PREFERENCE-AWARE RUBRIC LEARNING FOR PERSONALIZED EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As Large Language Models (LLMs) evolve from general-purpose assistants to user-centric agents, personalization has become central to aligning model behavior with individual preferences, making the evaluation of personalized alignment a critical bottleneck. Existing evaluation methods—ranging from automatic metrics to LLM-as-a-judge approaches—fail to capture subjective, user-specific preferences embedded in long-term interaction histories. We identify three essential principles for reliable and effective personalized evaluation: *Representativeness*, *User-Consistency*, and *Discriminativeness*. To address these principles, we introduce *Personalized Evaluation as Learning*, a paradigm that formulates personalized evaluation as a learning problem rather than a static judgment. Under this paradigm, we propose **PARL** (**P**reference-**A**ware **R**ubric **L**earning), a framework that learns to induce preference-aware evaluation rubrics directly from raw user histories and performs a self-validation mechanism to ensure consistency with the user's preferences. PARL integrates rubric induction with a discriminative reinforcement learning objective that contrasts user-authored responses against competitive personalized model outputs, enabling the learned rubrics to capture precise, user-specific decision boundaries. Experiments on real-world personalized text generation tasks show that PARL consistently induces high-fidelity rubrics that reliably identify user-aligned responses and generalize across users and tasks, while capturing stable stylistic preferences and fine-grained evaluative patterns. To ensure reproducibility, our code is available at `https://anonymous.4open.science/r/PARL-1909`.

1 INTRODUCTION

Large Language Models (LLMs) (Achiam et al., 2023; Guo et al., 2025a; Zhang et al., 2025c) have significantly advanced general-purpose intelligence and are now widely applied to reasoning, content generation, and decision-making tasks. However, their prevailing *"one-size-fits-all"* paradigm limits their ability to accommodate diverse and user-specific preferences. This limitation has motivated *LLM Personalization* (Chen et al., 2024; Zhang et al., 2024b), which aims to adapt model behavior to individual users across personalized tasks such as writing assistance and conversational interaction. Currently, LLM personalization has become an important and rapidly growing research direction, attracting increasing attention from both academia (Zhao et al., 2025c; Ryan et al., 2025; Wang et al., 2025) and industry (Jiang et al., 2025; Zhao et al., 2025a; Chen et al., 2026).

Existing research on LLM personalization has primarily focused on aligning model behavior with individual user preferences through improvements to the *generation process*, leading to a diverse range of approaches such as user-specific context injection (Qiu et al., 2025b; Salemi et al., 2024; Zhuang et al., 2024) and fine-tuning on historical user contexts (Tan et al., 2024; Liu et al., 2025a; Zhao et al., 2025c). In contrast, the *evaluation* of LLM personalization remains largely underexplored, despite being a critical bottleneck for the reliable design, comparison, and verification of personalization methods. This gap stems from a fundamental mismatch in evaluation assumptions: conventional evaluation methods rely on universal and objective ground truths, whereas personalized generation is inherently governed by subjective, user-specific criteria that are usually unobservable and only implicitly reflected in users' historical behaviors.

Existing evaluation methods for LLM personalization can be broadly grouped into three categories: (1) *Automatic metrics*, like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020). These metrics

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

generally assume a single objective reference and prioritize surface-level lexical overlap with the reference. As a result, they are inherently inadequate for personalized evaluation, failing to capture subjective, user-specific preferences and alternative but equally valid personalized responses. (2) *Human evaluation*, which is often regarded as the gold standard in general evaluation scenarios. Yet, it suffers from an inferential gap in personalization settings: external annotators lack direct access to a user's underlying preferences and decision criteria, making it difficult to faithfully assess alignment within the user's reference. (3) *LLM-as-a-judge* approaches (Liu et al., 2023; Fu et al., 2024), which partially alleviate these issues via LLMs' semantic capabilities. However, they rely on static judges and handcrafted prompts, restricting adaptation to individual users and task-specific preferences, and they typically yield coarse, holistic scores without fine-grained grounding, limiting the interpretability and verifiability of their judgments.

Given the limitations of existing approaches, we propose a shift toward a *Personalized Evaluation as Learning* paradigm for LLM personalization. Rather than relying on static and generic evaluation rules, this paradigm aims to adaptively and systematically learn user-specific evaluation criteria aligned with individual preferences. To realize this paradigm, we build upon the LLM-as-a-judge framework but depart from existing practices by instantiating it through *rubric-based evaluation* (Kim et al., 2024; Hashemi et al., 2024; Pathak et al., 2025), which decomposes holistic judgments into explicit, multi-dimensional criteria, referred to as rubrics. Such a formulation provides a natural foundation for learning-based personalized evaluation: rubrics can be adaptively induced from individual user historical contexts and applied across diverse scenarios. The core challenge, therefore, is to learn how to drive rubrics that faithfully capture individual user-specific decision boundaries.

To answer this, we identify and formalize three essential principles that any rigorous and reliable rubric-based framework for personalized evaluation must satisfy: (i) ***Representativeness***: Rubrics should faithfully distill the diverse and latent preferences embedded in a user's historical contexts, capturing the key dimensions that define their evaluative standards. (ii) ***User-Consistency***: Rubrics should reflect stable, user-specific preferences that persist across different historical contexts, rather than artifacts arising from isolated tasks or individual samples. (iii) ***Discriminativeness***: Rubrics should enable fine-grained distinction between authentic user-authored responses and generic yet high-quality model outputs, isolating true preference awareness from universal quality signals.

Guided by the three principles, we propose **PARL** (**P**reference-**A**ware **R**ubric **L**earning), a framework that learns to induce structured, multi-dimensional, preference-aware rubrics from raw user historical behaviors to support personalized LLM evaluation. PARL distills semantically rich, user-specific preferences from users' historical contexts into explicit rubric dimensions, ensuring *representativeness*. The induced rubrics are then validated across diverse historical contexts to enforce *user-consistency*. To further enhance *discriminativeness*, PARL incorporates a reinforcement learning stage that contrasts user-authored ground-truth responses with competitive personalized model outputs, encouraging learned rubrics to capture precise, user-specific decision boundaries.

We conduct comprehensive experiments on real-world personalized text generation tasks to systematically assess the learned rubrics. Empirical results demonstrate that PARL consistently induces rubrics that are representative, user-consistent, and discriminative. Further analysis reveals that the learned rubrics capture stable stylistic preferences and fine-grained criteria, reflecting distinct evaluative signatures. Moreover, the trained rubric generator generalizes well across contexts, inducing rubrics that are semantically rich, diverse, and tailored to individual users. Remarkably, these rubrics constitute a novel, reliable, and high-fidelity set of personalized evaluation metrics.

Our main contributions are summarized as follows:

- We introduce the *Personalized Evaluation as Learning* paradigm, which formalizes personalized LLM evaluation as a learnable process grounded in three essential criteria: *Representativeness*, *User-Consistency*, and *Discriminativeness*.
- We propose **PARL** (**P**reference-**A**ware **R**ubric **L**earning), a learning-based framework that induces preference-aware evaluation rubrics from user histories, optimized with a discriminative margin objective via RL.
- We conduct extensive experiments and analysis across real-world personalized generation tasks, demonstrating that PARL produces high-fidelity, reusable rubrics and offering new insights into the structure of user-specific evaluative preferences.

2 Related Work

LLM Personalization. The widespread use of LLMs has shifted alignment from *"one-size-fits-all"* objectives toward individual preferences, along two tracks: personalized generation (Xu et al., 2025b) and preference alignment (Xie et al., 2025; Zhang et al., 2024a). The former spans text, image, and audio (Li et al., 2024; Xu et al., 2025a; Guo et al., 2025b), with text as a foundational direction (Salemi et al., 2024; Zhao et al., 2025b; Chen et al., 2024): LaMP (Salemi et al., 2024) and LongLaMP (Kumar et al., 2024) benchmark short- and long-form settings, while HYDRA (Zhuang et al., 2024) and DEP (Qiu et al., 2025a) exploit user history via personalized reranking and soft-prompt distillation. The latter models user-specific reward functions (Jang et al., 2023; Chen et al., 2025b; Shen et al., 2025), *e.g.,* decomposing rewards into shared and user-specific components (Chen et al., 2025a), treating user context as latent variables (Poddar et al., 2024), restricting rewards to low-dimensional subspaces (Bose et al., 2025; Shenfeld et al., 2025), or employing multi-stage RL (Zhang et al., 2026). We instead shift from *learning* personalized rewards to *evaluating* personalized generation, inducing explicit history-grounded rubrics rather than implicit scalar rewards.

Evaluation of Personalized Text Generation. Automatic metrics based on n-gram overlap (Lin, 2004; Papineni et al., 2002; Banerjee & Lavie, 2005) or embeddings (Zhang et al., 2020; Yuan et al., 2021) presume a single objective reference, missing subjective preferences and multiple valid responses; human evaluation faces an inferential gap, as annotators lack access to the target user's latent preferences. LLM-based judges (Zheng et al., 2023; Kim et al., 2024; Gu et al., 2024; Zhang et al., 2025b) rely on static pre-trained models and handcrafted prompts, yielding coarse scores that are neither user-adaptive nor verifiable. Recent rubric-based evaluation adopts explicit, multi-dimensional criteria (Hashemi et al., 2024; Biyani et al., 2024; Liu et al., 2025b; Hong et al., 2026), and some studies apply structured judges to personalized tasks (Qiu et al., 2025b; Salemi et al., 2025a;b; Salemi & Zamani, 2025; Wang et al., 2023; Fu et al., 2025), yet still depend on static or heuristic rubrics without principled grounding in user histories—motivating our learning-based personalized evaluation framework.

3 Preliminary

In this section, we formally define the problem of personalized evaluation and introduce the rubric-based evaluation paradigm that serves as the foundation for our proposed framework.

3.1 Problem Formulation

Let $\mathcal{U}$ be the set of users. For each user $u \in \mathcal{U}$, we define their historical context as $\mathcal{H}_u = \{(x_i, y_i)\}_{i=1}^{n}$. Here, x_i denotes the task context (*e.g.,* a specific instruction or prompt) and y_i represents the ground-truth response authored by the user. This history $\mathcal{H}_u$ serves as the primary evidence for inducing u's implicit preferences.

In this work, we focus on the reference-free personalized evaluation setting, which mirrors real-world scenarios where a user's ground-truth response for a given task context may be unavailable. Given a new task input x and a model-generated response $\hat{y} = \mathcal{G}(x, \mathcal{H}_u)$, where $\mathcal{G}$ is a personalized generator (*e.g.,* an LLM), the objective is to assess how $\hat{y}$ aligns with the user's intrinsic preferences. Unlike general evaluation, this task measures whether $\hat{y}$ aligns with the specific user preferences as established in their history $\mathcal{H}_u$.

Consequently, we define a scoring function Φ to compute a personalized alignment score S:

$$S = \Phi(\hat{y}; x, \mathcal{H}_u). \tag{1}$$

The score S is normalized to $[0, 1]$, with higher values indicating stronger alignment with the user's preferences.

3.2 Rubric-based Evaluation

To enable a verifiable and granular assessment of personalized alignment, we introduce a rubric-based evaluation strategy that decomposes holistic user preferences into explicit, interpretable dimensions. Specifically, the assessment of $\hat{y}$ is mediated by a personalized rubric set $\mathcal{R}_u = \{r_1, r_2, \ldots, r_m\}$

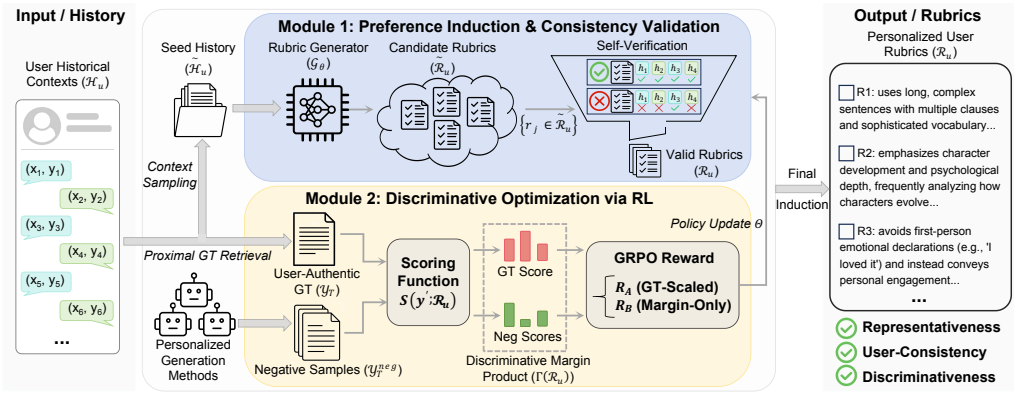

Figure 1: Overview of our proposed PARL framework for inducing personalized user rubrics for evaluating LLM personalization. PARL consists of two core modules: (1) Preference Induction & Consistency Validation; (2) Discriminative Optimization via RL. The final induced rubrics serve as high-fidelity evaluation metrics for personalized text generation.

for user u, where each rubric r_j represents a distinct analytical dimension of the user's preferences, distilled from $\mathcal{H}_u$. The final personalized alignment score S is computed as the normalized average of these binary judgments across all m dimensions for user u:

$$S = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}(\hat{y} \vDash r_j), \tag{2}$$

where $\vDash$ is a binary satisfaction operator. The indicator function $\mathbb{I}(\cdot)$ returns 1 when the response $\hat{y}$ satisfies rubric r_j, and 0 otherwise. By aggregating these binary outcomes, the resulting score $S \in [0, 1]$ provides a quantitative measure of alignment depth, where the binary nature of each $\mathbb{I}(\cdot)$ ensures that every dimension of the user's evaluative rubric is explicitly and transparently verified.

4 PARL (PREFERENCE-AWARE RUBRIC LEARNING)

In this section, we first provide an overview of our proposed **PARL** (**P**reference-**A**ware **R**ubric **L**earning) framework, followed by detailed introductions of its primary components.

4.1 OVERVIEW

The PARL framework redefines personalized evaluation by shifting from an opaque, monolithic scoring paradigm to a structured, learnable, and verifiable process. The primary objective of the framework is to transform unorganized, raw historical contexts $\mathcal{H}_u$ into user-specific rubrics $\mathcal{R}_u = \{r_1, r_2, \ldots, r_m\}$ for personalized evaluation that rigorously satisfy three fundamental principles: *Representativeness*, *User-Consistency*, and *Discriminativeness*.

