PREF: Reference-Free Evaluation of Personalised Text Generation in LLMs

Anonymous ACL submission

Abstract

Personalised text generation is essential for user-centric information systems, yet most evaluation methods overlook the individuality of users. We introduce **PREF**, a **P**ersonalised **R**eference-free **E**valuation **F**ramework that jointly measures general output quality and user-specific alignment without requiring gold personalised references. PREF operates in a three-step pipeline: (1) a *General-quality stage* uses a large language model (LLM) to generate a comprehensive, query-specific guideline covering universal criteria such as factuality, coherence, and completeness; (2) a *User-alignment stage* re-ranks and selectively augments these factors using the target user's profile, stated or inferred preferences, and context, producing a personalised evaluation rubric; and (3) a *scoring step* that assigns a score using the personmalised rubric. This separation of coverage from preference improves robustness, transparency, and reusability, and allows smaller models to approximate the personalised quality of larger ones. Experiments on the PrefEval benchmark, including implicit preference-following tasks, show that PREF achieves higher accuracy, better calibration, and closer alignment with human judgments than strong baselines. By enabling scalable, interpretable, and user-aligned evaluation, PREF lays the groundwork for more reliable assessment and development of personalised language generation systems.

1 Introduction

Large language models (LLMs) have propelled open-ended text generation to new heights (Brown et al., 2020; Achiam et al., 2023), enabling high-quality dialogue, code synthesis, and data-to-text narration at scale. Despite these successes, *evaluating* the outputs of such models remains an open problem. Traditional reference-based metrics, including BLEU and ROUGE (Papineni et al., 2002; Lin, 2004), which count n-gram overlap with one or more gold references, correlate weakly with human judgements on tasks that admit many equally valid answers (e.g., creative writing, recommendation, and advice). Embedding-based alternatives such as BERTScore (Zhang et al., 2020) mitigate surface mismatch by measuring semantic similarity in a latent space, yet they still require high-quality reference texts and ultimately reflect *generic* rather than *user-specific* desiderata.

A parallel line of research bypasses references altogether by treating a strong LLM as an automatic judge (Li et al., 2025; Zheng et al., 2023; Li et al., 2023). Benchmarks such as MT-Bench (Zheng et al., 2023) and AlpacaEval (Li et al., 2023) ask GPT-4 to rate or rank candidate answers and report impressive agreement with crowd workers. Although this *LLM-as-a-Judge* paradigm scales cheaply and reproducibly, its rubric is fundamentally universal—"a good answer is a good answer for everyone"—and therefore blind to individual preferences, constraints, or prior interactions that might radically alter a user's perception of quality.

Modern LLM-based applications increasingly interact directly with end users who expect outputs tailored to their tastes, goals, and contexts. Recent work has shown that LLMs can indeed adapt to user profiles, feedback signals, and stylistic norms (Zhang et al., 2024). However, evaluating personalised text generation is especially tricky because quality is now user-dependent, i.e., *an answer that delights one person might frustrate another*. Generic references or model-centric rubrics cannot capture this subjective dimension. On the other hand, human evaluation remains the gold standard but is expensive, time-consuming, and prone to annotator variance and bias, even when detailed guidelines are provided (Van Der Lee et al., 2019). Collecting user-aligned ratings for every system update or user cohort is therefore impractical, hindering rapid iteration on personalised experiences.

To fill this evaluation gap, we propose

PREF, a **P**ersonalised, **R**eference-free **E**valuation **F**ramework that scores generated text *without* gold personalised answers. Specifically, PREF performs personalised evaluation with two-stage rubrics:

1. **General-quality stage.** A general guideline enumerates the salient factors to check, such as truthfulness, coherence, and completeness, while*explicitly ignoring* user preferences to ensure baseline adequacy.

2. **User-alignment stage.** A personalised guideline is further synthesised from user information (profile attributes, past dialogue, stated preferences) and used to weight or re-rank the general factors, yielding a customised rubric that reflects what *this* user cares about.

The LLM then judges candidate answers against the composite rubric, producing a scalar score without ever consulting ground-truth references. In this way, PREF combines universal quality control with user-centric alignment while remaining scalable and reproducible.

Our contribution can be summarised as:

- We formulate a *reference-free* evaluation protocol for personalised generation that jointly considers task adequacy and user alignment.

- We instantiate this protocol with *two-stage automatic rubric construction* and LLM-based scoring, eliminating the need for gold references or repeated human ratings.

- Through extensive experiments on the PrefEval benchmark, we demonstrate that PREF's scores track human judgements of personalised quality more faithfully than existing baselines.

- We show that integrating PREF during development helps smaller models (e.g., LLaMA-3 8B) close much of the performance gap to larger counterparts, facilitating cost-effective deployment.

