PersRM-R1: Enhance Personalized Reward Modeling with Reinforcement Learning

Anonymous ACL submission

Abstract

Personalizing large language models (LLMs) to individual user preferences is essential in real-world application scenarios. Reward models (RMs), which are central to existing post-training methods, aim to align LLM outputs with human preference by providing feedback signals during fine-tuning. However, existing RMs struggle to capture nuanced, user-specific preferences, especially under limited data and across diverse domains. Thus, we introduce PersRM-R1, the first reasoning-based reward modeling framework specifically designed to identify and represent personal factors from only one or a few personal exemplars. To address challenges including limited data availability and the requirement for robust generalization, our approach combines synthetic data generation with a two-stage training pipeline consisting of supervised fine-tuning followed by reinforcement fine-tuning. Experimental results demonstrate that PersRM-R1 outperforms existing models of similar size and matches the performance of much larger models in both accuracy and generalizability, paving the way for more effective personalized LLMs.

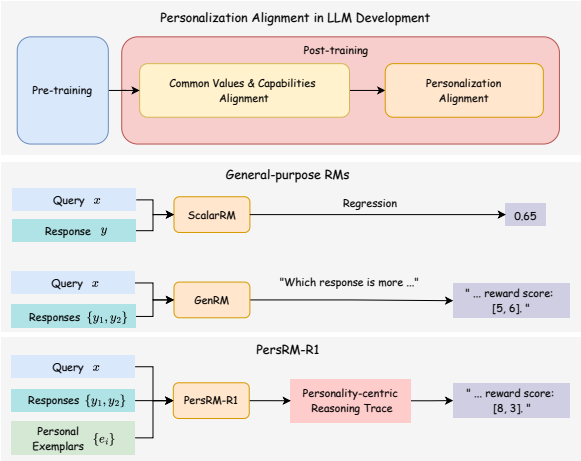

Figure 1: **Top**: Post-training for user-facing LLMs involves two key stages: 1) aligning the LLM with shared human values and endowing it with fundamental capabilities; 2) tailoring the LLM to individual preferences. **Middle**: General-purpose RMs, including both scalar and generative types, are trained on standard tasks to generate reward scores for given responses. **Bottom**: PersRM-R1 is capable of capturing subtle personality traits presented in personal exemplars, conducting personality-centric reasoning analyses, and generating reward scores that reflect how closely each response aligns with the style exhibited in the personal exemplars.

1 Introduction

The paradigm of pre-training followed by post-training has been widely adopted in both academia and industry for developing large language models (LLMs). In the post-training stage, common values, such as harmlessness, helpfulness, and honesty, that are shared among human beings, and common capabilities, such as chatting, reasoning, which are shared among various tasks, have been the optimization objective in model fine-tuning (Bai et al., 2022a,b; Ouyang et al., 2022). However, as LLMs are increasingly integrated into personalized applications, e.g., personal assistants, tutoring systems, and writing aids, there is a growing need for models that can not only follow specific user instructions but also align closely with individual preferences and communication styles. These differences, such as preference for conciseness, use of humorous expressions, and opinions on controversial topics, distinguish one person from another. Thus, *personalization alignment* that makes the user-facing LLM best fit for each individual is essential; however, it has not received much attention in the community(cf. Fig. 1, top panel).

Reinforcement learning (RL) methods based on reward models (RMs) are commonly applied in LLM post-training, where a reward model is trained to provide reward signals for the training of the policy model through reinforcement learning algorithms (Christiano et al., 2017; Ouyang et al.,

2022). The reliability of RMs lies at the core of the success of such methods (Li et al., 2023; Casper et al., 2023). RMs for common values and capabilities alignment are trained on large-scale crowd-sourcing data to represent homogeneous preferences for general purposes (cf. Fig. 1, middle panel). In contrast, RMs for individuals ought to be personalized, as studied in the literature of neuroscience, where the emission of rewarding signals, like dopamine, is influenced by the value component that reveals the brain's *subjective evaluation* of the effects of goal achievements (Schultz, 2015).

