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 · * Work done during Amazon internship

2026 SIGdial Atlanta, GA, USA

Read the Paper

LLM judges are trained to reproduce the **crowd's consensus** — but in open-ended evaluation, disagreement is **meaningful, not noise**. We show an LLM can simulate a **specific person's** preference judgments through in-context demonstrations built from that evaluator's own **labels, reasoning, and interface behavior**.

32	4,200	161K	2	+9.9pp	r=.73
trained annotators	human preference judgments	simulations (4×4×4 design)	complementary evaluator signals	best gain over zero-shot Base Judge	stable cross-task neutral trait

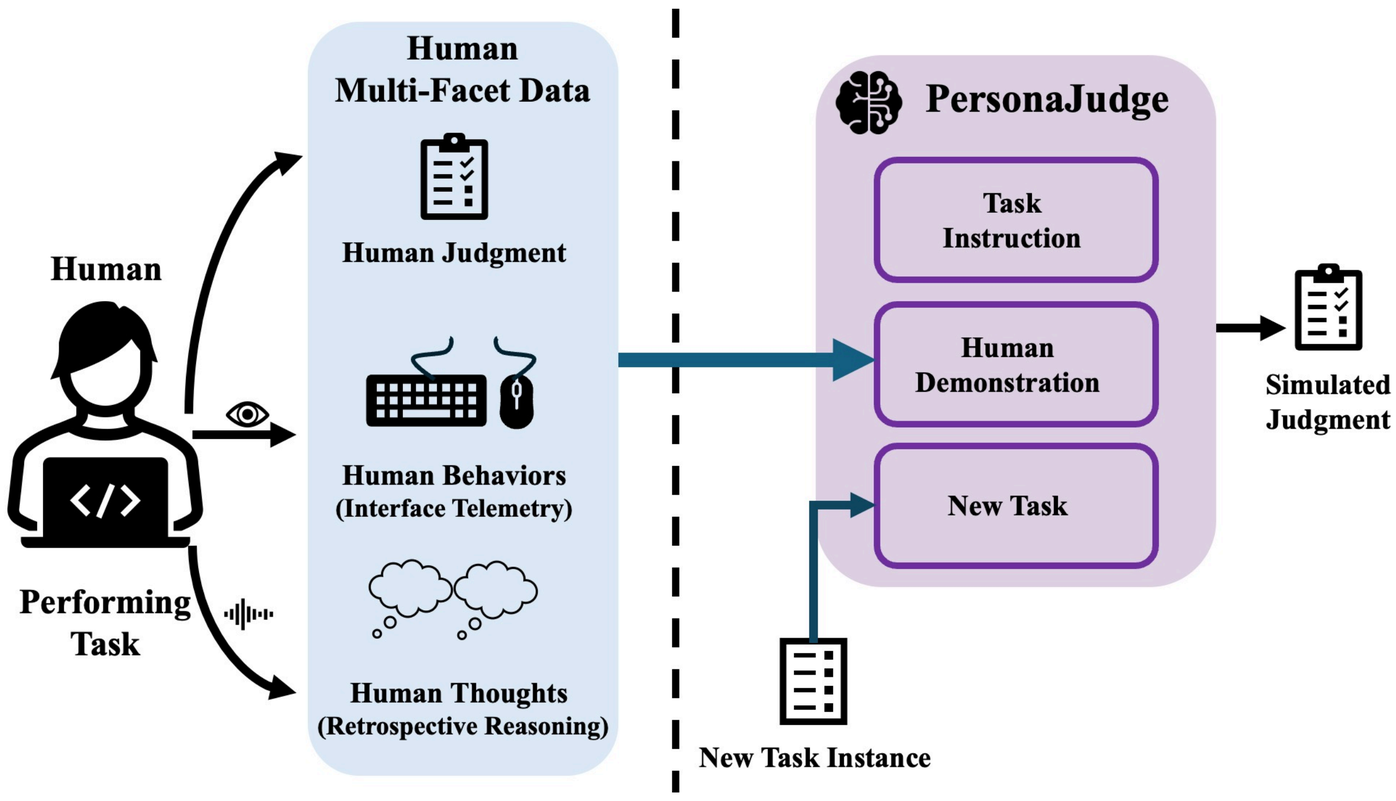
1 Consensus ≠ the Individual

LLM-as-Judge scores against pooled preferences that collapse across annotators — it simulates the crowd, not any one evaluator.

- In open-ended tasks a single standard may not exist; annotators apply **different criteria**, and disagreement — including **Neutral**, the modal label on Harmlessness (43.9%) — is **signal, not noise**.
- Persona / demographic proxies summarize a person — not their **own decision traces**.