As illustrated in Figure 1, we implement a hierarchical pipeline consisting of two core modules:

- **Preference Induction & Consistency Validation.** This module focuses on the transition from raw historical contexts to a compact set of preference-aware rubrics for each user. The framework first performs a multi-dimensional analysis of a user's diverse historical contexts, capturing a comprehensive spectrum of multifaceted preferences that define their unique evaluative identity. The resulting candidate rubrics are subsequently validated through the user history, and only those that demonstrate persistent satisfaction across diverse historical contexts are retained.
- **Discriminative Optimization via RL.** To better disentangle user-specific stylistic signatures from historical noise, we incorporate a specialized learning stage that transforms rubric induction into a learnable capability via reinforcement learning (RL). By leveraging outputs from a representative, competitive set of personalized methods as negative samples, the framework optimizes a margin-based objective to maximize the scoring gap between authentic user-authored responses and competitive alternatives, thereby sharpening the decision boundaries of the induced rubrics.

Ultimately, this approach enables learning rubrics that precisely delineate user-aligned decision boundaries. The resulting rubrics thus emerge as high-resolution, idiosyncratic evaluation metrics capable of capturing the subtle nuances of individual preferences.

4.2 Preference Induction & Consistency Validation

To obtain highly representative rubrics, the framework must capture the multi-dimensional nature of user preferences embedded in $\mathcal{H}_u$. We define a rubric generator $\mathcal{G}$ that induces a candidate set of rubrics $\tilde{\mathcal{R}}_u = \{r_1, r_2, \ldots, r_k\}$ from a subset of the user history. Unlike holistic summaries, each induced rubric r_j is constrained to be atomic and single-dimensional, capturing a specific stylistic, structural, or semantic invariant. Formally, we denote this process as:

$$\tilde{\mathcal{R}}_u = \mathcal{G}(\tilde{\mathcal{H}}_u), \tag{3}$$

where $\tilde{\mathcal{H}}_u \subset \mathcal{H}_u$ represents a diverse selection of user history as the seed history. This decoupled structure facilitates modular validation, enabling the precise identification and filtering of user-inconsistent candidate rubrics.

To explicitly ensure that the generated rubrics reflect enduring user traits rather than transient or task-specific artifacts, we subject the candidate set $\tilde{\mathcal{R}}_u$ to a self-verification process. This mechanism grounds rubrics in empirical evidence, reducing evaluative hallucinations. Since a genuine user preference should remain stable across diverse histories, we evaluate each r_j against every sample within the seed history $\tilde{\mathcal{H}}_u$. The final distilled rubric set $\mathcal{R}_u$ is defined as the subset of $\tilde{\mathcal{R}}_u$ that demonstrates universal satisfaction across the seed history:

$$\mathcal{R}_u = \left\{ r_j \in \tilde{\mathcal{R}}_u \mid \frac{1}{|\tilde{\mathcal{H}}_u|} \sum_{(x_i, y_i) \in \tilde{\mathcal{H}}_u} \mathbb{I}(y_i \vDash r_j) \geq \tau \right\}, \tag{4}$$

where τ serves as the consistency threshold. By setting a high τ (*e.g.*, 1.0), we enforce a strict requirement for temporal stability, ensuring that r_j is not a reaction to a specific task but a persistent evaluative standard for user u. This pruning process effectively filters out spurious patterns, yielding a stable and reliable evaluation benchmark that is consistent across diverse scenarios.

4.3 Discriminative Optimization via RL

While the distilled rubrics $\mathcal{R}_u$ capture the invariant core of user preferences, they still lack the resolution to disentangle idiosyncratic stylistic signatures from historical noise. The goal of this stage is to instill a discriminative capacity within the induced rubrics, enabling them to effectively differentiate between the user's authentic behavior and sophisticated AI imitations. We formulate this as an RL problem where the generator $\mathcal{G}_\theta$ is optimized to maximize the scoring margin between these two distributions. To this end, we introduce two variants of our framework, PARL-A and PARL-B, which differ in their reward formulation.

To ground the discriminative refinement, we pair the authentic response y_T from the most proximal interaction $(x_T, y_T) \in \mathcal{H}_u \setminus \tilde{\mathcal{H}}_u$ with a set of negative samples $\mathcal{Y}_T^{neg} = \{y_{T,1}^{neg}, \ldots, y_{T,m}^{neg}\}$ generated by representative methods. Unlike the binary satisfaction logic used for consistency filtering in Section 4.2, we define a scoring function $S(y'; \mathcal{R}_u)$ that provides a soft probabilistic measure of how well a response y' aligns with the generated rubrics. The total score is computed as the mean alignment across the valid rubric set: $S(y'; \mathcal{R}_u) = \frac{1}{|\mathcal{R}_u|} \sum_{r \in \mathcal{R}_u} S(y'; r)$. Utilizing this scoring paradigm, we further define the *Discriminative Margin Product* $\Gamma(\mathcal{R}_u)$ as the joint probability that the GT response is superior over all competitive negatives:

$$\Gamma(\mathcal{R}_u) = \prod_{y' \in \mathcal{Y}_T^{neg}} \sigma\left(\kappa \cdot (S(y_T; \mathcal{R}_u) - S(y'; \mathcal{R}_u))\right), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function and κ is a scaling factor that sharpens the discriminative boundary.

Reward Formulation: PARL-A vs. PARL-B.

We investigate two variants of reward computation, differing in whether the reward incorporates the absolute satisfaction of user-authored ground-truth response:

- PARL-A (GT-Scaled): $R_A = \lambda \cdot \mathcal{C} \cdot S(y_T; \mathcal{R}_u) \cdot \Gamma(\mathcal{R}_u)$. This variant integrates the relative margin with absolute alignment quality, favoring rubrics that are both strongly satisfied by the ground-truth and highly discriminative.
- PARL-B (Margin-Only): $R_B = \lambda \cdot \mathcal{C} \cdot \Gamma(\mathcal{R}_u)$. This variant isolates the discriminative signal by relying solely on the relative margin.

In both variants, $\mathcal{C}$ serves as a cardinality factor that normalizes the reward according to the number of valid rubrics, and λ is a global scaling constant to control reward magnitude.

Overall, the choice between PARL-A and PARL-B represents a trade-off: PARL-A balances discriminativeness with absolute preference fidelity, while PARL-B prioritizes pure contrastive sensitivity. In practice, both variants offer distinct utility, with PARL-A providing a robust, grounded benchmark for stable alignment and PARL-B enabling the high-resolution recovery of the most subtle, idiosyncratic signatures that define a user's unique evaluative identity.

Policy Optimization.

To optimize the rubric generator $\mathcal{G}_\theta$, we employ Group Relative Policy Optimization (GRPO), which allows the model to explore the complex space of evaluative criteria without the overhead of a separate value function. For each $\tilde{\mathcal{H}}_u$, we sample K rubric sets and maximize the following objective:

$$\mathcal{J}(\theta) = \mathbb{E}\Bigg[\frac{1}{K}\sum_{i=1}^{K}\min\Bigg(\frac{\mathcal{G}_\theta(\mathcal{R}_u^i|\tilde{\mathcal{H}}_u)}{\mathcal{G}_{\theta_{\text{old}}}(\mathcal{R}_u^i|\tilde{\mathcal{H}}_u)}A_i, \\ \text{clip}\left(\frac{\mathcal{G}_\theta(\mathcal{R}_u^i|\tilde{\mathcal{H}}_u)}{\mathcal{G}_{\theta_{\text{old}}}(\mathcal{R}_u^i|\tilde{\mathcal{H}}_u)}, 1-\epsilon, 1+\epsilon\right)A_i\Bigg)\Bigg] - \beta\mathbb{D}_{\text{KL}}(\mathcal{G}_\theta||\mathcal{G}_{\text{ref}}), \tag{6}$$

where ϵ and β are hyperparameters governing the clipping range and the strength of the KL regularization, respectively. The clip$(\cdot)$ operator enforces a trust-region to ensure stable policy updates, while the KL term $\mathbb{D}_{\text{KL}}$ regularizes the generator toward the reference policy $\mathcal{G}_{\text{ref}}$. The relative advantage A_k is computed by normalizing the reward $R(\mathcal{R}_u^k)$ within the sampled group: $A_i = (R_i - \bar{R})/\sigma_R$, where $R(\cdot) \in \{R_A, R_B\}$ is evaluated against the most proximal interaction (x_T, y_T). Here $\bar{R}$ and σ_R denote the mean and standard deviation of rewards within the sampled group, respectively.

4.4 Rubric Induction

During inference, the optimized rubric generator $\mathcal{G}_\theta$ induces preference-aware personalized rubrics for user u, conditioned on the user seed history $\tilde{\mathcal{H}}_u$. To further ensure these rubrics reflect enduring user preferences, we formalize the induction as a joint generation-verification process. The final rubric set $\mathcal{R}_u$ consists of rubrics that satisfy an empirical consistency threshold τ across the history:

$$\mathcal{R}_u = \left\{ r_j \in \mathcal{G}_\theta(\tilde{\mathcal{H}}_u) \mid \frac{1}{|\tilde{\mathcal{H}}_u|}\sum_{(x_i,y_i)\in\tilde{\mathcal{H}}_u}\mathbb{I}(y_i \vDash r_j) \geq \tau \right\}. \tag{7}$$

The resulting rubrics $\mathcal{R}_u$ satisfy the core requirements of *representativeness*, *user-consistency*, and *discriminativeness*. This yields a reliable, high-fidelity evaluation framework that effectively distinguishes authentic personalized resonance from AI outputs in downstream generation tasks.

5 Experiment

5.1 Experimental Setup

Datasets. To benchmark the rubrics induced by PARL across diverse personalized scenarios, we conduct experiments on three representative text generation tasks: *Amazon Review Generation* (Hou et al., 2024), *Reddit Topic Writing* (Völske et al., 2017), and *News Headline Generation* (Misra, 2022). These tasks were selected to provide a comprehensive testbed for assessing how different model-generated baselines align with rubrics induced by PARL, covering both short-form and long-form personalized text generation. Additional details and statistical analyses of datasets are provided in Appendix C.

Table 1: Evaluation results of induced rubrics across three personalized text generation tasks. Within each rubric evaluation group, the highest result is bolded, and the second-best is underlined. The symbol * indicates p-value < 0.05 in t-tests. The *user coverage* is reported as a percentage.

		Amazon Review Generation					*Reddit Topic Writing*					*News Headline Generation*				
		LM-8B	LM-235B	PARL-0	**PARL-A**	**PARL-B**	LM-8B	LM-235B	PARL-0	**PARL-A**	**PARL-B**	LM-8B	LM-235B	PARL-0	**PARL-A**	**PARL-B**
User-level Accuracy	**GT**	**0.931**$_{\pm 0.001}$	**0.794***$_{\pm 0.003}$	**1.000**$_{\pm 0.000}$	**0.931***$_{\pm 0.002}$	**0.930***$_{\pm 0.001}$	0.892$_{\pm 0.000}$	**0.840**$_{\pm 0.001}$	0.954$_{\pm 0.001}$	**0.848***$_{\pm 0.001}$	**0.860***$_{\pm 0.002}$	0.804$_{\pm 0.003}$	0.765$_{\pm 0.002}$	0.998$_{\pm 0.000}$	**0.832***$_{\pm 0.005}$	**0.804***$_{\pm 0.002}$
	Non	0.548$_{\pm 0.002}$	0.256$_{\pm 0.002}$	**1.000**$_{\pm 0.000}$	0.272$_{\pm 0.002}$	0.567$_{\pm 0.001}$	0.422$_{\pm 0.000}$	0.371$_{\pm 0.001}$	0.972$_{\pm 0.000}$	0.219$_{\pm 0.000}$	0.101$_{\pm 0.001}$	0.656$_{\pm 0.001}$	0.636$_{\pm 0.002}$	0.998$_{\pm 0.000}$	0.331$_{\pm 0.004}$	0.414$_{\pm 0.004}$
	RAG	0.624$_{\pm 0.002}$	0.344$_{\pm 0.002}$	0.998$_{\pm 0.000}$	0.325$_{\pm 0.002}$	0.616$_{\pm 0.001}$	0.615$_{\pm 0.000}$	0.531$_{\pm 0.001}$	**0.985**$_{\pm 0.000}$	0.267$_{\pm 0.001}$	0.234$_{\pm 0.001}$	0.740$_{\pm 0.000}$	0.714$_{\pm 0.003}$	0.998$_{\pm 0.000}$	0.468$_{\pm 0.004}$	0.569$_{\pm 0.005}$
	Non-Think	0.601$_{\pm 0.001}$	0.280$_{\pm 0.002}$	**1.000**$_{\pm 0.000}$	0.225$_{\pm 0.002}$	0.579$_{\pm 0.001}$	0.506$_{\pm 0.000}$	0.408$_{\pm 0.001}$	0.505$_{\pm 0.001}$	0.196$_{\pm 0.002}$	0.184$_{\pm 0.001}$	0.681$_{\pm 0.002}$	0.636$_{\pm 0.002}$	0.996$_{\pm 0.000}$	0.253$_{\pm 0.002}$	0.408$_{\pm 0.006}$
	RAG-Think	0.740$_{\pm 0.003}$	0.420$_{\pm 0.001}$	**1.000**$_{\pm 0.000}$	0.330$_{\pm 0.001}$	0.669$_{\pm 0.002}$	0.707$_{\pm 0.000}$	0.617$_{\pm 0.001}$	0.818$_{\pm 0.001}$	0.296$_{\pm 0.002}$	0.300$_{\pm 0.001}$	0.705$_{\pm 0.003}$	0.718$_{\pm 0.002}$	0.998$_{\pm 0.000}$	0.415$_{\pm 0.003}$	0.561$_{\pm 0.005}$
	SFT	0.827$_{\pm 0.002}$	0.642$_{\pm 0.003}$	**1.000**$_{\pm 0.000}$	0.877$_{\pm 0.002}$	0.880$_{\pm 0.001}$	0.803$_{\pm 0.001}$	0.764$_{\pm 0.001}$	0.971$_{\pm 0.001}$	0.760$_{\pm 0.004}$	0.790$_{\pm 0.002}$	**0.816**$_{\pm 0.001}$	**0.787**$_{\pm 0.002}$	**0.999**$_{\pm 0.000}$	0.814$_{\pm 0.004}$	0.750$_{\pm 0.004}$
	GRPO	0.930$_{\pm 0.001}$	0.787$_{\pm 0.001}$	**1.000**$_{\pm 0.000}$	0.832$_{\pm 0.002}$	0.918$_{\pm 0.001}$	0.891$_{\pm 0.001}$	0.798$_{\pm 0.003}$	0.982$_{\pm 0.000}$	0.268$_{\pm 0.001}$	0.484$_{\pm 0.001}$	0.783$_{\pm 0.002}$	0.745$_{\pm 0.003}$	0.998$_{\pm 0.000}$	0.496$_{\pm 0.004}$	0.591$_{\pm 0.008}$
	SFT+GRPO	0.917$_{\pm 0.003}$	0.759$_{\pm 0.001}$	**1.000**$_{\pm 0.000}$	0.905$_{\pm 0.003}$	0.927$_{\pm 0.001}$	0.899$_{\pm 0.001}$	0.836$_{\pm 0.001}$	0.971$_{\pm 0.000}$	0.770$_{\pm 0.002}$	0.824$_{\pm 0.000}$	0.773$_{\pm 0.001}$	0.774$_{\pm 0.002}$	**0.999**$_{\pm 0.000}$	0.775$_{\pm 0.004}$	0.742$_{\pm 0.010}$
	Max-Diff	+0.001	+0.007	−0.000	+0.026	+0.003	−0.007	+0.004	−0.031	+0.078	+0.036	−0.012	−0.022	−0.001	+0.018	+0.054
User Coverage		49.1%	84.6%	100.0%	99.3%	99.5%	75.0%	81.5%	100.0%	99.9%	99.4%	66.9%	87.1%	100.0%	99.0%	100.0%

Baselines. We compare against personalized generation methods spanning in-context learning (Non, RAG, Non-Think, RAG-Think), supervised fine-tuning (SFT), reinforcement learning (GRPO), and a hybrid approach (SFT+GRPO), with user-authored ground-truth (GT) as the gold standard. A broader range of baselines, along with their detailed descriptions, is provided in Appendix D.

