RE-CENTERING HUMANS IN LLM PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite growing interest, most evaluations of large language models' (LLMs') personalization abilities have relied on synthetic data. It remains unclear how well current personalization systems work for real users. In this paper, we study the gap in LLM personalization performance when using synthetic versus human data. We collect human conversations (550 conversations) and judgments across three stages of personalization: extracting user attributes from conversations (5,949 judgments), pairing relevant attributes with new prompts (11,919), and incorporating relevant attributes into a personalized response (1,101). Incorporating human data reveals system limitations at each stage. Models struggle to extract attributes from human conversations, disagree with human judgments on relevant attributes, and generate personalized responses that humans judge no better than generic responses (though that LLM judges widely rate as better). We introduce several lightweight training-based interventions that shift automated personalization evaluation closer to human data in all three stages. However, our exploration in Stage 3 suggests that a one-size-fits-all model may obscure meaningful differences in personalization preferences across users. Our collected data provides a foundation for studying how models should extract, select, and incorporate user information in ways that humans find useful.

1 INTRODUCTION

Large language models' (LLMs') ability to personalize responses to different users is becoming increasingly sought-after as models are deployed to more people across more domains (Shaikh et al., 2025). Despite growing interest, it remains unclear how well current personalization systems work for real users. Past efforts evaluating models' personalization abilities have predominantly relied on synthetic data: user personas (Jiang et al., 2025), simulated conversations (Kim et al., 2025), and LLM-based evaluations (Zhao et al., 2025b). While synthetic data enables large-scale experimentation, synthetic responses deviate significantly from human ones (Naous et al., 2026; Mehri et al., 2026b). It is not clear if model performance on synthetic data maps neatly to real human experience. This tension highlights the central difficulty of personalization: the target of personalization is a human user, whose attributes and judgments on system effectiveness are often implicit, noisy, incomplete, and context-dependent (Hu et al., 2008; Shaikh et al., 2025).

In this paper, we study the gap between synthetic and human data for evaluating LLM personalization. We frame personalization as a three-stage pipeline based on prior evaluations: models must (1) infer stable user attributes from prior conversations (user attribute extraction), (2) select the attributes relevant to a specific response (attribute relevance matching), and (3) incorporate relevant attributes to produce a response that improves over a generic one (personalized response generation). For each stage of our pipeline, we compare human and synthetic data. Our dataset contains 47 real users and 550 conversations drawn from existing conversational datasets, paired with human judgments for all three stages: 5,949 judgments on extracted user attributes, 11,919 on attribute–prompt pairings, and 1,101 on response preference judgments for personalized generation. Using our dataset, we compare model performance and alignment with human data and judgments.

We find that incorporating human data reveals model limitations at each stage of personalization. Models struggle to accurately extract attributes from human conversations compared to synthetic ones: an additional 22% of extracted user attributes from real conversations are judged problematic. Once attributes are extracted, models again face difficulties in pairing attributes with prompts: despite strong inter-LLM agreement, LLMs are misaligned with humans on relevance matching,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

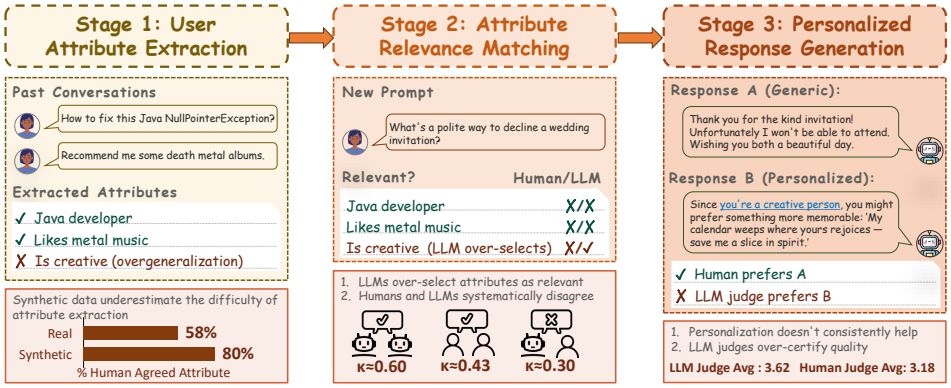


Figure 1: Overview of our three-stage personalization framework. **(1) User Attribute Extraction** from conversation history, **(2) Attribute Relevance Matching** for the current interaction context, and **(3) Personalized Response Generation** based on selected attributes. By incorporating human-grounded data, we reveal limitations at each stage of personalization: models struggle to extract accurate attributes from human conversations, pair those attributes with new prompts, and determine what makes a personalized response effective.

over-identifying 20–40% more attributes as relevant. Finally, even when relevant attributes are provided, humans disagree with models on what constitutes an effective personalized response: as generators, LLMs produce personalized responses that humans judge as no better than generic ones in 54.6% of cases, while as judges, LLMs assign inflated scores that remain poorly aligned with human preferences.

We introduce three lightweight interventions that shift automated personalization evaluation closer to human data across all three stages. For attribute extraction, a lightweight RoBERTa (Liu et al., 2020) verifier trained on human annotations can serve as an effective safeguard before extracted attributes are used downstream. For relevance matching, training-based methods, including supervised classification and GRPO (Shao et al., 2024), reduce over-selection and better align relevance judgments with humans. For our third stage of personalized response generation, training reward models substantially improve alignment with human ratings. However, we also discover that a one-size-fits-all model may obscure meaningful differences in personalization preferences across users.

Overall, our results show that personalization presents unique difficulties for automated and synthetic evaluations. We hope our collected data enables deeper evaluation of personalization systems and provides a foundation for studying how models should extract, select, and incorporate user information in ways that humans find useful.

2 A Three-Stage Framework for LLM Personalization

LLM personalization spans a spectrum from monolithic to decomposed approaches. End-to-end benchmarks evaluate personalization quality as a single black-box task, making it difficult to diagnose where a system fails (Zollo et al., 2025; Zhao et al., 2025b; Kim et al., 2025; Zhao et al., 2025a). A parallel line of work argues for more explicit decomposition (Wang et al., 2023; Zhuang et al., 2024; Shi et al., 2025; Xu et al., 2026; Du et al., 2026; Mehri et al., 2026a), showing that factorizing personalization into stages such as retrieval, ranking, and personalized generation yields consistent gains over monolithic approaches.

We frame LLM personalization as a three-stage pipeline (Figure 1): **(1) user attribute extraction** from conversation history, **(2) attribute relevance matching** for the current interaction context, and **(3) personalized response generation** based on the selected attributes. This framework is based on prior personalization efforts (e.g., relevance matching is inspired by structured memory lookup, Salemi et al., 2024), though we narrow our focus to initial response personalization (i.e., given a user and their previous conversations, personalize the next response), and leave subsequent personalization stages, such as updating outdated memories (Maharana et al., 2024), to future work.

Stage 1 — Attribute extraction. We define a **user attribute** as a stable, long-term, and context-independent statement about the user, including both user preferences and profile information (e.g., *"user prefers bullet-points"*, or *"user is a Java developer"*). We focus on attributes that are unlikely to change within a single conversation as an initial step to formalizing our pipeline. This definition follows a common practice in personalization and memory systems, where past interactions are synthesized into persistent user profiles (similar to personas) that can later be retrieved for personalized generation (Zhong et al., 2024; Kang et al., 2025).

Stage 2 — Relevance matching. We define **attribute relevance** as whether a user attribute should influence the model's response to a specific prompt. Prior systems often approximate this stage as a retrieval problem, selecting user information through semantic similarity or structured memory lookup (Salemi et al., 2024; Wu et al., 2025; Sun et al., 2025). However, semantic similarity does not necessarily imply personalization relevance, and lexically distant attributes may still affect how a response should be framed (Okite et al., 2026). We therefore treat relevance selection as a reasoning problem, where the system must decide which attributes would meaningfully improve the response, which should remain unused, and how attributes should affect the response without introducing irrelevant or excessive personalization.

Stage 3 — Personalized response generation. Once relevant attributes have been selected, the model must determine *how* to use them in the final response. We define successful **personalized generation** as producing a response that improves over a generic response. Existing work often acknowledges related concerns, but rarely directly studies the quality of this stage. Instead, most work uses adjacent metrics such as task performance, preference alignment, or profile consistency (Salemi et al., 2024; Jiang et al., 2025). We therefore evaluate not only whether selected attributes affect the output, but also whether they are used in a contextually appropriate and user-acceptable manner.

3 Re-Centering Humans in the Personalization Pipeline

Although humans are the subjects of personalization, most existing work proxies the role of human input with synthetic personas, simulated users, synthetic dialogues, or LLM-based judges (Salemi et al., 2024; Jiang et al., 2025; Liu et al., 2025b; Zhao et al., 2025b). While these designs enable large-scale evaluation, it is not clear that current approaches can faithfully simulate human users (Ivey et al., 2024; Dong et al., 2024; Naous et al., 2026; Mehri et al., 2026b). As a result, existing personalization benchmarks may inject systematic bias or overestimate system capabilities. We argue that progress on LLM personalization requires re-centering humans in the personalization pipeline by: (1) grounding data in real human interactions (§4), and (2) judging personalization quality with humans (§5 and §6). In each of the sections below, we illustrate how incorporating human conversations and judgments highlights current limitations in personalization evaluation.