Taken together, our findings position PREF as a practical and effective foundation for evaluating and ultimately improving user-aligned language generation systems.

2 Related Work

2.1 Foundations of Personalisation

Personalisation research has historically followed two complementary paradigms.

Explicit user modelling. Early adaptive hypermedia systems showed that maintaining a *structured user model*—encoding attributes such as background knowledge, goals, or learning style—enables real-time adaptation of content, sequencing, and navigation (Brusilovsky, 2001). These systems rely on explicitly specified user characteristics to guide system behaviour.

Implicit preference inference. In contrast, collaborative filtering approaches such as GroupLens infer preferences from *behavioural signals* at scale, exploiting patterns in user–item interactions to predict future agreement (Resnick et al., 1994). Modern LLM personalisation strategies largely inherit this dichotomy, combining *explicit* profile conditioning with *implicit* signal mining (e.g. clicks, dwell time, edits).

2.2 Personalised Text Generation with LLMs

Persona and attribute control. Early neural approaches such as Persona-Chat incorporated speaker embeddings into sequence-to-sequence models, improving persona consistency in dialogue (Li et al., 2016). Subsequent work on attribute and style control introduced mechanisms for steering generation toward user-specified properties. In particular, Plug-and-Play Language Models (PPLM) enabled gradient-based control of frozen LMs (Dathathri et al., 2020), while parameter-efficient methods (e.g. prefix-tuning, adapters, LoRA) achieved per-user or per-task specialisation with minimal trainable parameters (Li and Liang, 2021).

Personalised RLHF and prompt-based methods. More recently, Reinforcement Learning from Human Feedback (RLHF) has been adapted to heterogeneous users by learning multiple reward models or preference heads aligned with user clusters (Li et al., 2024). In parallel, prompt- and profile-based approaches replace opaque embeddings with editable natural-language profiles, improving transparency and controllability (Ramos et al., 2024a). Soft-prompt methods such as PeaPOD further integrate collaborative and individual signals without updating model parameters (Ramos et al., 2024b).

2.3 Evaluating Personalisation in LLMs

Limits of reference-based evaluation. Traditional metrics such as BLEU and ROUGE measure surface overlap with gold references, while semantic metrics like BERTScore still assume a *user-agnostic* notion of quality. As a result, these metrics correlate poorly with user-perceived quality in personalised settings.

LLM-as-judge evaluators. Reference-free evaluation has gained traction through LLM-as-a-judge approaches. Systems such as AuPEL score personalisation, relevance, and fluency jointly using a strong LLM (Wang et al., 2023), while PerSE conditions the evaluator on explicit user preferences, improving correlation with human judgements (Wang et al., 2024).

Preference-agnostic reference-free evaluators. Parallel work has developed general-purpose LLM evaluators that decompose quality into task-specific criteria. **G-Eval** prompts an LLM to reason over such criteria before scoring (Liu et al., 2023), and **CheckEval** enhances robustness by verifying outputs against automatically generated checklists (Lee et al., 2025). However, these methods assume a *universal* notion of quality: their criteria are fixed across users and cannot encode user-specific priorities or exclusions.

Benchmarks for personalisation. Several benchmarks target complementary aspects of personalised generation: LaMP and LongLaMP evaluate short- and long-form tasks (Salemi et al., 2023; Kumar et al., 2024); PersonaLens simulates multi-session user profiles (Zhao et al., 2025b); PersonaMem studies preference drift over time (Jiang et al., 2025); PersoBench focuses on persona consistency in dialogue (Afzoon et al., 2024); and PrefEval evaluates adherence to explicit and implicit preferences in conversational QA (Zhao et al., 2025a).

2.4 Positioning of This Work

Despite substantial progress, existing evaluators for personalised text generation typically suffer from at least one of three limitations: reliance on costly human annotations, assumption of a universal notion of quality, or narrow task-specific scope. PREF addresses these gaps through a principled redesign of reference-free evaluation.

Specifically, PREF (i) separates *general quality coverage* from *user-specific alignment* via a two-tier evaluation rubric, (ii) operates fully reference-free, enabling scalable and reproducible evaluation, and (iii) is model-agnostic, supporting diverse backbone LLMs and personalisation strategies. Unlike prior personalised evaluators that directly condition a single scorer on user preferences, or general-purpose judges that apply a fixed rubric, PREF makes user alignment an explicit, inspectable component of evaluation.

In the experiments that follow, we focus on open-domain question answering and use **PrefEval** as our primary benchmark, as it stresses both explicit and implicit preference-following and directly probes PREF's core design assumptions.

