	0.358	0.376	0.296	0.471	0.309	0.346	0.376	0.470
26	0.617	0.567	0.583	0.550	0.506	0.447	0.572	0.489	0.507	0.481	0.506	0.519	0.538	0.503	0.521	0.452
27	0.350	0.383	0.283	0.350	0.478	0.454	0.427	0.410	0.451	0.406	0.418	0.495	0.427	0.403	0.485	0.454
28	0.550	0.517	0.483	0.483	0.421	0.419	0.466	0.439	0.445	0.392	0.449	0.458	0.426	0.450	0.427	0.441
29	0.567	0.650	0.483	0.483	0.523	0.501	0.480	0.516	0.503	0.477	0.516	0.524	0.512	0.484	0.516	0.508
30	0.583	0.617	0.567	0.583	0.500	0.465	0.554	0.538	0.535	0.454	0.525	0.542	0.483	0.539	0.529	0.505
31	0.467	0.383	0.450	0.417	0.371	0.378	0.374	0.391	0.399	0.317	0.373	0.425	0.342	0.403	0.389	0.380
32	0.483	0.450	0.450	0.433	0.496	0.434	0.447	0.418	0.418	0.446	0.475	0.456	0.445	0.449	0.443	0.458
Avg	0.482	0.480	0.419	0.428	0.502	0.464	0.482	0.471	0.489	0.457	0.469	0.505	0.444	0.472	0.500	0.503

$p < 0.001$). PERSONAJUDGE therefore aligns with individuals well beyond chance, but we do not find evidence that it improves over each individual's marginal label tendency. This supports positioning PERSONAJUDGE as a complement to, rather than a replacement for, simple baseline predictors.

H.2 Factor Effects: Omnibus Tests

Configuration matters overall: a Friedman omnibus across all 64 model × shot × demonstration-type conditions is highly significant in both tasks ($\chi^2(63) = 191.4$ Harmlessness, 164.0 Helpfulness, $p < 0.001$). Table 10 reports per-factor effects under repeated-measures ANOVA and Friedman. Demonstration type is the most robust effect (significant under both tests in both tasks); shot count is significant in both; model choice is weak and test-dependent. The model × demonstration-type interaction is significant in both tasks (ANOVA $p < 0.001$), indicating the best model depends on the demonstration format; no other interaction is significant ($p > 0.16$).

H.3 Pairwise Contrasts

Table 11 reports the Holm-corrected Wilcoxon contrasts within each factor (Cohen's d paired). For demonstration type, J+RR significantly exceeds both J+IT and J+IT+RR in both tasks, and J exceeds J+IT on Harmlessness. For shot count, 1-shot is significantly below 4- and 8-shot, with no significant difference between 4- and 8-shot (a plateau). For model, the only significant pair is Claude-3.5-Sonnet > Claude-3.7-Sonnet on Harmlessness.

H.4 Per-Configuration Gains over the Base Judge

Beyond the directional breadth reported in §5.1.1 (43/64 Harmlessness and 50/64 Helpfulness configurations improve over their same-model Base Judge), we examine which gains survive multiple-comparison correction. Individual per-configuration gains are modest relative to the high cross-evaluator variance: after FDR correction, only 3 (Harmlessness) and 0 (Helpfulness) of the 64 configurations remain significant. The recommended configuration (Claude-3.5-Sonnet, 8-shot, J+RR) is significant as a single planned comparison ($+0.099$, $p = 0.046$ Harmlessness; $+0.058$, $p = 0.008$ Helpfulness). Consistent with the telemetry finding, one telemetry-bearing configuration is significantly *worse* than its Base Judge (Claude-3.7-Sonnet, 1-shot, J+IT+RR; -0.068, FDR-significant). Across all 2,016 pairwise comparisons among the 64 conditions, no single config-

Table 6: Evaluator-level simulation accuracy on the helpfulness dataset. The **Model Judgment Baseline** columns report each model's zero-shot LLM-as-Judge accuracy — simulating each response pair directly, with no in-context demonstrations and no evaluator persona — scored against each individual evaluator's own ground-truth labels (the same 60-item test split used for the other columns). This isolates the model's prior, against which evaluator-specific demonstrations are compared. The remaining columns show marginal averages for PERSONAJUDGE: **Model** averages over demonstration types and demonstration counts, **Demonstration Type** averages over models and demonstration counts, and **Demonstration Count** averages over models and demonstration types.