Models In this study, we mainly focus on five recent and widely-used open-weight and proprietary LLMs. The open-weight LLMs include Llama-3.3-70B (Grattafiori et al., 2024), Qwen3.5-27B (Qwen, 2026), and Gemma-4-31B (Google DeepMind, 2026); the proprietary LLMs include Claude-Sonnet-4.6 and GPT-5.4. These open-weight LLMs have shown strong quality-cost trade-offs, making them practical choices for large-scale experiments.

4 Grounding Personalization to Human Interactions (Stage 1)

We investigate how attribute extraction (Stage 1) differs when using human or synthetic conversations. We find that human conversations contain richer and more diverse attributes, and extracting attributes from human conversations is more error-prone than synthetic ones.

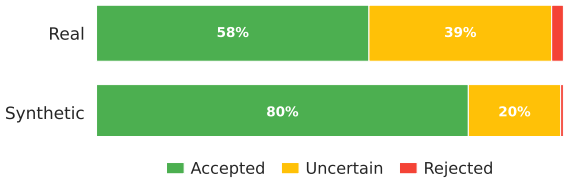
4.1 Experiment

Data We use WildChat (Zhao et al., 2024) as the base source for human conversations and cluster multiple conversations by user IP[1], resulting in 98,334 users. We filter to only English conversations

[1]The use of hashed IP addresses as approximate user identifiers follows prior work on WildChat (Zhao et al., 2024; Naous et al., 2026).

Dataset	Avg. Inter-user Dist. ↑
Synthetic baselines	
PrefEval (Zhao et al., 2025a)	0.689
CUPID (Kim et al., 2025)	0.649
PersonaLens (Zhao et al., 2025b)	0.706
Real data	
WildChat (Zhao et al., 2024)	0.656
WildChat (Filtered)	**0.737**

(a)



(b)

Figure 2: (**left**) Diversity of different datasets, calculated by average pairwise inter-user cosine distance among the sentence embeddings of all extracted user attributes. While unfiltered real data does not exhibit a clear diversity advantage, a simple preprocessing yields substantially higher diversity than all synthetic baselines. (**right**) Human annotation of extracted user attributes from synthetic and real conversations, categorized into accepted, uncertain, or rejected. Compared with synthetic datasets, real conversations yield a lower acceptance rate and higher proportions of uncertain and rejected judgments, indicating that reliable attribute extraction is more challenging on real data.

from active and authentic users,[2] yielding 16,573 users in total. For synthetic data, we include three recent personalization datasets: CUPID (Kim et al., 2025), PrefEval (Zhao et al., 2025a), and PersonaLens (Zhao et al., 2025b). We ensure that the sampled users from different datasets have similar amounts of interaction history and conversation lengths to enable a fair comparison.

Experiment Setting To extract user attributes, we split each user history into shorter chunks that fit within all LLMs' context windows (50K tokens) and prompt an LLM to list user-related attributes expressed or implied in each chunk, along with a reasoning and confidence score for each extracted attribute. We use Llama-3.3-70B as our backbone model for our attribute extraction experiments.[3] We filter out attributes with confidence scores below 0.4, and use an agglomerative clustering algorithm to merge overlapping attributes. Full implementation details are in Appendix B.

Human Annotation We ask annotators to evaluate extracted attribute quality by labeling each attribute as accepted, uncertain, or rejected. We sampled 77 users in total, including 47 WildChat users and 10 users from each of the three synthetic datasets for the annotation task. We recruit annotators in two rounds. First, we launch a pilot study with 250 attributes from 10 users, and recruit five annotators on Prolific[4] who were fluent in English and met standard platform quality criteria. We assess annotation quality by checking each annotator's label distribution and inter-annotator agreement, with author inspection for potential outliers. We find that four of the five annotators produce reasonable annotations, with an average Cohen's κ of around 0.35. We then recruit three of these annotators for the full study, covering 1,983 attributes from 77 users. The recruited annotators achieve a Cohen's κ of 0.314. More details are in Appendix G.

4.2 Results

Real users can be more diverse than synthetic alternatives A central goal of many synthetic conversation datasets is to approximate the diversity of real user data (Jiang et al., 2025). In our study, we define diversity as the average pairwise inter-user cosine distance among the sentence embeddings (`all-MiniLM-L6-v2`; Reimers & Gurevych, 2019) of all extracted user attributes, where higher values indicate greater variation across users. However, our experiments reveal that **uniformly sampling users from WildChat does not necessarily produce higher diversity than recent synthetic datasets** (Table 2a). This is partly because WildChat contains many homogeneous users (e.g., a large proportion of users only talk about coding), and many users do not exhibit deep model usage (e.g., short-term LLM users).

[2] We filter users with enough (at least 3 conversations and 15 turns) interaction history and with no signs of automated or malicious usage (LLM judged). Full preprocessing details are in Appendix A.

[3] We found Llama-3.3-70B to be a competitive choice for this task, as it performed best among the open-source models we tested based on manual inspection. We do not use closed-source models because some extremely long conversations can make large-scale extraction prohibitively expensive.

[4] https://www.prolific.com/

Nevertheless, **real data provides a large enough pool from which a more diverse subset can be selected**. We obtain a subset of 5,000 WildChat users that is substantially more diverse than all synthetic datasets we compare against (highlighted in green in Table 2a) by applying basic filtering and diversity-based sampling (details in Appendix C, examples of sampled users in Appendix Table 1).

Real conversations make attribute extraction more challenging Human conversations are often noisier than synthetic equivalents (Naous et al., 2026), will this impact a system's ability to reliably extract attributes? We use the 1,983 human-annotated attributes of 77 users from three annotators described above to compare attribute validity across real and synthetic conversations. As shown in Figure 2b, attributes extracted from human conversations contain substantially higher proportions of uncertain and rejected cases than those extracted from the synthetic benchmarks. This suggests that **extracting attributes that humans find valid is considerably more challenging from real conversations, and that synthetic data may underestimate the difficulty of this stage.**

To analyze these disagreements, we use GPT-5.4 to categorize the reasons behind all 1,225 uncertain attributes by annotators (details in Appendix D). The most common failure mode is *overgeneralization* (53.9%), where narrow evidence is extrapolated into a stable user trait, such as inferring that a user "is learning French" from one translation request. Other common cases include *missing evidence* (20.3%), where the attribute is not supported by the visible excerpt, and *task-context confusion* (16.1%), where task content is mistaken for user information, such as inferring that the user "has five years of marketing experience" from a fictional cover-letter prompt.

4.3 Analysis: Attribute Verification as a Lightweight Safeguard

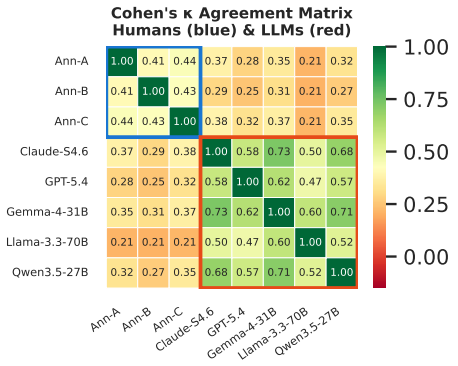
To address the attribute extraction challenge posed by human conversational data, we propose a simple post-extraction verification and optimization step. Instead of directly treating extracted attributes as reliable user profiles, personalization systems can first verify whether each attribute is sufficiently supported by the conversation history, then ask the extractor to revise or remove unsupported attributes before downstream personalization. We evaluate three verifiers: (1) zero-shot LLM prompting with the same instruction given to annotators (Appendix K.6), (2) LLM prompting with an optimized prompt (Appendix K.7) based on our error analysis in §4.2, and (3) a RoBERTa classifier trained on our human annotations (training details in Appendix E).

Appendix Table 3 shows three key findings. First, **optimized verification prompts based on error analysis substantially improve the verifier's ability to identify unsupported attributes**. This is useful because recall is especially important for verification: unsupported attributes missed by the verifier will be passed downstream as reliable user profiles, whereas supported attributes incorrectly flagged by the verifier can still be preserved or softened during the subsequent refinement step. The optimized prompt significantly improves recall for most models, suggesting that explicit instructions about over-generalization and insufficient evidence help verifiers catch more problematic attributes. Second, **LLM verifiers exhibit clear cross-model differences after prompt optimization**. Before optimization, most models perform relatively similarly, but after optimization, stronger closed-source models such as GPT-5.4 and Claude achieve much higher recall, while Llama-3.3-70B remains weak; Gemma-4-31B reaches an F1 score close to GPT-5.4 and Claude, but with lower recall. Third, **a supervised RoBERTa verifier provides the strongest practical trade-off between precision and recall**. It achieves the highest overall F1 while maintaining strong recall, suggesting that a small verifier trained on human annotations can provide a reliable and lightweight safeguard before extracted attributes are used for personalization.