Developing such models, however, is non-trivial, particularly in settings where only limited user-specific data is available. In this work, we investigate how to construct and train personalized reward models under realistic data constraints. We focus on two central technical challenges: (a) the *scarcity of individual-specific demonstrations* for effective model tuning, and (b) the base model's *insufficient sensitivity to nuanced personality traits* reflected in both exemplars and generated responses.

We introduce a novel approach to address these challenges, ultimately resulting in our RMs, namely PersRM-R1. The overall development pipeline of our method is illustrated in Fig. 3. Specifically, to tackle the first challenge, we design a synthetic data generation pipeline to enhance pairwise preference data (cf. Fig. 2, left panel). For the second challenge, we propose to incorporate a *personality-centric reasoning process* into reward modeling. This is motivated by our hypothesis that personality traits are inherently embedded within the text and must be explicitly extracted and analyzed to more accurately assess personality similarity between a personal exemplar and an arbitrary response. To this end, we generate synthetic reasoning traces for pairwise preference data (cf. Fig. 2, right panel), which are used to tune the base model in supervised fine-tuning (SFT) to enhance its foundational capability to reason about personality traits and to produce reward scores in a standardized format. Subsequently, reinforcement fine-tuning (RFT) is employed to further enhance its performance and generalizability, building on recent advances in reasoning-augmented LLMs (DeepSeek-AI et al., 2025; Liu et al., 2025).

The performance of PersRM-R1 is evaluated on the representative personalization alignment task of personal stylish writing, across diverse genres including email, essays, news articles, blogs, Tweet and more. Experimental results show that PersRM-R1 not only outperforms existing models of comparable size but also matches the accuracy and generalizability of much larger models. To the best of our knowledge, this work presents the first reasoning-based reward model tailored for personalization. We propose PersRM-R1, which integrates guided data augmentation with a two-stage fine-tuning pipeline to enable preference modeling from minimal user input. Beyond establishing a new direction for personalized alignment, our study systematically investigates alternative approaches such as in-context learning and conducts ablations to isolate the contributions of each pipeline component, providing a comprehensive foundation for future work in personalized reward modeling.

2 Related Work

LLM Personalization. Existing methods for personalizing LLMs can be categorized into two classes: 1) tuning-free methods, such as retrieval-augmented generation (RAG) (Salemi et al., 2024), prompt engineering (Park et al., 2023), and steering vector intervention (Konen et al., 2024; Cao et al., 2024); 2) tuning-based methods, which influence the manifested personality traits of LLMs through supervised fine-tuning (Shao et al., 2023), direct preference optimization (Li et al., 2024; Zeng et al., 2024), and RM-based approaches (Chen et al., 2025a). Tuning-free methods, while simple to implement, often induce extra token costs and sacrifice inference speed. Moreover, their performance can be inconsistent and sensitive to retrieved information, prompts, or steering vectors. In contrast, tuning-based methods that directly integrate personality traits into the model parameters offer a more robust and fundamental solution. A concurrent work by Chen et al. (2025a) focuses on developing personalized RMs by efficiently tuning parameters of the base model for each specific traits, a strategy that becomes impractical when considering the vast number of potential users. In contrast, our work focuses on a distinct path to personalized reward modeling, where we develop a unified RM that captures personal traits from individual exemplars and arbitrary responses, producing reward scores that reflect personality similarity.

Reasoning-enhanced Reward Modeling. RMs in current literature can be categorized into two classes: scalar and generative RMs (Lambert et al., 2024; Liu et al., 2025). Scalar RMs are trained to predict a scalar reward value given a pair of prompt

and response. In contrast, generative RMs offer a more general formulation for reward modeling that exhibits strong generalization potential. In addition, reasoning capability can be naturally incorporated into the rewarding process (Yang et al., 2024a; Chen et al., 2025b; Liu et al., 2025; Guo et al., 2025), which has attracted increasing attention recently, motivated by the success of incorporating long-reasoning process in LLMs in tasks of coding and math problem solving (DeepSeek-AI et al., 2025; OpenAI et al., 2024). Most of the existing work in this direction focus on tasks where ground-truth rule-based rewards can be obtained. In contrast, employing reinforcement learning in open domains is more challenging as reliable rewards are absent. While previous work (Liu et al., 2025) employs self-generated principles to generate rewards aimed at common value alignment, we demonstrate that, with appropriate modifications to address challenges such as identifying nuanced preference characteristics and coping with limited data, similar strategies can be adapted for personalized reward generation.