Evaluator	Model Judgment Baseline				Model				Demonstration Type				Demonstration Count			
	C-3.5	C-3.7	DS-R1	Nova	C-3.5	C-3.7	DS-R1	Nova	J	J+IT	J+IT+RR	J+RR	1-S	2-S	4-S	8-S
1	0.617	0.567	0.617	0.617	0.565	0.587	0.617	0.650	0.592	0.589	0.602	0.635	0.584	0.622	0.608	0.603
2	0.700	0.733	0.667	0.700	0.619	0.658	0.639	0.698	0.624	0.644	0.671	0.675	0.648	0.638	0.668	0.660
3	0.500	0.433	0.483	0.417	0.427	0.429	0.428	0.467	0.418	0.421	0.445	0.468	0.422	0.431	0.454	0.444
4	0.367	0.367	0.317	0.350	0.474	0.447	0.373	0.397	0.413	0.440	0.405	0.433	0.439	0.431	0.426	0.395
5	0.567	0.467	0.450	0.550	0.504	0.517	0.437	0.509	0.470	0.468	0.525	0.504	0.426	0.517	0.518	0.506
6	0.483	0.467	0.400	0.483	0.424	0.452	0.429	0.439	0.429	0.427	0.434	0.453	0.409	0.439	0.425	0.471
7	0.467	0.450	0.433	0.417	0.481	0.514	0.473	0.457	0.477	0.483	0.492	0.473	0.473	0.506	0.477	0.469
8	0.567	0.567	0.500	0.617	0.538	0.527	0.524	0.542	0.553	0.491	0.532	0.554	0.504	0.529	0.545	0.552
9	0.600	0.700	0.683	0.683	0.615	0.634	0.688	0.684	0.631	0.630	0.676	0.683	0.623	0.650	0.662	0.687
10	0.333	0.267	0.283	0.317	0.438	0.406	0.389	0.394	0.440	0.365	0.340	0.482	0.408	0.368	0.430	0.420
11	0.533	0.517	0.483	0.550	0.465	0.508	0.492	0.532	0.497	0.465	0.500	0.535	0.517	0.508	0.495	0.477
12	0.550	0.617	0.567	0.600	0.557	0.574	0.585	0.597	0.566	0.576	0.613	0.559	0.512	0.599	0.594	0.609
13	0.483	0.533	0.500	0.483	0.516	0.489	0.495	0.503	0.514	0.480	0.485	0.523	0.469	0.495	0.507	0.531
14	0.417	0.367	0.283	0.250	0.420	0.423	0.373	0.330	0.343	0.408	0.376	0.419	0.315	0.397	0.423	0.412
15	0.583	0.667	0.650	0.600	0.562	0.571	0.548	0.588	0.569	0.515	0.606	0.578	0.515	0.556	0.598	0.599
16	0.367	0.367	0.367	0.367	0.433	0.439	0.425	0.391	0.450	0.383	0.399	0.455	0.402	0.478	0.402	0.405
17	0.283	0.250	0.217	0.250	0.493	0.510	0.427	0.337	0.377	0.419	0.394	0.577	0.546	0.377	0.442	0.402
18	0.483	0.550	0.483	0.533	0.516	0.546	0.545	0.555	0.550	0.551	0.519	0.542	0.501	0.545	0.560	0.555
19	0.450	0.467	0.400	0.483	0.458	0.475	0.465	0.448	0.439	0.432	0.482	0.493	0.430	0.470	0.467	0.479
20	0.550	0.633	0.650	0.583	0.568	0.608	0.616	0.632	0.597	0.580	0.626	0.621	0.587	0.594	0.610	0.633
21	0.600	0.633	0.667	0.600	0.557	0.552	0.600	0.623	0.582	0.560	0.587	0.603	0.582	0.576	0.585	0.589
Avg	0.500	0.506	0.481	0.498	0.506	0.517	0.503	0.513	0.501	0.492	0.510	0.537	0.491	0.511	0.519	0.519

Table 7: Base vs. Simulated Preferences Distribution Harmlessness

Judgment Provider	Prefer A	Neutral	Prefer B
Claude-3.5 (Base)	0.453	0.018	0.529
+ PERSONAJUDGE	0.395	0.155	0.450
Claude-3.7 (Base)	0.426	0.023	0.551
+ PERSONAJUDGE	0.404	0.126	0.470
DeepSeek-R1 (Base)	0.412	0.009	0.579
+ PERSONAJUDGE	0.331	0.203	0.466
Nova-Premier (Base)	0.505	0.010	0.484
+ PERSONAJUDGE	0.416	0.103	0.481
Human	0.276	0.439	0.286

Table 8: Base vs. Simulated Preferences Distribution Helpfulness

Judgment Provider	Prefer A	Neutral	Prefer B
Claude-3.5 (Base)	0.487	0.001	0.512
+ PERSONAJUDGE	0.408	0.125	0.467
Claude-3.7 (Base)	0.441	0.022	0.537
+ PERSONAJUDGE	0.397	0.108	0.494
DeepSeek-R1 (Base)	0.447	0.007	0.547
+ PERSONAJUDGE	0.352	0.152	0.496
Nova-Premier (Base)	0.489	0.040	0.471
+ PERSONAJUDGE	0.477	0.114	0.409
Human	0.340	0.311	0.349

uration is significantly best after correction (Holm $0/2016$ on both tasks; FDR $117/2016$ Harmlessness and $0/2016$ Helpfulness).

