PREF: Reference-Free Evaluation of Personalised Text Generation

Anonymous ACL submission

Abstract

Personalised text generation is increasingly used in user-facing systems, but evaluating it remains difficult: a response must be not only generally correct and fluent, but also appropriate for a particular user's goals, constraints, and preferences. Human evaluation can capture this user-conditioned quality but is costly and difficult to scale, while existing automatic metrics and LLM judges largely assume a universal notion of quality. We introduce **PREF**, a **P**ersonalised **R**eference-free **E**valuation **F**ramework that explicitly separates general answer quality from user-specific alignment by first constructing query-specific evaluation criteria, then adapting them using the target user's profile, preferences, and context before applying an LLM judge for scoring. Across five personalised generation benchmarks, PREF aligns more closely with human preferences and makes more effective use of profile information than strong generic and personalised evaluation baselines. On the PrefEval benchmark, PREF even allows substantially smaller judge models to outperform stronger generic evaluation pipelines. Code is available anonymously.[1]

1 Introduction

Large language models (LLMs) are increasingly used for open-ended text generation in dialogue, recommendation, code synthesis, and data-to-text applications (Brown et al., 2020; Achiam et al., 2023), making reliable evaluation of their outputs increasingly important. Traditional reference-based metrics such as BLEU and ROUGE (Papineni et al., 2002; Lin, 2004) rely on overlap with gold references and are poorly suited to open-ended tasks where multiple responses may be equally valid. Embedding-based metrics such as BERTScore (Zhang et al., 2020) mitigate surface-form mismatch, but still depend on reference texts and primarily capture generic semantic similarity rather than holistic response quality. LLM-as-a-judge methods offer a scalable alternative by asking a strong model to rate or rank candidate responses (Li et al., 2025; Zheng et al., 2023; Li et al., 2023), and benchmarks such as MT-Bench (Zheng et al., 2023) and AlpacaEval (Li et al., 2023) show that such judges can align well with crowd preferences in general assistant settings.

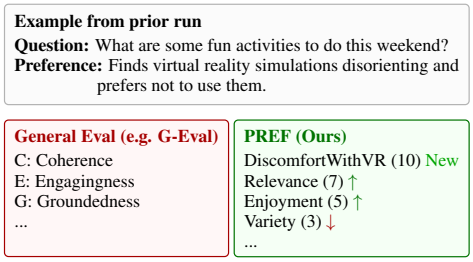

Figure 1: Key difference between general LLM-based evaluation and PREF. General evaluation uses a fixed rubric, while PREF reweights and augments a general rubric with the target preferences. Two detailed examples in Appendix F (Figure 12, 13)

Personalised generation, however, changes the evaluation target. A high-quality response must be not only useful for the task, but also appropriate for the specific user's goals, constraints, and preferences. These requirements do not always align. A travel recommendation, writing suggestion, or assistant action may be appropriate for one user yet unsuitable for another, while an answer to an objective factual question should not be rewarded merely because it reflects a user's preferences. Modern LLM applications increasingly rely on user profiles, histories, and feedback signals to adapt outputs to individuals (Zhang et al., 2024), but most evaluators still assume a universal notion of quality. Human evaluation can capture user-conditioned distinctions, but collecting per-

[1] https://anonymous.4open.science/r/pref_anon-D778/

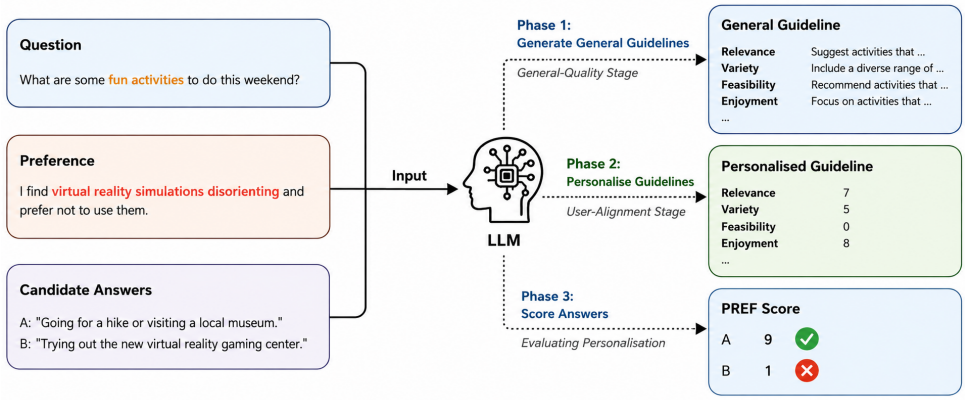

Figure 2: **The PREF scoring pipeline.** Coverage, preference, and scoring map (q, p, a) to a personalised score s_a.

sonalised judgements for every system, prompt, or user group is expensive and difficult to scale (Van Der Lee et al., 2019).

Existing approaches fail in opposite ways. Generic evaluators apply a fixed rubric regardless of the user, potentially overlooking user-specific constraints. Naively personalised judges, on the other hand, risk over-personalising cases where factuality, completeness, or task correctness should dominate. Consider a user who dislikes crowded places and prefers quiet cultural activities. For the query "What should I do in Paris this weekend?", a generic judge may reward a popular itinerary that ignores the profile. For "What is the capital of France?", a purely personalised judge should not prefer an answer that compromises factual correctness in favour of user preference. Mixed queries such as "Plan one afternoon in Paris for someone who dislikes crowds" require both objective usefulness and user alignment. Our key insight is that personalised evaluation should therefore not collapse quality into either a purely objective notion of correctness or a purely preference-matching notion of alignment.

We propose PREF, a **P**ersonalised **R**eference-free **E**valuation **F**ramework for scoring generated text with user-aware rubrics. PREF separates evaluation into three connected stages. First, a *general-quality stage* constructs a query-specific rubric capturing task-level criteria such as truthfulness, coherence, completeness, clarity, and specificity. Second, a *user-alignment stage* adapts this rubric using the user's profile, history, or stated preferences by reweighting existing criteria and introducing missing user-specific constraints where necessary. Finally, a scoring stage evaluates candidate answers against the resulting personalised rubric. This separation is the central design of PREF: user preferences can influence which factors matter most, while universal requirements remain explicit in the evaluation process. Figure 1 illustrates this distinction.

