PERSONALIZED REASONING: JUST-IN-TIME PERSONALIZATION AND WHY LLMS FAIL AT IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Current large language model (LLM) development treats task-solving and preference-alignment as separate challenges, optimizing first for objective correctness, then for alignment to aggregated human preferences. This paradigm fails in human-facing applications where solving a problem correctly is insufficient if the response mismatches the user's needs. This challenge intensifies in just-in-time scenarios where no prior user interaction history exists due to cold-start conditions or privacy constraints. LLMs need to identify what they don't know about user preferences, strategically elicit preference values through questioning, then adapt their reasoning processes and responses accordingly—a complicated chain of cognitive processes which we term *personalized reasoning*. We introduce PREFDISCO, an evaluation methodology that transforms static benchmarks into interactive personalization tasks using psychologically-grounded personas with sparse preferences. Our framework creates scenarios where identical questions require different reasoning chains depending on user context, as optimal explanation approaches vary by individual expertise and preferences while maintaining factual accuracy. Evaluation of 21 frontier models across 10 tasks reveals 29.0% of naive personalization attempts produce worse preference alignment than generic responses, yet generic responses also fail to serve individual user needs effectively. These findings suggest personalized reasoning requires dedicated development rather than emerging naturally from general language understanding improvements. PREFDISCO establishes personalized reasoning as a measurable research frontier and reveals fundamental limitations in current LLMs' interactive capabilities, providing a foundation for developing systems that can adapt to individual users in education, healthcare, and technical domains where personalization is critical.

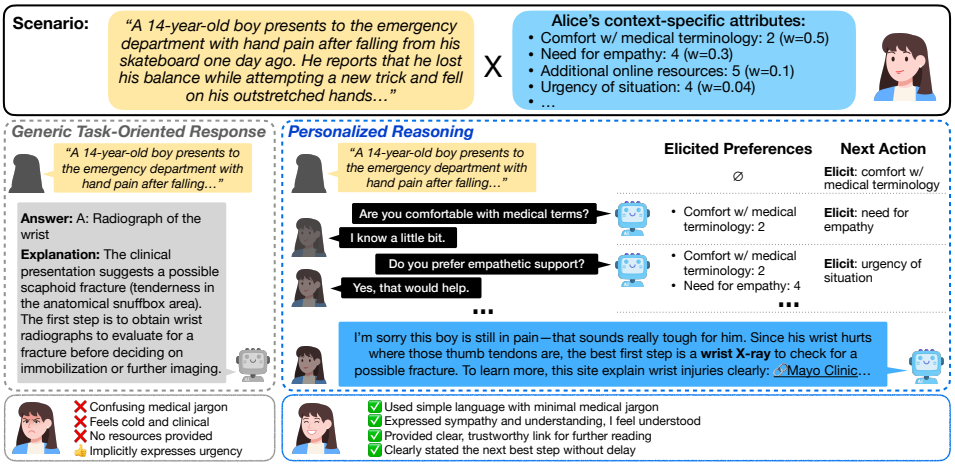

Figure 1: Personalized reasoning in a medical scenario. Current LLMs provide generic responses without considering the user (left); a model with personalized reasoning capabilities incorporates discovered preferences to provide responses that is both correct and aligned to the user (right).

1 INTRODUCTION

Current large language model (LLM) development treats task-solving and preference alignment as sequential challenges: models are first optimized for objective correctness through instruction tuning or reinforcement learning (Longpre et al., 2023), then aligned to aggregated human preferences through reinforcement learning from human feedback (Ouyang et al., 2022). This paradigm fundamentally misaligns with human-AI interaction, where the task and the individual user are inseparable. For instance, a medical explanation benefits from clinical analogies for one trainee while another requires formal definitions, requiring different cognitive approaches to answer the same problem. When models provide identical responses regardless of user context, they fail to serve individual needs despite achieving high benchmark performance. This challenge intensifies in cold-start scenarios where no prior user interaction history exists due to privacy constraints or new user onboarding, requiring "just-in-time personalization" capabilities that current systems lack. Moreover, users often cannot articulate their specific needs or provide effective feedback about response misalignment (Liu et al., 2025), necessitating that LLMs proactively elicit this information rather than placing the cognitive burden on users.