Backbone Models. In the main experiments, we use `Qwen3-8B` (Yang et al., 2025) and `Qwen3-235B-A22B-Instruct` as backbones for rubric induction. Specifically, `Qwen3-8B` is employed for both zero-shot rubric induction and as a learnable model for our proposed PARL framework. For the downstream personalized generation tasks, we use `Qwen3-8B` as the primary generator across all methods to ensure a consistent and fair comparison, following prior work (Jin et al., 2025). Additional experiments with more backbone models are provided in Appendix G.

Evaluation Details. To systematically analyze the quality of induced rubrics for users in personalized evaluation, we consider the following primary metrics:

- ***User-level Accuracy***. We implement a user-level evaluation protocol by employing a log-probability-based scoring scheme, where a strong evaluator LLM (`Qwen3-235B-A22B-Instruct`) is employed to score each rubric dimension by softmax-normalized "Yes/No" token probabilities, considering a dimension satisfied if $p(\text{Yes}) > 0.5$. We use an all-or-nothing aggregation, reporting the mean pass rate over test samples, where a sample scores 1 only if all personalized rubrics are satisfied.
- ***User Coverage***. This metric quantifies the breadth of applicability of our framework by measuring the proportion of users for whom it successfully generates a non-empty, valid set of rubrics.

To further quantify the evaluative resolution of the induced rubrics, we report *Max-Diff*, defined as the difference between the *user-level accuracy* of GT and that of the strongest baseline. For all *user-level accuracy* results, we report the mean and standard deviation over five independent runs. In addition, we perform paired t-tests to examine whether GT responses achieve significantly higher *user-level accuracy* than competitive baselines.

Implementation Details. We benchmark five rubric-induction variants: *LM-8B* and *LM-235B* (pre-trained `Qwen3-8B` and `Qwen3-235B-A22B-Instruct` as generators), and three learned variants optimized with the scoring function $S(y_i; \mathcal{R}_u)$ only (***PARL-0***), the GT-scaled reward R_A (***PARL-A***), or the margin-only reward R_B (***PARL-B***). Hyperparameters and prompts are detailed in Appendices F and I.

5.2 Main Results

We assess the quality of the generated personalized user rubrics by applying them as evaluation criteria for both GT and a suite of baseline outputs. As illustrated in Table 1, our analysis consists of a *user-level accuracy* evaluation across three personalized text generation datasets alongside a detailed assessment of *user coverage*, from which we draw the following observations:

- Static hand-crafted prompts over pre-trained LLMs (LM-8B, LM-235B) prove inadequate for personalized evaluation: GT *user-level accuracy* occasionally falls below generic baselines, revealing

a clear misalignment between induced rubrics and users' true preferences. This underscores the necessity of framing personalized evaluation as a learning objective.

- PARL-0, which excludes the Discriminative Margin Product, attains high *user-level accuracy* for both GT and personalized baselines yet retains virtually no discriminative power, yielding overly permissive rubrics that cannot distinguish authentic user responses from AI outputs. This validates the indispensable role of the Discriminative Margin Product in enforcing high-fidelity alignment.
- Across all three tasks, PARL-A and PARL-B achieve the highest GT *user-level accuracy* and consistently outperform all baselines with robust discriminative power, establishing a wide resolution margin between authentic user preferences and competitive alternatives. The induced rubrics are thus both empirically grounded and diagnostically faithful, highlighting their utility for rigorous personalized evaluation.
- In terms of practical viability, LM-8B and LM-235B exhibit low *user coverage* due to frequent unusable or hallucinated criteria, whereas our learned variants maintain near-complete *user coverage* of approximately 100%, enabling reliable deployment across diverse real-world users.

5.3 Ablation Studies

To further examine the contribution of each core design in our framework, we conduct ablation studies on the Amazon Review Generation task using PARL-A as the base configuration.

As shown in Table 2, removing any core component degrades the effectiveness of PARL. Removing Preference Induction (w/o PI) yields a negative M*ax-Diff* despite high GT *user-accuracy*, indicating that the rubrics become insufficiently personalized and fail to distinguish authentic user responses from strong alternatives. Removing Consistency Validation (w/o CV) reduces both *Max-Diff* and *user coverage*, showing the importance of filtering unstable or task-specific criteria. Finally, removing RL optimization (w/o RL) leads to near-zero *Max-Diff* and much lower *user coverage*, indicating that the rubrics lack sufficient discriminative power without margin-based optimization.

Table 2: Ablation studies of PARL-A on Amazon Review Generation.

Method	GT	Max-Diff ↑	User Coverage ↑
w/o PI	0.953	-0.018	81.1%
w/o CV	0.916*	0.011	90.5%
w/o RL	0.931	0.001	49.1%
PARL-A	0.931*	**0.026**	**99.3%**

Overall, these results demonstrate the complementary roles of the three modules: PI captures representative user-specific criteria, CV improves rubric reliability, and RL sharpens the decision boundary between authentic user responses and competitive alternatives.

5.4 In-Depth Analysis

5.4.1 Evaluation with LLM-as-a-Judge

To expose the limitations of holistic scoring in personalized contexts, we compare PARL against a standard LLM-as-a-judge baseline. For fairness, the judge operates reference-free on the same raw user history used for rubric generation, uses the same `Qwen3-235B-A22B-Instruct` backbone, and outputs a 1–5 score normalized to $[0, 1]$; the prompt is given in Appendix I.

As shown in Figure 2, the judge attains reasonable representativeness and discriminativeness on *Amazon Review* but proves inconsistent elsewhere: on *Topic Writing* it fails to prioritize personalized resonance, scoring user-authentic ground truth below generic baseline outputs, while on *News Headline* the scores show no discernible evaluative pattern. Thus, even when given user historical contexts, standard LLM-as-a-judge approaches struggle as reliable reference-free evaluators, underscoring the need for our *Personalized Evaluation as Learning* paradigm, where the evaluator is dynamically optimized to capture idiosyncratic preferences rather than relying on static, general-purpose inference.

5.4.2 Cross-Dataset Generalization Analysis

To assess cross-domain robustness, we perform an out-of-domain (OOD) evaluation by applying a rubric generator trained on *Amazon Review* to the distinct *CDs_and_Vinyl* category (details in Appendix C), where it must induce rubrics for users entirely unseen during training from their historical contexts alone.

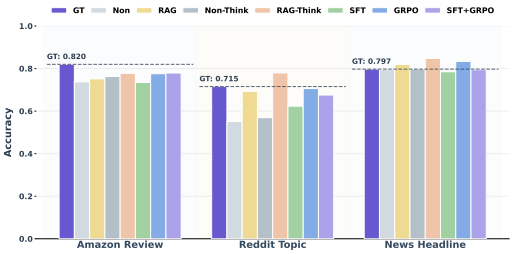

Figure 2: Results of LLM-as-a-judge evaluation scores across three datasets and various personalized generation methods.

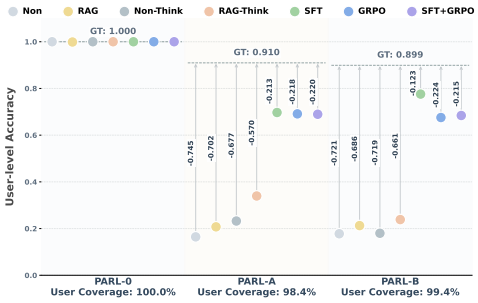

Figure 3: Evaluation of rubric induction variants on an out-of-domain dataset across various personalized generation methods.

As illustrated in Figure 3, PARL-0 attains high *user-level accuracy* and *user coverage* on GT data yet remains deficient in discriminative capability, confirming that without the Discriminative Margin Product the induced rubrics are overly permissive and generic. In contrast, both PARL-A and PARL-B simultaneously achieve strong representativeness and discriminativeness, indicating that the RL-driven induction distills persistent stylistic invariants rather than memorizing dataset-specific surface patterns. This supports our *Personalized Evaluation as Learning* paradigm: the rubric generator learns a transferable logic capturing the core of user identity.

5.4.3 Intrinsic Analysis: Semantic Diversity and Specificity

To further analyze the intrinsic quality of the induced rubrics, we vectorize them with `Qwen3-Embedding-8B` (Zhang et al., 2025d) and project the representations into a low-dimensional space to inspect their latent distribution and detect semantic collapse.

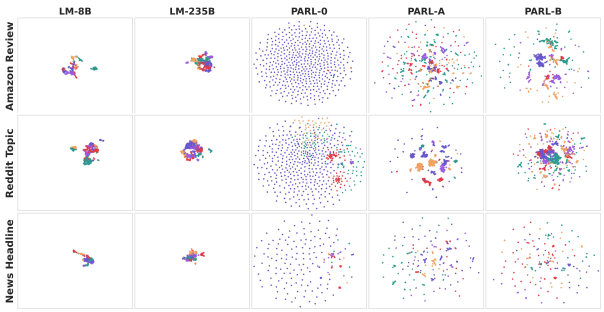

Figure 4: Intrinsic analysis of induced rubrics across five variants on three personalized generation tasks, revealing a shift from sparse, narrow clusters (LM-8B / LM-235B) and semantic collapse (PARL-0) to structured scattering (PARL-A / PARL-B).

As illustrated in Figure 4, the generators exhibit markedly different semantic distributions. Across all three tasks, LM-8B and LM-235B compress rubrics into sparse, narrow clusters, reflecting the limited capacity of pre-trained models to capture unique user traits beyond a few common patterns. PARL-0 instead shows clear semantic collapse: despite the high volume of rubrics, they form a dense, uniform sphere, confirming a tendency toward redundant, generic criteria that lack discriminative resolution across users. In contrast, PARL-A and PARL-B exhibit structured scattering, with rubrics broadly distributed yet forming clusters of distinct semantic features—indicating that our framework extracts idiosyncratic stylistic traits and induces multi-dimensional, tailored rubrics that establish a high-fidelity alignment boundary.

6 Conclusion

In this paper, we introduce the *Personalized Evaluation as Learning* paradigm, formulating personalized LLM evaluation as a learnable process grounded in three principles: *Representativeness*, *User-Consistency*, and *Discriminativeness*. Under this paradigm, **PARL** (**P**reference-**A**ware **R**ubric **L**earning) induces explicit, multi-dimensional rubrics from raw user histories via a discriminative RL objective, yielding rubrics that generalize across historical contexts while preserving user-specific decision boundaries. Experiments on real-world personalized generation tasks show that PARL consistently induces high-fidelity, reusable rubrics with strong user coverage, discriminative resolution, and semantic diversity, offering a foundational paradigm for future research on personalized evaluation and alignment.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code implementation, analysis and interpretation of experimental results, manuscript editing for readability, table and reference formatting, figure refinement, literature discovery, and polishing the phrasing of the paper's narrative. The core research contributions—including the formulation of the *Personalized Evaluation as Learning* paradigm, the three evaluation principles (*Representativeness*, *User-Consistency*, *Discriminativeness*), the design of the PARL framework and its discriminative reward objectives, and the experimental methodology—were developed by the authors. The authors reviewed all AI-assisted work, checked the implementation for correctness, and verified the analyses and interpretations against the experimental evidence. We take responsibility for the final content of this work, including text, claims, code, and figures produced with the aid of generative AI.

ETHICS STATEMENT

Our research involves the use of personalized user behavioral histories. To protect user privacy, we use open-source datasets, ensuring full compliance with the original datasets' MIT license. All data utilized in our experiments, including training sets for the rubric generator and the evaluation benchmarks, have been strictly de-identified.

While personalized evaluation aims to align with individual preferences, we acknowledge the risk that Large Language Models (LLMs) may inadvertently amplify social biases present in user data or pre-training corpora. To mitigate this, our induced rubrics are designed to be transparent and interpretable, allowing for human auditing of the criteria. We encourage users of our framework to implement additional safety filters to ensure that personalized criteria do not promote discriminatory or harmful content. Besides, we strongly advocate for the responsible deployment of personalized evaluation frameworks. The intended use of PARL is to improve the helpfulness and alignment of AI assistants, and we strictly prohibit its use for generating manipulative or malicious outputs.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the preference induction and consistency validation procedures in Section 4.2, and the discriminative reward objectives and reinforcement learning formulation in Section 4.3. Section 5.1 specifies the backbone models, evaluation protocol, and statistical reporting procedures. We also provide dataset sources and temporal splits (Appendix C), baseline descriptions (Appendix D), additional evaluation details (Appendix E), and implementation settings, including training hyperparameters, decoding configurations, and computational resources (Appendix F). The prompts used for rubric induction and evaluation are provided in Appendix I. Our source code is available at `https://anonymous.4open.science/r/PARL-1909`.

REFERENCES

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

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Param Biyani, Yasharth Bajpai, Arjun Radhakrishna, Gustavo Soares, and Sumit Gulwani. Rubicon: Rubric-based evaluation of domain-specific human ai conversations. In *Proceedings of the 1st ACM International Conference on AI-Powered Software*, pp. 161–169, 2024.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. Lore: Personalizing LLMs via low-rank reward modeling. In *Second Conference on Language Modeling*, 2025.

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, 2025a.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42, 2024.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. PAD: personalized alignment of llms at decoding-time. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, 2025b.