We evaluate PREF across settings that probe both *human preference alignment* and broader *personalised generation*. PrefEval (Zhao et al., 2025a) serves as our primary benchmark because it provides multiple candidate answers with user-conditioned human preference annotations, enabling direct comparison between evaluators. To test whether PREF generalises beyond this multiple-choice setting, we also evaluate on PerMPST (Wang et al., 2024), LongLaMP (Kumar et al., 2024), PersonaLens (Zhao et al., 2025b), and LaMP-QA (Salemi et al., 2023), covering scalar personalised ratings, long-form generation, task-oriented assistant personalisation, and personalised question answering. Across these settings, we ask whether PREF better matches human-conditioned preferences than existing evaluators and whether its explicit user-aware rubrics make more effective use of profile information. Our contributions are:

- We introduce PREF, a reference-free evaluation framework for personalised generation that separates general task adequacy from user-specific alignment through rubric construction, user-conditioned adaptation, and LLM-based scoring.
- Across multiple personalised generation benchmarks, PREF aligns more closely with human preferences and makes more effective use of profile information than strong generic and personalised baselines.
- We analyse robustness and interpretability

through prompt variants, stronger-judge audits, factor-ranking correlations, and qualitative rubric analysis.

Taken together, these results position PREF as a practical and interpretable framework for evaluating user-aligned language generation.

2 Related Work

2.1 Personalised Text Generation with LLMs

Personalisation in information systems has long combined *explicit user models* with *implicit preference inference*, from adaptive hypermedia and collaborative filtering (Brusilovsky, 2001; Resnick et al., 1994) through early neural persona modelling (Li et al., 2016) and controllable generation via plug-and-play or parameter-efficient methods (Dathathri et al., 2020; Li and Liang, 2021). Modern LLM-based personalisation builds on this lineage: Reinforcement Learning from Human Feedback has been extended to heterogeneous users via per-cluster reward or preference models (Li et al., 2024), while prompt- and profile-based approaches replace opaque embeddings with editable natural-language profiles or soft prompts that combine collaborative and individual signals (Ramos et al., 2024a,b). Our focus is not on these generation methods themselves, but on how their outputs should be *evaluated* when correctness depends on the individual user.

2.2 LLM-based and Reference-free Evaluation

Reference-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020) measure overlap or semantic similarity with gold references and assume a *user-agnostic* notion of quality, leading to poor correlation with user-perceived quality in personalised settings. Reference-free LLM-as-a-judge methods avoid gold references by prompting a strong model to reason about quality (Zheng et al., 2023; Li et al., 2023, 2025). **G-Eval** decomposes evaluation into task-specific criteria and asks an LLM to reason before scoring (Liu et al., 2023), and **CheckEval** improves robustness by verifying outputs against automatically generated checklists (Lee et al., 2025). However, both apply a *fixed* rubric derived from the task description and cannot represent user-specific priorities or exclusions, which is precisely what personalised evaluation requires.

2.3 Personalised Evaluation and Benchmarks

A second line of work targets personalisation directly. Reference-free personalised evaluators score *personalisation*, relevance, and fluency jointly with a strong LLM (Wang et al., 2023), or condition the evaluator on explicit user preferences to improve correlation with human ratings (Wang et al., 2024). Complementary benchmarks evaluate models across formats: LAMP and LONGLAMP cover short- and long-form personalised tasks (Salemi et al., 2023; Kumar et al., 2024); PERSONALENS simulates rich multi-session user profiles for task-oriented assistants (Zhao et al., 2025b); PERSONAMEM probes preference drift across sessions (Jiang et al., 2025); PERSOBENCH measures persona consistency in dialogue (Afzoon et al., 2024); and PREFEVAL evaluates adherence to explicit and implicit preferences in conversational QA (Zhao et al., 2025a). These benchmarks largely focus on whether *generators* personalise well; they do not by themselves provide a general reference-free *evaluator* that separates baseline adequacy from user-specific alignment.

2.4 Positioning of This Work

Existing evaluators for personalised 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: it (i) separates *general quality coverage* from *user-specific alignment* via a two-tier rubric, (ii) operates fully reference-free, and (iii) is model-agnostic. Unlike prior personalised evaluators that condition a single scorer directly on user preferences, or general-purpose judges that apply a fixed rubric, PREF makes user alignment an explicit, inspectable component of evaluation. PrefEval (Zhao et al., 2025a) is our primary human-preference benchmark, providing direct evaluator-to-human comparison against zero-shot, reminder, G-Eval (Liu et al., 2023), and CheckEval (Lee et al., 2025) baselines. We further test whether the same evaluator behaviour generalises across broader personalised generation settings on Per-MPST (Wang et al., 2024) (scalar personalised ratings against PERSE and GPT-4 baselines), and on LongLaMP (Kumar et al., 2024), PersonaLens (Zhao et al., 2025b), and LaMP-QA (Salemi et al., 2023), where we measure how each eval-

uator uses profile information across long-form, task-oriented assistant, and personalised QA tasks.

3 A Framework for Personalised Evaluation

Algorithm 1 PREF Scoring Pipeline

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 Technical Challenge and Design Intuition

Reference-free evaluation of personalised generation must separate two related questions: whether an answer is generally adequate for the query, and whether it prioritises what matters to a particular user. A direct LLM-as-a-judge prompt that simply appends the user profile can entangle these questions: the judge may over-reward surface-level preference matching while overlooking factuality or task completion, or it may apply a generic rubric and ignore the profile. Such direct scoring is also difficult to inspect because the role of the profile is hidden inside a single judgement.