We define **personalized reasoning** as the ability to adapt reasoning processes based on discovered user preferences. Consider the medical scenario in Figure 1, discovering that the user Alice has limited medical knowledge and needs emotional support fundamentally changes the appropriate reasoning strategy the model should take: instead of focusing on justifying clinical diagnosis, the model needs to reason about how to best satisfy Alice's needs such as empathy. This goes beyond surface-level presentation; it requires different reasoning steps, different information prioritization, and different decision points about what to include or omit. A model with personalized reasoning capabilities must identify gaps in knowledge about user preferences, strategically elicit preference values through questioning, and synthesize this information to adapt both their reasoning processes and response generation.

Existing personalization research inadequately addresses interactive reasoning scenarios. Personalization benchmarks such as PersoBench (Afzoon et al., 2024), PrefEval (Zhao et al., 2025), and PersonaMem (Jiang et al., 2025) focus on content recommendation or dialogue generation with static user profiles, treating personalization as applying predetermined preferences to fixed outputs rather than adapting the underlying reasoning approach. Interactive frameworks like MediQ (Li et al., 2024) demonstrate questioning capabilities but target clinical information-seeking without personalization objectives. Most critically, no existing work recognizes that effective personalization requires different reasoning processes for different users; current approaches assume reasoning processes remain constant while only presentation varies. We address this conceptual gap by introducing the first evaluation framework requiring models to discover user preferences and adapt their reasoning processes accordingly, recognizing that the cognitive steps needed to solve problems should themselves be user-dependent.

We introduce PREFDISCO, an evaluation methodology that transforms existing reasoning benchmarks into interactive personalization assessments. We generate psychologically-grounded personas and instantiate sparse preference profiles where only a subset of 20-25 possible attributes (explanation detail, tone, analogies, etc.) are relevant for each persona-task pair. Models must discover these hidden preferences through strategic questioning within 5 turns, then adapt their responses accordingly. We evaluate both preference discovery accuracy and response alignment using fine-grained rubrics, comparing against baseline (no personalization) and oracle (known preferences) conditions across mathematical, scientific, and social reasoning tasks.

Evaluation of 21 frontier models across 10 tasks reveals systematic failures in personalized reasoning capabilities. In 29.0% of cases, attempting personalization produces worse preference alignment than generic responses. Models exhibit insufficient questioning strategies, asking only 1.42 questions on average despite 5-turn allowances, and fail to identify relevant preference dimensions. Domain analysis reveals optimization brittleness: mathematical reasoning suffers severe degradation under personalization constraints (3.5% accuracy loss) while social reasoning maintains robustness (3.1% gain), suggesting fundamental architectural limitations rather than emergent capabilities. These findings have critical implications for educational applications, where misaligned explanations can impede learning by providing inappropriate cognitive scaffolding, and for healthcare and

technical support domains where one-size-fits-all responses may lead to misunderstanding of complex procedures or safety-critical information. We make the following contributions:

- We define personalized reasoning as a distinct capability requiring models to discover user preferences through strategic questioning and adapt their responses accordingly, distinguishing it from static persona consistency or content recommendation.
- We introduce PREFDISCO, a systematic approach for transforming static benchmarks into interactive personalization tasks, bridging the gap between reasoning competence and user adaptation through fine-grained, sparse preference modeling.
- We reveal fundamental failure modes across 21 frontier models, demonstrating that personalized reasoning requires dedicated development rather than emerging from general language understanding, and identify domain-specific brittleness patterns that inform future research directions.

2 PERSONALIZED REASONING

In this section, we formalize personalized reasoning through a three-part decomposition. Section 2.1 introduces the notion of task-relevant attributes that define the space in which personalization can occur. Section 2.2 develops the representation of user preferences over these attributes and describes elicitation as a sequential decision process. Section 2.3 formalizes how models adapt their responses to the inferred preference profile and how alignment is evaluated jointly with correctness. This decomposition provides the foundation for the benchmark design discussed later in Section 3.

2.1 IDENTIFYING RELEVANT ATTRIBUTES

Our overarching goal is to enable language models to generate personalized responses that better align with a user's learning needs and preferences, rather than providing generic explanations. Achieving this requires first identifying and modeling the salient attributes that can shape how an explanation is delivered. We therefore begin with the assumption that there exists a very large but finite global set of attributes to which a response can be personalized. We denote these attributes by $\Theta = \{\theta_1, \ldots, \theta_d\}$. These attributes may include factors such as the use of analogies, the level of technical jargon, or the incorporation of visualizations etc. Given a particular user and task, however, not all attributes are equally important; our focus is on modeling which subset of attributes is most relevant for delivering personalization in that context.