3 From Personalisation to Evaluation

Algorithm 1 PREF Scoring Pipeline (Prompts and Samples in appendix A and B)

Require: Question set $Q = \{q_n\}_{n=1}^{N}$,
Preference set $P = \{p_n\}_{n=1}^{N}$,
Answer set $A = \{a_n\}_{n=1}^{N}$,
Coverage LLM $\mathcal{M}_{\text{cov}}$,
Preference LLM $\mathcal{M}_{\text{pref}}$,
Scoring LLM $\mathcal{M}_{\text{score}}$
Ensure: Personalised scores $S = \{s_n\}_{n=1}^{N}$
Phase 1: Generate General Guidelines
for each $q_n \in Q$ **do**
$g_n \leftarrow \textsc{GenerateGuideline}(\mathcal{M}_{\text{cov}}, q_n)$
▷ g_n lists factors $\{f_{n,1}, \ldots, f_{n,K_n}\}$
end for
Phase 2: Personalise Guidelines
for each index n from 1 to N **do**
$g_n^* \leftarrow \textsc{Personalise}(\mathcal{M}_{\text{pref}}, q_n, p_n, g_n)$
▷ Re-rank / augment factors to obtain personalised guideline
end for
Phase 3: Score Answers
for each index n from 1 to N **do**
$s_n \leftarrow \textsc{ScoreAnswer}(\mathcal{M}_{\text{score}}, q_n, p_n, g_n^*, a_n)$
end for
return S

3.1 Overview and Design Intuition

We study personalisation in the context of evaluation, where the quality of a generated answer depends on the needs and preferences of a particular user [1]. Following Kirk et al. (2024), we view personalisation as operating within a system's general capabilities: many behaviours may be possible in principle, but only a subset will be useful for a given user in a given context.

PREF is motivated by a simple observation about user attention. If users had unlimited atten-

[1] https://dictionary.cambridge.org/dictionary/english/personalization

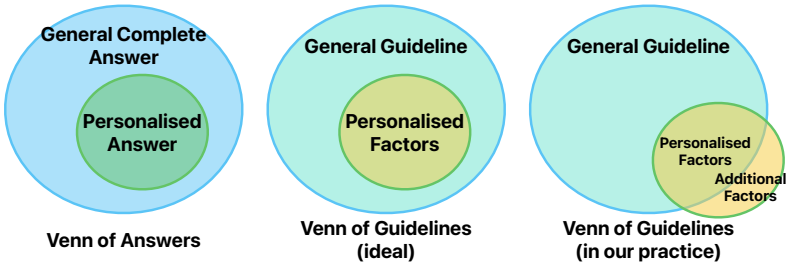

Figure 1: **PREF at a glance.** Venn diagrams illustrating: (**left**) the relationship between an *ideal, exhaustive* answer and a *personalised* answer to the same query; (**centre**) the relationship between a *general* guideline and a *user-specific* selection (or ordering) of evaluation factors; and (**right**) the practical regime in PREF, where the preference stage may add factors when the general guideline omits user-salient constraints.

tion and patience, a single exhaustive answer would suffice, making explicit personalisation largely unnecessary. In practice, attention is limited, and users therefore prefer answers that prioritise information aligned with their preferences, even if less relevant details are omitted.

Figure 1 illustrates this perspective. A personalised answer can be understood as a prioritised slice of an ideal complete answer, and a personalised evaluation rubric as a user-specific prioritisation of general quality factors. In practice, a general guideline may omit criteria that become salient under a particular user profile; PREF therefore permits the preference stage to add targeted factors to prevent such omissions from propagating to the final score.

3.2 The PREF Pipeline

PREF operationalises the above intuition through a modular, three-step pipeline consisting of two rubric-construction stages followed by scoring (Figure 2).

General-quality stage. Given a query q, a *coverage LLM* generates a general evaluation guideline g that enumerates salient factors $\mathcal{F} = \{f_1, \ldots, f_n\}$, such as factuality, coherence, and completeness. This stage is query-driven and explicitly ignores user preferences, ensuring baseline adequacy.

User-alignment stage. Conditioned on a user profile p, a *preference LLM* transforms the general guideline into a personalised guideline g^*. Concretely, it induces either an ordering π over $\mathcal{F}$ or non-negative weights $\mathbf{w} \in \mathbb{R}_{\geq 0}^{|\mathcal{F}|}$. When necessary, it may augment $\mathcal{F}$ with additional user-salient factors. The resulting rubric reflects what matters most for this user.

Scoring step. Finally, a *scoring LLM* evaluates a candidate answer a against the personalised guideline g^* in the context of (q, p), producing a scalar score $s_a \in \mathbb{R}$, written as

$$s_a = \text{Score}(a \mid q, p, g^*).$$

Interpretation. Conceptually, the general guideline g provides *coverage*, specifying what aspects of an answer should be checked, while the personalised guideline g^* provides *alignment*, specifying which aspects matter most for a particular user.