PREF addresses this challenge by making evaluation criteria explicit: personalisation should act as a constrained prioritisation in which a personalised answer is a focused slice of an ideal complete answer, and a personalised rubric is a user-specific prioritisation of general quality factors. We therefore first ask what a high-quality answer should cover, and only then ask which parts of that coverage should matter most for the target user. This intuition motivates a modular pipeline, summarised in Figure 2 and Algorithm 1, in which coverage is established before user-specific alignment is applied.

3.2 The PREF Pipeline

PREF maps a query q, user profile p, and candidate answer a to a personalised score through three steps (Figure 2; Algorithm 1).

Coverage guideline generation. Given q, a coverage LLM generates a general guideline g with factors $\mathcal{F} = \{f_1, \ldots, f_n\}$, such as factuality, coherence, completeness, and task-specific constraints. This guideline is intentionally user-agnostic: it describes what should be checked for a strong answer before considering individual preferences. This prevents the evaluator from using the profile to obscure basic quality failures and provides a stable reference point for later ablations.

Preference-guided rubric personalisation. Given p and g, a preference LLM transforms the general guideline into a personalised guideline g^*. The transformation may rank factors, assign relative weights, or augment g with user-salient criteria that were missing from the initial coverage stage. For example, two users may ask the same question but differ in whether they value concision, safety constraints, implementation details, or domain-specific assumptions. By representing these differences as changes to the rubric, PREF makes explicit where personalisation enters the evaluation.

Personalised answer scoring. Finally, a scoring LLM evaluates a against g^* in the context of (q, p):

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

Scoring against g^* is more controlled than directly asking the judge to "score for this user", because the personalised criteria are supplied as an explicit object rather than left implicit in the prompt. For simplicity, our experiments instantiate the coverage, preference, and scoring roles with a single backbone model, but the roles are conceptually modular and can be replaced independently.

3.3 Why Decompose Personalised Evaluation?

The decomposition is the main methodological design choice in PREF. First, it improves inspectability: g shows the task-level criteria, while g^* shows how the user profile changes those criteria. Second, it supports controlled comparisons, since the

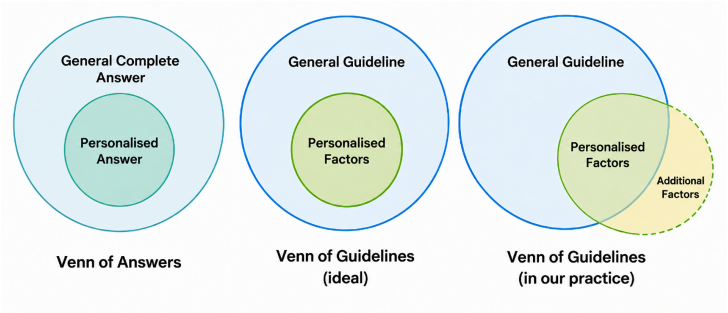

Figure 3: **PREF at a glance.** A personalised answer preserves task adequacy while prioritising the target user's needs. PREF mirrors this in evaluation: it builds a general rubric, then *reweights* and *augments* it with user-specific criteria before LLM scoring.

same general guideline can be reused across different users, and personalised and non-personalised rubrics can be compared directly. Third, it reduces omission risk: when the coverage stage misses a user-critical constraint, the preference stage can add it rather than forcing the scorer to infer it implicitly.

We keep two guardrails to avoid turning personalisation into unconstrained preference chasing. User alignment does not excuse factual errors or violations of core task requirements, and newly added factors should not contradict general-quality criteria without explicit justification. These constraints preserve the reliability of reference-free evaluation while still allowing PREF to capture user-specific trade-offs.

4 Experiments

We evaluate PREF through three research questions that mirror the paper's contributions:

RQ1: Does explicit rubric decomposition improve human-preference alignment and calibration on PrefEval? (Section 5.1)

RQ2: Does PREF generalise to external personalised-generation benchmarks and use profile information constructively? (Section 5.2)

RQ3: Are the intermediate rubrics interpretable and robust? (Section 5.3)

Benchmarks We evaluate PREF in two complementary regimes. First, PrefEval (Zhao et al., 2025a) provides direct human preference labels over four candidate answers, enabling accuracy, calibration, and ranking evaluation. Second, external personalised-generation benchmarks test broader forms of user conditioning: Per-MPST (Wang et al., 2024) provides reviewer histories and scalar movie ratings, while LongLaMP (Kumar et al., 2024), PersonaLens (Zhao et al., 2025b), and LaMP-QA (Salemi et al., 2023) provide personalised targets or references for profile-sensitivity analysis. Full benchmark coverage and metric definitions are given in Appendix A.

LLMs We instantiate PREF with multiple backbone LLMs on PrefEval and use GPT-4.1 Mini as the default judge for the external-benchmark experiments. All reported scoring runs use temperature 0 for deterministic evaluation. Appendix B lists the exact models, judge settings, and stronger-judge audit protocol.

Baselines and Protocols On PrefEval, we compare PREF with two prompting baselines (ZERO-SHOT and REMINDER) (Zhao et al., 2025a), G-Eval (Liu et al., 2023), and CheckEval (Lee et al., 2025). On LongLaMP, PersonaLens, and LaMP-QA, we compare each evaluator with and without access to profile information; the main quantity of interest is the signed score change after adding the profile. On Per-MPST, we follow the PERSE setting with three annotated reviews ($k = 3$) as reviewer history and report Pearson, Spearman, and Kendall correlations with human ratings. Additional protocol details are provided in Appendix A.

5 Results

5.1 Preference Alignment and Calibration

This subsection answers **RQ1** by testing whether explicit rubric decomposition improves preference alignment, score calibration, and personalised answer identification on PrefEval.

PrefEval threshold accuracy. Figure 4 compares three strategies on the PrefEval implicit multiple-choice benchmark: ZERO-SHOT (default

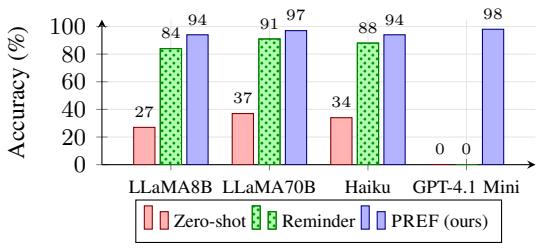

Figure 4: PrefEval accuracy per backbone. Zero-shot and Reminder are from (Zhao et al., 2025a); zero-shot and reminder results are not available for GPT-4.1 Mini. Confidence intervals are in Appendix D.