3 Methodology

In this section, we introduce the three key phases in developing personalization reward models: data curation (cf. Sec. 3.1), supervised fine-tuning (SFT, cf. Sec. 3.2), and reinforcement fine-tuning (RFT, cf. Sec. 3.3). Each subsection herein addresses distinct challenges introduced by RM in the context of personalization alignment, while preliminaries on SFT and RFT are provided in Appendix A. An overview of the full training pipeline is shown in Figure 3.

3.1 Personalization Data Augmentation

To address the challenge of lacking user-specific preference data for reward modeling, we employ data augmentation techniques using LLMs to generate a synthetic dataset conditioned on a limited user corpus $\mathcal{D}_{\text{expl}} = \{e_i\}$. First, LLMs are prompted to produce both aligned and divergent responses conditioned on user exemplars. Second, deeper reasoning is elicited to re-evaluate the auto-labeled responses generated in the first stage, enabling more faithful scoring. Together, these two stages distill the LLMs' understanding of user-specific preferences, producing high-quality synthetic data sufficient for fine-tuning the reward model.

Pairwise Preference Data Construction. Given a query/problem x, we prompt an LLM to generate content that is close (y^+), or divergent (y^-) to the user's personality traits exhibited in the exemplars/context e, respectively, resulting in a synthetic collection $\mathcal{D}_{\text{syn}} = \{x, e, y^+, y^-\}$.

- To generate positive samples y^+ that preserve user-specific stylistic characteristics, we employ two strategies. The first, *intra-author retrieval*, involves selecting alternative responses authored by the same user for different queries, thereby ensuring authentic and consistent personalized style. The second, *lexical perturbation*, creates controlled variants of demonstrations through (up to six) synonym substitutions while maintaining the original grammatical structure and sentence order.

- Negative samples y^- are obtained via three complementary approaches. *Cross-author retrieval* sources responses from different users, introducing clear stylistic and preference divergences. *Random sampling* employs large language models to generate loosely related or off-topic responses, serving as weak negative examples. Finally, *confounding sampling* leverages LLMs to generate responses that approximate the stylistic features present in the user's exemplars. While the current LLMs lack the ability to faithfully capture personalized styles, a gap this work seeks to address, these samples nonetheless serve as strong adversarial negatives, encouraging the reward model to attend to fine-grained stylistic cues.

This comprehensive sampling framework spans a continuum of difficulty levels, which is essential for developing robust and generalizable reward models (Li et al., 2025).

Reasoning Trace Generation. Reward models augmented with explicit reasoning traces have demonstrated superior capabilities in delivering fine-grained and reliable preference assessments (Li et al., 2025; Lee et al., 2024). To further enhance alignment with user-specific preferences, we elicit comparative reasoning from LLMs to assess the stylistic match or mismatch between generated responses.

Specifically, given a quadruple (x, e, y^+, y^-) from the synthetic dataset $\mathcal{D}_{\text{syn}}$, we prompt the LLM to evaluate the similarity between each response (y^+ and y^-) and the user exemplars e respectively. The model is constrained to output a

