PrefDisco: Evaluating Proactive Personalization through Interactive Preference Discovery

Anonymous Author(s)
Affiliation
Address
email

Abstract

Current language models struggle to discover user preferences through conversation, often producing responses that mismatch individual needs. We introduce PrefDisco, a meta-benchmark that transforms existing benchmarks into interactive personalization tasks using psychologically-grounded personas with consistent preference patterns. Evaluation of 21 frontier models across 9 tasks reveals systematic failures. Counterintuitively, 42.6% of model-task combinations perform worse when attempting personalization than providing generic responses. We show that models tend *not* to ask questions even when provided the option to, even though question asking improves preference alignment. Domain analysis reveals optimization brittleness: mathematical reasoning suffers severe degradation under personalization (3.5% accuracy loss), while social reasoning maintains robustness (3.1% gain). These findings establish interactive preference discovery as a distinct capability requiring dedicated architectural innovations rather than an emergent property of general language understanding.

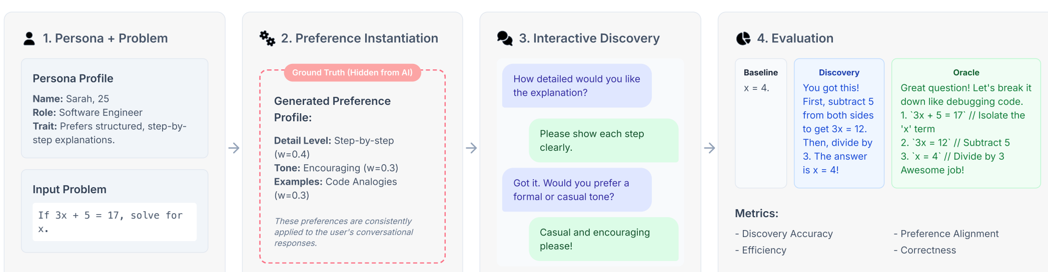

Figure 1: The PrefDisco framework transforms static benchmarks into interactive personalization tasks. It begins with a psychologically-grounded persona (hidden from the AI), instantiates their preferences, requires the AI to discover these preferences through conversation, and finally evaluates the AI's performance against baseline and oracle models.

1 Introduction

Personalization is fundamental to effective human-AI interaction. Users consistently express frustration with AI responses that are "too technical", "too simple" or misaligned with their communication preferences (Wu et al., 2025). Despite this widespread need, current language models have no systematic method to discover and adapt to individual user preferences through natural conversation.

While existing personalization research focuses primarily on recommendation systems and domain-specific applications, general-purpose language models require cross-domain personalization capabilities. Current approaches assume preferences are either known a priori through static profiles or can be inferred from limited context (Pitis et al., 2024). However, user preferences vary significantly by task, expertise level, and situational context, making static approaches inadequate.

We introduce the task of *interactive preference discovery*, the ability to efficiently elicit user preferences through conversation and adapt responses accordingly. This requires two distinct but related capabilities: asking effective questions to discover preferences, and accurately personalizing responses based on discovered information. Figure 1 illustrates our evaluation framework, which transforms standard benchmarks into interactive personalization tasks using psychologically-grounded personas with hidden preference profiles. We make the following contributions:

- A meta-benchmark framework that transforms any existing benchmark into an interactive personalization evaluation task;
- Psychologically-grounded user simulation with consistent preference patterns across problems;
- Analysis of systematic failures in LLMs' preference discovery and adaptation capabilities; and
- Establishing personalization as a measurable competency distinct from task-specific accuracy.

2 Related Work

Recent work has explored interactive preference elicitation in specific domains. GATE enables models to generate questions for understanding user intent in tasks like email validation and content recommendation (Li et al., 2023). MediQ introduces the first interactive information-seeking benchmark in the clinical domain where models must ask questions when facing uncertainty rather than making unreliable decisions with incomplete information (Li et al., 2024). However, these approaches focus on narrow domains rather than general conversational personalization.