Method	Model	Threshold Acc. ↑	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 1: PrefEval performance. C/E/G/N denote fixed-aspect variants of the baseline evaluators: coherence, engagingness, groundedness, and naturalness. PREF constructs a personalised rubric for each query–user pair. Notably, PREF with smaller backbone (LLaMA 8B) outperforms GPT-4.1-Mini-based G-Eval and CheckEval on threshold accuracy and nDCG. Confidence intervals are in Appendix D.

prompting), REMINDER (adding a short instruction to consider the user preference), and PREF, our full framework. Zero-shot and Reminder results are taken from the PrefEval paper (Zhao et al., 2025a).

Across all backbone models, PREF achieves the highest threshold accuracy. Compared with Reminder, PREF yields absolute gains of roughly +9–14 points on models where Reminder results are available; compared with Zero-shot, it improves accuracy by more than +60 points. This shows that preference awareness is not simply a matter of adding more instructions: the evaluator benefits from explicitly constructing and applying a personalised rubric.

Calibration and ranking quality. To further answer **RQ1**, we evaluate whether PREF's scores are calibrated, not merely whether the preferred answer is identified. We report mean-squared error (MSE) and normalised discounted cumulative gain (nDCG). Table 1 shows that PREF combines strong accuracy with better score calibration than G-Eval and CheckEval, which often assign high scores to distractors when generic quality correlates only weakly with the user's preference. GPT-4.1 Mini and LLaMA 3 70B reach near-perfect accuracy and the lowest MSE, while smaller backbones also identify the preferred answer in most cases. Notably, PREF with LLaMA 3 8B outperforms GPT-4.1-Mini-based G-Eval and CheckEval on all three metrics.

LLM	Method	Accuracy ↑	MSE ↓	nDCG ↑
GPT-4.1 Mini	PREF$_{base}$	0.97	2.20	**0.9998**
	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
	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
	PREF$_{w/S1}$	0.86	9.21	0.9748
	PREF	**0.94**	**4.02**	**0.9907**

Table 2: Ablation of PREF stages. $base$ uses direct scoring; $w/S1$ uses only the coverage stage. Confidence intervals are in Appendix D.

Ablating rubric decomposition. The ablation explains why PREF answers **RQ1** positively. Table 2 isolates the contribution of general-quality coverage (Stage 1) and user alignment (Stage 2). Coverage alone can improve over direct scoring, especially for LLaMA 3 8B, but it gives less reliable scores than the full framework. Adding user alignment yields the lowest MSE for every backbone and improves accuracy for the smaller open-weight and Haiku judges. General coverage is therefore useful but incomplete: it defines what makes an answer adequate, while the user-alignment stage determines which adequate answer best fits the individual user.

Overall, the answer to **RQ1** is positive: PREF improves personalised answer identification, produces better calibrated scores, and helps smaller backbones through structured evaluation.

5.2 Generalisation to External Benchmarks

This subsection answers **RQ2**: whether PREF generalises beyond PrefEval and uses profile information constructively in external personalised-generation settings. Per-MPST (Wang et al., 2024) provides direct human ratings for personalised movie-review evaluation, while LongLaMP (Kumar et al., 2024), PersonaLens (Zhao et al., 2025b), and LaMP-QA (Salemi et al., 2023) let us study how evaluators respond when profile information

Method	Pearson	Spearman	Kendall
LLaMA-2-7B	0.15	0.12	0.09
LLaMA-2-13B	0.17	0.18	0.15
LLaMA-2-70B	0.21	0.23	0.18
GPT-4	0.32	0.31	0.25
PERSE-7B	0.31	0.33	0.26
PERSE-13B	0.35	0.37	0.29
PREF (OURS)	**0.45**	**0.44**	**0.34**

Table 3: Correlation with human ratings on Per-MPST ($k = 3$). Baselines are from the existing work (Wang et al., 2024); tests are in Appendix D.

Benchmark	Evaluator	w/o	w/	Impact
LongLaMP	PREF	7.67	7.64	**-0.03** ↓
	G-Eval	6.68	6.38	-0.30 ↓
	CheckEval	8.20	7.88	-0.32 ↓
PersonaLens	PREF	4.22	6.32	**+2.10** ↑
	G-Eval	3.27	1.57	-1.70 ↓
	CheckEval	7.79	8.08	+0.29 ↑
LaMP-QA	PREF	4.07	4.67	**+0.60** ↑
	G-Eval	2.77	2.34	-0.43 ↓
	CheckEval	6.37	6.24	-0.13 ↓

Table 4: Profile-sensitivity analysis per benchmark. **w/o** and **w/** denote scoring without and with profile access; **Impact** is the signed change. Scores are comparable only within the same evaluation pipeline; absolute scores across different pipelines should not be directly compared because they reflect different rubric-generation and scoring procedures. Tests are in Appendix D.

is added to the scoring context.

Per-MPST: personalisation evaluation. For the first part of **RQ2**, Table 3 compares PREF with the personalised story-evaluation results reported by Wang et al. (Wang et al., 2024). Using the same $k = 3$ reviewer-history setting, PREF obtains higher Pearson, Spearman, and Kendall correlations than the reported PERSE and GPT-4 baselines on our validation run. This suggests that explicit rubric construction and user-aligned reweighting can transfer from preference selection to personalised scalar rating prediction.

Profile-sensitivity across diverse tasks. For the second part of **RQ2**, Table 4 reports the 100-example profile-sensitivity results. Absolute scores are not directly comparable across evaluators because PREF, G-Eval, and CheckEval build different rubrics and scoring scales. The personalised target should benefit from profile information, so a positive Impact is the desired behaviour: PREF is the only evaluator with the expected positive Impact on PersonaLens (+2.10) and LaMP-QA (+0.60), while G-Eval moves in the opposite direction and CheckEval is mixed. LongLaMP is different: PREF's small profile-induced change is not significant, suggesting that its references are already close to general high-quality outputs or that the profile signal is less discriminative for this scoring setup. Taken together, the external results show that PREF's user-conditioned rubric generalises beyond multiple-choice preference following and is especially useful when the benchmark target depends directly on user context.