Figure 2 and Algorithm 1 summarise the full pipeline. Given a query q, user profile p, and candidate answer a, the coverage LLM produces g, the preference LLM derives g^*, and the scoring LLM outputs s_a. For simplicity, we instantiate all three roles using a single backbone model.

3.3 Design Rationale and Constraints

Why separate coverage from preference? Separating coverage from preference yields three practical benefits. **(i) Robustness to omission:** coverage is query-driven, while the preference stage can introduce missing but user-critical constraints. **(ii) Transparency:** the guidelines g and g^* form a human-auditable rubric that explains why an answer received a given score. **(iii) Reusability:** the same general guideline can be reused across users, and the same user profile can be applied across related queries, reducing recomputation and enabling controlled ablations.

Constraints and guardrails. The scorer balances universal desiderata (e.g., factuality, completeness, clarity) with user-specific ones (e.g., exclusions, tone, budget). We enforce two invariants: (i) user alignment does not excuse factual

errors, and (ii) newly added factors should not contradict coverage factors without explicit justification. These guardrails preserve reliability while allowing user-specific trade-offs.

4 Experiments

4.1 Dataset

To assess the performance of PREF, we adopt the **PrefEval** benchmark (Zhao et al., 2025a), which supplies *question–preference–answer* triples tailored for personalised text generation. We focus on the *implicit multiple-choice* subset of PrefEval (1000 questions), where the link between user preferences and the "ideal" answer is not stated explicitly. For instance, if a user dislikes fish, a question asking for a recommended Japanese dish is implicitly challenging because many Japanese dishes contain fish, even though the question never mentions it. Each example offers four candidate answers, only one of which a human annotator marks as satisfactory. We restrict our evaluation to PrefEval because it is, to our knowledge, the only publicly available benchmark that provides explicit user preference context together with multiple candidate answers and human preference annotations conditioned on those preferences.

4.2 LLMs

To probe the flexibility of PREF, we pair it with four different backbone language models: **Claude Haiku** (claude-3-haiku-20240307), **GPT-4.1 Mini** (gpt-4.1-mini), **LLaMA 8B** (llama3-8b-instruct), and **LLaMA 70B** (llama3-70b-instruct). For every model, we set the sampling temperature to 0 to guarantee deterministic generation and full reproducibility.

4.3 Baselines

We compare PREF against four baselines. Two baselines, ZERO-SHOT and REMINDER, are adopted directly from the PrefEval paper (Zhao et al., 2025a), using the original prompting strategies and reported results. In addition, we include two widely used automatic evaluation frameworks, G-EVAL (Liu et al., 2023) and CHECKEVAL (Lee et al., 2025), which assess responses by aggregating scores over four dimensions: Naturalness, Coherence, Engagingness, and Groundedness.

Method	LLaMA 8B	LLaMA 70B	Haiku	GPT-4.1 Mini
Zero-shot	27	37	34	-
Reminder	84	91	88	-
PREF (ours)	**94**	**97**	**94**	98

Table 1: Accuracy (%) on the PrefEval *implicit multiple-choice* subset. **Bold** numbers denote the best score for each backbone model. Zero-shot and Reminder results are taken from the PrefEval paper (Zhao et al., 2025a).

5 Results and Discussion

5.1 Evaluating Personalisation

Baselines and overall accuracy. Table 1 compares the preference on the *PrefEval* implicit multiple-choice benchmark: ZERO-SHOT, where the backbone LLM answers with its default prompt; REMINDER, which prepends a single instruction to consider the user preference; and PREF, our full two-stage framework. Zero-shot and Reminder results are taken from the PrefEval paper (Zhao et al., 2025a)

Across all backbone models, PREF achieves the highest accuracy. Compared to the Reminder baseline, PREF yields absolute gains of $+6$–10 points ($\approx 6\%$–12% relative) on all models where Reminder results are available. Relative to the Zero-shot setting, PREF improves accuracy by more than $+60$ absolute points across backbones. These results indicate that while lightweight preference cues provide some benefit, explicitly constructing and weighting a personalised evaluation rubric is essential for accurate personalised evaluation, regardless of backbone model capacity.

Beyond accuracy: calibration and ranking quality. Accuracy alone captures only whether the correct answer is top-ranked. To assess score calibration and ranking quality, we additionally report mean-squared error (MSE) and normalised discounted cumulative gain (nDCG), mapping gold answers to 10 and distractors to 0.