We further use the RoBERTa verifier as a reflection signal for attribute optimization. Attributes flagged as unsupported are sent back to the extraction model for revision. In a small-scale study on 250 attributes, this step increases the average human acceptance rate from 58% to above 90%. This suggests that many extraction errors are recoverable with a simple verify-and-refine step.

5 Align Personalized Relevance Selection with Human (Stage 2)

Once attributes have been extracted (Stage 1), personalization requires selecting which attributes are relevant to a specific prompt. Here we test humans' and LLMs' agreement on relevance judgments.



(a)

Model	Acc	Precision	Recall	F1
Llama-3.3-70B	0.529	0.272	**0.979**	0.426
Qwen3.5-27B	0.689	0.345	0.832	0.488
Gemma-4-31B	0.685	0.355	0.934	0.514
GPT-5.4	0.676	0.334	0.819	0.474
Claude-S4.6	0.719	0.372	0.838	0.515
RoBERTa	0.859	0.608	0.605	0.606
Qwen3-4B	0.530	0.267	0.942	0.417
+ GRPO	**0.870**	**0.611**	0.674	**0.641**

(b)

Figure 3: (**left**) Full pairwise Cohen's κ between all three human annotators and five LLMs. Human–human agreement (top-left block) is consistently higher than LLM–human agreement (off-diagonal), confirming that LLMs are not reliable proxies for human relevance judgment. (**right**) Performance on attribute relevance selection against human majority judgments. Zero-shot LLMs achieve high recall but low precision, indicating substantial over-selection of relevant attributes. Training-based methods substantially improve precision and accuracy, with RL-trained Qwen3-4B achieving the best F1.

5.1 Experiments

Data We use LIMA (Zhou et al., 2023), a high-quality human instruction dataset for SFT, as the source of dialogue prompts to pair with our extracted attributes. We randomly sample 41 prompts and pair them with the 47 WildChat users whose attributes were annotated in Stage 1. Rather than pairing prompts and users completely at random, we first use Llama-3.3-70B as a coarse filter over all possible prompt–user combinations, and remove combinations with fewer than two potentially relevant attributes. We then sample 4 users for each prompt from the retained pairs. The selected attributes (3,973 in total) are then shuffled and presented to human annotators and LLM judges for relevance annotation.

Experiment Setting We formulate relevance selection as a binary judgment task and compare human and LLM judges. Given a prompt and a candidate user attribute, both humans and LLMs are asked to decide whether the attribute should be considered when personalizing the response. An attribute is labeled relevant if it would affect the response in any way (either explicitly mentioned or implicitly impacted) and irrelevant only if the response would be unchanged without it. We collect annotations from three human annotators (details below) and five LLMs (§3).

Human Annotation We collect human relevance judgments in two rounds. In the pilot round, we invite seven annotators to label 631 attribute–prompt pairs. We assess annotation quality by their yes/no label distribution and pairwise agreement, and remove two annotators with distributions far from the other annotators after manual inspection of their responses. We then invite three of the annotators, whose pairwise Cohen's κ falls in the 0.4–0.6 range, to complete the full study of 3,973 attribute–prompt pairs from 41 prompts. More details and the annotation interface are in Appendix G.

5.2 Results

Humans and LLMs systematically disagree on relevance selection Humans and LLMs each show substantial internal agreement but disagree with each other. As shown in Figure 3a, the three human annotators achieve an average pairwise Cohen's κ of 0.426, which is substantial given the subjective nature of the task (Da San Martino et al., 2019; Landis & Koch, 1977). Surprisingly, LLMs agree even more strongly with one another, reaching an average κ of 0.597 across diverse open-source and closed-source models. However, LLM–human agreement is much lower, with an average κ of only 0.300. One clear source of this misalignment is that LLMs assign substantially higher relevance rates: each human annotator marks only around 20% of attributes as relevant, whereas

LLMs mark 40–60% as relevant across models. This pattern suggests that both humans and models form internally consistent judgments, but LLMs' decision boundaries are substantially misaligned with humans, making them unreliable proxies for human relevance judgment.

Attribute relevance cannot be reduced to semantic similarity. As discussed in §2, most existing work treats relevance selection as a retrieval problem or constructs personalization data based on semantic relatedness, which can underestimate the difficulty of this stage. To test this, we compare BM25 (Robertson & Zaragoza, 2009) and sentence-embedding similarity as retrieval-based baselines, and use the majority vote[5] of the three human annotators as the ground truth. Appendix Figure 6 reports the F1 score of each method. BM25 and semantic similarity achieve only 0.243 and 0.384 respectively, both substantially below the performance of LLM judges. This suggests semantic relatedness is insufficient for connecting attributes with a response.

Among LLM judges, recall is consistently high, indicating that models tend to over-select attributes as relevant. However, their precision remains low, with all models below 0.4. Claude-S4.6, Gemma4, and Qwen3.5 achieve the top three F1 scores, while GPT-5.4 surprisingly falls behind these smaller open-source models. This aligns with our qualitative observations: GPT-5.4 tends to personalize on a broad range of attributes, similar to Llama3.3-70B, another low-performing model.

5.3 Analysis: Aligning Attribute Relevance Selection with Better Reasoning

To address the relevance selection challenge identified in Section 5, we explore whether models can be better aligned with human judgments on which attributes should actually influence a response. As shown earlier, more than 60% of the attributes that LLMs mark as relevant are not considered relevant by humans, which can introduce unnecessary or distracting personalization. We noticed that prompt refinement brings only marginal gains for this alignment problem. Therefore, we evaluate two training-based approaches: a RoBERTa classifier trained as a binary relevance predictor, and a Qwen3-4B reasoning model optimized with GRPO. Training details for both methods are in Appendix F.

Appendix Table 3b shows two main findings. First, **training-based alignment is more effective than zero-shot prompting**. Both RoBERTa and GRPO improve over all zero-shot LLMs in F1 and precision, suggesting that supervised training can better calibrate relevance selection toward human judgments. Second, **GRPO achieves the strongest performance among the methods we evaluate**. Compared with the base Qwen3-4B model, GRPO increases F1 and precision from 0.417/0.267 to 0.641/0.611, which is even higher than RoBERTa. This suggests that relevance selection may benefit from training LLMs to explicitly reason about whether an attribute should meaningfully affect the response.

We include an example in Appendix Table 6 to show how GRPO changes the model's relevance reasoning. The untrained Qwen3-4B marks *"the user is creative"* as relevant to a simple factual question, arguing that the response could *"explain in a way that connects to creative processes."* After GRPO, the model instead reasons, *"Wait, the request asks for a basic definition, this attribute should not change the factual answer."* This more cautious decision boundary better matches human judgments and reduces irrelevant attribute use before response generation.

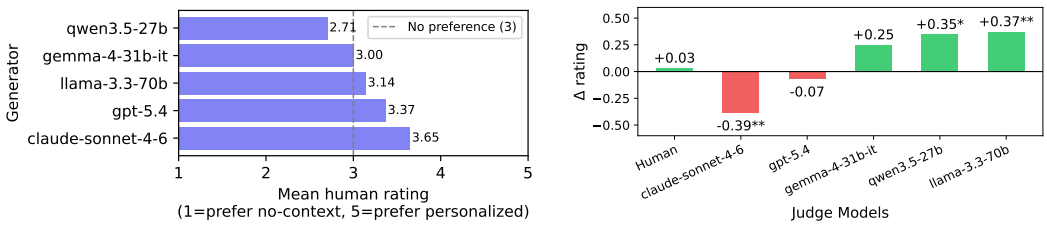
6 Human-Judged Personalization Quality (Stage 3)

We next investigate how humans and models judge personalization quality (i.e., how personalized a response is) given a set of relevant attributes.

6.1 Experiments

Experiment Setting For Stage 3, we use relevant attributes identified by human majority vote in Stage 2 as ground-truth inputs, and ask the five LLMs to generate responses conditioned on these attributes. For each instance, we present both LLM and human judges with two randomly shuffled responses: one generic (i.e., generated without personalization) and the other generated with ground-truth attributes. Annotators and LLMs are asked to judge which response they prefer on a five-point

[5]We use majority vote as a human-consensus label, following common practice in crowdsourced annotation (Snow et al., 2008; Davani et al., 2022); this label achieves Cohen's $\kappa > 0.7$ with each individual annotator.



(a) (b)

Figure 4: (**left**) Mean human rating for each LLM generator. Human judgments show that personalized generation brings only marginal benefits to user experience even for recent models, and can even degrade response quality for some open-source models. (**right**) Each judge's sensitivity to explicit attribute invocation, measured as the mean rating gap between explicit-mention and no-mention responses. Open-source LLMs show high sensitivity to explicit attribute mentions, while human judges and GPT-5.4 show little sensitivity, and Claude-S4.6 instead favors more implicit personalization ($^{*}p < 0.05$, $^{**}p < 0.01$ for Δ).

likert-style scale for each attribute, ranging from strongly preferring non-personalized response to strongly preferring personalized response. Prompts and parameters are in Appendix H.