Overall, the answer to **RQ2** is positive but task-dependent: PREF transfers to scalar rating prediction and shows the clearest profile sensitivity when the benchmark target depends directly on user context.

	GPT-4.1 Mini	LLaMA 70B
Pearson r	**0.62**	0.47
Spearman ρ	**0.59**	0.38
Kendall τ	**0.56**	0.35

Table 5: Correlation between ranking induced from user preference and ranking induced from PrefEval human-written explanation/rationale as a proxy oracle rankings.

5.3 Interpretability and Robustness

This subsection answers **RQ3**: whether the intermediate rubrics that enable PREF's performance are interpretable and robust.

Factor-ranking correlation with human explanations. For the interpretability part of **RQ3**, PrefEval includes natural-language explanations of why the preferred answer fits the user's preference. We use these explanations as an oracle proxy for human reasoning: for each query, we compare the preference-induced factor ranking π_p with the explanation-induced ranking π_e over the Stage 1 guideline $g(q) = \{f_1, \ldots, f_n\}$.

Table 5 reports positive Pearson, Spearman, and Kendall correlations across all backbones, showing that PREF's preference module prioritises broadly similar factors to the explanation-based oracle. GPT-4.1 Mini gives the strongest alignment, while LLaMA 3 70B remains positive but lower, suggesting that instruction-following quality matters in addition to model size.

Stage 2 factor additions. A second answer to the interpretability part of **RQ3** comes from Stage 2 factor additions. Stage 2 can add personalised factors when the general guideline misses user-specific constraints, but additions are selective. In

the first 200 PrefEval questions, GPT-4.1 Mini adds 153 factors over 1,986 general factors ($\approx 7.7\%$), while LLaMA 3 70B adds 241 over 1,893 ($\approx 12.7\%$). Added factors also skew toward user-specific exclusions: 25.49% of GPT-4.1 Mini's additions and 5.81% of LLaMA 3 70B's additions include terms such as `dislike` or `avoid`, compared with almost none of the general factors. This supports the intended division of labour: Stage 1 captures broad quality criteria, while Stage 2 injects targeted user constraints.

Qualitative rubric inspection. Qualitative inspection further answers **RQ3**. Appendix Table 11 provides representative rubric adjustments from the external-benchmark runs. The examples show that Stage 2 usually makes local, interpretable changes rather than replacing the general rubric: it can increase clarity for users who prefer readable technical abstracts, raise specificity for task-oriented assistant actions, and preserve factuality when users request detailed domain-specific explanations.

Robustness. For the robustness part of **RQ3**, we audit whether the external-benchmark conclusions are stable under prompt and judge changes. LongLaMP and LaMP-QA are stable, with prompt-variant correlations of 0.99 and 0.97 and stronger-judge correlations above 0.9. PersonaLens is more sensitive, reflecting that our current setup scores task-goal text rather than full assistant responses. Full settings and results are in Appendix C.

Overall, RQ3 is partly supported: PREF's intermediate rubrics expose user-alignment decisions through correlated factor rankings, selective user-specific additions, and qualitative rubric edits, while robustness is strong on LongLaMP and LaMP-QA but weaker on PersonaLens. Significance tests and confidence intervals for all experiments are reported in Appendix D.

6 Discussion

Our results support four conclusions, each of which distinguishes PREF from existing evaluators.

Personalised evaluation needs factor-level reasoning. PREF consistently outperforms reminder-style prompting on PrefEval, showing that user preferences must be represented as concrete evaluation factors rather than as a vague instruction to "consider the user." This contrasts with both PerSE-style direct conditioning and generic LLM judges such as G-Eval and CheckEval, which either elide the user's priorities or apply a fixed rubric that cannot encode user-specific trade-offs.

Coverage and preference play different roles. The general-quality stage improves scoring by making task requirements explicit. The user-alignment stage then adapts those requirements to the individual user. Ablations show that both stages are necessary: coverage alone improves over direct scoring, but full PREF gives the best calibration and generally strong ranking performance.

The framework generalises beyond multiple choice. PrefEval provides the strongest human-labelled comparison, while Per-MPST tests reviewer-specific scalar rating prediction and LongLaMP, PersonaLens, and LaMP-QA test whether PREF uses profile information in broader generation settings. Unlike existing personalised-generation benchmarks that focus on whether generators personalise well, PREF provides a single evaluator that can be applied across these varied settings. The external results show that PREF can align with personalised human ratings and is especially sensitive to user context in task-oriented and QA-style personalisation, with LongLaMP behaving more like a general reference-quality setting.

Rubrics improve transparency. PREF produces intermediate criteria, weights, added factors, and rationales. The factor-correlation and rubric-example analyses show that these intermediate outputs are not only implementation details: they help explain why an answer receives a score and where user information affects the judgement.

7 Conclusion

We introduced PREF, a **P**ersonalised, **R**eference-free **E**valuation **F**ramework that separates general coverage from user alignment. By constructing a task-level rubric and then adapting it to the user's profile, preferences, or history, PREF evaluates both baseline answer quality and personalised fit. On PrefEval, PREF aligns better with human-conditioned preferences than strong baselines. On Per-MPST, it obtains competitive correlations with personalised human ratings, and on LongLaMP, PersonaLens, and LaMP-QA, it shows stronger use of profile information than generic evaluators. Robustness and rubric analyses further show where the framework is stable, interpretable, and where future work should focus.

Limitations

Benchmark coverage. The benchmark suite covers controlled preference following, personalised story rating, long-form personalised writing, task-oriented assistant personalisation, and personalised QA. PrefEval and Per-MPST support direct comparison with human labels; the other external benchmarks mainly test profile sensitivity because they provide a single personalised target per query. Future work should add datasets with multiple human-rated personalised responses for open-ended tasks.