Yuxin Chen, Yi Zhang, Zhengzhou Cai, Yaorui Shi, Zhiyuan Yao, Chenhang Cui, Jingnan Zheng, Yaqi Huo, Xi Su, Qi Gu, Xunliang Cai, Xiang Wang, An Zhang, and Tat-Seng Chua. Vitabench 2.0: Evaluating personalized and proactive agents in long-term user interactions. *arXiv preprint arXiv:2605.27141*, 2026.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576, 2024.

Xiao Fu, Hossein A Rahmani, Bin Wu, Jerome Ramos, Emine Yilmaz, and Aldo Lipani. Pref: Reference-free evaluation of personalised text generation in llms. *arXiv preprint arXiv:2508.10028*, 2025.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Zhiliang Guo, Teng Tu, Yunshan Ma, and Xun Yang. Vibemus: Proactive agentic system for music personalization. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pp. 1–3, 2025b.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, 2024.

Yihan Hong, Huaiyuan Yao, Bolin Shen, Wanpeng Xu, Hua Wei, and Yushun Dong. Rulers: Locked rubrics and evidence-anchored scoring for robust llm evaluation. *arXiv preprint arXiv:2601.08654*, 2026.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

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

Song Jin, Juntian Zhang, Yong Liu, Xun Zhang, Yufei Zhang, Fei Jiang, Guojun Yin, Wei Lin, and Rui Yan. Tagging the thought: Unlocking personalization reasoning via reinforcement learning. *arXiv preprint arXiv:2509.23140*, 2025.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pp. 611–626, 2023.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference 2024*, pp. 3367–3378, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Jiahong Liu, Wenhao Yu, Quanyu Dai, Zhongyang Li, Jieming Zhu, Menglin Yang, Tat-Seng Chua, and Irwin King. Exploring personalization shifts in representation space of llms. In *Knowledgeable Foundation Models at ACL 2025*, 2025a.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. *arXiv preprint arXiv:2510.07743*, 2025b.

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

Sichun Luo, Guanzhi Deng, Jian Xu, Xiaojie Zhang, Hanxu Hou, and Linqi Song. Reasoning meets personalization: Unleashing the potential of large reasoning model for personalized generation. *arXiv preprint arXiv:2505.17571*, 2025.

Rishabh Misra. News category dataset. *arXiv preprint arXiv:2209.11429*, 2022.

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

Aditya Pathak, Rachit Gandhi, Vaibhav Uttam, Arnav Ramamoorthy, Pratyush Ghosh, Aaryan Raj Jindal, Shreyash Verma, Aditya Mittal, Aashna Ased, Chirag Khatri, et al. Rubric is all you need: Improving llm-based code evaluation with question-specific rubrics. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V. 1*, pp. 181–195, 2025.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, 2018.

Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. Latent inter-user difference modeling for LLM personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10610–10628, 2025a.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. Measuring what makes you unique: Difference-aware user modeling for enhancing LLM personalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21258–21277, 2025b.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*, 2023.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8045–8078, 2025.

Alireza Salemi and Hamed Zamani. LaMP-QA: A benchmark for personalized long-form question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1139–1159, 2025.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024.

Alireza Salemi, Julian Killingback, and Hamed Zamani. ExPerT: Effective and explainable evaluation of personalized long-form text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17516–17532, 2025a.

Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, Tao Chen, Zhuowan Li, Michael Bendersky, and Hamed Zamani. Reasoning-enhanced self-training for long-form personalized text generation. *arXiv preprint arXiv:2501.04167*, 2025b.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jingyan Shen, Jiarui Yao, Rui Yang, Yifan Sun, Feng Luo, Rui Pan, Tong Zhang, and Han Zhao. MiCRo: Mixture modeling and context-aware routing for personalized preference learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17458–17474, 2025.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. In *Second Conference on Language Modeling*, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025*, pp. 1279–1297, 2025.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6476–6491, 2024.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, 2017.

Chengbing Wang, Yang Zhang, Wenjie Wang, Xiaoyan Zhao, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Think-while-generating: On-the-fly reasoning for personalized long-form generation. *arXiv preprint arXiv:2512.06690*, 2025.

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

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, et al. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*, 2025.

Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. Personalized image generation with large multimodal models. In *Proceedings of the ACM on Web Conference 2025*, pp. 264–274, 2025a.

Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. Personalized generation in large model era: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24607–24649, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pp. 27263–27277, 2021.

Kepu Zhang, Teng Shi, Weijie Yu, and Jun Xu. Prlm: Learning explicit reasoning for personalized rag via contrastive reward optimization. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5484–5488, 2025a.

Pinyi Zhang, Ting-En Lin, Yuchuan Wu, Jingyang Chen, Zongqi Wang, Hua Yang, Ze Xu, Fei Huang, Kai Zhang, and Yongbin Li. P-genrm: Personalized generative reward model with test-time user-based scaling. *arXiv preprint arXiv:2602.12116*, 2026.

Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. Crowd comparative reasoning: Unlocking comprehensive evaluations for LLM-as-a-judge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5059–5074, 2025b.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

Yang Zhang, Juntao You, Yimeng Bai, Jizhi Zhang, Keqin Bao, Wenjie Wang, and Tat-Seng Chua. Causality-enhanced behavior sequence modeling in llms for personalized recommendation. *arXiv preprint arXiv:2410.22809*, 2024a.

Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. Collm: Integrating collaborative embeddings into large language models for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2025c.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025d.

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

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

Xiaoyan Zhao, Ming Yan, Yilun Qiu, Haoting Ni, Yang Zhang, Fuli Feng, Hong Cheng, and Tat-Seng Chua. Steerx: Disentangled steering for llm personalization. *arXiv preprint arXiv:2510.22256*, 2025b.

Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. Nextquill: Causal preference modeling for enhancing llm personalization. *arXiv preprint arXiv:2506.02368*, 2025c.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

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. 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*, 2023.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. HYDRA: model factorization framework for black-box LLM personalization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024*, 2024.

A Discussion: Advantages over Existing Evaluation Paradigms

Existing approaches to personalized evaluation span automatic metrics, human evaluation, and LLM-as-a-judge frameworks. Automatic metrics rely on surface-level matching and fixed references; human evaluation suffers from limited access to users' true preferences; and LLM-as-a-judge approaches rely on static pre-trained models as judgers and heuristic prompts. As a result, these approaches lack interpretability, adaptability, or principled grounding in long-term user behavioral history.

In contrast, our PARL framework provides a reference-free and explicit evaluation framework that induces reusable, multi-dimensional rubrics directly grounded in user behavior. Unlike reward modeling methods, which entangle preference representation with policy optimization, PARL decouples evaluation from generation, enabling principled benchmarking across models, tasks, and training regimes. Moreover, by formulating rubric induction as a learning objective, PARL adapts dynamically to user-specific decision boundaries rather than relying on fixed or heuristic criteria, offering a robust and extensible foundation for personalized evaluation.

B Limitations

The effectiveness of our rubric generator depends heavily on the quality and quantity of available user behavioral history. In cold-start scenarios, where historical signals are sparse, the model may struggle to induce sufficiently detailed and stable criteria, constraining its performance despite robust results for active users. Moreover, our current framework assumes user preferences to be static signatures captured at a single point in time, without explicitly modeling their temporal evolution. Since human preferences and stylistic tendencies are inherently dynamic, future work could explore cross-user knowledge transfer, meta-learning strategies, or adaptive update mechanisms to enable rapid initialization for new users and continual refinement as preferences evolve.

Our investigation focuses on personalized text generation evaluation, while empirical exploration of personalized alignment and extensions to multi-modal settings remain open directions for future research. More fundamentally, we acknowledge a philosophical constraint: only the individual user ultimately defines their own preferences, and our framework serves as a principled proxy for measurement rather than a perfect reflection of human subjectivity. Although PARL achieves promising results, it has not yet attained a fully ideal state—where GT user-level accuracy reaches 1.0 and user coverage remains consistently 100%. This suggests that the nuanced preferences of inconsistent users remain difficult to capture, highlighting the need for future research to refine preference-aware strategies within this emerging paradigm.

C Dataset Details

In this section, we provide more detailed descriptions and statistical analyses of our three representative and widespread evaluation tasks:

- *Amazon Review Generation*[1] (Hou et al., 2024): The objective of this task is to generate a detailed item review text conditioned on the input and the user profile. The input includes the item title and description, the user's rating for this item, and the title for this review. The user profile consists of the user's past reviews, including the corresponding item title and description, the rating for the item, the review title, and the review text. Specifically, we utilize the *Movies_and_TV* and *Books* categories for the main experiments in Section 5.2, while reserving *CDs_and_Vinyl* as an out-of-domain (OOD) dataset for the analysis in Section 5.4.2. This task was processed by DPL (Qiu et al., 2025b).
- *Reddit Topic Writing*[2] (Völske et al., 2017): This task involves generating the full content of a Reddit post, based on the post's summary written by the author, and conditioned on a user profile consisting of the author's previous summary-content pairs. This task was processed by LongLaMP (Kumar et al., 2024).

[1] https://huggingface.co/datasets/SnowCharmQ/DPL-main
[2] https://huggingface.co/datasets/LongLaMP/LongLaMP

Table 3: Statistical data overview of three personalized text generation tasks.

Task	# user	Dataset	# data	len(data)
Amazon Review	2242	*train*	16166	1379.65±1058.33
		val	2242	1604.38±1529.67
		test	2242	1632.54±1660.65
Reddit Topic	2452	*train*	19337	1341.36±1126.73
		val	2452	1886.86±1759.52
		test	2452	1443.47±1286.57
News Headline	827	*train*	5169	73.38±20.34
		val	827	53.91±20.77
		test	827	54.39±20.13

- *News Headline Generation*[3] (Misra, 2022): This task requires the model to generate a headline for a given input news article, conditioned on a user profile that consists of the author's historical article-headline pairs. This task was processed by LaMP (Salemi et al., 2024).

In our experiments, we follow the temporal settings of these three datasets to configure the validation and test data splits. Specifically, each test instance corresponds to a user's most recent writing. The detailed statistics of these datasets are summarized in Table 3.

D Baseline Details

In this section, we provide more detailed descriptions of baselines benchmarked by PARL. To comprehensively evaluate the effectiveness and robustness of our framework, we extend our analysis beyond methods featured in Section 5.1 to include a broader suite of competitive baselines. These methods are categorized into four paradigms: ICL (In-Context Learning), SFT (Supervised Fine-Tuning), RL (Reinforcement Learning), and Hybrid (SFT + RL).

ICL:

- **Non**: This method generates the target content without leveraging any user-specific information. The input to the model includes only the content-specific input.
- **RAG** (Salemi et al., 2024): This method enhances the personalized generation process by retrieving relevant historical user data from the user's profile, which is then formatted and included in the prompt context alongside the input.
- **PAG** (Richardson et al., 2023): Building upon RAG, this method first summarizes the most recent historical instances from the user's profile into a compact personalized profile. The generated profile, along with the retrieved records, is included in the input to the LLM, allowing it to generate personalized content.
- **Non-Think**: This method enhances the basic Zero-Shot approach by enabling the thinking mode of LLM to analyze the input before generating the final content. No user history is leveraged.
- **RAG-Think**: This method enhances the RAG approach by generating an intermediate reasoning step that analyzes user preferences based on the retrieved history, and then leverages this reasoning to guide the personalized generation.
- **PAG-Think**: Building upon PAG, this method generates a reasoning path based on the summarized profile and retrieved records, guiding the subsequent generation of the personalized output.
- **R2P** (Luo et al., 2025): This method enhances model reasoning for personalization by incorporating a Hierarchical Reasoning Thought Template, Reasoning Process Intervention, and Self-Referencing Module to guide the LLM toward structured and user-aligned outputs.

[3] https://lamp-benchmark.github.io/

SFT:

- **SFT** (Salemi et al., 2024): This method directly trains the LLM on the full training dataset using Supervised Fine-Tuning (SFT), where the input prompt for each instance is augmented with the user history.
- **NextQuill** (Zhao et al., 2025c): This is an SFT-based method grounded in Causal Preference Modeling. It enhances personalization by aligning model-side causal preference effects with ground-truth data, ensuring the model focuses only on preference-driven tokens.

RL:

- **GRPO** (Shao et al., 2024): This method applies the Group Relative Policy Optimization (GRPO) RL framework to explicitly enhance the model's ability to generate effective personalization reasoning paths. All input prompts are augmented with the user's historical data. The reward used for GRPO training is defined as the sum of ROUGE-1 and METEOR scores.
- **GSPO** (Zheng et al., 2025): Similar to GRPO, this method employs the Group Sequence Policy Optimization (GSPO) RL framework. The input data are the same as those used for GRPO. The reward used for GSPO training is defined as the sum of ROUGE-1 and METEOR scores.
- **PrLM** (Zhang et al., 2025a): This is an RL-based method that trains the LLM for personalized RAG to explicitly reason over retrieved user profiles. It uses a contrastively trained personalization reward model to guide GRPO learning.

Hybrid:

- **TagPR** (Jin et al., 2025): A framework that enhances personalization reasoning by "tagging the thought". It employs a synergistic SFT-RL training strategy, guided by a composite reward signal that integrates tag-based structural constraints and a Personalization Reward Model with User Embeddings.
- **SFT+GRPO**: This method combines SFT and GRPO in a two-stage pipeline, first capturing general user styles via SFT and then refining the model using GRPO to better align with user preferences. The reward used during the RL stage is defined as the sum of ROUGE-1 and METEOR scores.

E Evaluation Details

In addition to the metrics described in Section 5.1, we introduce a more granular measure to evaluate the fine-grained alignment of the generated rubrics:

- ***Rubric-level Accuracy***. This metric provides a high-resolution assessment of rubric satisfaction by evaluating individual dimensions independently rather than using all-or-nothing aggregation. It is calculated as the mean satisfaction rate across the entire pool of generated rubrics, where each specific rubric is scored using the same log-probability-based scheme ($p(\text{Yes}) > 0.5$) to determine if the response fulfills that particular stylistic constraint.

F Implementation Details

We provide more details about our implementation in this section.

Rubric Induction:

To balance historical representativeness with inference efficiency, the rubric generator is conditioned on up to 10 of the latest interactions from $\mathcal{H}_u$. For the RL stage, we optimize the policy using GRPO with a group size $K = 5$, a clipping threshold $\epsilon = 0.2$, and a KL-penalty coefficient $\beta = 0.01$. In the reward function, the scaling factor κ is set to 20 to sharpen the discriminative margin, the reward scaling constant λ is set to 10, and the reward is modulated by a cardinality factor $\mathcal{C} = \min(1, |\mathcal{R}_u|/2)$.

We train the rubric generation model using the `verl` (Sheng et al., 2025) framework with a GRPO-based (Shao et al., 2024) optimization objective. Training is conducted with a batch size of 8, a maximum prompt length of 10240 tokens, and a maximum response length of 2048 tokens.