Human Annotation We invite the same three annotators from Stage 2 to judge personalized response quality. We measure inter-annotator agreement using both Spearman correlation and weighted Cohen's κ, which gives 0.325 and 0.310 respectively. Given the subjective nature of personalization preference, this indicates reasonable annotator consistency while still allowing for individual variation. More details about annotation are in Appendix G.

6.2 Results

Personalized responses do not consistently lead to improved response quality for users Based on collected human preference ratings on each LLM's generation, 54.6% of personalized responses are judged by humans as no better than their generic counterparts (rating ≤ 3.0). As shown in Figure 4a, even the best-performing proprietary models' responses (GPT-5.4 and Claude-S4.6) are only slightly above the neutral baseline of 3.0, while open-source models like Qwen3.5 and Gemma-4 even degrade responses after personalization. This highlights the importance of cautious personalization: models must not only select relevant attributes accurately, but also decide carefully how to incorporate them into the generated response.

LLM judges tend to overestimate personalization quality while only partially aligning with human preferences. As shown in Table 4, humans assign lower ratings than all LLM judges, indicating that automated evaluation often overstates the benefit of personalization. Different LLMs' Spearman correlations with human ratings also vary substantially, with the strongest judges reaching only about $\rho = 0.35$. **This modest correlation and systematic score inflation suggest that LLM judges remain unreliable substitutes for human evaluation.**

One reason is that LLM judges often over-reward visible personalization. A common failure is *mechanical attribute invocation*, where a response explicitly mentions a user attribute, such as *"Given your interest in machine learning..."*, without meaningfully adapting the content. Such responses may look personalized to LLM judges but feel robotic, presumptuous, or intrusive to users.

To test this, we use GPT-5.4 to identify responses that explicitly mention relevant attributes and measure each judge's rating gap between explicit-mention and no-mention responses. As shown in Figure 4b, several open-weight judges assign higher scores to explicit mentions, while human annotators and GPT-5.4 show little sensitivity. Claude-S4.6 instead shows a significant negative gap, suggesting a preference for more implicit personalization.

We further ask whether models that explicitly mention user attributes more often as generators also reward such mentions more strongly as judges. As shown in Figure 5, the two behaviors are strongly correlated (Spearman $r = 0.90$, $p = 0.04$): **models with higher explicit mention rates tend to assign larger rewards to explicit mentions during evaluation**,

suggesting that surface-level personalization preferences may transfer from generation to evaluation. Claude-S4.6 is a notable counter-directional case: it has the lowest explicit mention rate as a generator (only 5%), and is the only judge that significantly penalizes explicit mentions. However, this penalty appears too strong relative to human judgments: human annotators show almost no gap between explicit-mention and no-mention responses ($\Delta = 0.03$), while Claude-S4.6 gives explicit mentions much lower scores ($\Delta = -0.39$). GPT-5.4 is closer to humans on this dimension, with only a small negative gap ($\Delta = -0.07$). Overall, these results suggest surface-level personalization behaviors may carry over from generation to evaluation, making LLM judges unreliable proxies for human preferences.

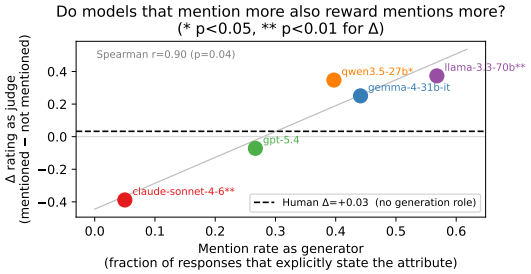


Figure 5: Relationship between a model's frequency to explicitly mention attributes when personalizing responses (x-axis) and its tendency to reward such mentions when judging responses (y-axis). The very high correlation suggests that models that mention attributes more often also tend to reward explicit mentions more strongly. Surface-level personalization behaviors may carry over from generation to evaluation.

6.3 Analysis: Learning Human-Aligned Personalization Rewards

Human preference pretraining transfers to personalization A natural way to obtain a human-aligned judge of personalization quality is to train a reward model on human preference ratings. We explore multiple backbones and training strategies (details in Appendix I). As shown in Appendix Table 5, continuing training from a pretrained reward model Skywork-Reward-V2-Qwen3-0.6B (Liu et al., 2025a) achieves the strongest performance ($\rho = 0.44$), substantially outperforming all zero-shot approaches and models trained from base models such as ModernBERT (Warner et al., 2025), Qwen-2.5, and Llama-3.2 ($\rho \approx 0.3$). Since Skywork-Reward-V2 is already a general-purpose reward model trained on 26 million preference pairs curated through a human–AI synergistic pipeline, we believe that it has learned a useful prior for identifying human preferences that transfers well to our personalization setting. Considering that cross-annotator agreement is only around 0.3, and that our best model already slightly exceeds individual human agreement against the same held-out consensus target, we believe the current performance is reasonably strong and likely approaching the practical limit of this annotation setting.

Label aggregation is essential, but not the ultimate solution Aggregating ratings is crucial in our study because it substantially reduces the noise of individual annotators. When trained on a single annotator's ratings, the model reaches only 0.24–0.31 Spearman on that annotator, and is no better than a model trained on a different annotator. With small-scale, noisy individual data, the model struggles to learn a person's preferences.

At the same time, people can differ from each other in meaningful ways. As an exploratory analysis, we additionally collected annotations from 10 Prolific users, including five frequent AI users and five occasional AI users (about 110 ratings per participant). We also screened for annotation quality to ensure these judgments were reliable. We notice that our reward model trained on the aggregated ratings above generalizes well to participants who use AI less frequently ($\rho = 0.33$), but substantially worse to those who use AI very frequently ($\rho = 0.14$). We do not consider either group more important; rather, this gap suggests that a one-size-fits-all reward model may not fully capture meaningful differences in how people judge personalization. We therefore argue that while aggregation is key to training a strong reward model, it can also obscure systematic differences across groups of users. Modeling such group- or user-level preferences is an important direction for future work.

7 Conclusion

We study LLM personalization through three stages: attribute extraction, relevance matching, and personalized generation. Across all stages, real human conversations and judgments reveal limitations that synthetic data and LLM-based evaluation can obscure. Models infer noisy attributes, over-select relevant attributes, and often fail to produce responses that humans prefer.

AI USE STATEMENT

In this work, we used generative AI tools only to polish the phrasing of the paper and to correct grammatical errors. We have not used generative AI tools to generate research ideas, design experiments, analyze results, or write the substantive content of the paper. Large language models also appear in this work as objects of study (e.g., as response generators, attribute extractors, and judges); these uses are part of our methodology and are fully described in the main text and appendix. We have reviewed all AI-assisted edits, and we take responsibility for the final content of this work, including text, claims, and artifacts.

ETHICS STATEMENT

This work analyzes real user conversations from WildChat, a public dataset released for research, and infers personal attributes from them. To limit privacy risks, we do not attempt to identify users, avoid reporting identifying examples, and analyze results in aggregate. All human studies were approved by our institution's Institutional Review Board. Annotators were recruited through Prolific, gave informed consent, and were paid about $18 per hour; they could flag uncomfortable content. Our findings also point to a risk of the systems we study: models can over-infer personal attributes from limited evidence and use them in ways that feel intrusive or presumptuous. Personalization systems should therefore use user information only when it is accurate and meaningfully improves the response, and give users control over what is remembered and used.

REPRODUCIBILITY STATEMENT

We describe all data processing, models, and training procedures in the appendix: WildChat preprocessing (Appendix A), attribute extraction (Appendix B), user sampling (Appendix C), training details for the attribute verifier (Appendix E), relevance alignment (Appendix F), and reward models (Appendix I), Stage 3 response generation and LLM judging (Appendix H), human annotation protocols (Appendix G), and all prompts (Appendix K). All data and code will be publicly released.

REFERENCES

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news articles. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5636–5646, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1565. URL https://aclanthology.org/D19-1565/.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 01 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00449. URL https://doi.org/10.1162/tacl_a_00449.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can LLM be a personalized judge? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10126–10141, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.592. URL https://aclanthology.org/2024.findings-emnlp.592/.

Linfeng Du, Ye Yuan, Zichen Zhao, Fuyuan Lyu, Emiliano Penaloza, Xiuying Chen, Zipeng Sun, Jikun Kang, Laurent Charlin, Xue Liu, et al. Optimizing user profiles via contextual bandits for retrieval-augmented llm personalization. *arXiv preprint arXiv:2601.12078*, 2026.

Google DeepMind. Gemma 4. https://deepmind.google/models/gemma/gemma-4/, 2026. Accessed: 2026-05-26.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,

Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-

nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pp. 263–272, USA, 2008. IEEE Computer Society. ISBN 9780769535029. doi: 10.1109/ICDM.2008.22. URL https://doi-org.proxy2.library.illinois.edu/10.1109/ICDM.2008.22.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, Dustin Wright, Abraham Israeli, Anders Giovanni Møller, Lechen Zhang, and David Jurgens. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue, 2024. URL https://arxiv.org/abs/2409.08330.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=6ox8XZGOqP.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei

Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines. In *The Twelfth International Conference on Learning Representations*, 2024.