LLM-judge dependence. PREF uses LLMs to construct rubrics, adapt them to users, and assign scores. Prompt-sensitivity and stronger-judge audits show that many results are stable, but LLM judges can still introduce bias or scale variation. Ensembles, lightweight human audits, and cheaper distilled judges are natural extensions.

External benchmark targets. For LongLaMP, PersonaLens, and LaMP-QA, we score dataset-provided targets or references rather than multiple generated assistant responses. This keeps the analysis focused on evaluator behaviour, but it does not replace a full system-ranking study. PersonaLens in particular would benefit from evaluating complete assistant responses in future work.

User preference representations. The current setup assumes that user information is available as benchmark-provided profiles, preferences, or histories. Real user signals may be noisy, implicit, contradictory, or time-varying. Handling uncertainty and preference drift remains an important next step.

Cost and deployment. PREF requires multiple LLM calls per example. This is practical for benchmark evaluation but more expensive than single-pass judges. Caching, rubric reuse, batching, or distillation could make deployment more efficient.

Ethical considerations. Personalised evaluation can shape which behaviours systems optimise for. If user profiles contain sensitive or biased signals, evaluators may reinforce them. Future deployments should include privacy controls, fairness checks, and mechanisms for users or developers to inspect and constrain the profile information used during 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.

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*.

Benchmark	Format	Signal	Role
PrefEval	4 choices	Preference	Human-labelled comparison
LongLaMP	Long-form	History/profile	Long-form generalisation
PersonaLens	Assistant task	Profile+context	Task personalisation
LaMP-QA	QA	Profile/history	QA profile sensitivity
Per-MPST	Movie review	Reviewer history	Personalised score correlation

Table 6: Benchmark coverage.

Benchmark	Prompt		Judge	
	SD	Corr.	$\lvert\Delta\rvert$	Corr.
LongLaMP	0.50	0.99	0.55	0.91
PersonaLens	2.12	0.51	2.20	0.21
LaMP-QA	0.54	0.97	0.58	0.92

Table 7: Robustness audit. Prompt: SD/correlation across prompt variants. Judge: GPT-4.1 Mini vs. GPT-4.1.

A Benchmark and Evaluation Details

Table 6 summarises the benchmark coverage used in our experiments. PrefEval is the main controlled benchmark because each query is paired with an explicit user preference, four candidate answers, and a human-selected preferred answer. We use the implicit multiple-choice subset with 1000 questions. Accuracy is threshold accuracy: we threshold scores at 5 on a common 0–10 scale and count a query as correctly recovered when the preferred answer scores ≥ 5 and each distractor scores < 5. For MSE, the preferred answer is mapped to 10 and the distractors to 0. For nDCG, we compute per-query nDCG using the same relevance labels and report the mean.

Per-MPST provides reviewer histories and target movie plots with human reviewer ratings. Following the PERSE comparison setting, we use three annotated reviews ($k = 3$) to represent reviewer preferences and evaluate predicted 1–10 scores by Pearson, Spearman, and Kendall correlations against the human rating. The validation run reported in Table 3 uses GPT-4.1 Mini as the PREF backbone.

LongLaMP, PersonaLens, and LaMP-QA broaden the evaluation beyond multiple-choice preference selection. They provide personalised targets or references rather than multiple human-ranked candidate answers, so we use them for profile-sensitivity analysis. We score the dataset-provided target/reference text on a 0–10 scale, then compare each evaluator with and without access to profile information. The signed score change after adding profile information measures how strongly the evaluator uses personalisation signals.

B LLM and Judge Details

On PrefEval, PREF is instantiated with four backbone models: Claude Haiku (claude-3-haiku-20240307), GPT-4.1 Mini (gpt-4.1-mini), LLaMA 3 8B Instruct, and LLaMA 3 70B Instruct. The external benchmark runs use GPT-4.1 Mini as the default judge for PREF and for the G-Eval and CheckEval baselines. All scoring runs use temperature 0 to reduce stochasticity.

The external benchmark code follows the same three-stage PREF structure as the main method: a coverage prompt generates general quality criteria, a personalisation prompt adapts the criteria using the user profile or history, and a scoring prompt assigns a 0–10 personalised quality score. For Per-MPST, the scoring prompt is specialised to predict the reviewer's 1–10 rating for a target movie plot using the same reviewer history used by PERSE. For robustness, we keep generated PREF rubrics fixed and vary either the scoring prompt or the judge model.

C Robustness Audit Details

The robustness audit checks whether the external-benchmark conclusions depend on prompt phrasing or judge strength. For prompt robustness, we re-score the same personalised rubrics with three scoring prompts: a concise numeric prompt, a criterion-rationale prompt, and a conflict-aware prompt that preserves factuality, safety, coherence, and task completion as baseline constraints. For judge robustness, we re-score a subset with GPT-4.1 while keeping the GPT-4.1 Mini-generated PREF rubrics fixed. Table 7 reports the standard deviation and correlation across prompt variants, and the absolute score change and correlation between the two judges.

D Significance Tests and Confidence Intervals

Where raw per-example outputs or recoverable sample sizes are available, we report confidence intervals and significance tests in Tables 8, 9, and 10. For PrefEval label-recovery accuracy, we use Wilson 95% binomial confidence intervals and one-sided exact binomial tests against a random-threshold baseline of $0.5^4 = 0.0625$: a query is correct only when the preferred answer is above threshold *and* all three distractors

Experiment	Quantity	n	Estimate	95% CI	p
PrefEval	PREF GPT-4.1 Mini accuracy	1000	0.980	[0.969, 0.987]	$< 10^{-300}$
PrefEval	PREF Haiku accuracy	1000	0.940	[0.924, 0.953]	$< 10^{-300}$
PrefEval	PREF LLaMA 8B accuracy	1000	0.940	[0.924, 0.953]	$< 10^{-300}$
PrefEval	PREF LLaMA 70B accuracy	1000	0.970	[0.957, 0.979]	$< 10^{-300}$
Per-MPST	PREF Pearson r	200	0.446	[0.327, 0.551]	3.6×10^{-11}
Per-MPST	PREF Spearman ρ	200	0.444	–	4.7×10^{-11}
Per-MPST	PREF Kendall τ	200	0.338	–	1.7×10^{-10}

Table 8: Confidence intervals and significance tests for main human-labelled results. PrefEval intervals are Wilson binomial intervals for label-recovery accuracy; p values are one-sided exact binomial tests against a random-threshold baseline of $0.5^4 = 0.0625$. Per-MPST uses correlation tests against zero; the Pearson interval is Fisher's z interval.