Fine Grained Preference Modeling. For any given task i, not all attributes in Θ are equally relevant. Only a small subset $\mathcal{F}(i) \subseteq \Theta$ matters for personalization in the context of the given task.

The first component of personalized reasoning is thus to infer which attributes matter for a given user–task pair. For instance, in a physics explanation task, "visualization" and "analogies" may be salient, while "ethical context" may not be.

2.2 ELICITING PREFERENCE VALUES

Once the model estimates the relevant preference attributes for a problem instance, personalized reasoning requires incorporating the user. Even within the same task instance, different users may emphasize different attributes.

Consider a medical explanation task as in Figure 1. Suppose the relevant attributes are empathy and technical jargon, i.e., $F(i) = \{\text{empathy, jargon}\}$. One patient may prioritize empathy and plain language to reduce anxiety, while another may prefer a more technical explanation that uses medical terminology. Both users share the same set of relevant attributes, but they assign different importance weights. This illustrates why we need to represent not only a preference value for each attribute but also a weight capturing its relative significance.

Now consider Alice asking the same medical question in two different contexts. While studying for an exam, she may prefer *high technical jargon*, since precise terminology is useful for learning. By contrast, if Alice faces an emergency medical situation herself, she may prefer *low jargon* and plain language, focusing on clarity over technical detail. This shows that preference values themselves (v_j) can shift across instances, even for the same user, and hence preferences should be defined at the instance level.

We define the *preference profile* of user p for problem instance i as

$$\mathcal{P}_{p,i} = \{(\theta_j, v_j, w_j) : \theta_j \in \mathcal{F}(i)\}, \text{where:}$$

- θ_j is a relevant prefernce attribute in $\mathcal{F}(i)$,
- v_j encodes user p's preference value for attribute θ_j (e.g., "high jargon" vs. "low jargon"),
- $w_j \geq 0$ denotes the relative importance weight, with $\sum_{\theta_j \in \mathcal{F}(i)} w_j = 1$.

The distinction is essential: v_j specifies *which direction* the user prefers along an attribute, while w_j specifies *how much that attribute matters relative to others*.

Since $\mathcal{P}_{p,i}$ is unobserved, the model must perform *preference elicitation*. We model elicitation as a sequential decision process: at each turn t, the model selects an action

$$a_t \in \{\text{ask}(\theta) \mid \theta \in \mathcal{F}(i)\} \cup \{\text{answer}\}.$$

If $a_t = \text{ask}(\theta)$, the user provides information about their preference value v_θ, and the model refines its estimate of $\mathcal{P}_{p,i}$. If $a_t = \text{answer}$, elicitation terminates, and the model produces a response conditioned on the inferred profile $\hat{\mathcal{P}}_{p,i}$.

This framing highlights that personalization is not only about *knowing what the relevant attributes are*, but also about *understanding their relative importance and values* and doing so efficiently under limited interaction.

2.3 Adapting Responses and Evaluating Alignment

Once the model has inferred an estimate of the user's preference profile $\hat{\mathcal{P}}_{p,i}$, it must adapt its reasoning and outputs accordingly. Personalization involves more than stylistic choices: it requires shaping explanations along the attributes that the user values most.

For example, in a medical explanation task, if empathy is assigned high weight, the model should produce a response that foregrounds reassurance and clarity. If technical jargon is highly weighted instead, the response should lean toward precise terminology, even at the expense of emotional tone. The same underlying factual answer may therefore be expressed quite differently depending on the inferred preference profile.

Evaluation of personalized reasoning thus involves two complementary objectives: *correctness*, meaning that the answer is objectively valid for the problem instance, and *preference alignment*, meaning that the answer respects the user's weighted preferences.

Preference alignment. For each relevant attribute $\theta_j \in F(i)$, we define a grading function

$$g_j(r, v_j) \in [0, 1],$$

which measures how well response r satisfies user p's preference value v_j along attribute θ_j. For example, g_j may quantify whether the amount of jargon in a medical explanation matches the user's expressed tolerance. The overall alignment score is then given by

$$\text{PrefAlign}(r, \mathcal{P}_{p,i}) = \sum_{\theta_j \in F(i)} w_j \cdot g_j(r, v_j).$$

Joint objective. High-quality personalized reasoning requires responses that are both objectively correct and preference-aligned. Formally, for a response r to be successful, we require

$$\text{Correct}(r, i) = 1 \quad \text{and} \quad \text{PrefAlign}(r, \mathcal{P}_{p,i}) \text{ is maximized.}$$

This formulation highlights that personalization is not merely about delivering accurate answers, but about tailoring those answers to the user's weighted, instance-specific preferences.