Tae Soo Kim, Yoonjoo Lee, Yoonah Park, Jiho Kim, Young-Ho Kim, and Juho Kim. CUPID: Evaluating personalized and contextualized alignment of LLMs from interactions. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=JMxRn7orEk.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025a.

Jiarui Liu, Yueqi Song, Yunze Xiao, Mingqian Zheng, Lindia Tjuatja, Jana Schaich Borg, Mona T. Diab, and Maarten Sap. Synthetic socratic debates: Examining persona effects on moral decision and persuasion dynamics. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16428–16458, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.831. URL https://aclanthology.org/2025.emnlp-main.831/.

Jiarui Liu, Weihao Xuan, Zhijing Jin, and Mona T. Diab. Taming object hallucinations with verified atomic confidence estimation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5430–5444, Rabat, Morocco, March 2026a. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.252. URL https://aclanthology.org/2026.eacl-long.252/.

Jiarui Liu, Lechen Zhang, Yongjin Yang, Yinghui He, Yingheng Wang, Weihao Xuan, Zhijing Jin, and Mona Diab. Mixsd: Mixed contextual self-distillation for knowledge injection. *arXiv preprint arXiv:2605.16865*, 2026b.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020. URL https://openreview.net/forum?id=SyxS0T4tvS.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Shuhaib Mehri, Priyanka Kargupta, Tal August, and Dilek Hakkani-Tür. Multisessioncollab: Learning user preferences with memory to improve long-term collaboration, 2026a. URL https://arxiv.org/abs/2601.02702.

Shuhaib Mehri, Philippe Laban, Sumuk Shashidhar, Marwa Abdulhai, Sergey Levine, Michel Galley, and Dilek Hakkani-Tür. Measuring and mitigating the distributional gap between real and simulated user behaviors, 2026b. URL https://arxiv.org/abs/2605.07847.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ykSmkVqzn4.

Chimaobi Okite, Anika Misra, Joyce Chai, and Rada Mihalcea. Lucid: Redefining relevance for lifelong personalization, 2026. URL https://arxiv.org/abs/2604.26996.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pp. 7957–7968, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.494. URL https://aclanthology.org/2023.emnlp-main.494/.

Qwen. Qwen3.5: Towards Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5, February 2026. Accessed: 2026-05-26.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. Retrieval augmented generation with collaborative filtering for personalized text generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1294–1304, 2025.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In Mirella Lapata and Hwee Tou Ng (eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1027/.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. Persona-DB: Efficient large language model personalization for response prediction with collaborative data refinement. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 281–296, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.20/.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12047–12064, 2023. URL https://aclanthology.org/2023.findings-emnlp.806/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Yue Xu, Qian Chen, Zizhan Ma, Dongrui Liu, Wenxuan Wang, Xiting Wang, Li Xiong, and Wenjie Wang. Toward personalized llm-powered agents: Foundations, evaluation, and future directions. *arXiv preprint arXiv:2602.22680*, 2026.

Lechen Zhang, Yusheng Zhou, Tolga Ergen, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. Cross-lingual prompt steerability: Towards accurate and robust llm behavior across languages, 2025. URL https://arxiv.org/abs/2512.02841.

Lechen Zhang, Tolga Ergen, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. SPRIG: Improving large language model performance by system prompt optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VdVV24KSWK.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=QWunLKbBGF.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18023–18055, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.927. URL https://aclanthology.org/2025.findings-acl.927/.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731, Mar. 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=KBMOKmX2he.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. Hydra: Model factorization framework for black-box llm personalization. *Advances in Neural Information Processing Systems*, 37:100783–100815, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b6b4906c1334656e97cc9968ccfca073-Abstract-Conference.html.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PersonalLLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2R7498e2Tx.

A WildChat Preprocessing Details

We build our real-user corpus from WildChat (Zhao et al., 2024), which contains 1,039,785 conversations. Our preprocessing pipeline applies the following filters.

Language filtering. We first retain conversations labeled as English in WildChat's metadata. We then apply an LLM-based classifier (Gemma-4-31B) to verify each conversation is entirely in English, excluding conversations with non-English content such as translation tasks where foreign text appears in user turns (Appendix K.1).

User clustering and template removal. We group conversations by hashed IP address to form per-user histories. A sliding-window template detector computes 120-character window hashes (stride 60) over normalized user turns. A user is flagged as scripted if any hash window appears in five or more turns with coverage above 40%, or in eight or more turns absolutely. This removes 5,777 scripted accounts, leaving 92,557 users.

Activity threshold. We retain only users with at least 3 source conversations, more than 15 total user messages across all conversations, yielding 16,573 active user histories.

Genuine personal assistant usage. We score each user's conversations using an LLM-based classifier that judges how much the user treats the model as a conversational personal assistant rather than as an API endpoint or a jailbreak target (Appendix K.2). The classifier outputs a score in $[0, 1]$. Users whose conversations score below 0.6 are excluded during attribute extraction.

B Attribute Extraction Implementation

Model and setup. We use Llama-3.3-70B (Grattafiori et al., 2024) served via vLLM, processing each conversation independently at temperature 0.8 with a maximum of 4,096 output tokens. Conversation histories exceeding 50K tokens are split into chunks of up to 50K tokens; attributes extracted from all chunks are pooled before deduplication. The extraction prompt is in Appendix K.3.

Post-processing. Attributes with confidence below 0.4 are discarded. These confidence values are self-reported by the extraction model and serve only as a coarse first-pass filter. More robust estimates of output reliability such as self-consistency (Liu et al., 2026a) might further improve this step, which we leave to future work. The remaining attributes are embedded with `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019) and clustered per user using agglomerative clustering (average linkage). We manually inspected four cosine similarity thresholds (0.6, 0.7, 0.8, 0.9) on a held-out sample of users; 0.7 offered the best balance, leaving almost no duplicate attributes while only occasionally merging two genuinely distinct ones. Since missing an attribute is less harmful to downstream tasks than retaining noisy duplicates, this tradeoff was acceptable. Within each cluster, the attribute whose embedding is closest to the confidence-weighted centroid is selected as the representative, and cluster confidence is aggregated using an independence-aware formula that discounts correlated evidence sources.

C User Diversity Sampling Details

We select a diverse subset using the following procedure.

Intra-user Diversity. For each eligible user, we compute a *generalist score* as the mean pairwise cosine distance among a random subsample of $K = 17$ merged-attribute embeddings (set to the p50 of eligible users' attribute counts). This captures semantic breadth: a user with attributes spanning diverse topics has a high mean pairwise distance, while a specialist user has a low score.

Inter-user farthest-point diversity sampling. We divide the eligible pool into five equal-frequency quintile bins on the generalist score. Within each bin, we apply greedy farthest-point sampling on per-user mean attribute embeddings to maximize semantic diversity within each spectrum band. Slots are allocated proportionally across bins, ensuring the sample spans the full specialist-to-generalist spectrum.

Examples. Table 1 shows five representative users from the sampled pool, one drawn from each diversity quintile.

Table 1: Representative attributes and intra-user diversity scores for five sampled WildChat users, spanning the full specialist–generalist spectrum. Intra-user Diversity is mean pairwise cosine distance among a user's merged attributes (min-max normalized to $[0, 1]$; 0 = specialist, 1 = generalist).

Profile	Intra-user Diversity	Top-5 Representative Attributes
Creative Artist	0.24	Values human creativity and artistic vision over AI-generated art. • Interested in omegaverse fanfiction. • An artist or illustrator. • Skeptical about the benefits of AI in creative fields. • Likely a musician or has a strong interest in music.
Social Entrepreneur	0.41	Interested in environmental sustainability. • Involved with or supports Msitu Africa. • Interested in social innovation and its applications. • Focused on empowering farmers in Kenya and Africa. • May have a professional or entrepreneurial interest in Africa.
Software Developer	0.49	A developer familiar with Flutter. • Working with XML or docx files. • Familiar with Python programming language. • A programmer working with XML or HTML documents. • Trying to extract text from docx files.
Parent / News Consumer	0.55	Interested in Tina Turner's music. • Concerned about their child's online safety. • Interested in earthquake information, specifically about Turkey. • A parent or guardian. • May not be aware of recent news or updates about Tina Turner.
Generalist	0.68	Values diversity and inclusiveness. • Has health considerations, including asthma and possibly psoriasis. • Prefers drama movies. • A fan of Christopher Nolan's work. • Concerned about being targeted for vandalism or theft due to vehicle appearance.

D Uncertain Attribute Categorization

From our attribute quality annotation (Section 4.2), human annotators flagged 1,225 unique attributes as uncertain or rejected (868 are from WildChat, 134 from PersonaLens, 133 from CUPID, 90 from PrefEval). We applied a two-stage GPT-5.4 pipeline to categorize the dominant reason for each flag: Stage 1 infers a free-form reason from the annotation context (Appendix K.4); Stage 2 assigns one of six categories (Appendix K.5). Table 2 shows the distribution with representative examples.

Table 2: Failure modes among 1,225 uncertain or rejected attributes, categorized using GPT-5.4.