Table 2 reveals two consistent patterns. The largest backbone (*LLaMA 3 70B*) achieves both high accuracy and the lowest MSE, indicating more fine-grained and better-calibrated scoring when paired with PREF. In contrast, smaller models (*Claude 3 Haiku* and *LLaMA 3 8B*) attain near-perfect nDCG despite substantially higher MSE: they reliably rank the correct answer first but tend to over-score distractors. Notably, *GPT-4.1 Mini* achieves the highest accuracy and the second-lowest MSE, suggesting that despite its smaller

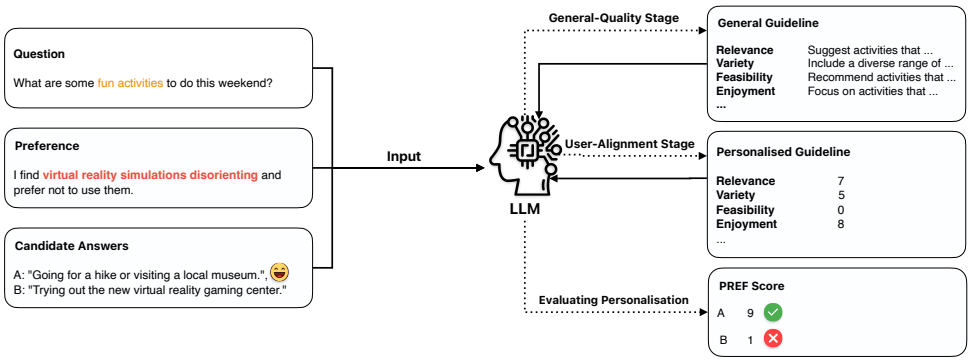

Figure 2: **The PREF scoring pipeline.** A three-step process—coverage, preference (user profiles), and scoring—maps a query q, user profile p, and candidate answer a to a personalised score s_a (shown with an example from PrefEval).

Method	Model	Accuracy ↑	MSE ↓	nDCG ↑
PREF	GPT4.1 Mini	**0.98**	2.11	**0.9992**
PREF	Haiku	0.94	4.02	0.9907
PREF	LLaMA 8B	0.94	4.77	0.9894
PREF	LLaMA 70B	0.97	**1.87**	0.9980
CheckEval	GPT4.1 Mini	0.48	21.36	0.9852
CheckEval$_C$		0.56	20.80	0.9530
CheckEval$_E$		0.77	23.36	0.9381
CheckEval$_G$		0.79	20.72	0.9805
CheckEval$_N$		0.25	20.87	0.8904
G-Eval	GPT4.1 Mini	0.76	24.27	0.8469
G-Eval$_C$		0.76	24.81	0.8012
G-Eval$_E$		0.75	24.89	0.7546
G-Eval$_G$		0.75	25.00	0.7273
G-Eval$_N$		0.74	22.59	0.8207

Table 2: **Performance on the *PrefEval* benchmark**. Higher values are better for Accuracy and nDCG, while lower is better for MSE. **Boldface** marks the best score for each metric. CheckEval and G-Eval are introduced as a comparison, here C for coherence, E for engagingness, G for groundedness and N for naturalness.

size, its more recent architecture yields strong calibration and ranking performance.

Comparison with prior reference-free evaluators. Table 2 also reveals a substantial gap between PREF and prior reference-free evaluators such as *CheckEval* and *G-Eval*. While both baselines perform reasonably in generic settings, their accuracy and calibration degrade sharply on personalised evaluation. CheckEval variants exhibit particularly low accuracy and high MSE, indicating difficulty separating generically plausible from user-misaligned answers. G-Eval is more stable but still lags behind PREF, especially in nDCG. These results suggest that task-centric, user-agnostic criteria are insufficient for personalised evaluation.

Effect of the user-alignment stage. To isolate the contribution of Stage 2, we ablate the user-alignment stage and score answers using only general guidelines. As shown in Table 4, removing Stage 2 consistently degrades performance across all backbones, confirming that explicit user alignment is a critical component of PREF.

In summary, PREF's two-stage rubric provides three key benefits: **(i) robust accuracy** through explicit preference integration; **(ii) improved calibration** by discouraging inflated scores on user-irrelevant answers; and **(iii) capacity amplification**, enabling smaller backbones to approach the performance of much larger models. Together, these results position PREF as a scalable and model-agnostic framework for evaluating personalised language generation.

5.2 Explainability

Leveraging PrefEval's answer explanations. Unlike many personalisation benchmarks, *PrefEval* supplies not only a gold (personalised) answer for each query but also a *natural-language explanation* describing *why* that answer is appropriate. This additional signal allows us to probe a different capability of PREF: its ability to identify and prioritise the *right* evaluation factors.