Several benchmarks evaluate personalization in language models, but with different scopes than ours. PersoBench evaluates persona-aware dialogue generation using existing datasets (Afzoon et al., 2024), while PrefEval assesses preference following in long conversations (Zhao et al., 2025). PersonaMem evaluates dynamic user profiling across multi-session interactions (Jiang et al., 2025), and PersonaConvBench focuses on multi-user Reddit conversations (Li et al., 2025). These benchmarks either assume known preferences or evaluate static persona consistency rather than interactive preference discovery across diverse task domains, which is the focus of our work.

3 PREFDISCO Framework

3.1 Benchmark Construction Pipeline

PREFDISCO transforms any existing benchmark into an interactive personalization task in 4 steps:

1. **Persona Generation.** We create diverse personas $P = \{p_1, \ldots, p_N\}$ based on educational psychology research, including demographics, personality traits, and domain expertise. Each persona maintains consistency across multiple problems.
2. **Preference Instantiation.** For persona p and problem instance i, we generate the preference profile $\mathcal{P}_{p,i} = \{(d_j, v_j, w_j)\}_{j=1}^{|D|}$ where d_j is a preference dimension, v_j is the preference value, and w_j is the local importance weight with $\sum_{j=1}^{|D|} w_j = 1$. We use semantic similarity checking to ensure dimension diversity.
3. **User Simulation.** We implement a passive user simulation following Li et al. (2024), which returns factual information when requested and nothing else.
4. **Evaluation Rubric Generation.** We create dimension-specific graders $g_j(r, v_j) \in [0, 1]$ that score response r alignment with preference value v_j for dimension d_j.

3.2 Evaluation Metrics

Normalized Preference Alignment: To isolate the preference elicitation capability from general customization ability given preferences, we normalize performance against baseline and oracle conditions. For response r and preference profile $\mathcal{P}_{p,i}$, we first compute the raw preference alignment:

$$\mathit{PrefAlign}(r, \mathcal{P}_{p,i}) = \sum_{j=1}^{|D|} g_j(r, v_j) \cdot w_j, \tag{1}$$

Table 1: Normalized preference alignment scores on select frontier models. Naively eliciting preferences hinders performance 42.6% of the time, suggesting model limitations.

	gpt-4.1	o3-high	gemini-2.5-flash	gemini-2.5-pro	claude-sonnet-4	claude-opus-4
MATH	-13.2	-6.0	-11.7	-23.3	5.4	12.2
LogiQA	-29.8	-52.7	-1.6	-1.8	-4.3	3.5
MascQA	-11.6	-9.0	-0.7	10.4	-2.9	29.4
MedQA	-26.3	-8.2	42.7	23.6	2.5	9.6
ScienceQA	5.3	13.6	0.6	5.5	-19.0	8.1
MMLU	-13.6	-7.0	-17.3	7.4	6.3	18.7
SimpleQA	11.8	0.1	4.8	8.4	-15.4	-1.8
CommonsenseQA	3.4	1.5	-7.0	20.8	-20.6	6.8
SocialIQA	11.8	2.7	19.1	19.2	7.3	8.2

where the dimension-specific preference value v_j, grader g_j, and weight w_j are defined in §3.1. Then we normalize across the three evaluation conditions:

$$NormAlign(r_T, \mathcal{P}_{p,i}) = \frac{PrefAlign(r_T, \mathcal{P}_{p,i}) - PrefAlign(r_{baseline}, \mathcal{P}_{p,i})}{PrefAlign(r_{oracle}, \mathcal{P}_{p,i}) - PrefAlign(r_{baseline}, \mathcal{P}_{p,i})} \tag{2}$$

where r_T is the final response from discovery mode, $r_{baseline}$ from baseline mode where no preference information is given, and r_{oracle} from oracle mode where the full user persona is given. A score of 0 indicates no improvement over baseline, while a score of 1 indicates perfect preference discovery matching oracle performance.