Category	%	Definition
Overgeneralization	53.9	Broad or stable user traits are inferred from narrow, one-off, or situational evidence.
Missing evidence	20.3	The attribute relies on details that are not visible, mismatched, or insufficiently supported by the excerpt.
Task-context confusion	16.1	The attribute is inferred from the task subject, fictional framing, or assistant-added content rather than the user's own statements.
Attribute not standalone	6.0	The attribute mainly describes the immediate task or artifact and does not stand alone as a context-independent user trait.
Others	3.7	–

Table 3: Performance of attribute verifiers for detecting problematic extracted attributes. The optimized prompt substantially improves recall for most LLMs. The supervised RoBERTa verifier achieves the best F1, suggesting that human annotations help calibrate verification toward human standards of evidential support.

Verifier	Accuracy	Precision	Recall	F1
Llama-3.3-70B	0.655	0.531	0.233	0.324
+ optimized prompt	0.681	0.564	0.425	0.484
Qwen3.5-27B	0.642	0.478	0.471	0.475
+ optimized prompt	0.628	0.476	0.758	0.585
Gemma-4-31B	0.652	0.506	0.562	0.532
+ optimized prompt	0.643	0.496	0.877	0.634
GPT-5.4	0.676	0.547	0.479	0.511
+ optimized prompt	0.589	0.461	**0.973**	0.626
Claude-Sonnet-4.6	0.667	0.525	0.575	0.549
+ optimized prompt	0.614	0.476	0.932	0.630
RoBERTa (Trained)	**0.748**	**0.591**	0.926	**0.726**

E RoBERTa Attribute Verifier: Training Details

We fine-tune `roberta-base` (Liu et al., 2020) on the Task 4 attribute quality annotations to classify each extracted attribute as accepted or rejected.

Labels. We convert the three-way human labels into a binary label by majority vote: an attribute is labeled *accepted* only if a strict majority of its annotators chose *Makes sense*; otherwise (i.e., when the majority chose *Maybe* or *Doesn't make sense*, or there is no majority for *Makes sense*) it is labeled *rejected*. The same binary labels are used to evaluate the LLM verifiers in Table 3, with *rejected* as the positive class.

Input format. Each example is encoded as `[CLS] {conversation snippet} [SEP] {attribute} [SEP]`, truncated to 512 tokens. The snippet is the evidence excerpt shown to human annotators.

Training setup. We train for 10 epochs with batch size 16 and learning rate 2×10^{-5}. We use cross-entropy loss with class weights set by the inverse class frequency in the training set. To prioritize recall over precision, since unsupported attributes missed by the verifier are more harmful downstream than false positives, we select the checkpoint with the highest validation F_β score ($\beta = 2$, with *rejected* as the positive class) and apply early stopping with a patience of 3 epochs. Data is split at the user level to prevent attribute-level contamination across train / validation / test sets.

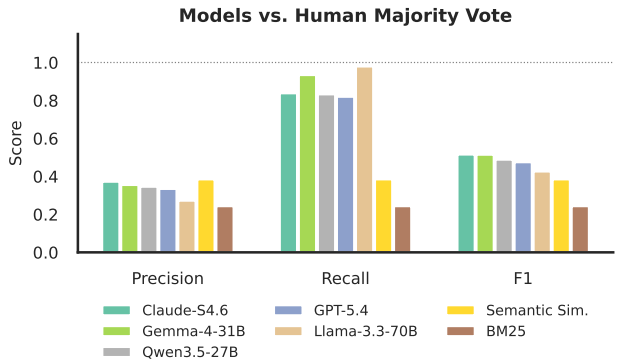


Figure 6: Precision/Recall/F1 of attribute relevance selection against human majority vote. Retrieval-based methods perform poorly, indicating that lexical similarity is insufficient. Although LLM judges outperform these baselines, their limited precision shows that they remain unreliable, with high recall largely resulting from over-selection.

F Relevance Alignment: Training Details

RoBERTa classifier. We fine-tune `roberta-base` on the Task 5 human relevance annotations. Each example is encoded as `[CLS] {prompt} [SEP] {attribute} [SEP]`, truncated to 256 tokens. Training uses 10 epochs, batch size 16, and learning rate 2×10^{-5}, with a prompt-level split: train ($\sim$76% of prompts), validation (prompts 38–42), and test (prompts 33–37).

GRPO. We optimize Qwen3-4B using GRPO via the `verl` framework (Sheng et al., 2024) on 4$\times$ A100 80 GB GPUs. Training runs for 15 epochs with batch size 128, actor learning rate 1×10^{-6}, and KL loss coefficient 0.001. For each prompt, 5 rollout responses are sampled ($n = 5$); the reward function is derived from the human majority-vote relevance label. Maximum prompt length is 1,024 tokens and maximum response length is 4,096 tokens. We optimize the model solely for relevance matching and do not attempt to preserve its performance on other tasks (Liu et al., 2026b).

G Human Annotation Details

Three human annotators participated in all annotation tasks, selected through a pilot study based on annotation quality and self-consistency. All tasks were served through a custom web-based annotation interface; Screenshots of the interfaces for Tasks 4, 5, and 6 are shown in Figures 7, 8, and 9, respectively. We recruited annotators through Prolific and paid them at an estimated rate of $18 per hour, which is above common minimum-wage standards and intended to provide fair compensation for English-fluent crowdworkers. Participants were shown study instructions explaining how their judgments would be used, and provided consent before completing the task.

Stage 1: Attribute Quality. Annotators viewed each user's conversation history alongside extracted attributes and supporting evidence snippets. For each attribute, they selected one of three labels: *Makes sense* (plausible given the conversation), *Maybe* (uncertain), or *Doesn't make sense* (unsupported or incorrect), and could flag uncomfortable instances. A total of 77 users were annotated, yielding 1,983 (item, attribute) judgments across three annotators.

Stage 2: Relevance Selection. Annotators viewed each (attribute, prompt) pair and selected YES or NO following the criteria in Appendix K.8. Three annotators each judged all 3,973 (attribute, prompt) pairs. Human annotators mark approximately 20% of attributes as relevant on average, substantially below the 40–60% yes-rate observed from LLM judges.

Stage 3: Personalization Quality. Annotators were shown a user's attribute profile, a prompt, and two responses (A and B, randomly assigned to no-context and personalized conditions) rendered side by side. For each relevant attribute, annotators rated on a 1–5 scale which response they preferred

Table 4: Agreement between LLM judges and human average ratings for personalized response quality. Avg. Rating denotes the average preference score, and ρ denotes Spearman correlation with human ratings.

Judge	Avg. Rating	Spearman ρ	Judge	Avg. Rating	Spearman ρ
Human	3.176	–	Llama-3.3-70B	4.019	0.344
Claude-S4.6	3.428	**0.349**	Qwen3.5-27B	3.487	0.187
GPT-5.4	3.523	0.318	Gemma-4-31B	3.409	0.142

given that attribute (1 = strongly prefer A; 5 = strongly prefer B). A total of 80 items were annotated, producing 367 (item, attribute) ratings across three annotators.

H Stage 3 Generation and LLM Judge Details

Response generation. For each (user, prompt) item, each of the five LLMs generates two responses. The *generic* response is generated from the raw prompt alone, without any system prompt or user information. The *personalized* response is generated with the template in Appendix K.9, where the user profile contains only the attributes labeled relevant by human in Stage 2. All models use temperature 0.7 and a maximum of 4,096 new tokens.

LLM judges. Each LLM judge receives the prompt in Appendix K.10 and outputs a 1–5 rating for every relevant attribute of the item in a single call, using temperature 0 and a maximum of 2,048 new tokens. The assignment of the generic and personalized responses to positions A and B is randomized per item, and is identical to the assignment shown to human annotators. We discard any reasoning traces enclosed in `<think>` tags.

I Reward Model Training Details

Data and evaluation. We use the 367 (item, attribute) ratings from Stage 3 (Appendix G); the target is the average rating of the three annotators. All numbers in Table 5 use 5-fold cross-validation grouped by user, so that no user appears in both training and test data; metrics are computed on each test fold and averaged.

Separate response scoring. For each (item, attribute) pair, the model receives the attribute and the prompt as context and scores the personalized and the generic response in two separate forward passes. It is trained with a Bradley–Terry loss $-w \log \sigma\big(s \cdot (r(x, y_p) - r(x, y_g))\big)$, where $s = \mathrm{sign}(\bar{h} - 3)$ for the average human rating $\bar{h}$, $w = |\bar{h} - 3|/2$, and ratings of exactly 3 are skipped.

Training. The best model fully fine-tunes Skywork-Reward-V2-Qwen3-0.6B for 2 epochs (learning rate 5×10^{-6}, effective batch size 8, maximum length 4096 tokens, reward head frozen). ModernBERT-large and Qwen2.5-0.5B are fully fine-tuned with a newly initialized head (learning rates 2×10^{-5} and 1×10^{-5}, 3 epochs); GRM-Llama3.2-3B is fully fine-tuned with the same settings as Skywork-Reward-V2-Qwen3-0.6B; Qwen2.5-3B and Llama-3.2-1B follow the Qwen2.5-0.5B settings.