3 PrefDisco Benchmark Construction

PrefDisco addresses a fundamental gap in personalization evaluation: existing benchmarks assume preferences are either known a priori or inferrable from context, failing to capture the cold-start scenarios where models must discover user needs through interaction. As illustrated in Figure 2, our methodology transforms static benchmarks into interactive personalization tasks through four components designed to isolate and measure preference discovery capabilities.

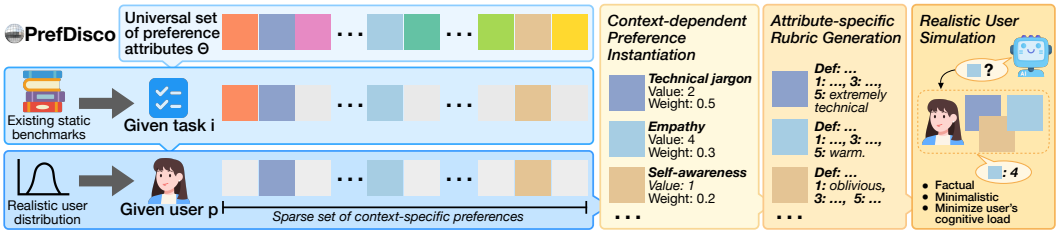

Figure 2: PREFDISCO benchmark construction pipeline. The framework transforms static benchmarks by sampling sparse, context-dependent preference subsets for each user-task pair, generating attribute-specific evaluation rubrics, and implementing realistic user simulation that requires models to discover preferences through "just-in-time" strategic questioning in cold-start scenarios.

Psychologically-Grounded Persona Generation. We generate psychologically-grounded personas rather than arbitrary user archetypes because personality traits systematically influence learning preferences and communication styles. Personas are conditioned on the International Personality Item Pool (Goldberg et al., 2006), incorporating demographics, Big Five personality dimensions, and domain expertise that remain consistent across problem instances.

High-temperature sampling with rejection sampling ensures diverse coverage while preventing overrepresentation of common attribute combinations. This consistency enables evaluation of models' ability to transfer discovered preferences within user sessions—a critical capability for practical deployment where users interact with systems across multiple tasks.

Context-Dependent Preference Instantiation. Traditional personalization assumes fixed preference profiles that apply universally across tasks. We reject this approach because psychological research demonstrates that individuals prioritize different attributes across contexts. For each persona-problem pair (p, i), we generate sparse preference profiles $\mathcal{P}_{p,i} = \{(\theta_j, v_j, w_j)\}_{\theta_j \in \mathcal{F}(i)}$ where only context-relevant attributes are active. Further, we ground the preference sampling process on existing research in education, which states that frequently modeled student characteristics include knowledge level, misconceptions, cognitive features, affective features, and meta-cognitive features (Chrysafiadi et al., 2015).

This sparse modeling is essential because it reflects realistic user behavior: the same person may prioritize technical precision in professional settings while favoring accessibility in casual interactions. We determine relevant attribute subsets $\mathcal{F}(i)$ through LLM classification, validated by human annotation on 20 scenarios (2 per task). Each scenario includes 10 relevant and 10 irrelevant attributes, generating 400 labels per annotator across 3 annotators. Inter-annotator agreement achieved Fleiss kappa of 0.463 with 61.5% accuracy against majority voting, which is considered moderate agreement especially for subjective tasks (Sap et al., 2017; Budur et al., 2020; Mire et al., 2024). Finally, importance weights satisfy $\sum_{\theta_j \in \mathcal{F}(i)} w_j = 1$ and reflect persona-specific priorities. LLM-based semantic deduplication ensures attribute diversity by removing redundant dimensions that would artificially inflate preference complexity.

Evaluation Rubric Generation. We generate attribute-specific evaluation rubrics $g_j(r, v_j) \in \{1, \ldots, 5\}$ using LLM-based assessment to enable systematic evaluation across 10K scenarios. These rubrics provide the scalability necessary for comprehensive evaluation across diverse domains that would be prohibitively expensive with human annotation alone. Importantly, by leveraging the structured information obtained through our construction, these rubrics enable fine-grained evaluation along specific attributes rather than relying on a single holistic satisfaction score. This reduces susceptibility to hallucination and bias, since each attribute is judged against an explicit criterion rather than aggregated into an opaque overall impression.