Optimization is carried out using Adam (Kingma & Ba, 2015) with a learning rate of 1×10^{-5} and a cosine learning rate scheduler. For rollout generation, we employ `vLLM` (Kwon et al., 2023) with nucleus sampling ($p = 0.95$), temperature 0.6, and top-k sampling ($k = 20$), producing 5 samples per prompt. For rubric generation, we adopt the same `vLLM`-based deployment configuration. The selected baselines for computing *Discriminative Margin Product* include: *Non*, *RAG*, *SFT*, *GRPO*, *SFT+GRPO*.

Personalized Text Generation:

For downstream personalized text generation implementation, we note that several baselines do not have official open-source implementations and are not built upon the same backbone model (Luo et al., 2025; Jin et al., 2025). We therefore reproduce these methods and ensure that all baselines are evaluated under unified training and inference settings to guarantee a fair comparison. For baselines that require user history, we adopt a BM25 (Robertson et al., 2009) retriever and consistently retrieve the top-4 user history items.

For SFT-based methods, we set the learning rate to 5×10^{-6} and the weight decay to 0.025, with a cosine learning rate scheduler and a warmup ratio of 0.01. For RL-based methods, we set the learning rate to 1×10^{-6} with a cosine learning rate scheduler, and we sample 5 responses for each prompt during the policy rollout stage. For hybrid methods, we use the same settings as the SFT- and RL-based methods in the respective stages.

Training is performed on 8 H800 GPUs, while generation is conducted using vLLM on a single H800 GPU with nucleus sampling ($p = 0.95$), temperature 0.6, and top-k sampling ($k = 20$).

G Additional Experiments

G.1 Cardinality Analysis

To analyze the impact of the cardinality constraint $\mathcal{C}$ during training, we compare our primary threshold-based implementation on the *Amazon Review* task, defined as $\mathcal{C} = \min(1, |\mathcal{R}_u|/2)$, against a linear variant where $\mathcal{C} = |\mathcal{R}_u|/10$. This linear implementation is designed to incentivize the rubric generator to produce the maximum possible number of valid rubrics for evaluation. Here, we utilize the reward function from PARL-A, and we report the *User-level Accuracy*, *User Coverage*, *Max-Diff*, and the average number of valid rubrics induced per user as *Avg. Num. / User*.

Table 4: Comparison of different cardinality constraints during rubric generator training.

Paradigm	***Threshold***	***Linear***
GT	**0.9313**$^{*}_{\pm 0.0020}$	$\underline{0.9349}_{\pm 0.0010}$
Non	$0.2718_{\pm 0.0018}$	$0.3543_{\pm 0.0015}$
RAG	$0.3252_{\pm 0.0024}$	$0.3689_{\pm 0.0032}$
Non-Think	$0.2250_{\pm 0.0017}$	$0.2956_{\pm 0.0013}$
RAG-Think	$0.3299_{\pm 0.0013}$	$0.3447_{\pm 0.0017}$
SFT	$0.8773_{\pm 0.0017}$	$\mathbf{0.9414}_{\pm 0.0012}$
GRPO	$0.8322_{\pm 0.0018}$	$0.8318_{\pm 0.0009}$
SFT+GRPO	$\underline{0.9053}_{\pm 0.0027}$	$0.9140_{\pm 0.0007}$
Max-Diff	$+0.0260$	-0.0065
User Coverage	99.33%	99.96%
Avg. Num. / User	4.01	6.56

As shown in Table 4, while the Linear paradigm significantly increases the average number of valid rubrics induced per user, it fails to yield the highest GT *user-level accuracy*. This indicates that simply maximizing the number of rubrics during training is suboptimal, suggesting that an excessive quantity of rubrics may introduce redundancy or generic criteria, diluting evaluative precision. Instead, focusing on a reasonable volume of high-utility, discriminative rubrics is essential for effective personalized evaluation.

G.2 Average Number of Rubrics per User

We report the average number of rubrics induced per user across various generator variants, with the results summarized in Table 5. The data indicate that standard pre-trained models generally yield a limited number of valid rubrics. This constrained output highlights their relatively weak capacity to capture representative user signatures from raw historical contexts.

While PARL-0 generates a significantly higher volume of rubrics, these tend to be overly general and lack the discriminative resolution necessary for high-fidelity evaluation. In contrast, PARL-A

Table 5: Average number of valid rubrics induced per user across three datasets.

Dataset	LM-8B	LM-235B	LM-30B-IT	LM-30B-Think	GPT-4o-mini	PARL-0	PARL-A	PARL-B
Amazon Review	0.71	1.81	0.44	2.27	2.11	10.00	4.01	3.40
Reddit Post	1.20	1.46	0.66	2.65	2.64	9.26	4.55	4.60
News Headline	1.04	2.02	1.01	6.04	2.40	9.99	4.98	5.34

and PARL-B maintain a more moderate and reasonable rubric count. This balance suggests that our learning-based process successfully filters out redundant or generic noise to distill a concise yet potent set of user-specific preferences.

G.3 Effect of History Number

We further analyze how the number of historical contexts affects rubric induction on the Amazon Review Generation task using PARL-A as the base configuration. As shown in Table 6, using no historical context results in a significantly negative *Max-Diff*, confirming that user history is essential for inducing personalized rubrics. With only 2 historical contexts, the generator still fails to obtain a positive margin, suggesting that limited history provides insufficient evidence for stable preference extraction. As the history size increases to 5 and 8, *Max-Diff* and *user coverage* improve consistently, with 8 histories approaching the full setting of 10 histories. These results indicate that PARL does not require extensive user histories, but a moderately sized seed history is important for inducing reliable and representative personalized rubrics.

Table 6: Effect of history number of PARL-A on Amazon Review Generation.

Method	**GT**	**Max-Diff ↑**	**Coverage ↑**
#His=0	0.953	-0.018	81.1%
#His=2	0.930	-0.007	91.3%
#His=5	0.918	0.002	96.7%
#His=8	0.921*	0.019	99.1%
PARL-A	0.931*	**0.026**	**99.3%**

G.4 Experimental Results on User-level Accuracy

In this section, we report results for all introduced baselines in Appendix D. Besides, we employ a broader set of pre-trained backbone models, including the following: `Qwen3-30B-A3B-Instruct` (***LM-30B-It***), `Qwen3-30B-A3B-Thinking` (***LM-30B-Think***), and `GPT-4o-mini`. The complete experimental results on *user-level accuracy* are shown in Table 7.

The results further confirm the overall superiority of our proposed PARL framework: authentic user-authored GT responses consistently achieve the highest *user-level accuracy*, demonstrating statistical significance in alignment with true preferences. Moreover, our framework establishes a clear discriminative margin across diverse baselines, effectively distinguishing high-fidelity personalization from AI outputs. In contrast, existing pre-trained models fail to meet the required evaluative standards, often yielding inconsistent or overly permissive *user-level accuracy*. This gap further underscores the structural advantage of our learning-driven approach in capturing the nuances of LLM personalization.

G.5 Experimental Results on Rubric-level Accuracy

We provide the complete experimental results on *rubric-level accuracy* as shown in Table 8.

Our framework, PARL, consistently ensures that authentic user-authored GT responses achieve the highest *rubric-level accuracy* across all three personalized text generation tasks. Beyond absolute accuracy, PARL demonstrates robust discriminative power, effectively establishing a clear evaluative margin between GT and competitive baselines. These results confirm that the induced rubrics are both well-grounded and highly diagnostic for personalized assessment.

G.6 Experimental Results on Other Metrics

We further provide experimental results of benchmarked baselines on existing evaluation metrics: ROUGE-1 (Lin, 2004), ROUGE-L, BLEU[4] (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), BERTScore[5] (Zhang et al., 2020), and EmbScore[6].

From the experimental results presented in Table 9, we draw the following observations:

- ICL-based methods that utilize historical user data outperform non-personalized baselines across almost all evaluated metrics. Its performance gain is highly consistent with the trends observed in our proposed PARL metric, further validating that historical context is indispensable for aligning model outputs with individual user characteristics.
- Both SFT and RL strategies effectively enhance the alignment between model outputs and actual user ground-truth responses. Notably, the hybrid SFT+RL paradigm demonstrates the most comprehensive improvements across the majority of metrics.
- A more granular analysis reveals notable discrepancies in evaluation trends, where high performance in lexical-based metrics (e.g., ROUGE-1, ROUGE-L, and BLEU) does not consistently translate to superior results in semantic-based metrics (e.g., METEOR, BERTScore, and EmbScore). These metrics often exhibit divergent trends across different methods, suggesting they capture distinct and potentially conflicting facets of text quality. This inconsistency prompts a necessary reassessment of traditional evaluation protocols, underscoring the limitations of these metrics in providing a holistic measure of personalized generation.

H Case Study

In this section, we present a case study from the *Amazon Review* task to qualitatively illustrate how our induced rubrics evaluate personalized content, as shown in Figure 5. By comparing an authentic user-authored response (GT) with an AI-generated output, we demonstrate the framework's capacity to distinguish fine-grained stylistic nuances and personal preferences. This analysis highlights the interpretability and diagnostic power of the rubrics in identifying the subtle misalignments often present in generic AI outputs.

I Overview of Prompts

In this section, we illustrate the prompt design used in our framework PARL.

We present prompts used for generating personalized user rubrics in Table 10, Table 11, and Table 12. To ensure the systematic and reliable generation of personalized user rubrics, our prompt design follows a structured hierarchical logic. Initially, the model is contextualized with its specific role and input data schema. Subsequently, we define the core task under a rigorous set of drafting rules, encompassing atomicity, consistency, determinism, specificity, agnosticism, multi-dimensionality, and personalization. Finally, the output is mandated in a structured JSON format to facilitate seamless downstream evaluation and programmatic parsing.

We present prompts used to evaluate the generated text and user-authored ground truth using personalized user rubrics generated by our framework in Table 13, Table 14, and Table 15.

As shown in Table 16, Table 17, and Table 18, we also provide prompts used in the comparison LLM-as-a-judge experiments in Section 5.4.1 for reference.

[4]We use the standard open-source `SacreBLEU` (Post, 2018) library to calculate the BLEU score: `https://github.com/mjpost/sacrebleu`.

[5]We adopt the `led-base-16384` (Beltagy et al., 2020) model to obtain embeddings.

[6]We use two state-of-the-art embedding models (`Qwen3-Embedding-0.6B` (Zhang et al., 2025d) and `Qwen3-Embedding-4B`) to encode both user-authored ground-truth texts and generated outputs, compute the cosine similarity for each model, and report the average as the score here.

Table 7: Detailed evaluation results of induced rubrics across three personalized text generation tasks on *user-level accuracy*.

	Amazon Review		LM-8B	LM-235B	LM-30B-IT	LM-30B-Think	GPT-4o-mini	PARL-0	**PARL-A**	**PARL-B**
User-level Accuracy		**GT**	**0.9307**$_{\pm 0.0011}$	**0.7943***$_{\pm 0.0030}$	**0.9256***$_{\pm 0.0017}$	0.9300$_{\pm 0.0013}$	0.9393$_{\pm 0.0015}$	**1.0000**$_{\pm 0.0000}$	0.9313$_{\pm 0.0020}$	**0.9295***$_{\pm 0.0006}$
		Non	0.5476$_{\pm 0.0023}$	0.2557$_{\pm 0.0017}$	0.4359$_{\pm 0.0035}$	0.6856$_{\pm 0.0007}$	0.6281$_{\pm 0.0016}$	**1.0000**$_{\pm 0.0000}$	0.2718$_{\pm 0.0018}$	0.5668$_{\pm 0.0011}$
		RAG	0.6236$_{\pm 0.0021}$	0.3442$_{\pm 0.0020}$	0.5306$_{\pm 0.0015}$	0.7561$_{\pm 0.0013}$	0.6724$_{\pm 0.0016}$	0.9979$_{\pm 0.0002}$	0.3252$_{\pm 0.0024}$	0.6164$_{\pm 0.0014}$
		PAG	0.6133$_{\pm 0.0017}$	0.3534$_{\pm 0.0011}$	0.5162$_{\pm 0.0040}$	0.7444$_{\pm 0.0024}$	0.6167$_{\pm 0.0020}$	0.9987$_{\pm 0.0000}$	0.2573$_{\pm 0.0023}$	0.5410$_{\pm 0.0028}$
	ICL	Non-Think	0.6011$_{\pm 0.0013}$	0.2800$_{\pm 0.0016}$	0.4623$_{\pm 0.0015}$	0.6961$_{\pm 0.0011}$	0.6219$_{\pm 0.0007}$	**1.0000**$_{\pm 0.0000}$	0.2250$_{\pm 0.0017}$	0.5790$_{\pm 0.0010}$
		RAG-Think	0.7396$_{\pm 0.0029}$	0.4203$_{\pm 0.0011}$	0.5942$_{\pm 0.0030}$	0.7892$_{\pm 0.0006}$	0.7333$_{\pm 0.0012}$	0.9997$_{\pm 0.0002}$	0.3299$_{\pm 0.0013}$	0.6694$_{\pm 0.0020}$
		PAG-Think	0.7100$_{\pm 0.0026}$	0.4266$_{\pm 0.0010}$	0.5628$_{\pm 0.0010}$	0.7875$_{\pm 0.0012}$	0.6368$_{\pm 0.0019}$	**1.0000**$_{\pm 0.0000}$	0.2397$_{\pm 0.0048}$	0.5614$_{\pm 0.0021}$
		R2P	0.7640$_{\pm 0.0016}$	0.4473$_{\pm 0.0007}$	0.6633$_{\pm 0.0018}$	0.8138$_{\pm 0.0008}$	0.7552$_{\pm 0.0016}$	**1.0000**$_{\pm 0.0000}$	0.3538$_{\pm 0.0016}$	0.7100$_{\pm 0.0019}$
	SFT	SFT	0.8269$_{\pm 0.0023}$	0.6416$_{\pm 0.0033}$	0.8124$_{\pm 0.0015}$	0.9057$_{\pm 0.0016}$	0.8671$_{\pm 0.0018}$	**1.0000**$_{\pm 0.0000}$	0.8773$_{\pm 0.0017}$	0.8800$_{\pm 0.0012}$
		NextQuill	0.6733$_{\pm 0.0015}$	0.5449$_{\pm 0.0014}$	0.7124$_{\pm 0.0019}$	0.8116$_{\pm 0.0027}$	0.7316$_{\pm 0.0019}$	**1.0000**$_{\pm 0.0000}$	0.8673$_{\pm 0.0011}$	0.7880$_{\pm 0.0011}$
	RL	GRPO	0.9304$_{\pm 0.0012}$	0.7869$_{\pm 0.0007}$	0.9182$_{\pm 0.0030}$	**0.9588**$_{\pm 0.0017}$	**0.9617**$_{\pm 0.0009}$	**1.0000**$_{\pm 0.0000}$	0.8322$_{\pm 0.0018}$	0.9182$_{\pm 0.0010}$
		GSPO	0.9276$_{\pm 0.0012}$	0.7740$_{\pm 0.0020}$	0.9197$_{\pm 0.0017}$	0.9516$_{\pm 0.0015}$	0.9475$_{\pm 0.0018}$	**1.0000**$_{\pm 0.0000}$	0.8006$_{\pm 0.0021}$	0.9191$_{\pm 0.0011}$
		PrLM	0.9176$_{\pm 0.0027}$	0.7579$_{\pm 0.0016}$	0.8922$_{\pm 0.0009}$	0.9525$_{\pm 0.0013}$	0.9431$_{\pm 0.0010}$	**1.0000**$_{\pm 0.0000}$	0.7665$_{\pm 0.0044}$	0.8965$_{\pm 0.0020}$
	Hybrid	TagPR	0.9265$_{\pm 0.0013}$	0.7625$_{\pm 0.0007}$	0.9063$_{\pm 0.0029}$	0.9482$_{\pm 0.0011}$	0.9424$_{\pm 0.0020}$	**1.0000**$_{\pm 0.0000}$	0.8658$_{\pm 0.0031}$	0.9118$_{\pm 0.0009}$
		SFT + GRPO	0.9165$_{\pm 0.0025}$	0.7592$_{\pm 0.0009}$	0.9005$_{\pm 0.0021}$	0.9552$_{\pm 0.0019}$	0.9512$_{\pm 0.0019}$	**1.0000**$_{\pm 0.0000}$	0.9053$_{\pm 0.0027}$	0.9272$_{\pm 0.0009}$
	User Coverage		49.06%	84.61%	35.24%	78.64%	94.11%	100.00%	99.33%	99.58%