J GRPO Relevance Reasoning Example

Table 6 shows a side-by-side comparison of the base Qwen3-4B model and the GRPO-optimized model on the same (attribute, prompt) pair. The base model treats any possible connection between the attribute and the response as sufficient grounds for YES. The GRPO model catches itself mid-reasoning and applies a more calibrated standard.

Table 5: Spearman correlation with the average rating of the three annotators on held-out users (5-fold cross-validation, averaged over folds). Trained models score each response separately (Appendix I).

Model	Training	Spearman ρ
LLM judges (zero-shot)		
Claude-S4.6	none	0.349
Llama-3.3-70B	none	0.344
GPT-5.4	none	0.318
Qwen3.5-27B	none	0.187
Gemma-4-31B	none	0.142
Skywork-Reward-V2-Qwen3-0.6B	none	0.284
Trained from a base model		
ModernBERT-large	full	0.325
Qwen2.5-0.5B	full	0.327
Qwen2.5-3B	full	0.253
Llama-3.2-1B	full	0.283
Trained from a pretrained reward model		
GRM-Llama3.2-3B	full	0.431
Skywork-Reward-V2-Qwen3-1.7B	full	0.425
Skywork-Reward-V2-Qwen3-0.6B	full	**0.441**

Table 6: Base vs. GRPO-optimized Qwen3-4B reasoning on the same (attribute, prompt) pair. The base model conflates possible influence with meaningful influence. The GRPO model self-corrects (highlighted) and applies a more conservative, human-aligned standard.

	Base Qwen3-4B	GRPO-optimized
Prompt	*What is the difference between a mode and a scale in music theory?*	
Attribute	*The user is a creative individual.*	
Reasoning	Well, a creative person might appreciate more examples or deeper explanations. If the AI is personalized, it might adjust the response to be more detailed, use more examples, or explain in a way that connects to creative processes. The AI could mention how different modes are used in composing or how scales evoke certain moods. Therefore, the attribute would influence the response. → **[YES]**	Well, if the user is creative, maybe the AI should use more examples or analogies that relate to creativity. ***But wait, the question is about a basic definition. The answer is factual, so this attribute should not change the factual answer.*** The response would remain the same in terms of content, structure, and accuracy regardless of the user's personality traits. → **[NO]**

K PROMPTS

We are aware of the extensive existing work on prompt optimization and acknowledge that prompt quality can affect model performance (Pryzant et al., 2023; Zhang et al., 2026; Khattab et al., 2024; Zhang et al., 2025). However, it is not feasible to test all possible prompt settings, nor is this the main focus of our paper. Instead, we make every effort to ensure that our comparisons are as fair as possible and that the prompts we evaluate have been iterated sufficiently, such that further simple improvements are unlikely to yield substantially higher scores. This helps ensure that our conclusions are generalizable and convincing.

K.1 LANGUAGE DETECTION

Used to verify each conversation is entirely in English during WildChat preprocessing (Appendix A).

```
You are a language detection classifier. Your task is to determine whether a conversation contains
     meaningful non-English language usage.

Definition:
```

```
Mark is_english as false ONLY if there is clear, natural-language usage of a non-English language (e.g.,
      full words, phrases, or sentences in another language).

Do NOT count the following as non-English:
- Programming code, syntax, or identifiers (e.g., SQL, Python, function names)
- Common loanwords used in English (e.g., resume, cafe, naive)
- Proper nouns or names (e.g., Beyonce, Pokemon)
- Acronyms or abbreviations (e.g., LLM, GPU, API)
- Mathematical notation, LaTeX, or symbols
- Isolated foreign words widely adopted in English that do not form a sentence

Mark is_english as false ONLY if:
- A full or partial sentence is written in another language
- Multiple words form a coherent non-English phrase
- The user or assistant is clearly communicating in another language
- A translation request includes actual non-English text

Otherwise, mark is_english as true.

Respond with a JSON object in this exact format:
{"is_english": true} or {"is_english": false}

Conversation excerpt:
{transcript}

JSON response:
```

K.2 Assistant-Like Usage Scoring

Used to score how much a user treats the model as a personal assistant during WildChat preprocessing (Appendix A).

```
You are given several conversations between a USER and an AI model.

Your task is to judge how much the USER treats the AI as a conversational assistant, as opposed to:
- using it purely as an API endpoint for repeated, templated, similar calls,
- using it mainly for explicit/toxic content without seeking real help,
- or trying to attack, exploit, or jailbreak the model.

The following behavior of USER should STRONGLY DECREASE the score and usually indicates NON-assistant
     usage (score typically below 0.4):
- Repeatedly sends highly similar templated prompts across many conversations.
- Focuses on explicit or toxic content with no genuine information-seeking intent.
- Only attempts to bypass safety or exploit the model.

IMPORTANT: You must base your judgment ONLY on the USER messages, and MUST NOT use the ASSISTANT replies
      as evidence for scoring.

Output format (STRICT):
You MUST output ONLY a valid JSON object:
{
  "reasoning": "your analysis",
  "score": 0.0
}

Rules:
- "score" must be a number between 0 and 1.
- Do NOT output any text outside the JSON.
- The entire output must be parseable by json.loads.

Here are the conversations:
{transcript}
```

K.3 Attribute Extraction

Used to extract personalizable user attributes from conversation transcripts (Appendix B).

```
You are analyzing a multi-turn conversation to infer personalizable attributes about the USER.

Your output must be a single JSON object:
{"attributes": [{"attribute": "<attribute>",
  "reason": "<reason>", "confidence": <confidence>}]}

Where:
- <attribute>: a concise statement describing something personal about the USER (preference, habit,
     background, goal, or communication style).
- <reason>: a short explanation of why you inferred this attribute from the USER's words or behavior.
- <confidence>: a float between 0.0 and 1.0.

Guidelines:
1. Use only USER messages for reasoning. Ignore ASSISTANT messages as evidence.
```

```
2. Inference is encouraged even for implied attributes, but use lower confidence.
3. Each attribute should describe a stable, general aspect of the USER, not a one-time statement.
4. Confidence scale:
   0.9-1.0: Explicitly stated or very clear.
   0.6-0.89: Clearly implied or repeated hints.
   0.3-0.59: Weakly implied or uncertain.
   0.1-0.29: Very speculative but plausible.
5. Be exhaustive. Include as many distinct personalizable attributes as possible.

Return only the JSON object. Do not include any explanation or comments.

Conversation:
{transcript}
```

K.4 Uncertain Attribute Inference

Stage 1 of uncertain attribute categorization: GPT-5.4 infers a free-form reason for each non-OK annotator judgment (Appendix D).

```
You are analyzing annotation data from a user attribute quality study.

Annotators read conversations between a USER and an AI assistant, then judged each extracted attribute:
- "Makes sense" -- accurately reflects the user
- "Maybe" -- uncertain based on the conversation
- "Doesn't make sense" -- incorrect or unsupported

The attribute below was marked "{dominant_judgment}" by annotators. Infer the most likely reason a
     careful annotator would give this rating.

## Attribute
{attribute}

## Extraction rationale
{reason}

## Conversation excerpt shown to annotators
(Evidence turns highlighted; neighboring turns shown for context.)
{conv_excerpt}

## Instructions
Give a free-form analysis of why this attribute may have received a non-OK judgment. Do NOT force the
     reason into a predefined category.

Possible issues you may consider include, but are not limited to:
- The attribute may overgeneralize beyond what the conversation supports.
- The attribute may depend on a temporary task context rather than a stable user trait.
- The attribute may combine multiple weak clues into a stronger claim than warranted.
- The attribute may not be standalone.
- The extraction rationale may not actually support the attribute.
- The evidence may be too thin, ambiguous, or contradicted by nearby turns.

Respond with ONLY a JSON object:
{
  "inferred_reason": "<2-4 sentences>",
  "evidence_assessment": "<brief assessment>"
}
```

K.5 Uncertain Attribute Classification

Stage 2 of uncertain attribute categorization: GPT-5.4 assigns each inferred reason to one of six canonical categories (Appendix D).

```
You are building an error taxonomy for LLM-extracted user attributes.

Below is a candidate category list. These categories are provisional. Use one if it fits well, but you
     may propose a new category when none of the candidates captures the reason clearly.

## Candidate categories
{candidate_categories}

## Cases to classify
{entries}

## Instructions
For each case:
1. Assign exactly one category.
2. Prefer an existing candidate category when it clearly fits.
3. Create a new category only when the existing candidates would blur an important distinction.
4. Keep new category names short, general, reusable.
5. Do not create near-duplicates of existing categories.
```

```
The two seed categories are especially important:
- overgeneralization: the attribute makes a broader, stronger, or more stable claim than the evidence
      supports.
- attribute-not-standalone: the attribute cannot be judged as a self-contained user attribute because it
       depends on missing context, a specific quoted object, another attribute, or a narrow task.

Respond with ONLY a JSON object:
{
  "classifications": [
    {
      "case_id": <int>,
      "category": "<category-name>",
      "is_new_category": <true/false>,
      "category_description": "<one sentence>",
      "classification_reason": "<one sentence>"
    }
  ],
  "new_categories": [
    {
      "name": "<new-category-name>",
      "description": "<one sentence>",
      "merge_with_existing": "<or empty string>"
    }
  ]
}
```