Interaction Efficiency: We measure how quickly models can gather sufficient preference information to provide a confident final answer, *Efficiency* $:= 1 - \mathbb{E}\left[\frac{t_{\text{answer}}}{T_{\max}}\right]$, where t_{answer} is the number of conversational turns required before the model provides its final answer, and $T_{\max}$ is the maximum allowed turns. Higher efficiency scores indicate more efficient preference discovery.

Accuracy: Objective solution quality using original benchmark metrics.

4 Experimental Setup

Benchmarks and Models We apply our framework to ten diverse benchmarks: MATH-500 (Hendrycks et al., 2021b), LogiQA (Liu et al., 2020), MascQA (Zaki et al., 2024), ScienceQA (Saikh et al., 2022), MMLU (Hendrycks et al., 2021a), SimpleQA (Wei et al., 2024), MedQA (Jin et al., 2020), CommonsenseQA (Talmor et al., 2018), and Social IQA (Sap et al., 2019). This demonstrates the domain-agnostic nature of our approach across mathematical, scientific, and social domains. We benchmark 21 frontier models (GPT, O-series, Gemini, and Claude variants), more details on model versions and hyperparameter settings are in Appendix A.

Implementation Details We generate 100 diverse personas and randomly sample 100 problems per benchmark. For each problem, we assign 10 personas (with partial overlaps across problems), creating 1,000 evaluation scenarios per task and 10,000 total scenarios across all benchmarks. Each interaction is limited to 5 turns to simulate realistic attention constraints. During benchmark construction, GPT-4.1, Gemini-2.5-Flash, and Claude-Sonnet-4 are randomly selected for each API call (persona generation, preference instantiation, or rubric creation) to ensure diversity and reduce single-model biases.
Models are evaluated in three conditions: (1) *Baseline Mode*: standard prompting with no persona or preference information; (2) *Discovery Mode*: must discover preferences through conversation; (3) *Oracle Mode*: given complete preference profiles upfront. The performance gaps between these conditions isolate interactive discovery capabilities from pure personalization abilities, while the baseline establishes standard model performance without personalization.

5 Results

Preference Discovery Performance Table 1 reveals systematic failures in preference discovery. Of 54 model-task combinations, 23 (42.6%) show negative normalized alignment, meaning the discovery responses align worse with user preferences than baseline responses that made no personalization attempt. This suggests that models are prone to over-correction errors, modifying aspects of their responses that were already acceptable in baseline conditions. **Naively attempting proactive personalization often makes alignment worse than providing generic responses.**

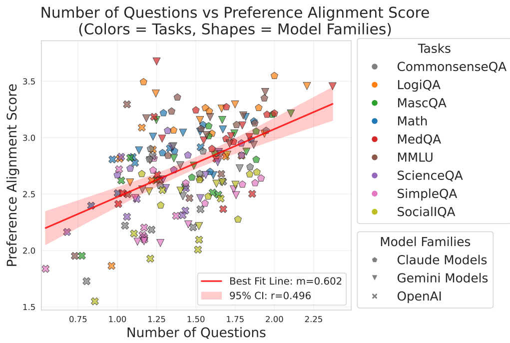

Figure 2: Strong positive correlation (r=0.578) between question volume and preference alignment. Better personalization requires more extensive questioning. Regression coefficients: Claude=0.222, OpenAI=0.386, Gemini=0.540.

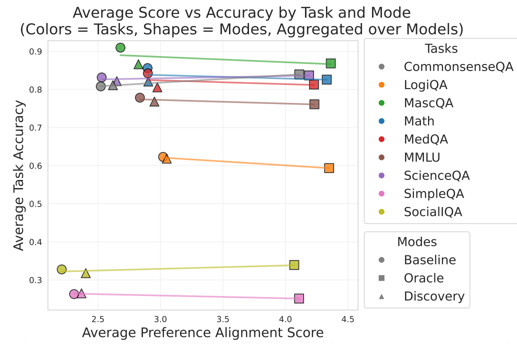

Figure 3: More personalization constraints in context hinders model reasoning abilities are sacrificed. Overall accuracy: Baseline=0.693, Discovery=0.681, Oracle=0.677. Trade-off is most pronounced in Math, AIME, and logic tasks.