	Reddit Topic		LM-8B	LM-235B	LM-30B-IT	LM-30B-Think	GPT-4o-mini	PARL-0	**PARL-A**	**PARL-B**
User-level Accuracy		**GT**	0.8916$_{\pm 0.0004}$	**0.8395**$_{\pm 0.0028}$	0.8557$_{\pm 0.0017}$	0.8260$_{\pm 0.0028}$	0.8584$_{\pm 0.0021}$	0.9542$_{\pm 0.0007}$	**0.8480***$_{\pm 0.0014}$	**0.8604***$_{\pm 0.0018}$
		Non	0.4218$_{\pm 0.0002}$	0.3714$_{\pm 0.0007}$	0.4201$_{\pm 0.0023}$	0.4175$_{\pm 0.0014}$	0.2846$_{\pm 0.0011}$	0.9724$_{\pm 0.0002}$	0.2190$_{\pm 0.0003}$	0.1013$_{\pm 0.0006}$
		RAG	0.6150$_{\pm 0.0003}$	0.5307$_{\pm 0.0012}$	0.5275$_{\pm 0.0015}$	0.6241$_{\pm 0.0015}$	0.5406$_{\pm 0.0009}$	0.9849$_{\pm 0.0004}$	0.2666$_{\pm 0.0007}$	0.2340$_{\pm 0.0013}$
		PAG	0.6404$_{\pm 0.0006}$	0.5817$_{\pm 0.0020}$	0.5185$_{\pm 0.0009}$	0.6719$_{\pm 0.0021}$	0.5520$_{\pm 0.0024}$	0.9852$_{\pm 0.0002}$	0.1918$_{\pm 0.0015}$	0.1929$_{\pm 0.0012}$
	ICL	Non-Think	0.5055$_{\pm 0.0002}$	0.4075$_{\pm 0.0012}$	0.4027$_{\pm 0.0021}$	0.4552$_{\pm 0.0021}$	0.3978$_{\pm 0.0012}$	0.5046$_{\pm 0.0011}$	0.1958$_{\pm 0.0022}$	0.1841$_{\pm 0.0008}$
		RAG-Think	0.7070$_{\pm 0.0004}$	0.6170$_{\pm 0.0012}$	0.5799$_{\pm 0.0028}$	0.6806$_{\pm 0.0007}$	0.6463$_{\pm 0.0020}$	0.8184$_{\pm 0.0006}$	0.2965$_{\pm 0.0015}$	0.2998$_{\pm 0.0013}$
		PAG-Think	0.6898$_{\pm 0.0008}$	0.6294$_{\pm 0.0011}$	0.5405$_{\pm 0.0020}$	0.6914$_{\pm 0.0021}$	0.6063$_{\pm 0.0010}$	0.7845$_{\pm 0.0011}$	0.2637$_{\pm 0.0013}$	0.3049$_{\pm 0.0013}$
		R2P	0.7986$_{\pm 0.0002}$	0.6736$_{\pm 0.0007}$	0.7305$_{\pm 0.0006}$	0.6879$_{\pm 0.0021}$	0.7320$_{\pm 0.0020}$	0.9472$_{\pm 0.0005}$	0.5296$_{\pm 0.0027}$	0.4817$_{\pm 0.0026}$
	SFT	SFT	0.8028$_{\pm 0.0007}$	0.7642$_{\pm 0.0011}$	0.8343$_{\pm 0.0014}$	0.7577$_{\pm 0.0020}$	0.7248$_{\pm 0.0013}$	0.9706$_{\pm 0.0007}$	0.7600$_{\pm 0.0044}$	0.7899$_{\pm 0.0016}$
		NextQuill	0.6879$_{\pm 0.0016}$	0.6612$_{\pm 0.0038}$	0.7927$_{\pm 0.0030}$	0.6635$_{\pm 0.0018}$	0.5626$_{\pm 0.0018}$	0.9803$_{\pm 0.0002}$	0.7365$_{\pm 0.0018}$	0.7705$_{\pm 0.0006}$
	RL	GRPO	0.8909$_{\pm 0.0011}$	0.7985$_{\pm 0.0026}$	0.8577$_{\pm 0.0011}$	0.8602$_{\pm 0.0009}$	0.8727$_{\pm 0.0011}$	0.9821$_{\pm 0.0004}$	0.2678$_{\pm 0.0011}$	0.4841$_{\pm 0.0009}$
		GSPO	0.8948$_{\pm 0.0009}$	0.8038$_{\pm 0.0015}$	0.8462$_{\pm 0.0025}$	0.8552$_{\pm 0.0016}$	0.8728$_{\pm 0.0013}$	**0.9933**$_{\pm 0.0002}$	0.3533$_{\pm 0.0014}$	0.4739$_{\pm 0.0007}$
		PrLM	0.7677$_{\pm 0.0021}$	0.6137$_{\pm 0.0012}$	0.7375$_{\pm 0.0017}$	0.7612$_{\pm 0.0009}$	0.6683$_{\pm 0.0014}$	0.9646$_{\pm 0.0003}$	0.2896$_{\pm 0.0026}$	0.4936$_{\pm 0.0009}$
	Hybrid	TagPR	**0.9067**$_{\pm 0.0012}$	0.8006$_{\pm 0.0015}$	0.8343$_{\pm 0.0025}$	**0.8875**$_{\pm 0.0009}$	**0.8902**$_{\pm 0.0008}$	0.9851$_{\pm 0.0002}$	0.4192$_{\pm 0.0018}$	0.6071$_{\pm 0.0010}$
		SFT + GRPO	0.8988$_{\pm 0.0009}$	0.8359$_{\pm 0.0008}$	**0.8757**$_{\pm 0.0014}$	0.8741$_{\pm 0.0015}$	0.8700$_{\pm 0.0006}$	0.9711$_{\pm 0.0004}$	0.7705$_{\pm 0.0018}$	0.8243$_{\pm 0.0004}$
	User Coverage		74.96%	81.53%	51.43%	85.73%	98.12%	100.00%	99.88%	99.43%

	News Headline		LM-8B	LM-235B	LM-30B-IT	LM-30B-Think	GPT-4o-mini	PARL-0	**PARL-A**	**PARL-B**
User-level Accuracy		**GT**	0.8043$_{\pm 0.0029}$	0.7647$_{\pm 0.0019}$	0.8184$_{\pm 0.0000}$	0.7198$_{\pm 0.0006}$	0.7653$_{\pm 0.0010}$	0.9976$_{\pm 0.0000}$	**0.8325***$_{\pm 0.0053}$	**0.8039***$_{\pm 0.0022}$
		Non	0.6561$_{\pm 0.0007}$	0.6364$_{\pm 0.0016}$	0.6893$_{\pm 0.0000}$	0.5939$_{\pm 0.0008}$	0.6192$_{\pm 0.0015}$	0.9976$_{\pm 0.0000}$	0.3311$_{\pm 0.0037}$	0.4135$_{\pm 0.0040}$
		RAG	0.7396$_{\pm 0.0000}$	0.7136$_{\pm 0.0027}$	0.7987$_{\pm 0.0000}$	0.6712$_{\pm 0.0008}$	0.7058$_{\pm 0.0006}$	0.9976$_{\pm 0.0000}$	0.4679$_{\pm 0.0043}$	0.5693$_{\pm 0.0053}$
		PAG	0.7707$_{\pm 0.0029}$	0.7425$_{\pm 0.0023}$	0.8201$_{\pm 0.0009}$	0.5961$_{\pm 0.0005}$	0.7187$_{\pm 0.0012}$	0.9964$_{\pm 0.0000}$	0.5541$_{\pm 0.0081}$	0.6784$_{\pm 0.0049}$
	ICL	Non-Think	0.6810$_{\pm 0.0018}$	0.6356$_{\pm 0.0024}$	0.6937$_{\pm 0.0000}$	0.5747$_{\pm 0.0006}$	0.6137$_{\pm 0.0005}$	0.9964$_{\pm 0.0000}$	0.2535$_{\pm 0.0018}$	0.4082$_{\pm 0.0062}$
		RAG-Think	0.7052$_{\pm 0.0032}$	0.7183$_{\pm 0.0022}$	0.7681$_{\pm 0.0000}$	0.6012$_{\pm 0.0000}$	0.6772$_{\pm 0.0000}$	0.9976$_{\pm 0.0000}$	0.4151$_{\pm 0.0035}$	0.5608$_{\pm 0.0048}$
		PAG-Think	0.7649$_{\pm 0.0011}$	0.7197$_{\pm 0.0042}$	0.7681$_{\pm 0.0000}$	0.5870$_{\pm 0.0015}$	0.6663$_{\pm 0.0010}$	0.9976$_{\pm 0.0000}$	0.5053$_{\pm 0.0034}$	0.6839$_{\pm 0.0049}$
		R2P	0.7928$_{\pm 0.0014}$	0.7536$_{\pm 0.0023}$	0.8249$_{\pm 0.0000}$	0.7323$_{\pm 0.0005}$	0.7646$_{\pm 0.0000}$	0.9964$_{\pm 0.0000}$	0.4449$_{\pm 0.0066}$	0.6537$_{\pm 0.0049}$
	SFT	SFT	0.8156$_{\pm 0.0011}$	**0.7867**$_{\pm 0.0021}$	0.8048$_{\pm 0.0009}$	**0.7482**$_{\pm 0.0005}$	0.7463$_{\pm 0.0013}$	**0.9988**$_{\pm 0.0000}$	0.8139$_{\pm 0.0039}$	0.7497$_{\pm 0.0040}$
		NextQuill	0.7935$_{\pm 0.0027}$	0.7550$_{\pm 0.0029}$	0.8206$_{\pm 0.0000}$	0.7318$_{\pm 0.0010}$	0.7395$_{\pm 0.0005}$	**0.9988**$_{\pm 0.0000}$	0.8034$_{\pm 0.0020}$	0.7378$_{\pm 0.0055}$
	RL	GRPO	0.7834$_{\pm 0.0018}$	0.7450$_{\pm 0.0026}$	0.8337$_{\pm 0.0000}$	0.6299$_{\pm 0.0006}$	0.7319$_{\pm 0.0005}$	0.9976$_{\pm 0.0000}$	0.4960$_{\pm 0.0039}$	0.5906$_{\pm 0.0077}$
		GSPO	0.8108$_{\pm 0.0018}$	0.7722$_{\pm 0.0018}$	**0.8446**$_{\pm 0.0000}$	0.7399$_{\pm 0.0008}$	0.7790$_{\pm 0.0006}$	0.9964$_{\pm 0.0000}$	0.5587$_{\pm 0.0039}$	0.6788$_{\pm 0.0066}$
		PrLM	0.8130$_{\pm 0.0022}$	0.7767$_{\pm 0.0031}$	0.8293$_{\pm 0.0000}$	0.6989$_{\pm 0.0013}$	**0.7851**$_{\pm 0.0009}$	0.9976$_{\pm 0.0000}$	0.5111$_{\pm 0.0068}$	0.6455$_{\pm 0.0083}$
	Hybrid	TagPR	**0.8174**$_{\pm 0.0034}$	0.7792$_{\pm 0.0020}$	0.8096$_{\pm 0.0000}$	0.6626$_{\pm 0.0008}$	0.7749$_{\pm 0.0005}$	0.9964$_{\pm 0.0000}$	0.4847$_{\pm 0.0061}$	0.5555$_{\pm 0.0042}$
		SFT + GRPO	0.7732$_{\pm 0.0009}$	0.7742$_{\pm 0.0023}$	0.8140$_{\pm 0.0000}$	0.7394$_{\pm 0.0013}$	0.7661$_{\pm 0.0009}$	**0.9988**$_{\pm 0.0000}$	0.7753$_{\pm 0.0035}$	0.7420$_{\pm 0.0100}$
	User Coverage		66.87%	87.06%	55.26%	98.55%	95.53%	100.00%	99.03%	100.00%

Table 8: Detailed evaluation results of induced rubrics across three personalized text generation tasks on *Rubric-level Accuracy*.