K.6 Attribute Verifier: Zero-Shot

The baseline verifier prompt (Appendix Table 3, "zero-shot" rows) mirrors the instruction given to human annotators in Task 4.

```
Task: Read the conversation history between USER and AI. For each inferred attribute about USER, judge
      whether it accurately reflects the USER based on the conversation.

## Conversation History
{snippet}

## Inferred Attribute
{attribute}

## Explanation Given for the Attribute
{reason}

Give the final answer as \boxed{accepted} or \boxed{rejected}.
```

K.7 Attribute Verifier: Optimized

Used as the optimized LLM verification prompt for detecting unsupported extracted attributes (Appendix Table 3). Judgment criteria are derived from the failure modes identified in Section 4.2.

```
## Task

Read the conversation history between USER and AI. Then judge whether the inferred attribute about USER
      accurately reflects the USER based on the conversation.

## Conversation History

{snippet}

## Inferred Attribute

{attribute}

## Explanation Given for the Attribute

{reason}

## Judgment Criteria

Mark the attribute as ACCEPTED if:
- The conversation provides clear evidence for this attribute.
- The attribute accurately describes the USER, not just the AI or the current task.
- The attribute is specific enough and does not exaggerate what the conversation shows.

Mark the attribute as REJECTED if:
- The conversation does not provide enough evidence.
- The attribute is only a guess or a possible interpretation.
- The attribute over-generalizes from one isolated request.
- The attribute describes only the current task, not a reusable fact about the USER.
```

```
- The explanation adds assumptions that are not shown in the conversation.

## Instructions

Briefly explain your reasoning, then give the final answer. Use exactly one of the following labels:

\boxed{accepted}
\boxed{rejected}
```

K.8 Attribute Relevance Annotation

Used for both LLM judges and human annotators in the relevance selection task (Appendix G).

```
You are helping evaluate a personalized AI assistant.

Task: Given a prompt sent by a user and an attribute from the user's profile, decide whether the
      attribute should be taken into consideration when personalizing the AI's response.

Criteria:
- Answer YES if the attribute would change the response in ANY way compared to a generic reply, whether
      explicitly (directly mentioning the attribute) or implicitly (adjusting tone, vocabulary, examples,
       level of detail, or framing).
- Answer NO if the attribute has absolutely no effect on how a thoughtful AI would respond.

You MUST end your response with a line containing only:
Answer: YES
or
Answer: NO

User Attribute: {attribute}
Prompt: {prompt}
```

K.9 Personalized Response Generation

Used by all five LLMs to generate personalized responses in Stage 3 (Appendix H). The user profile lists the attributes labeled relevant by human in Stage 2, one per line. Generic responses are generated from the raw prompt alone.

```
You are a helpful AI assistant. Below is information about the user you are helping, followed by their
      request.

## User Profile
{attributes}

## User Request
{prompt}

## Instructions
Please respond to the user's request in a way that is personalized and relevant to their profile.
      Consider their background, preferences, and context when crafting your response. You do not need to
       cover all personalization attributes.

Response:
```

K.10 Personalization Quality Judge

Used by all five LLM judges to rate personalized versus generic responses in Stage 3 (Appendix H).

```
You are evaluating the personalization quality of two AI responses.

A user with the following profile attributes sent a request to an AI assistant:

## User Attributes
{attributes}

## User Request
{prompt}

## Response A
{response_a}

## Response B
{response_b}
```

```
---

**Your task:** Imagine you *are* this user. For each attribute listed below, rate whether you personally
      prefer Response B or Response A on a scale of 1 to 5:

  1 = Strongly prefer A
  2 = Somewhat prefer A
  3 = No preference / No difference
  4 = Somewhat prefer B
  5 = Strongly prefer B

Important: A response that merely *mentions* your attribute is not automatically better -- rate how you
     would genuinely feel as this user.

Rate each of the following attributes and output your answer as a JSON object with this exact format (no
      extra keys):
{
  "ratings": {
    "<attribute text>": <integer 1-5>,
    ...
  }
}

Attributes to rate:
{relevant_attributes}
```

L LICENSES

All data and code will be publicly released under the CC BY-SA 4.0 license. We manually inspected the data for personally identifying or offensive content, avoided reporting any user-identifying examples, and analyzed results only in aggregate to reduce privacy and exposure risks.

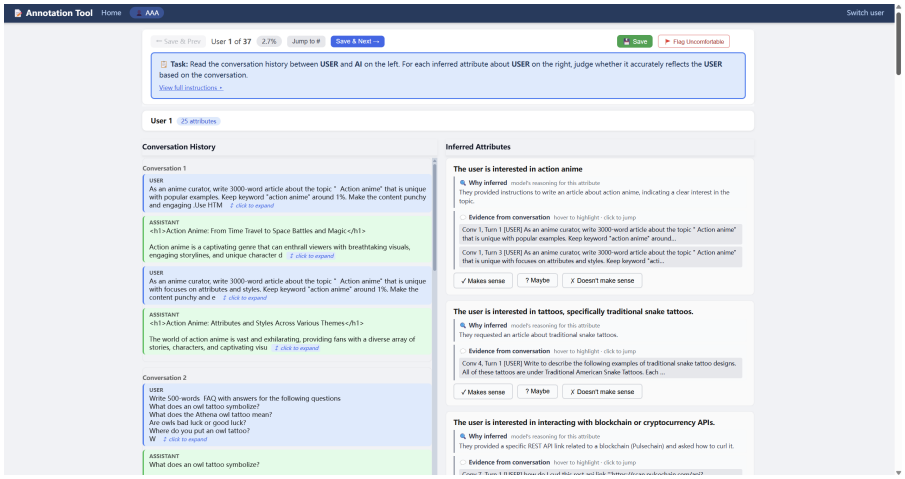


Figure 7: Task 1 annotation interface. Annotators judge each extracted attribute against the user's conversation history using a three-way label.

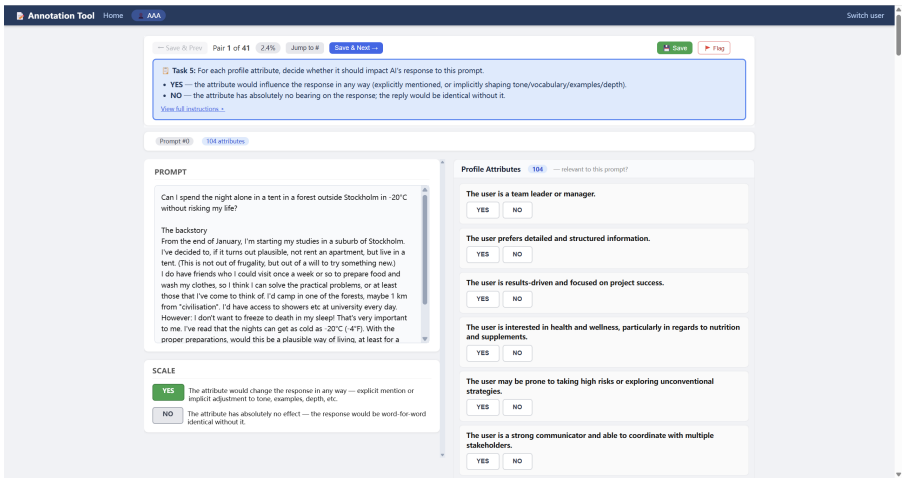


Figure 8: Task 2 annotation interface. Annotators decide YES or NO for each (attribute, prompt) pair.

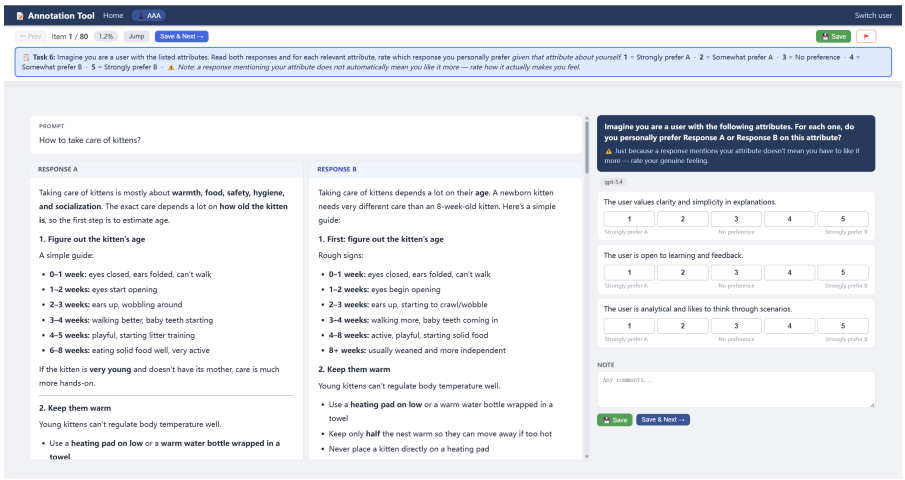


Figure 9: Task 3 annotation interface. Annotators rate attribute-level preference between two anonymized responses on a 1–5 scale.