Concretely, we treat the explanation as an "oracle" preference and ask: *Does the ranking produced by* PREF *(when conditioned on the user's preference) agree with the ranking obtained when we instead condition on the explanation?* Formally, let e be the accompanying explanation supplied by *PrefEval* and $g(q) = \{f_1, \ldots, f_n\}$ be the general guideline produced in Stage 1.

We then execute the following steps: PREFERENCE-BASED RANKING. Feed p

	GPT-4.1 Mini	LLaMA 70B
Pearson r	**0.6164**	0.4693
Spearman ρ	**0.5906**	0.3811
Kendall τ	**0.5612**	0.3532

Table 3: Rank–rank correlation between the factor ordering induced by the user preference and the oracle ordering derived from *PrefEval* answer explanations. Higher values indicate closer agreement.

into the preference LLM to derive an ordering π_p (or weight vector $\mathbf{w}_p$) over $g(q)$. EXPLANATION-BASED RANKING. Replace p with the explanation e and obtain a second ordering π_e (or $\mathbf{w}_e$). We regard π_e as an oracle since it reflects the human rationale behind the gold answer. COMPARE THE RANKINGS. Measure correlation coefficients between π_p and π_e with a suitable correlation metric (e.g., Kendall's τ, Spearman's ρ and pearson's p). High correlation indicates that the preference-conditioned rubric surfaces *the same* factors humans deem decisive.

This procedure isolates the *factor-ranking* competence of PREF: even if the final scores differ, we can verify whether the framework attends to the criteria that matter most for a high-quality personalised answer.

Table 3 reports the correlation between the factor ranking produced by PREF when conditioned on the *user preference* and the "oracle" ranking obtained from *PrefEval*'s human-written explanations. All three coefficients—Pearson's r, Spearman's ρ, and Kendall's τ—are *positive and statistically substantial*, confirming that both backbones identify broadly the same criteria that humans deemed decisive.

GPT-4.1 Mini shows the stronger alignment, achieving a Pearson r of 0.62 and a Kendall τ of 0.56, whereas LLaMA 3 70B trails at 0.47 and 0.35, respectively. In practical terms, GPT-4.1 Mini not only places the correct factors near the top but preserves their relative ordering more faithfully. The gap suggests that, despite its larger parameter count, LLaMA 3 70B still benefits from the second stage (accuracy gains in Table 1) yet requires further tuning to internalise nuanced preference cues.

Overall, the moderate–high correlations indicate that PREF's preference module successfully surfaces the factors humans reference when justifying a good personalised answer, but there remains headroom—especially for open-source models—to tighten that alignment.

5.3 Additional Factors in Stage Two

In Stage 2, the preference model may *augment* the general guideline. Two concise observations from the first 200 questions:

(1) Additions are few. With *GPT-4.1 Mini*, we observe 1,986 general factors vs. 153 added ($\approx 7.7\%$); on *LLaMA 3 70B*, 1,893 vs. 241 ($\approx 12.7\%$). Per question, this is 9.93 general and 0.765 additional factors on average. Thus, coverage already captures most criteria; Stage 2 makes targeted fixes.

(2) Additions encode exclusions. Keyword filtering (`dislike`, `avoid`) shows that 25.49% of added factors (vs. 0% general) on *GPT-4.1 Mini* and 5.81% (vs. 0.01%) on *LLaMA 3 70B* express user-specific "blacklists."

In summary, augmentation is *selective* and primarily used to inject user-dependent constraints that the preference-agnostic coverage stage omits.

5.4 Ablation Study

We conduct ablation experiments to isolate the contributions of PREF's two components—general-quality coverage (Stage 1) and user alignment (Stage 2)—and to validate the design choice of separating coverage from preference.

Effect of removing user alignment. We first ablate Stage 2 and score answers using only the general guideline from Stage 1 (*w/S1*). As shown in Table 4, removing user alignment consistently degrades performance across all backbones. Accuracy and nDCG decrease, while MSE increases substantially, indicating poorer calibration. These results show that explicit preference weighting is essential for distinguishing generically plausible but user-misaligned answers, particularly for smaller models.

Contribution of each stage. We further compare three variants: **base** (direct scoring without guidelines), **w/S1** (coverage only), and full PREF. Two trends emerge. First, the coverage stage alone yields large gains over the base setting, especially for LLaMA 3 8B, demonstrating that explicit factor enumeration provides a strong inductive bias. Second, adding user alignment on top of coverage further improves performance, most notably calibration, yielding the lowest MSE for GPT-4.1 Mini and LLaMA 3 8B.