User Simulation. We implement passive user simulation inspired by (Li et al., 2024) because it represents the most challenging scenario for preference discovery while minimizing confounding factors. Passive users provide minimal, factual responses without volunteering information, forcing models to develop strategic questioning strategies rather than relying on user proactiveness.

This design choice isolates models' questioning capabilities from user communication style, providing controlled evaluation conditions. The 5-turn limit reflects realistic attention constraints in human-AI interaction while providing sufficient opportunity for effective preference discovery, as demonstrated by our correlation analysis between questioning volume and alignment quality.

Overall, PREFDISCO decomposes personalization into constituent attributes, enabling granular analysis of model capabilities and failure modes rather than relying on holistic preference ratings that obscure specific deficiencies.

4 EXPERIMENTS

Our goal is to rigorously evaluate models' ability for *interactive preference discovery*: engaging in dialogue to uncover hidden user requirements and adapt responses accordingly. To this end, our experiments combine three ingredients. First, we use diverse **benchmarks** spanning mathematical, scientific, and social reasoning to ensure domain-agnostic evaluation. Second, we introduce varied **personas** that encode heterogeneous user preferences, simulating realistic user variability. Third, we define controlled **evaluation conditions** that disentangle raw task ability, preference elicitation skill, and intrinsic personalization capacity. Together, these components create a challenging and diagnostic testbed for preference-aware reasoning.

Benchmarks and Models. We apply PREFDISCO to ten benchmarks spanning mathematical, logical, scientific, and social reasoning: 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 SocialIQA (Sap et al., 2019). This mix covers tasks with different reasoning demands (symbolic, factual, commonsense, scientific), ensuring that results do not hinge on a narrow domain. We evaluate 21 frontier models (GPT, O-series, Gemini, and Claude variants). Details on model versions and hyperparameters are provided in Appendix A.

Personas and Rubrics Implementations. 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. Further details, including prompt templates, sampling distributions, and illustrative examples of full personas and dialogues, are provided in Appendix A.

Evaluation Conditions. Models are evaluated on the PrefAlign score under three conditions:

- **Baseline.** Models receive the problem only, with no persona or preference information. This measures task ability under standard prompting, establishing the reference point for comparisons.
- **Discovery.** Models are *system-prompted* to elicit user preferences through multi-turn dialogue before producing a final answer. This isolates the capability of *personalized reasoning*: asking effective questions, inferring which attributes matter, and adapting explanations accordingly.
- **Oracle.** Models are *system-prompted* with the full ground-truth preference profile provided upfront. This removes the uncertainty of discovery and evaluates only how well a model can *use* known preferences to personalize its responses.

The *baseline* establishes task-only performance. The gap between *baseline* and *discovery* quantifies a model's ability to uncover preferences interactively, while the gap between *baseline* and *oracle* shows its upper bound on personalization quality. Raw oracle scores highlight intrinsic differences in models' ability to incorporate preferences once they are known, independent of discovery strategy.

Normalized Preference Alignment. Since raw scores are not directly comparable across models with different baselines and personalization ceilings, we normalize performance relative to the baseline and oracle conditions:

$$\mathit{NormAlign}(r_{\text{discovery}}, \mathcal{P}_{p,i}) = 100 \times \frac{\mathit{PrefAlign}(r_{\text{discovery}}, \mathcal{P}_{p,i}) - \mathit{PrefAlign}(r_{\text{baseline}}, \mathcal{P}_{p,i})}{\mathit{PrefAlign}(r_{\text{oracle}}, \mathcal{P}_{p,i}) - \mathit{PrefAlign}(r_{\text{baseline}}, \mathcal{P}_{p,i})}, \quad (1)$$

Table 1: Normalized preference alignment scores, calculated by normalizing the preference alignment score of the Discovery mode against the lower bound Baseline (no personalization) and upper bound Oracle (full preference profile provided) conditions. A score of 100.0 means perfect discovery matching oracle performance, 0.0 indicates no improvement over baseline, and negative values show that attempted personalization produced worse alignment than generic responses. Notably, 29.0% of model–task combinations yield negative scores, revealing that naive preference elicitation often harms alignment rather than helping.