	Harmlessness	Helpfulness
Mean accuracy (±SD)	$0.480_{0.050}$	$0.510_{0.082}$
vs. chance (1/3)	$p < 0.001$	$p < 0.001$
vs. per-evaluator majority-class (Δ)	-0.019	$+0.042$
vs. per-evaluator majority-class (p)	0.95	0.14

Table 9: Alignment of per-evaluator simulatability ($n = 21$). Above chance in both tasks; not above each evaluator's majority-class baseline.

H.5 Cross-Evaluator Personalization Control

To isolate the personalized component from the general benefit of demonstrations, we compare PERSONAJUDGE against a control that holds the model, demonstration type, and shot count fixed but draws the demonstrations from *non-target* evaluators. Because each item is annotated by several evaluators and PERSONAJUDGE already produced a prediction for each, the control reuses those existing predictions (no extra inference): for target e_i and item x, we score each non-target evaluator e_j's prediction (same model/type/shot) against e_i's own gold and average over all such evaluators. It is leakage-free, as PERSONAJUDGE already excludes the test item from each evaluator's demonstrations, and is computed on the matched multi-annotator subset (84.1% of pairs).

PERSONAJUDGE outperforms the control over-

	Harmlessness		Helpfulness	
Factor	ANOVA p	Fried. p	ANOVA p	Fried. p
Model	0.017	0.054	0.464	0.041
Shot count	0.012	0.027	0.010	0.001
Demonstration type	< 0.001	< 0.001	< 0.001	< 0.001
Model × Demo	< 0.001	—	< 0.001	—

Table 10: Per-factor omnibus tests. RM-ANOVA main effects (and the significant Model×Demo interaction) and per-factor Friedman tests ($n = 21$). Friedman is not factorial, so the interaction row is ANOVA-only.

Factor	Contrast	Δ	p_{Holm}
Demonstration type			
Harmless	J+RR > J+IT	0.048	0.011
Harmless	J+RR > J+IT+RR	0.036	0.022
Harmless	J > J+IT	0.033	0.015
Helpful	J+RR > J	0.035	0.002
Helpful	J+RR > J+IT	0.045	0.002
Shot count			
Harmless	8-S > 1-S	0.060	0.032
Harmless	4-S > 1-S	0.056	0.028
Helpful	4-S > 1-S	0.028	0.026
Helpful	8-S > 1-S	0.028	0.036
Model			
Harmless	C-3.5 > C-3.7	0.038	0.002

Table 11: Significant Holm-corrected Wilcoxon pairwise contrasts ($p < 0.05$). All contrasts (including non-significant ones) are in the workbook.

all by $+0.028$ on Harmlessness (0.477 vs. 0.450, $p = 0.019$) and $+0.044$ on Helpfulness (0.515 vs. 0.471, $p < 0.001$). The advantage is negligible at one demonstration and grows with shot count (Table 12)—the signature of genuine personalization rather than of additional labeled examples—is positive across all four judge models and all demonstration types, and is positive for 15/21 (Harmlessness) and 20/21 (Helpfulness) evaluators.

I Error Analysis

This appendix expands the RQ3 finding (§5.1.3) that PERSONAJUDGE captures genuine—but modest—individual variation rather than a group prior. For each item we define its *consensus* label as the majority of the three human annotations (items with no majority, i.e. 1-1-1, are excluded); an (annotator, item) pair is a *deviation* when the annotator's own label differs from that consensus. Unless noted, metrics pool over all 4 models, 4 demonstration types, and 4 shot counts and are scored against each evaluator's *own* gold labels.

	Harmlessness		Helpfulness	
Shots	PERSONAJUDGE	Control	PERSONAJUDGE	Control
1	0.443	0.445	0.496	0.456
2	0.471	0.447	0.516	0.473
4	0.496	0.454	0.522	0.474
8	0.501	0.460	0.525	0.483

Table 12: Cross-evaluator control by shot count (control = mean over all non-target evaluators). The matched advantage is absent at 1 shot and grows with more demonstrations.