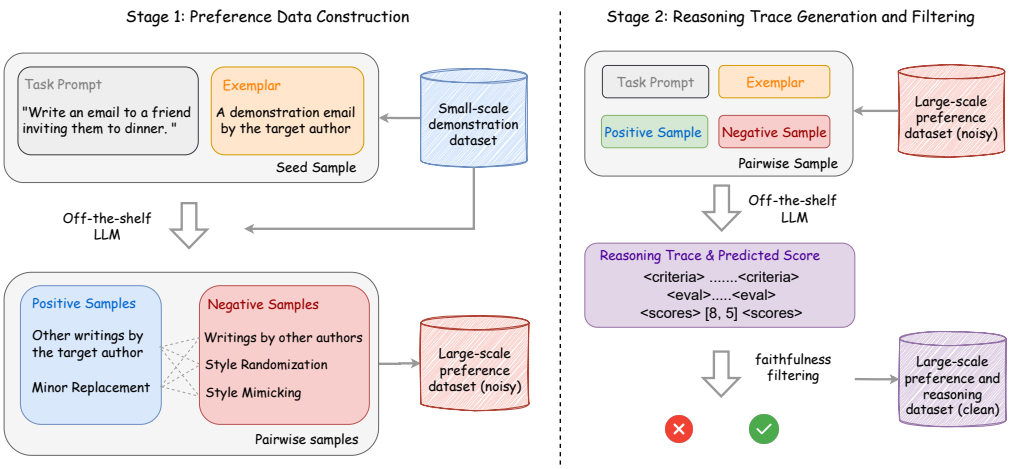

Figure 2: Pairwise preference data generation pipeline includes: LEFT) personalization data argumentation with contrastive prompting (Stage 1), and RIGHT) reasoning traces collection and post-hoc data filtering (Stage 2).

reasoning-based evaluation tuple $\mathcal{V} = (\tau, r^+, r^-)$, where:

- τ is a step-by-step reasoning trace explaining the comparative judgment, including an analysis of stylistic alignment, tone, phrasing, or semantic intent;
- r^+ and r^- are scalar reward scores, typically normalized within a bounded range (e.g., [1, 10], reflecting the model's confidence in the stylistic compatibility of y^+ and y^- with the user exemplars.

This encourages the model to articulate fine-grained distinctions and builds interpretability into reward modeling by exposing the rationale behind preference decisions.

Faithful Reasoning Trace Filtering. To ensure alignment between reasoning traces and the intended preference signal, we apply a filtering step that retains only *faithful* reasoning outputs, i.e. those that are consistent with the preference implied by the contrastive generation stage (i.e., y^+ should be stylistically more aligned with the exemplars e than y^-). Reasoning traces that contradict this assumption are excluded from the dataset, as they may introduce conflicting supervision and hinder effective training. This filtering ensures coherence between the scalar reward scores and the reasoning trace τ, resulting in the dataset $\mathcal{D}_{\text{SFT}} = \{x, e, y^+, y^-, \mathcal{V}\}$, which contains contrastive response pairs with faithful justifications. This dataset is used for subsequent supervised fine-tuning (SFT) to elicit reasoning justifications that improve the reward model's ability to assess preferences with increased accuracy and interpretability.

3.2 Supervised Fine-Tuning

With the curated dataset $\mathcal{D}_{\text{SFT}}$ in place, we first perform supervised fine-tuning on the RM to internalize the knowledge extracted through guided prompting. This results in a warm-started RM, denoted as PersRM-SFT, that can better assess user-aligned preferences in a generative way. The training objective is to maximize the conditional log-likelihood

$$\max_{\theta} \ \mathbb{E}_{(x,e,y^+,y^-,\mathcal{V})\sim\mathcal{D}_{\text{SFT}}} \ \log p_\theta(\mathcal{V} \mid x, y^+, y^-, e),$$

where $\mathcal{V}$ is the structured reasoning-augmented evaluations and θ denotes the trainable parameters of the reward model. This SFT phase serves to align the model's outputs with high-quality, faithful supervision signals, preparing it for subsequent reinforcement learning.

3.3 Reinforcement Fine-Tuning

Following SFT, we further optimize the reward model through RL to promote exploration of diverse reasoning patterns and improve robustness in preference assessments. While SFT provides high-quality supervision signals, it remains limited to the imitation of static patterns seen during data curation. In contrast, RFT enables the model to generate novel reasoning traces that generalize beyond the curated examples, allowing for more adaptive and discriminative preference modeling.