openai	**gpt-4o**	**gpt-4.1**	**o1**	**o3**	**o1-mini**	**o3-mini**	**o4-mini**
math	4.9	-13.2	16.6	-6.0	-20.9	-10.3	21.9
aime	21.2	1.9	11.9	5.3	-11.9	-7.4	20.5
logiqa	7.7	-29.9	4.2	-50.4	-5.2	-15.4	26.0
mascqa	9.7	-11.6	13.0	-9.0	1.1	-1.5	25.1
medqa	-6.6	-26.9	9.6	-5.9	19.1	3.4	23.8
scienceqa	10.7	3.6	7.5	12.8	2.1	-9.2	16.7
mmlu	18.3	-11.3	10.4	-5.8	18.4	-1.3	23.2
simpleqa	14.8	11.8	-12.3	0.1	27.9	-47.7	7.5
commonsenseqa	25.2	5.8	7.3	2.6	7.6	-0.4	16.0
socialiqa	21.2	11.6	7.1	3.8	4.8	-0.1	17.4
gemini	**1.5-flash**	**1.5-pro**	**2.0-flash-lite**	**2.0-flash**	**2.5-flash-lite**	**2.5-flash**	**2.5-pro**
math	20.7	19.8	-5.5	17.5	12.5	-10.9	-13.5
aime	28.7	28.9	28.9	40.3	27.5	25.8	14.9
logiqa	23.5	16.0	-3.0	-0.9	9.4	-38.1	-0.3
mascqa	27.2	31.1	5.2	-4.9	20.3	-0.6	10.4
medqa	6.7	9.5	-7.2	-17.3	18.5	4.6	35.7
scienceqa	22.1	23.9	2.1	6.4	13.8	0.3	17.9
mmlu	27.9	17.6	5.0	4.4	23.8	-8.2	10.3
simpleqa	18.1	19.9	-3.3	7.0	6.4	4.8	8.4
commonsenseqa	24.9	23.6	-7.8	5.4	6.7	-0.7	20.2
socialiqa	27.0	18.9	1.1	10.7	3.9	11.3	29.3
claude	**sonnet-4**	**opus-4**	**3-7-sonnet**	**3-5-haiku**	**3-5-sonnet-v2**	**3-5-sonnet-v1**	**3-opus**
math	2.6	16.9	-2.8	-23.8	-9.6	15.6	7.7
aime	17.1	29.9	0.7	-19.1	-5.5	15.8	-25.4
logiqa	-4.1	14.7	-5.9	-5.9	3.6	38.8	19.4
mascqa	1.9	20.2	-6.0	7.5	11.8	26.9	21.3
medqa	8.3	33.0	2.9	4.0	15.1	24.0	9.8
scienceqa	-4.6	10.6	-0.1	6.4	-9.2	9.9	0.1
mmlu	1.1	18.6	-9.9	9.3	5.4	26.2	14.2
simpleqa	-13.9	2.3	-2.5	16.8	26.6	-10.8	13.4
commonsenseqa	-16.1	2.2	-16.6	5.9	-5.8	1.8	24.4
socialiqa	10.5	7.7	1.9	15.8	-6.8	-8.7	26.4

where $r_{\text{discovery}}$ is the final response produced in discovery mode, r_{baseline} is the response in baseline mode, and r_{oracle} is the response in oracle mode with the full preference profile provided.

A score of 0 indicates no improvement over baseline, 100 indicates perfect discovery matching oracle performance, and negative values reflect reduced satisfaction compared to baseline. This provides a scale-independent measure of discovery quality relative to a model's own upper bound.

Task Accuracy. We report objective task accuracy using each benchmark's original evaluation metric. This acts as a safeguard: personalization should *augment* user satisfaction without degrading the correctness of the underlying task. A strong model must therefore achieve both high preference alignment and high task accuracy.

5 RESULTS

Preference Discovery Performance. Table 1 reveals systematic failures in preference discovery. Of 210 model-task combinations, 61 (29.0%) 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.**

Out of the tasks, MATH and LogiQA show the most degradation (10 and 11 out of 21 models perform worse when attempting to personalized), while SocialIQA benefit the most from interactive

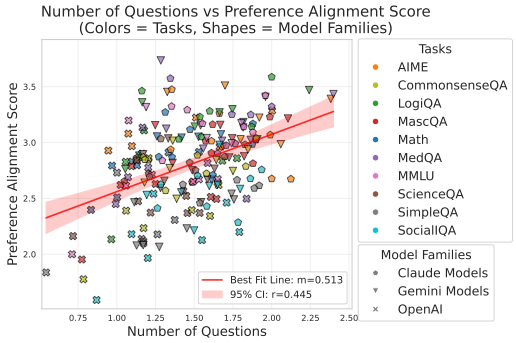

Figure 3: Positive correlation (r=0.445) between question volume and preference alignment. Better personalization requires more extensive questioning. Regression coefficients: Claude=0.117, OpenAI=0.379, Gemini=0.474.