	Amazon Review		LM-8B	LM-235B	LM-30B-IT	LM-30B-Think	GPT-4o-mini	PARL-0	**PARL-A**	**PARL-B**
Rubric-level Accuracy		**GT**	0.9447	**0.8825**	0.9341	0.9717	0.9680	**1.0000**	**0.9766**	**0.9768**
	ICL	Non	0.5971	0.3927	0.4468	0.8473	0.7758	**1.0000**	0.5652	0.8013
		RAG	0.6675	0.5011	0.5593	0.8722	0.7965	0.9998	0.6100	0.8289
		PAG	0.6776	0.5299	0.5532	0.8711	0.7806	0.9998	0.5410	0.7961
		Non-Think	0.6593	0.4422	0.4863	0.8555	0.7855	**1.0000**	0.4994	0.8086
		RAG-Think	0.7957	0.6061	0.6353	0.9086	0.8609	**1.0000**	0.6055	0.8628
		PAG-Think	0.7775	0.6085	0.6130	0.9076	0.8126	**1.0000**	0.5134	0.8115
		R2P	0.8158	0.6373	0.6981	0.9184	0.8761	**1.0000**	0.6402	0.8838
	SFT	SFT	0.8567	0.7687	0.8278	0.9617	0.9257	**1.0000**	0.9595	0.9533
		NextQuill	0.6958	0.6632	0.7285	0.9137	0.8116	**1.0000**	0.9574	0.9115
	RL	GRPO	**0.9485**	0.8780	0.9331	**0.9839**	**0.9826**	**1.0000**	0.9423	0.9736
		GSPO	0.9428	0.8719	**0.9352**	0.9823	0.9727	**1.0000**	0.9301	0.9747
		PrLM	0.9371	0.8667	0.9108	0.9819	0.9729	**1.0000**	0.9138	0.9652
	Hybrid	TagPR	0.9390	0.8606	0.9189	0.9798	0.9699	**1.0000**	0.9580	0.9728
		SFT + GRPO	0.9397	0.8620	0.9149	0.9809	0.9763	**1.0000**	0.9687	0.9752

	Reddit Topic		LM-8B	LM-235B	LM-30B-IT	LM-30B-Think	GPT-4o-mini	PARL-0	**PARL-A**	**PARL-B**
Rubric-level Accuracy		**GT**	0.9180	**0.8942**	0.8806	0.9313	0.9324	0.9933	**0.9543**	**0.9559**
	ICL	Non	0.4501	0.4743	0.4265	0.6828	0.4975	0.9965	0.6024	0.3508
		RAG	0.6546	0.6427	0.5557	0.8175	0.7373	0.9980	0.6198	0.4193
		PAG	0.6871	0.6721	0.5471	0.8476	0.7589	**0.9984**	0.5471	0.3567
		Non-Think	0.5273	0.5006	0.4166	0.7094	0.6042	0.9374	0.5578	0.3292
		RAG-Think	0.7501	0.7012	0.5840	0.8455	0.8090	0.9773	0.6184	0.4418
		PAG-Think	0.7392	0.7107	0.5557	0.8496	0.8007	0.9720	0.5827	0.4298
		R2P	0.8293	0.7557	0.7422	0.8576	0.8595	0.9935	0.8058	0.6748
	SFT	SFT	0.8507	0.8458	0.8566	0.9035	0.8575	0.9957	0.9306	0.9154
		NextQuill	0.7491	0.7557	0.8160	0.8561	0.7481	0.9973	0.9257	0.9141
	RL	GRPO	0.9231	0.8657	0.8769	0.9458	0.9418	0.9977	0.7178	0.6699
		GSPO	0.9204	0.8696	0.8609	0.9427	0.9411	0.9993	0.7848	0.7358
		PrLM	0.8320	0.7409	0.7735	0.9032	0.8362	0.9959	0.7381	0.6301
	Hybrid	TagPR	**0.9292**	0.8702	0.8517	**0.9553**	**0.9488**	0.9982	0.7684	0.7997
		SFT + GRPO	0.9238	0.8920	**0.8948**	0.9525	0.9400	0.9959	0.9356	0.9396

	News Headline		LM-8B	LM-235B	LM-30B-IT	LM-30B-Think	GPT-4o-mini	PARL-0	**PARL-A**	**PARL-B**
Rubric-level Accuracy		**GT**	0.8345	0.8679	0.8725	0.9397	0.8821	0.9998	**0.9461**	**0.9520**
	ICL	Non	0.6678	0.7525	0.7592	0.9029	0.7748	0.9998	0.6510	0.7824
		RAG	0.7743	0.8279	0.8629	0.9257	0.8383	0.9998	0.7341	0.8652
		PAG	0.8299	0.8697	0.8856	0.9105	0.8645	0.9996	0.8049	0.9120
		Non-Think	0.6875	0.7519	0.7580	0.8907	0.7758	0.9996	0.5555	0.7765
		RAG-Think	0.7569	0.8314	0.8367	0.9007	0.8343	0.9998	0.6945	0.8742
		PAG-Think	0.8206	0.8422	0.8427	0.9091	0.8348	0.9998	0.7665	0.9154
		R2P	0.8264	0.8542	0.8820	0.9445	0.8715	0.9996	0.7181	0.9023
	SFT	SFT	0.8542	**0.8840**	0.8915	**0.9487**	0.8640	**0.9999**	0.9256	0.9294
		NextQuill	0.8345	0.8703	0.8951	0.9469	0.8730	**0.9999**	0.9256	0.9362
	RL	GRPO	0.8426	0.8631	0.8987	0.9187	0.8660	0.9998	0.7575	0.8821
		GSPO	**0.8657**	0.8781	**0.9023**	0.9453	0.8806	0.9996	0.7861	0.9138
		PrLM	0.8553	0.8781	0.8939	0.9379	0.8892	0.9998	0.7580	0.8995
	Hybrid	TagPR	0.8611	0.8804	0.8808	0.9247	**0.8917**	0.9996	0.7142	0.8568
		SFT + GRPO	0.8206	0.8769	0.8808	0.9457	0.8710	**0.9999**	0.9152	0.9357



[Item Title]: Justice League (BD)
[Review Rating]: 5.0
[Review Title]: great movie for the collection and comes with digital code.

✓	[R1] Uses casual, conversational language with frequent use of contractions (e.g., 'it's', 'don't', 'I'll') and informal expressions (e.g., 'rocked the roll', 'cool as heck', 'whaalllah').	✗
✓	[R2] Writes in short to medium-length sentences with a tendency toward loose, run-on structures connected by conjunctions or commas rather than formal punctuation.	✗
✓	[R3] Tends to evaluate movies based on entertainment value, pacing, and family appropriateness rather than deep analysis of plot, direction, or cinematography.	✓

Authentic User-Authored GT	AI Model-Generated Output
anything with Superman is a favorite of mine. I thought this was done well, of course, there's a bunch of jerks who say differently. I'm from an era of black and white tv's with dials on them so it impressed me. A lot of action and a good story. Pick this one up and the thirteen dollar price it's a steal.	Great movie for the collection and comes with digital code. As someone who enjoys superhero films, I really enjoyed this Blu-ray release of Justice League. The story is engaging, with a strong cast that brings their characters to life. The action sequences are well-choreographed and visually stunning, making for an immersive experience. I appreciate the inclusion of the digital code, which adds value to the purchase. This is a must-have for any DC fan's collection.

Figure 5: Qualitative analysis of induced rubrics on the *Amazon Review* task. R1, R2, and R3 denote the induced rubrics for the target user. We evaluate whether the authentic user-authored GT and the AI model-generated output satisfy these rubrics.

Table 9: Experimental results of measured methods on six other metrics: ROUGE-1 (R-1), ROUGE-L (R-L), BLEU (BL.), METEOR (ME.), BERTScore (BS.), and EmbScore (EmbS.).

Category	Method	Amazon Review			Reddit Topic			News Headline		
		R-1	R-L	BL.	R-1	R-L	BL.	R-1	R-L	BL.
ICL	Non	0.2381	0.1305	1.2345	0.1951	0.1044	0.6381	0.1413	0.1247	3.0168
	RAG	0.2731	0.1426	2.0540	0.2671	0.1298	1.8512	0.1563	0.1413	4.1086
	PAG	0.2723	0.1400	2.0441	0.2737	0.1267	1.4741	0.1441	0.1280	3.5684
	Non-Think	0.2227	0.1220	1.0618	0.1928	0.1003	0.6875	0.1367	0.1186	2.7902
	RAG-Think	0.2579	0.1358	1.7980	0.2339	0.1122	1.0177	0.1361	0.1205	3.2062
	PAG-Think	0.2534	0.1319	1.5694	0.2405	0.1126	0.9469	0.1398	0.1237	3.0538
	R2P	0.2734	0.1459	2.6768	0.2375	0.1189	1.1350	0.1630	0.1493	4.6484
SFT	SFT	0.3473	0.1724	7.1323	0.3257	0.1667	4.9135	0.2023	0.1869	6.4891
	NextQuill	0.3241	**0.1831**	**9.1712**	0.2918	0.1658	4.7346	**0.2150**	**0.2005**	6.8295
RL	GRPO	0.3735	0.1740	8.6424	0.3429	0.1610	4.0917	0.1877	0.1686	5.2834
	GSPO	0.3700	0.1722	8.4999	0.3443	0.1485	2.9119	0.1863	0.1714	5.6436
	PrLM	0.3654	0.1688	8.1936	0.3256	0.1486	2.9488	0.1914	0.1714	5.5231
Hybrid	TagPR	0.3760	0.1724	8.6316	0.3424	0.1536	3.7206	0.1917	0.1735	5.7818
	SFT+GRPO	**0.3829**	0.1775	8.9420	**0.3648**	**0.1739**	**5.2343**	0.2120	0.1974	**6.9968**

Category	Method	Amazon Review			Reddit Topic			News Headline		
		ME.	BS.	EmbS.	ME.	BS.	EmbS.	ME.	BS.	EmbS.
ICL	Non	0.1361	0.4717	0.3144	0.1071	0.4364	0.4873	0.1185	0.5166	0.6564
	RAG	0.1719	0.4785	0.3229	0.1820	0.4578	0.4345	0.1319	0.5281	0.6640
	PAG	0.1816	0.4740	0.3130	0.2016	0.4517	0.4231	0.1292	0.5172	0.6537
	Non-Think	0.1295	0.4596	0.2973	0.1129	0.4341	0.4867	0.1217	0.5110	0.6546
	RAG-Think	0.1623	0.4695	0.3247	0.1564	0.4431	0.4498	0.1212	0.5154	0.6540
	PAG-Think	0.1665	0.4630	0.3079	0.1699	0.4398	0.4385	0.1296	0.5075	0.6490
	R2P	0.1786	0.4856	0.3289	0.1455	0.4558	0.5115	0.1362	0.5324	0.6664
SFT	SFT	0.2691	0.5369	0.4372	0.2250	0.5064	0.5756	0.1700	0.5647	0.6816
	NextQuill	0.2689	**0.5531**	**0.4443**	0.2000	0.4998	0.5731	0.1709	0.5694	**0.6909**
RL	GRPO	0.3476	0.5451	0.4035	0.2773	0.4959	0.5078	0.1600	0.5364	0.6792
	GSPO	0.3451	0.5418	0.3956	0.2743	0.4836	0.5664	0.1518	0.5395	0.6765
	PrLM	0.3469	0.5364	0.4004	0.2750	0.4698	0.4397	0.1620	0.5474	0.6893
Hybrid	TagPR	0.3505	0.5410	0.3907	0.2748	0.4816	0.5578	0.1709	0.5499	0.6847
	SFT+GRPO	**0.3523**	0.5491	0.4277	**0.2920**	**0.5166**	**0.6016**	**0.1749**	**0.5722**	0.6887

Table 10: The prompt used to induce personalized user rubrics for the *Amazon Review* dataset.

Rubrics Induction - *Amazon Review*

[Instruction]
Role
You are an expert in Personalization Analysis and Rubric Generation. Your job is to generate a self-contained set of personalized evaluation rubrics based on a user's historical reviews. These rubrics should be used to evaluate whether a newly generated review text matches the user's preferences.

Input Data
- **User History:** A collection of historical review texts written by the target user.

Task
Analyze the user's historical reviews from multiple dimensions (e.g., writing style, linguistic features, structure, tone, content focus, expression patterns) and generate 10 highly personalized rubrics that cover different aspects of the user's preferences.

Drafting Rules
- **Atomicity:** Each rubric should be a single, clear and self-contained criterion that can be used to evaluate a review text independently. It should not require context from other rubrics or external information to be meaningful. Avoid combining multiple criteria into one rubric, as this makes evaluation ambiguous and reduces clarity.
- **Consistency**: Each rubric should reflect a consistent pattern identified across all the user's historical reviews. Each rubric must be satisfied when applied to all reviews in the user's profile. Avoid rubrics that only match a subset of the user's historical reviews, as they do not represent consistent patterns across the user's preferences.
- **Determinism:** Each rubric should be unambiguous and objectively evaluable, allowing for clear yes/no judgments rather than vague or subjective assessments. When applied to a review text, evaluators should be able to determine whether the rubric is satisfied without relying on personal interpretation or probabilistic reasoning.
- **Specificity:** Each rubric should be specific to a concrete, observable feature. Instead of vague statements like "matches writing style", provide precise characteristics such as "uses short, fragmented sentences without punctuation" or "prefers active voice over passive voice".
- **Item-Agnosticism:** Each rubric should be applicable to *any* future review of *any* item, not just the specific items in the user's history. Focus on *how* the user writes reviews (their writing style, structure, tone, linguistic patterns) rather than *what* items they reviewed (specific items, events, or subject matter). The rubrics should capture transferable writing characteristics that persist across different review contexts.
- **Multi-Dimensionality:** The set of rubrics should collectively cover multiple dimensions of the user's writing characteristics, such as writing style (sentence structure, paragraph organization), linguistic features (vocabulary choice, grammar patterns), tone (formality, emotional expression), content focus (what aspects are emphasized), and expression patterns (how ideas are conveyed). Avoid generating rubrics that all focus on the same dimension; instead, ensure diversity across different aspects of the user's preferences.
- **Personalization:** Each rubric should be personalized to the target user's historical reviews, capturing their unique writing characteristics and preferences. Each rubric should reflect something distinctive about this specific user's writing style that distinguishes them from other users. Generic rubrics that could apply to any user should be avoided; the rubrics must be tailored to reveal this particular user's preferences.

Output Format
Example:
[
    {"rule": "The review uses long, complex sentences with multiple clauses and sophisticated vocabulary, avoiding simplistic or fragmented phrasing."},
    {"rule": "The review maintains a formal, essay-like tone with structured paragraphs, each focusing on a distinct analytical point or thematic thread."},
    {"rule": "The review is structured with clearly labeled sections using uppercase headings followed by colons (e.g., 'THE STORY:', 'THE COOL THINGS:', 'BEST SCENES:', 'THE VERDICT:', 'GRADES:')."},
    {"rule": "The structure follows a discursive, essay-like format with clear thematic paragraphs, each advancing a distinct analytical point without rigid section headings."},
    {"rule": "The language includes informal but articulate expressions of personal judgment (e.g., 'wears thin VERY quickly', 'cool stuff!!!', 'decent as always') that convey strong subjective opinion without overly casual slang."},
    {"rule": "The review employs a formal yet conversational tone with complex sentence structures, minimal contractions, and careful attention to grammar and punctuation."},
    {"rule": "The review must contain a paragraph focused on the lead and supporting cast, highlighting the actors' performances with specific character names and roles, and often comparing the actor to their other well-known parts or assessing how they embody their character's moral or psychological complexity."},
    {"rule": "The review explicitly mentions the source of the complimentary book copy (e.g., 'I received a complimentary copy from...') and specifies the publisher and review context (e.g., blog tour, NetGalley)."},
    {"rule": "The review ends with a clear recommendation statement using the phrase 'I will recommend this book to others' or a close variant (e.g., 'will certainly recommend it')."},
    {"rule": "The review concludes with a reflective, often uplifting or contemplative statement that underscores the lasting impact or significance of the work."},
]

Provide a JSON array of objects. Each object must contain exactly one key: `rule`.
Do not output markdown code blocks or extra text.