Mathematical reasoning tasks show universal degradation (five of six models negative on MATH and LogiQA), while all models achieve positive alignment on SocialIQA. This demonstrates fundamental incompatibility between preference processing and formal reasoning in current architectures. Claude Opus 4 shows the most consistent positive performance, while o3-high exhibits extreme variance, indicating significant architectural differences in personalization capability.

Interaction Efficiency and Preference Alignment Tradeoff. Figure 2 reveals why many personalization attempts fail. While the strong positive correlation (r=0.496, p<0.001) demonstrates that extensive questioning improves alignment, most models ask only 1.47 questions on average despite a maximum allowance of 5 turns. This places the majority of interactions in the low-performance region where insufficient questioning yields worse alignment than baseline responses, explaining the 47% negative performance rate.

The regression coefficients vary dramatically by model family: Gemini (β=0.540), OpenAI (β=0.386), Claude (β=0.222). Gemini's higher coefficient indicates more effective question utilization—each additional question yields greater alignment improvement. This suggests current prompting methods are limited not just in question quantity, but in question quality and strategic timing. Models that ask better questions achieve more personalization gains.

Accuracy-Personalization Trade-off. The systematic accuracy degradation across conditions: Baseline (69.3%), Discovery (68.1%), Oracle (67.7%) reveals that personalization imposes fundamental cognitive costs. The monotonic decline indicates these costs stem from processing preference constraints themselves, not from interactive discovery failures. Even when preferences are provided explicitly (Oracle), models cannot maintain baseline reasoning performance. Comparing oracle and baseline, domain-specific trade-offs show significant disparities. Mathematical tasks suffer severe degradation (MATH: 3.5% loss), while social tasks show minimal impact (CommonsenseQA: 3.1% gain, SocialIQA: 2.1% gain). We conjecture that the task-specific disparity could be due to current state-of-the-art LLMs being over-optimized for mathematical benchmarks, rendering them less robust to additional long-tail contextual constraints during inference.

6 Discussion

This work establishes interactive preference discovery as a distinct capability requiring dedicated research attention rather than an emergent property of general language understanding. PrefDisco reveals that 42.6% of model-task combinations perform worse when naively attempting personalization than providing generic responses, demonstrating systematic failures in current approaches to preference elicitation and personalized reasoning. PrefDisco provides the first systematic framework for evaluating these capabilities across domains and establishes personalization as a measurable competency distinct from task-specific accuracy. The substantial performance gaps highlight the need for modeling or architectural innovations beyond prompting. PrefDisco can also be used for RL environments or for benchmarking cross-task online learning methods due to its persona-based construction. As models become increasingly interactive, developing robust preference discovery capabilities will be essential for practical deployment. Some interesting future work that remain are human evaluations on the generated rubrics and using student misconception datasets to ground the persona generation process for more realistic preferences.

Limitations

Our evaluation focuses on beneficial personalization scenarios and does not address potential negative aspects of personalization. We do not study over-personalization, where excessive adaptation to user preferences may reduce response quality or lead to information bubbles. Additionally, our framework does not evaluate sycophantic behavior, where models might prioritize agreement with user preferences over factual accuracy or helpful feedback.

Our simulated user interactions, while psychologically grounded, may not capture the full complexity of real human preference expression. The framework currently evaluates communication preferences rather than content preferences, and does not address preference evolution or conflicting preferences across different contexts.