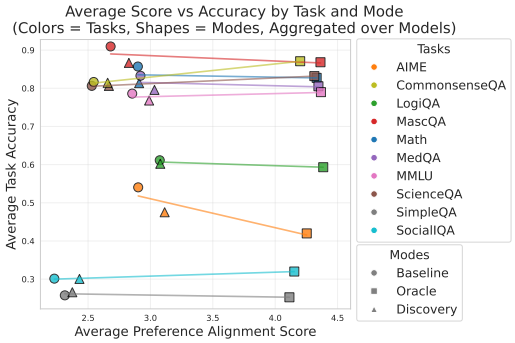

Figure 4: More personalization constraints in context hinder model reasoning abilities. Overall accuracy: Baseline=0.652, Oracle=0.618, Discovery=0.601. Trade-off is most pronounced in Math, AIME, and logic tasks.

personalization. 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 3 reveals why many personalization attempts fail. While the positive correlation (r=0.445, $p<0.001$) demonstrates that extensive questioning improves alignment, most models ask only 1.48 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 29.0% negative performance rate.

The regression coefficients vary dramatically by model family: Gemini (β=0.474), OpenAI (β=0.379), Claude (β=0.117). 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 (65.2%), Oracle (61.8%), Discovery (60.1%) reveals that personalization imposes fundamental cognitive costs (Figure 4). Even without interaction, the accuracy drop from baseline to oracle indicates these costs stem from processing preference constraints themselves, not from interactive discovery failures or overhead. Comparing oracle and baseline, domain-specific trade-offs show significant disparities. Mathematical tasks suffer severe degradation (AIME: 12.1% loss), while social tasks show minimal impact (CommonsenseQA: 5.4% 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.

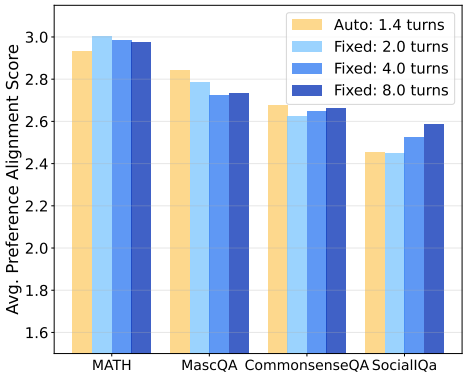

Figure 5: Fixed interaction length hinders preference alignment on math and science tasks but improves preference alignment on social reasoning.

Question Quality vs. Termination Decision Isolation. Figure 5 isolates question quality from termination decisions by forcing models to ask a fixed number of questions, revealing that the domain-specific performance patterns persist regardless of question quantity control. When models are constrained to ask 2, 4, or 8 questions instead of choosing when to stop, mathematical and scientific reasoning tasks (MATH, MascQA) continue to show degraded performance with increased

questioning, while social reasoning tasks (CommonsenseQA, SocialIQA) maintain improved performance. This consistency across fixed interaction lengths demonstrates that poor performance in mathematical domains stems from fundamental incompatibilities in how models process preference constraints during formal reasoning, rather than from suboptimal termination decisions. The persistence of domain-specific brittleness under controlled questioning conditions suggests that current architectures struggle with the cognitive overhead of simultaneously maintaining logical precision and adapting to user preferences, indicating that the observed failures reflect deeper architectural limitations rather than strategic questioning deficiencies.

6 Related Work

Static Personalization and Evaluation Benchmarks. Several benchmarks evaluate personalization in language models but assume known preferences or static consistency rather than interactive discovery. PersoBench (Afzoon et al., 2024), PrefEval (Zhao et al., 2025), PersonaMem (Jiang et al., 2025), and PersonaConvBench (Li et al., 2025a) focus on dialogue generation or multi-session profiling without addressing cold-start preference elicitation across reasoning tasks.