Our question: given an evaluator's history and decision traces, can an LLM predict how **that person** judges a new instance?

2 PersonaJudge: Multi-Facet In-Context Learning



Each judgment pairs a **primary judgment (J)** with **two complementary signals (IT, RR)** — the evaluator-specific in-context demonstrations:

① Categorical Judgment (J)
Prefer A / Neutral / Prefer B

② Interface Telemetry (IT)
How they inspected the task
e.g., clicks, dwell, reveal order.

③ Retrospective Reasoning (RR)
Why they decided
A post-task think-aloud.

An LLM turns **task instruction + evaluator-specific demonstrations + a new item** into a **simulated judgment**, via a **two-round cascade** (preference vs. neutral → direction A / B).

3 Study Setup

DATA COLLECTION

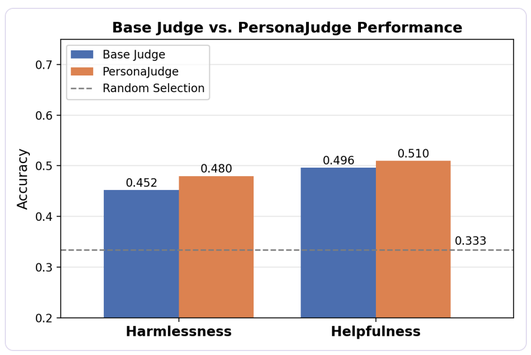
- **Data:** Anthropic HH — Helpfulness & Harmlessness, 700 conversations each.
- **People:** 32 trained annotators — 21 per task (10 did both) × 100 judgments → 2,100 per task, 4,200 total.
- **Two stages:** provide judgment (interaction logged), then replay & think aloud.

STUDY DESIGN

- **4×4×4 factorial:** demo type {J, J+IT, J+RR, J+IT+RR} × shots {1, 2, 4, 8} × models {Claude 3.5 Sonnet, Claude 3.7 Sonnet, DeepSeek-R1, Nova Premier}.

Metric: 3-class accuracy vs. each person's *own* label; baselines Random & per-model **Base Judge** (zero-shot).

4 RQ1 — Personalization Beats the Base Judge



Averaged over all 64 configurations. Best single configuration below.

Best config (Claude 3.5 Sonnet, 8-shot J+RR): **+9.9 pp** Harmlessness · **+5.8 pp** Helpfulness over Base Judge — significant on both.

Cross-evaluator control: demos from a *different* person help far less → genuine **personalization**, not just more examples.

5 RQ2 — Reasoning Helps; Telemetry Can Hurt

Avg. accuracy by demonstration type (all models & shots):

Demonstration type	Harmlessness	Helpfulness	
J + RR	**0.505**	**0.537**	▲ best
J only	**0.489**	**0.501**	
J + IT + RR	**0.469**	**0.510**	
J + IT	**0.457**	**0.492**	▼ below J
Base Judge	0.452	0.496	zero-shot

- **Reasoning > telemetry** — robust & significant.
- **Cost twist:** reasoning costs **~5×** more time than telemetry — yet telemetry *reduces* fidelity.
- **Scaling:** gains **plateau after 4 shots**; model choice is minor.

6 RQ3 — Who Is Hard to Simulate?

- **Wide variation:** per-evaluator accuracy 0.375–0.565 (Harmlessness), 0.386–0.655 (Helpfulness).
- **Works where consensus can't:** on **deviation items** (~20–23%, where group predictors score 0 by construction), PersonaJudge recovers the person's exact label **~36%** of the time.
- **But modest:** it doesn't beat a per-person majority baseline → **complements consensus judges, doesn't replace them.**
- **Predicts difficulty:** heavier **Neutral usage** (r = −.89 Helpfulness) and **divergence from consensus** (r = −.68 Helpfulness) make an evaluator harder to simulate.

What transfers across tasks: not simulation fidelity (r = 0.18) but the evaluator's **Neutral-usage tendency** (r = 0.73) — the stable trait is judgment *style*.

Key Takeaways & Implications

01 · Feasible
Individual-level simulation is **real & meaningful**: evaluator-specific demos improve over non-personalized judges.

02 · Reasoning wins
Collect **why**, not just **how**. More signals ≠ better — cheap telemetry can **hurt**.

03 · Who's hard to model
People who often pick **"Neutral"** or **disagree with the group** are the hardest to simulate.

04 · A shift
LLM-as-Judge moves from **pooled labels** → simulating **how individuals evaluate**.

Limitations: pairwise HH only · 32 annotators · in-context learning (no fine-tuning) · retrospective reports are post-hoc accounts.

PennState · Columbia University in the City of New York