Claude 3 Haiku shows a slight degradation with coverage alone, suggesting that generic rubrics can

Method	Accuracy ↑	MSE ↓	nDCG ↑
GPT4.1 Mini			
PREF_{base}	0.97	2.20	**0.9998**
$\text{PREF}_{w/S1}$	0.97	3.01	0.9997
PREF	**0.98**	**2.11**	0.9992
LLaMA 8B			
PREF_{base}	0.85	14.36	0.9329
$\text{PREF}_{w/S1}$	0.92	5.71	**0.9889**
PREF	**0.94**	**4.77**	0.9894
Claude Haiku			
PREF_{base}	0.90	6.77	0.9794
$\text{PREF}_{w/S1}$	0.86	9.21	0.9748
PREF	**0.94**	**4.02**	**0.9907**

Table 4: **Ablation on the two stages of PREF**. Here, $base$ means prompting the model directly (no General-quality stage and User-alignment stage); $w/S1$ means prompting the model with only the answers from the Coverage stage.

occasionally conflict with model heuristics; this effect disappears once user alignment is introduced.

Overall, the ablations confirm that (i) coverage and preference play complementary roles, (ii) coverage alone is insufficient for personalised evaluation, and (iii) explicit user alignment is particularly beneficial for capacity-limited backbones, empirically justifying PREF's two-stage design.

6 Discussion

Our experimental results yield four key insights.

Fine-grained personalisation is critical. Evaluating personalised generation requires reasoning at the level of individual factors rather than coarse, global criteria. By explicitly modelling factor-level priorities, PREF improves accuracy by $6\% - 12\%$ over the one-line *Reminder* baseline (Table 1), demonstrating that lightweight preference cues are insufficient.

Separating coverage from preference is essential. Ablation results show that removing the user-alignment stage consistently degrades calibration (higher MSE) and ranking quality (lower nDCG), with the largest impact observed for the 8B model (Table 4). While general coverage provides a strong inductive bias, it cannot capture user-specific trade-offs on its own.

Evaluation can compensate for limited model capacity. When paired with PREF, *LLaMA 3 8B* approaches the performance of substantially larger backbones such as *Claude 3 Haiku*. This suggests that improved evaluation and guidance—not merely increased parameter counts—can substantially narrow the quality gap in personalised settings.

Learned factor rankings align with human reasoning. Rankings induced by user preferences correlate moderately to strongly with "oracle" rankings extracted from human-written explanations (Table 3). This alignment indicates that PREF surfaces evaluation criteria that humans themselves consider decisive when judging personalised quality.

7 Conclusion

We introduced PREF, a **P**ersonalised, **R**eference-free **E**valuation **F**ramework that explicitly disentangles *coverage* from *preference*. By first ensuring baseline adequacy and then tailoring evaluation criteria to individual users, PREF delivers accurate, interpretable, and scalable assessment without requiring gold personalised references. Experiments on the *PrefEval* benchmark show that PREF (i) aligns closely with human judgements, (ii) improves calibration and ranking quality across backbone models, and (iii) substantially raises the performance ceiling of smaller models.

Code and evaluation scripts will be released upon publication. We hope that PREF helps catalyse research into genuinely user-centred language generation—and into evaluation methodologies capable of keeping pace with that ambition.

8 Limitations

Evaluator reliability. PREF relies on LLMs to construct evaluation guidelines, personalise factor weightings, and assign final scores. Consequently, it inherits known limitations of LLM-based evaluators, including sensitivity to prompt phrasing, latent biases, and occasional reasoning failures. Although we mitigate variance through deterministic decoding (temperature $= 0$), systematic biases may still affect both factor construction and scoring. Hybrid approaches—such as ensembling multiple judges, incorporating lightweight human audits, or combining LLM judgments with symbolic or statistical checks—represent promising directions for improving robustness.

User preference representations. Our framework assumes access to relatively concise and coherent user preference signals, expressed either explicitly or implicitly through benchmark-provided

context. In real-world settings, however, preferences are often noisy, incomplete, evolving over time, or inferred indirectly from behavioural signals. PREF does not explicitly model preference uncertainty or temporal drift, and its performance may degrade when preferences are sparse, contradictory, or outdated. Extending the user-alignment stage to handle uncertain, implicit, or dynamic preferences remains an open challenge.

Task and domain generality. Our empirical evaluation focuses on open-domain question answering using the PrefEval benchmark. While this setting captures core challenges in personalised evaluation, it does not cover other important generation tasks such as long-form summarisation, creative writing, code generation, or multimodal outputs. Applying PREF to these domains may require richer coverage representations, task-specific factor schemas, or alternative aggregation strategies.