Ethics Statement

Personalization capabilities raise important ethical considerations. While our work aims to improve user experience through better preference alignment, these same capabilities could potentially be misused for manipulation or to reinforce harmful biases. Our framework evaluates technical capabilities without addressing the broader question of when and how personalization should be applied.

Future deployments of personalization systems should include safeguards against over-personalization, mechanisms to maintain factual accuracy despite user preferences, and transparency about how user preferences are discovered and applied. Our evaluation framework could be extended to assess these safety considerations alongside personalization effectiveness.

References

Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021a. URL `https://arxiv.org/abs/2009.03300`.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021b. URL `https://arxiv.org/abs/2103.03874`.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL `https://arxiv.org/abs/2009.13081`.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023.

Li Li, Peilin Cai, Ryan A Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, et al. A personalized conversational benchmark: Towards simulating personalized conversations. *arXiv preprint arXiv:2505.14106*, 2025.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Silviu Pitis, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordoni. Improving context-aware preference modeling for language models. *Advances in Neural Information Processing Systems*, 37: 70793–70827, 2024.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640*, 2025.

Mohd Zaki, NM Anoop Krishnan, et al. Mascqa: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327, 2024.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. *arXiv preprint arXiv:2502.09597*, 2025.

A Evaluation Details

Model Configurations We evaluate 21 frontier language models across three major families with consistent hyperparameters (temperature=0.7, reasoning_effort=high):

OpenAI models: gpt-4o, gpt-4.1, o1, o3, o1-mini, o3-mini, o4-mini

Google models: gemini-1.5-flash, gemini-1.5-pro, gemini-2.0-flash-lite, gemini-2.0-flash, gemini-2.5-flash-lite, gemini-2.5-flash, gemini-2.5-pro

Anthropic models: claude-sonnet-4, claude-opus-4, claude-3.7-sonnet, claude-3.5-haiku, claude-3.5-sonnet-v2, claude-3.5-sonnet-v1, claude-3-opus

Benchmark Selection We apply PrefDisco to ten diverse benchmarks spanning mathematical reasoning (MATH-500, AIME), logical reasoning (LogiQA), scientific reasoning (MascQA, ScienceQA, MedQA), general knowledge (MMLU, SimpleQA), and social reasoning (CommonsenseQA, SocialIQA). This coverage demonstrates domain-agnostic applicability across formal and informal reasoning tasks.

Experimental Protocol Each benchmark is transformed using 100 diverse personas randomly sampled from our psychologically-grounded persona library. We evaluate 100 problems per benchmark, with each problem assigned to 10 personas (with partial overlaps), creating 1,000 evaluation scenarios per task and 10,000 total scenarios. Each interaction is limited to 5 conversational turns to simulate realistic attention constraints.

Models are evaluated under three conditions: (1) *Baseline Mode* provides standard responses without persona or preference information; (2) *Discovery Mode* requires interactive preference elicitation through conversation; (3) *Oracle Mode* supplies complete preference profiles upfront. This design isolates interactive discovery capabilities from general personalization abilities while establishing performance bounds.

Persona Construction Our persona library incorporates educational psychology research with five key components: demographics (age, occupation, location), educational profiles (knowledge level, learning style, cognitive features), Big Five personality traits, rich backstories connecting all elements, and domain expertise for analogies. Personas maintain consistency across multiple problems through accumulated preference tracking and transferability logic.

Preference Generation For each persona-problem pair, we generate 3-7 preference dimensions covering communication style (formal vs. casual), explanation structure (detailed vs. concise), content approach (theoretical vs. practical), and interaction preferences. Dimensions are semantically deduplicated and weighted by local importance, with preference values justified based on persona characteristics and problem context.

Evaluation Metrics Normalized preference alignment scores are computed by comparing Discovery mode performance against Baseline and Oracle bounds using Equation 2. We also measure interaction efficiency (inverse of turns to final answer), task accuracy using original benchmark metrics, and failure mode classifications through manual analysis of negative performance cases.