[Inputs]
Given the following user historical reviews, please generate a self-contained set of personalized evaluation rubrics for the user.
<User Historical Reviews>
{historical_reviews}
</User Historical Reviews>

Table 11: The prompt used to induce personalized user rubrics for the *Reddit Post* dataset.

Rubrics Induction - *Reddit Post*

[Instruction]
Role
You are an expert in Personalization Analysis and Rubric Generation. Your job is to generate a self-contained set of personalized evaluation rubrics based on a user's historical reddit posts. These rubrics should be used to evaluate whether a newly generated reddit post content matches the user's preferences.

Input Data
- **User History:** A collection of historical reddit post contents written by the target user.

Task
Analyze the user's historical reddit posts from multiple dimensions (e.g., writing style, linguistic features, structure, tone, content focus, expression patterns) and generate 10 highly personalized rubrics that cover different aspects of the user's preferences.

Drafting Rules
- **Atomicity:** Each rubric should be a single, clear and self-contained criterion that can be used to evaluate a reddit post content independently. It should not require context from other rubrics or external information to be meaningful. Avoid combining multiple criteria into one rubric, as this makes evaluation ambiguous and reduces clarity.
- **Consistency**: Each rubric should reflect a consistent pattern identified across all the user's historical reddit posts. Each rubric must be satisfied when applied to all reddit post contents in the user's profile. Avoid rubrics that only match a subset of the user's historical reddit posts, as they do not represent consistent patterns across the user's preferences.
- **Determinism:** Each rubric should be unambiguous and objectively evaluable, allowing for clear yes/no judgments rather than vague or subjective assessments. When applied to a reddit post content, evaluators should be able to determine whether the rubric is satisfied without relying on personal interpretation or probabilistic reasoning.
- **Specificity:** Each rubric should be specific to a concrete, observable feature. Instead of vague statements like "matches writing style", provide precise characteristics such as "uses short, fragmented sentences without punctuation" or "prefers active voice over passive voice".
- **Topic-Agnosticism:** Each rubric should be applicable to *any* future post of *any* topic, not just the specific posts in the user's history. Focus on *how* the user writes posts (their writing style, structure, tone, linguistic patterns) rather than *what* topics they posted (specific topics, events, or subject matter). The rubrics should capture transferable writing characteristics that persist across different post contexts.
- **Multi-Dimensionality:** The set of rubrics should collectively cover multiple dimensions of the user's writing characteristics, such as writing style (sentence structure, paragraph organization), linguistic features (vocabulary choice, grammar patterns), tone (formality, emotional expression), content focus (what aspects are emphasized), and expression patterns (how ideas are conveyed). Avoid generating rubrics that all focus on the same dimension; instead, ensure diversity across different aspects of the user's preferences.
- **Personalization:** Each rubric should be personalized to the target user's historical reddit post contents, capturing their unique writing characteristics and preferences. Each rubric should reflect something distinctive about this specific user's writing style that distinguishes them from other users. Generic rubrics that could apply to any user should be avoided; the rubrics must be tailored to reveal this particular user's preferences.

Output Format
Provide a JSON array of objects. Each object must contain exactly one key: `rule`.
Example:
[
{"rule": "The post begins with a personal disclosure or contextual backstory that establishes the user's current emotional state, situation, or relevant history before introducing the main topic."},
{"rule": "The writing uses lowercase letters at the beginning of sentences inconsistently, often starting sentences with lowercase 'i' for first-person pronouns and occasionally neglecting standard capitalization at the beginning of sentences."},
{"rule": "The sentence structure must favor complete, grammatically standard sentences with consistent punctuation, avoiding fragmented prose, bullet points, or meta-commentary within the story itself."},
{"rule": "The post structures thoughts in a stream-of-consciousness style with minimal paragraph breaks, allowing one idea to flow into the next without formal transitions."},
{"rule": "The post must present biomechanical or anatomical reasoning to justify training recommendations, explicitly connecting muscle structure or joint mechanics to exercise selection or technique."},
{"rule": "The post balances emotional content with analytical reasoning, expressing personal feelings without becoming overly sentimental or dramatic."},
{"rule": "Writes in a direct, no-nonsense tone with minimal embellishment, prioritizing clarity and factual accuracy over emotional expression or storytelling."},
{"rule": "The post uses a conversational and informal tone with frequent use of first-person perspective (e.g., 'I', 'my', 'me') to express personal experiences, opinions, or questions."},
{"rule": "Vocabulary is plainspoken and accessible, favoring common idioms and colloquial expressions over technical jargon or literary flourishes, even when discussing sensitive or complex topics."},
{"rule": "The post concludes with a minimal or absent summary statement, instead ending directly on a question, emotional plea, or unresolved tension, leaving resolution open to community input."},
]

Do not output markdown code blocks or extra text.

[Inputs]
Given the following user historical reddit posts, please generate a self-contained set of personalized evaluation rubrics for the user.
<User Historical Reddit Posts>
{historical_reddit_posts}
</User Historical Reddit Posts>

Table 12: The prompt used to induce personalized user rubrics for the *News Headline* dataset.

Rubrics Induction - *News Headline*

[Instruction]
Role
You are an expert in Personalization Analysis and Rubric Generation. Your job is to generate a self-contained set of personalized evaluation rubrics based on a user's historical news. These rubrics should be used to evaluate whether a newly generated news headline matches the user's preferences.

Input Data
- **User History:** A collection of historical news contents and their corresponding news headlines written by the target user.

Task
Analyze how user transforms news content into news headline and generate 10 highly personalized rubrics that cover different aspects of the user's preferences.

Drafting Rules
- **Atomicity:** Each rubric should be a single, clear and self-contained criterion that can be used to evaluate a news headline independently. It should not require context from other rubrics or external information to be meaningful. Avoid combining multiple criteria into one rubric, as this makes evaluation ambiguous and reduces clarity.
- **Consistency**: Each rubric should reflect a consistent pattern identified across all the user's historical news headlines. Each rubric must be satisfied when applied to all news headlines in the user's profile. Avoid rubrics that only match a subset of the user's historical news headlines, as they do not represent consistent patterns across the user's preferences.
- **Determinism:** Each rubric should be unambiguous and objectively evaluable, allowing for clear yes/no judgments rather than vague or subjective assessments. When applied to a news headline, evaluators should be able to determine whether the rubric is satisfied without relying on personal interpretation or probabilistic reasoning.
- **Specificity:** Each rubric should be specific to a concrete, observable feature. Instead of vague statements like "matches writing style", provide precise characteristics such as "uses short, fragmented sentences without punctuation" or "prefers active voice over passive voice".
- **News-Agnosticism:** Each rubric should be applicable to *any* future headline of *any* news, not just the specific news in the user's history. Focus on *how* the user writes news headline (their writing style, structure, tone, linguistic patterns) rather than *what* items they reviewed (specific items, events, or subject matter). The rubrics should capture transferable writing characteristics that persist across different news contexts.
- **Multi-Dimensionality:** The set of rubrics should collectively cover multiple dimensions of the user's writing characteristics, such as writing style (sentence structure, paragraph organization), linguistic features (vocabulary choice, grammar patterns), tone (formality, emotional expression), content focus (what aspects are emphasized), and expression patterns (how ideas are conveyed). Avoid generating rubrics that all focus on the same dimension; instead, ensure diversity across different aspects of the user's preferences.
- **Personalization:** Each rubric should be personalized to the target user's historical news, capturing their unique writing characteristics and preferences. Each rubric should reflect something distinctive about this specific user's writing style that distinguishes them from other users. Generic rubrics that could apply to any user should be avoided; the rubrics must be tailored to reveal this particular user's preferences.

Output Format
Provide a JSON array of objects. Each object must contain exactly one key: `rule`.
Do not output markdown code blocks or extra text.

[Inputs]
Given the following user historical news, please generate a self-contained set of personalized evaluation rubrics for the user.
<User Historical News>
{historical_news}
</User Historical News>

Table 13: The prompt used to evaluate the model-generated outputs and user-authored ground truth using personalized user rubrics on *Amazon Review* dataset.

Rubrics Evaluation - *Amazon Review*

[Instruction]
Role
You are an expert Personalization Evaluator. Your job is to evaluate the quality of a review text based on a personalized evaluation rubric for the user. Given the metadata of the item and the review to be evaluated, the review text, and a personalized evaluation rubric, please judge whether the review text satisfies the given personalized evaluation rubric.

Input Data
- **Review Metadata:** The metadata of the item and the review to be evaluated.
- **Review Text:** The review text to be evaluated.
- **Personalized Evaluation Rubric:** The personalized evaluation rubric for the user to evaluate the review text.

Task
Judge whether the review text satisfies the given personalized evaluation rubric.

Output Format
Only output the answer of whether the review text satisfies the given personalized evaluation rubric. The answer should be `yes` or `no`.

[Inputs]
<Review Metadata>
{metadata}
</Review Metadata>

<Review Text>
{response}
</Review Text>

<Personalized Evaluation Rubric>
{rubric}
</Personalized Evaluation Rubric>

Table 14: The prompt used to evaluate the model-generated outputs and user-authored ground truth using personalized user rubrics on *Reddit Post* dataset.

Rubrics Evaluation - *Reddit Post*

[Instruction]
Role
You are an expert Personalization Evaluator. Your job is to evaluate the quality of a reddit post content based on a personalized evaluation rubric for the user. Given the reddit post summary, the reddit post content to be evaluated, and a personalized evaluation rubric, please judge whether the reddit post content satisfies the given personalized evaluation rubric.

Input Data
- **Reddit Post Summary:** The summary of the reddit post.
- **Reddit Post Content:** The content of the reddit post to be evaluated.
- **Personalized Evaluation Rubric:** The personalized evaluation rubric for the user to evaluate the reddit post content.

Task
Judge whether the reddit post content satisfies the given personalized evaluation rubric.

Output Format
Only output the answer of whether the reddit post content satisfies the given personalized evaluation rubric. The answer should be `yes` or `no`.

[Inputs]
<Reddit Post Summary>
{summary}
</Reddit Post Summary>

<Reddit Post Content>
{content}
</Reddit Post Content>

<Personalized Evaluation Rubric>
{rubric}
</Personalized Evaluation Rubric>

Table 15: The prompt used to evaluate the model-generated outputs and user-authored ground truth using personalized user rubrics on *News Headline* dataset.

Rubrics Evaluation - *News Headline*

[Instruction]
Role
You are an expert Personalization Evaluator. Your job is to evaluate the quality of a news headline based on a personalized evaluation rubric for the user. Given the news content, the news headline to be evaluated, and a personalized evaluation rubric, please judge whether the news headline satisfies the given personalized evaluation rubric.

Input Data
- **News Content:** The content of the news.
- **News Headline:** The news headline to be evaluated.
- **Personalized Evaluation Rubric:** The personalized evaluation rubric for the user to evaluate the news headline.

Task
Judge whether the news headline satisfies the given personalized evaluation rubric.

Output Format
Only output the answer of whether the news headline satisfies the given personalized evaluation rubric. The answer should be `yes` or `no`.

[Inputs]
<News Content>
{content}
</News Content>

<News Headline>
{response}
</News Headline>

<Personalized Evaluation Rubric>
{rubric}
</Personalized Evaluation Rubric>

Table 16: The prompt used to evaluate the model-generated outputs and user-authored ground truth using LLM-as-a-judge on *Amazon Review* dataset.

LLM-as-a-judge Evaluation - *Amazon Review*

[Instruction]
Role
You are an expert Personalization Evaluator. Your job is to evaluate the quality of a review text based on the user histories. Given the user historical reviews, the metadata of the item and the review to be evaluated, the review text, please judge how well the review text matches the user's preferences.

Input Data
- **User History:** A collection of historical review texts written by the target user.
- **Review Metadata:** The metadata of the item and the review to be evaluated.
- **Review Text:** The review text to be evaluated.

Output Format
Only output the score of the review text. The score should be a number between 1 and 5.
Do not output markdown code blocks or extra text.

[Inputs]
<User Historical Reviews>
{past_reviews}
</User Historical Reviews>

<Review Metadata>
{meta_info}
</Review Metadata>

<Review Text>
{review}
</Review Text>

Table 17: The prompt used to evaluate the model-generated outputs and user-authored ground truth using LLM-as-a-judge on *Reddit Post* dataset.

LLM-as-a-judge Evaluation - *Reddit Post*

[Instruction]
Role
You are an expert Personalization Evaluator. Your job is to evaluate the quality of a reddit content based on the user histories. Given the user historical reddit post contents, the reddit post summary, and the reddit post content to be evaluated, please judge how well the reddit content matches the user's preferences.

Input Data
- **User History:** A collection of historical reddit post contents written by the target user.
- **Reddit Post Summary:** The summary of the reddit post.
- **Reddit Post Content:** The content of the reddit post to be evaluated.

Output Format
Only output the score of the reddit content. The score should be a number between 1 and 5.
Do not output markdown code blocks or extra text.

[Inputs]
<User Historical Reddit Posts>
{past_posts}
</User Historical Reddit Posts>

<Reddit Post Summary>
{summary}
</Reddit Post Summary>

<Reddit Post Content>
{content}
</Reddit Post Content>

Table 18: The prompt used to evaluate the model-generated outputs and user-authored ground truth using LLM-as-a-judge on *News Headline* dataset.

LLM-as-a-judge Evaluation - *News Headline*

[Instruction]
Role
You are an expert Personalization Evaluator. Your job is to evaluate the quality of a news headline based on the user histories. Given the user historical news, the news content, and the news headline to be evaluated, please judge how well the news headline matches the user's preferences.

Input Data
- **User History:** A collection of historical news contents and their corresponding news headlines written by the target user.
- **News Content:** The content of the news.
- **News Headline:** The news headline to be evaluated.

Output Format
Only output the score of the news headline. The score should be a number between 1 and 5.
Do not output markdown code blocks or extra text.

[Inputs]
<User Historical News>
{past_news}
</User Historical News>

<News Content>
{content}
</News Content>

<News Headline>
{headline}
</News Headline>