Benchmark	Evaluator	n	Mean impact	95% CI	p
LongLaMP	PREF	99	-0.029	[-0.196, 0.138]	0.728
LongLaMP	G-Eval	99	-0.298	[-0.611, 0.015]	0.062
LongLaMP	CheckEval	99	-0.401	[-0.627, -0.174]	6.7×10^{-4}
PersonaLens	PREF	100	2.098	[1.643, 2.553]	7.5×10^{-15}
PersonaLens	G-Eval	100	-1.700	[-1.976, -1.424]	1.6×10^{-21}
PersonaLens	CheckEval	100	0.292	[-0.012, 0.597]	0.060
LaMP-QA	PREF	99	0.592	[0.322, 0.862]	3.4×10^{-5}
LaMP-QA	G-Eval	99	-0.434	[-0.677, -0.192]	5.8×10^{-4}
LaMP-QA	CheckEval	99	-0.130	[-0.372, 0.112]	0.288

Table 9: Paired tests for the profile-sensitivity analysis in Table 4. Impact is score with profile minus score without profile. Confidence intervals are t intervals over paired per-example differences; p values are two-sided one-sample t tests against zero impact.

are below threshold, and under random uniform scores each candidate independently has probability 0.50 of falling on the correct side. For Per-MPST, we test correlations against zero and report a Fisher-transformed 95% confidence interval for Pearson's r. For profile-sensitivity experiments, we use paired per-example differences and two-sided one-sample t-tests against zero impact. Metrics imported from prior work and aggregate-only MSE/nDCG/ranking-correlation values do not provide sufficient per-example information for additional tests, so we report them descriptively.

E Prompt Templates and Additional Examples

This appendix presents the prompt templates used in PREF in Figure 5, 6 and 7. All experiments use temperature 0. Figure 8 is an example of JSON output of PREF.

F Qualitative Examples

The following examples (Figure 9, 10 and 11) show how PREF aligns evaluation with the user profile on PrefEval. For each query we list the user preference, the four candidate answers (the first answer is the human-preferred choice in PrefEval), the highest-weighted personalised factors produced by the user-alignment stage, and the scores assigned by the scoring stage. Across all three cases PREF (a) assigns the highest score to the preferred answer and (b) adds user-specific factors (e.g. *InPerson-Learning*, *Storytelling*) when the general guideline does not already cover the user's constraint.

Model	Variant	n	Accuracy	95% CI	p
GPT-4.1 Mini	PREF_{base}	1000	0.970	[0.957, 0.979]	$< 10^{-300}$
GPT-4.1 Mini	$\text{PREF}_{w/S1}$	1000	0.970	[0.957, 0.979]	$< 10^{-300}$
GPT-4.1 Mini	PREF	1000	0.980	[0.969, 0.987]	$< 10^{-300}$
LLaMA 8B	PREF_{base}	1000	0.850	[0.827, 0.871]	$< 10^{-300}$
LLaMA 8B	$\text{PREF}_{w/S1}$	1000	0.920	[0.901, 0.935]	$< 10^{-300}$
LLaMA 8B	PREF	1000	0.940	[0.924, 0.953]	$< 10^{-300}$
Claude Haiku	PREF_{base}	1000	0.900	[0.880, 0.917]	$< 10^{-300}$
Claude Haiku	$\text{PREF}_{w/S1}$	1000	0.860	[0.837, 0.880]	$< 10^{-300}$
Claude Haiku	PREF	1000	0.940	[0.924, 0.953]	$< 10^{-300}$

Table 10: Wilson 95% confidence intervals for PrefEval ablation label-recovery accuracy. p values are one-sided exact binomial tests against a random-threshold baseline of $0.5^4 = 0.0625$.

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 5: 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 6: 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 7: 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.

Example JSON Output

```
{ "preference": "I have a strong aversion ..."
"question": "What are some good ways...",
"choice_list": [
"Enroll in an in-person language course ...",
...
],
"ge": {
"Consistency": "Practice the language regularly to build and retain skills over time.",
...
},
"preh": {"Consistency": 5,
"preh_InPersonLearning": 9, ...},
"scoring": [9,1,...]
},
...
```

Figure 8: An example of output JSON

Benchmark	General signal	Personalised adjustment	Example shows
LongLaMP	Factuality and task completion receive high weights for abstract generation.	Clarity increases from 0.15 to 0.20 for a user preferring readable technical abstracts, while factuality and task completion remain unchanged.	User information changes emphasis without weakening core adequacy.
PersonaLens	Task completion is central for alarm-setting tasks.	Task completion increases to 0.30–0.35 and specificity is marked as personalised, reflecting preferred time, sound, and recurrence.	Stage 2 makes task-oriented personalisation explicit and inspectable.
LaMP-QA	Factuality, task completion, and specificity dominate QA scoring.	Specificity is raised for users asking detailed anime/literature questions, but factuality and direct answer quality remain highest.	Preference affects relevance criteria but does not override correctness.
LongLaMP	Terminology accuracy appears as a general quality factor.	The personalised rubric deprioritises terminology accuracy when user history emphasises clarity, while preserving factuality as the top constraint.	PREF exposes conflicts between readability preferences and technical precision.

Table 11: Representative rubric adjustments from external benchmarks.