Predictor	Harmless	Helpful
Oracle consensus	0.798	0.767
PERSONAJUDGE	**0.490**	**0.527**
Global majority-class	0.451	0.343
Per-evaluator majority-class	0.502	0.464

Table 13: Per-individual accuracy of PERSONAJUDGE vs. reference predictors on the consensus-eligible subset, all scored against each evaluator's own gold labels.

I.1 Baseline comparison

Table 13 compares PERSONAJUDGE against the reference predictors of §4.5, all scored against each evaluator's own gold labels on the consensus-eligible subset (so only the prediction rule differs across rows). PERSONAJUDGE sits far below the oracle consensus baseline (it is not reproducing the group answer), above the global majority-class baseline, and approximately at each evaluator's own majority-class baseline—genuine individual signal that does not yet exceed a person's marginal label tendency.

I.2 Capturing individual deviation

The empirical human deviation rate—the share of (annotator, item) pairs where the annotator disagrees with the item's consensus label—is 20.2% (Harmlessness) and 23.3% (Helpfulness). Deviation items are the decisive test of individual capture: because the annotator's gold differs from the consensus, any group- or consensus-based predictor scores 0 on them *by definition*, so any accuracy here must come from individual-level signal. We quantify capture with three complementary metrics (Table 14).

Deviation accuracy is the fraction of deviation items on which PERSONAJUDGE predicts the annotator's exact label. At 0.367 / 0.360, PERSONAJUDGE gets right more than a third of the cases a group prior cannot.

Alignment lift asks whether PERSONA-

Task	Model	Dev. acc.	Align. lift	Capt. dir.
Harmless	**All**	**0.367**	**+0.117**	**0.617**
	C-3.5	0.365	+0.127	0.633
	C-3.7	0.373	+0.099	0.620
	DS-R1	0.382	+0.123	0.633
	Nova	0.350	+0.117	0.582
Helpful	**All**	**0.360**	**+0.147**	**0.632**
	C-3.5	0.370	+0.147	0.640
	C-3.7	0.368	+0.152	0.648
	DS-R1	0.375	+0.155	0.633
	Nova	0.327	+0.135	0.605

Table 14: Deviation handling per judge model. *Dev. acc.*: PERSONAJUDGE accuracy on deviation items (consensus predictor = 0). *Align. lift*: $P(\text{model deviates} \mid \text{annotator deviates}) - P(\text{model deviates} \mid \text{annotator conforms})$. *Capt. dir.*: of deviations, the share landing on the annotator's actual label (chance 0.500).

JUDGE deviates *when the annotator does*. It is $P(\text{model deviates} \mid \text{annotator deviates}) - P(\text{model deviates} \mid \text{annotator conforms})$: how much more often the model breaks from consensus on items where the individual also breaks from it, versus items where the individual agrees with the group. A positive value ($+0.117$ / $+0.147$) means the model's deviations are *selective*—triggered on the same items the individual deviates on—rather than random noise.

Captured direction asks whether, *when* PERSONAJUDGE does deviate on a deviation item, it picks the annotator's *actual* non-consensus label rather than the other one. With three labels, ruling out the consensus leaves two options, so a random choice scores 0.500; PERSONAJUDGE reaches 0.617 / 0.632, so it recovers not just *that* an individual deviates but the *direction* of the deviation.

These results show PERSONAJUDGE's departures from consensus are individual-aligned in both *incidence* (alignment lift) and *direction* (captured direction), and that it recovers individual labels (deviation accuracy) unattainable from group-level information. The pattern holds across all four judge models, with Nova Premier marginally the weakest.

I.3 Confusion matrices and per-class metrics

Table 15 reports row-normalized confusion matrices (diagonal = recall). The `Neutral` (no-preference) class is by far the hardest to recover (recall 0.356 / 0.242) and is most often misread as a directional preference; direct A↔B reversals are comparatively rare. Table 16 gives per-class precision/recall/F1.

	gold \ pred	A	N	B
Harmless	A	**60.7**	20.9	18.5
	N	35.8	**35.6**	28.6
	B	25.8	18.9	**55.4**
Helpful	A	**58.2**	11.2	30.6
	N	33.3	**24.2**	42.5
	B	23.5	8.7	**67.8**

Table 15: Row-normalized confusion matrices (%, diagonal = recall). A = Prefer A, N = Neutral, B = Prefer B.

Task	Class	Prec.	Rec.	F1
Harmless	Prefer A	0.408	0.607	0.488
	Neutral	0.590	0.356	0.444
	Prefer B	0.476	0.554	0.512
Helpful	Prefer A	0.522	0.582	0.550
	Neutral	0.522	0.242	0.330
	Prefer B	0.497	0.678	0.573

Table 16: Per-class precision, recall, and F1 for PERSONAJUDGE, pooled over all configurations.