Scalability and computational cost. Although PREF avoids expensive human evaluation and gold personalised references, it requires multiple LLM calls per instance, introducing non-trivial computational overhead compared to single-pass evaluators. While smaller backbone models benefit substantially from PREF, large-scale deployment may require caching, guideline reuse, or further optimisation to reduce inference costs.

Ethical considerations. Personalised evaluation frameworks risk reinforcing filter bubbles or amplifying sensitive user attributes if deployed without safeguards. Although PREF is an evaluation framework rather than a user-facing generation system, its outputs may influence model development and optimisation in ways that indirectly affect user experiences. Future work should incorporate fairness diagnostics, privacy-aware preference handling, and mechanisms that allow users or developers to inspect, constrain, or override inferred preferences.

Overall, these limitations highlight important directions for future research while underscoring that PREF is intended as a modular and extensible foundation rather than a complete solution to personalised evaluation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Peter Brusilovsky. 2001. Adaptive hypermedia. *User modeling and user-adapted interaction*, 11(1):87–110.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations (ICLR 2020)*. ArXiv:1912.02164.

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

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15782–15809, Suzhou, China. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of

LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL–IJCNLP 2021)*, pages 4582–4597.

Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jerome Ramos, Hossen A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024a. Transparent and scrutable recommendations using natural language user profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 13971–13984.

Jerome Ramos, Bin Wu, and Aldo Lipani. 2024b. Peapod: Personalized prompt distillation for generative recommendation. *arXiv preprint arXiv:2407.05033*.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th international conference on natural language generation*, pages 355–368.

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

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025a. Do llms recognize your preferences? evaluating personalized preference following in llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. 2025b. Personalens: A benchmark for personalization evaluation in conversational ai assistants. *arXiv preprint arXiv:2506.09902*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Prompts

This appendix presents the prompt templates used in PREF in Figure 3, 4 and 5. All For reproducibility, we use temperature $= 0$.

Stage 1: General-quality Prompt

Given the following question:
{question}

Please provide points for what makes a good answer to this question. Each point should be a pair of a keyword and a sentence. for example:

"Goals":"Understand the user's goals and What the user wants to achieve"

output as python dict:

Figure 3: Stage 1 (General-quality) prompt used in PREF to elicit a query-specific, user-agnostic evaluation guideline. The prompt instructs the LLM to enumerate key factors defining a good answer, independent of any user preferences.

Stage 2: User-alignment Prompt

Given the following question:
{question}

Here is a list of factors that from an expert that should be considered to answer this question:
{ge}

Here is a statement of the user's preference:
{preference}

When answering the question, what factors are highly related to the user's preference? Reply with the full list of factors where highly related factors with the score from 0 to 10, and the factors that are not addressed with 0. if there are factors don't exist in the list of factors, add new factors with the keywords start with prefix 'preh_'. Remenber most of the factors are not highly related.

output as json dict:

Figure 4: Stage 2 (User-alignment) prompt used in PREF. Given a general evaluation guideline and a user preference statement, the LLM re-weights existing factors and introduces new preference-specific factors when necessary, producing a personalised evaluation rubric.

Step 3: Scoring Prompt

Given the following question:
{question}

Here is the user's preference:
{preference}

Here is a list of candidate answers to the question:
{choice_list}

Here is a list of factors that should be considered to answer this question:
{factor_list}
(a high score means the factor is highly related to the ideal answer)

Consider the user's preference and the list of factors, score each choice from 0 to 10. output as a json list, each item is the score of the corresponding candidate answer. The list should only include scores, no other text.

Figure 5: Stage 3 (Scoring) prompt used in PREF. The LLM scores candidate answers by jointly considering the user preference and the personalised factor list, outputting scalar scores without relying on gold references.

B Example

```
{
"preference": "I have a strong aversion to digital or online learning environments
and prefer traditional, in-person classroom settings.",
"question": "What are some good ways to learn a new language?",
"choice_list": [
    "Enroll in an in-person language course ...",
    "Sign up for an online language ...",
    "Use a language learning app ...",
    "Join an online language
    exchange community..."
],
"ge": {
    "Consistency": "Practice the language regularly to build and retain
    skills over time.",
    "Immersion": "Surround yourself with the language through media, conversations,
    and environments where it is spoken.",
...
    "Cultural Understanding": "Learn about the culture associated with the language
    to gain context and motivation."
},
"preh": {
    "Consistency": 5,
    "Immersion": 7,
...
    "Cultural Understanding": 5,
    "preh_InPersonLearning": 9
},
"scoring": [
    9,
    1,
    0,
    0
]
},
...
```

Figure 6: Illustrative JSON example showing a complete PREF instance, including the user preference, query, candidate answers, general evaluation factors (`ge`), personalised factor weights (`preh`), and final scores assigned to each candidate answer.