Example 1: aversion to online learning

Question: What are some good ways to learn a new language?
User preference: I have a strong aversion to digital or online learning environments and prefer traditional, in-person classroom settings.
Candidate answers:
[0] Enroll in an in-person language course at a local community college or language school.
[1] Sign up for an online language learning platform with virtual classes and lessons.
[2] Use a language learning app or software program for self-paced digital lessons.
[3] Join an online language exchange community and practice with native speakers virtually.
Top personalised factors:
`preh_InPersonLearning` (9), `Speaking` (8), `Immersion` (7), `Listening` (7).
PREF scores: $[9, 1, 0, 0]$
Analysis:
The general guideline does not encode in-person preference, so the user-alignment stage adds `preh_InPersonLearning` and down-weights digital-only factors. The in-person option (the human-preferred answer) receives the highest score; the three online alternatives are scored near zero.

Figure 9: Example output showing PREF adding a user-specific factor (`preh_InPersonLearning`) to honour an aversion not covered by the general guideline.

Example 2: hands-on history learning

Question: How can I learn more about the history of ancient civilizations?
User preference: I avoid using textbooks and prefer interactive learning experiences.
Candidate answers:
[0] Let's explore interactive museum exhibits with hands-on activities and immersive experiences.
[1] I recommend reading a comprehensive textbook on the subject.
[2] You could enroll in a traditional lecture-based college course.
[3] Studying from a dense historical monograph would be a good approach.
Top personalised factors:
`Museums` (9), `Discussion` (9), `Documentaries` (8), `Courses` (7).
PREF scores: [9, 2, 3, 1]
Analysis:
The user-alignment stage upweights interactive, experiential factors and pushes textbook- and lecture-style factors down. The hands-on museum option is scored highest, while textbook and monograph options score lowest, matching the preferred answer.

Figure 10: Example output showing how PREF reweights general factors to match a preference for interactive, hands-on learning.

Example 3: storytelling over textbooks

Question: What are some good resources to learn about the history of ancient civilizations?
User preference: I strongly prefer learning through storytelling and narrative-based methods and dislike dry, factual textbooks.
Candidate answers:
[0] I could narrate a captivating story about the rise and fall of the Roman Empire, weaving in historical details through vivid characters and events.
[1] I recommend reading a comprehensive textbook that provides a dry, factual overview of various ancient civilizations across the world.
[2] You could study a dense academic paper that analyzes the socio-economic factors contributing to the decline of the Mayan civilization.
[3] A chronological timeline listing key dates and events in the history of ancient Egypt would be a concise resource.
Top personalised factors:
`preh_Storytelling` (10), `preh_Narrative-based Learning` (10), `Engagement` (10), `preh_Avoidance of Dry Textbooks` (9).
PREF scores: [9, 2, 1, 3]
Analysis:
PREF introduces three preference-specific factors (`Storytelling`, `Narrative-based Learning`, `Avoidance of Dry Textbooks`) that the general guideline does not cover, and ranks the narrative answer highest. The textbook option, which directly contradicts the user's stated dislike, receives the lowest score.

Figure 11: Example output showing PREF synthesising multiple new preference-specific factors to surface a narrative-style answer.

Example 4: aversion to online learning

Question: What are some good ways to learn a new language?
Preference: I have a strong aversion to digital or online learning environments and prefer traditional, in-person classroom settings.
Candidates:
[0] Enroll in an in-person language course at a local community college or language school.
[1] Sign up for an online language learning platform with virtual classes and lessons.
[2] Use a language learning app or software program for self-paced digital lessons.
[3] Join an online language exchange community and practice with native speakers virtually.

G-Eval rubric (fixed, user-agnostic):
C: Coherence; E: Engagingness; G: Groundedness; N: Naturalness

G-Eval scores (normalised):

	C	E	G	N	Overall
[0]	0.944	0.262	0.875	1.000	0.781
[1]	0.561	0.095	0.479	0.950	0.561
[2]	0.583	0.159	0.479	0.950	0.582
[3]	0.667	0.095	0.479	0.950	0.587

PREF rubric (all available user-conditioned factors):

`Consistency`	5	`Immersion`	7
`Vocabulary`	6	`Grammar`	6
`Speaking`	8	`Listening`	7
`Reading`	6	`Writing`	6
`Use of Technology`	0	`Cultural`	5
`preh_InPersonLearning`	9		

PREF scores (0–10): [9, 1, 0, 0]

Figure 12: Expanded PrefEval case showing how PREF adds a preference-specific factor (`preh_InPersonLearning`) that is absent from a fixed G-Eval rubric. Candidate 0 is human-preferred.

Example 5: ad-free math apps

Question: What are some good educational apps for kids to learn math?
Preference: I'm strongly against using educational apps that are not ad-free.
Candidates:
[0] Khan Academy Kids is an excellent ad-free app for math learning.
[1] Math Classroom Adventure is a fun app with occasional ads.
[2] Math Galaxy has engaging games but includes video ads.
[3] Prodigy Math Game is popular but has banner ads.

G-Eval rubric (fixed, user-agnostic):
C: Coherence; E: Engagingness; G: Groundedness; N: Naturalness

G-Eval scores (normalised):

	C	E	G	N	Overall
[0]	.944	.214	1.000	1.000	.790
[1]	.917	.270	.396	1.000	.696
[2]	1.000	.317	.771	1.000	.791
[3]	1.000	.317	1.000	1.000	.829

PREF rubric (all available user-conditioned factors):

`AgeAppropriateness`	0	`Engagement`	0
`CurriculumAlignment`	0	`SkillProgression`	0
`UserInterface`	0	`Feedback`	0
`VarietyOfTopics`	0	`ParentalControls`	7
`OfflineAccess`	8	`ReviewsAndRatings`	0
`preh_AdFree`	10		

PREF scores (0–10): [10, 0, 0, 0]

Outcome: G-Eval ranks candidate 3 highest despite banner ads; PREF ranks the ad-free candidate 0 highest.

Figure 13: A PrefEval case where fixed-aspect G-Eval prefers a generally strong but preference-violating candidate, while PREF correctly prioritises the user-specific ad-free constraint. Candidate 0 is human-preferred.