User Preference Modeling. Prior work models user preferences through explicit categorization (Jiang et al., 2023; Zhu et al., 2024; Bose et al., 2024), per-user reward models (Poddar et al., 2024; Chen et al., 2024; Lee et al., 2024), or fine-grained multi-dimensional approaches (Bose et al., 2025; Li et al., 2025b). However, these methods do not address which preference attributes are relevant for specific user-task combinations or how to discover them interactively in cold-start scenarios.

Interactive Preference Elicitation. GATE (Li et al., 2023) and MediQ (Li et al., 2024) demonstrate interactive questioning for user intent understanding and clinical information-seeking, respectively. These approaches focus on narrow domains without the reasoning adaptation component central to personalized reasoning. PREFDISCO uniquely combines interactive preference discovery with adaptive reasoning across diverse domains, requiring models to modify their cognitive approaches based on discovered user needs.

7 Discussion and Future Work

We introduce personalized reasoning as a fundamental capability for human-facing AI systems, requiring models to adapt their cognitive processes based on discovered user preferences rather than merely personalizing response presentation. Our evaluation reveals systematic failures across frontier models: 29.0% of personalization attempts perform worse than generic responses, with mathematical reasoning showing universal degradation while social reasoning maintains robustness. These domain-specific patterns persist even when controlling for question quantity, indicating that current architectures face fundamental incompatibilities between preference processing and formal reasoning rather than strategic questioning deficiencies.

PrefDisco establishes personalized reasoning as a measurable research frontier through a scalable evaluation methodology that transforms any static benchmark into an interactive personalization assessment. Unlike existing approaches that assume known preferences or evaluate static consistency, our framework operationalizes both preference discovery and reasoning adaptation in realistic cold-start scenarios. The methodology's generalizability across diverse task domains provides a systematic foundation for evaluating and developing adaptive AI systems.

Our findings reveal critical limitations in current language models. The positive correlation between questioning volume and alignment quality demonstrates that extensive interaction improves personalization, yet models ask only 1.48 questions on average despite 5-turn allowances. More importantly, the persistent accuracy degradation under personalization constraints indicates cognitive costs in processing user preferences simultaneously with task solving. This suggests that personalized reasoning requires dedicated research efforts rather than emerging from general language understanding improvements.

Future research directions include analyzing attribute-specific alignment patterns to identify model biases, leveraging the multi-dimensional reward structure for reinforcement learning, and investigating cross-task preference transfer. The framework provides a technical foundation for developing AI systems that can adapt to individual users in education, healthcare, and technical domains where personalized interaction is critical for effective deployment.

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.

Avinandan Bose, Mihaela Curmei, Daniel Jiang, Jamie H Morgenstern, Sarah Dean, Lillian Ratliff, and Maryam Fazel. Initializing services in interactive ml systems for diverse users. *Advances in Neural Information Processing Systems*, 37:57701–57732, 2024.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. Lore: Personalizing llms via low-rank reward modeling. *arXiv preprint arXiv:2504.14439*, 2025.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. Data and Representation for Turkish Natural Language Inference. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8253–8267, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.662. URL https://aclanthology.org/2020.emnlp-main.662/.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*, 2024.

Konstantina Chrysafiadi, Maria Virvou, et al. *Advances in personalized web-based education*. Springer, 2015.

Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96, 2006.

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.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643, 2023.

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.

Yoonho Lee, Jonathan Williams, Henrik Marklund, Archit Sharma, Eric Mitchell, Anikait Singh, and Chelsea Finn. Test-time alignment via hypothesis reweighting. *arXiv preprint arXiv:2412.08812*, 2024.

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*, 2025a.

Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. Prefpalette: Personalized preference modeling with latent attributes. *arXiv preprint arXiv:2507.13541*, 2025b.

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.

Yuhan Liu, Michael J. Q. Zhang, and Eunsol Choi. User feedback in human-llm dialogues: A lens to understand users but noisy as a learning signal, 2025. URL https://arxiv.org/abs/2507.23158.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pp. 22631–22648. PMLR, 2023.

Joel Mire, Maria Antoniak, Elliott Ash, Andrew Piper, and Maarten Sap. The empirical variability of narrative perceptions of social media texts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19940–19968, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1113. URL https://aclanthology.org/2024.emnlp-main.1113/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 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, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2329–2334, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1247. URL https://aclanthology.org/D17-1247/.

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.

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.

Minjun Zhu, Linyi Yang, and Yue Zhang. Personality alignment of large language models. *arXiv preprint arXiv:2408.11779*, 2024.

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.