Sampling and Format Validation. Given an input quadruple (x, e, y^+, y^-), we sample reasoning-based evaluations $\mathcal{V} \sim p_\theta(\mathcal{V} \mid x, y^+, y^-, e)$ from

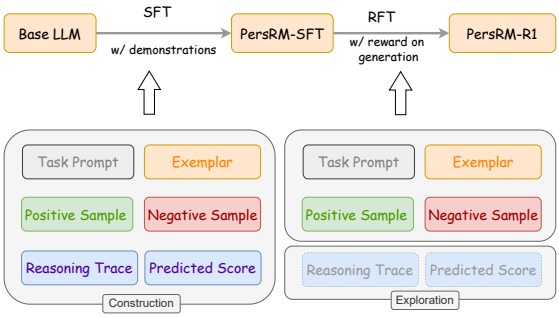

Figure 3: Training pipeline of PersRM-R1. The process begins with personalization data augmentation (see Sec. 3.1), followed by supervised fine-tuning (Sec. 3.2), and concludes with reinforcement fine-tuning (Sec. 3.3) to explore effective reasoning traces.

the current reward model. To ensure valid learning signals, we first perform format validation using a deterministic procedure $\texttt{fmt}(\mathcal{V})$, which checks whether the generated output conforms to the expected structured format. If the format is valid, we parse the reasoning trace and scores as $(\hat{\tau}, \hat{r}^+, \hat{r}^-) \leftarrow \mathcal{V}$; otherwise, the response is considered invalid and penalized.

Reward Function Design. To guide the model toward both faithful reasoning and correct preference assessments, we define a sparse reward function based on *format correctness* and *consistency with contrastive supervision*:

$$r(\mathcal{V}, x, y^+, y^-, e) = \begin{cases} -1, & \neg\texttt{fmt}(\mathcal{V}) \\ 0, & \texttt{fmt}(\mathcal{V}) \wedge \hat{r}^+ \leq \hat{r}^- \\ +1, & \texttt{fmt}(\mathcal{V}) \wedge \hat{r}^+ > \hat{r}^-. \end{cases} \quad (1)$$

This reward signal encourages the generation of well-structured outputs that faithfully reflect the intended preference signal, where the preferred response y^+ should receive a higher score than y^-. Outputs that violate formatting constraints or misalign with the contrastive signal receive a penalty, thereby discouraging hallucinated or unreliable evaluations.

Policy Optimization. The reward model is then updated using standard RL algorithms with the objective of maximizing expected reward:

$$\max_{\theta} \; \mathbb{E}_{(x,e,y^+,y^-)\sim\mathcal{D}_{\text{syn}},\mathcal{V}\sim p_\theta} \; r(\mathcal{V}, x, y^+, y^-, e).$$

This process results in a reinforcement-tuned personalized reward model, denoted as PersRM-R1, which benefits from both SFT pretraining and RL-guided exploration. To assess the necessity of SFT, we include an ablation variant, denoted as PersRM-RFT, which applies RL directly on the pretrained base model without SFT initialization. This cold-start configuration allows us to evaluate whether reinforcement alone suffices to capture user-specific preferences or if the curated supervision from SFT is critical for performance.

4 Experiments

4.1 Experimental Setup

Dataset. Our experiments use two datasets centered on writing styles: the **CCAT** dataset, which contains news articles by 50 authors (Lewis et al., 2004), and **CMCC** dataset, which includes multi-genre writings such as emails, blogs, essays by 21 authors (Goldstein-Stewart et al., 2008). Our evaluation protocol is designed to test generalization to both unseen authors and unseen genres, necessitating a strict, non-overlapping partition of data into training, validation, and test sets. We employ a *strictly author-disjoint split*, ensuring authors used for training, validation, and testing are entirely separate. This allows our evaluation to assess the model's true capabilities in personality trait analysis and scoring, rather than relying on memorization.

Additionally, we develop a challenging cross-domain test set to assess the generalizability of models in measuring personality similarity across different genres. An ideal model should perform well in this scenario, as we hypothesize that personal traits remain consistent cross domains. This test set is constructed exclusively from CMCC, which features multiple genres for the same author, whereas CCAT contains only news articles.

- **Training Set**: We construct the training set using our data curation pipeline based on news articles, emails, and essays of a large group of authors, comprising *45 authors from the CCAT corpus* and *18 authors from the CMCC corpus*, which results in approximately 17.2k pairwise preference and reasoning samples.

- **Validation Set**: The validation set is built using a held-out group of authors: *2 authors from CCAT* and *1 author from CMCC*. This set consists of 200 samples.

- **Standard Test Set**: The test set is composed of a separate group of held-out authors: the *remaining 3 authors from CCAT* and the *remaining 2 authors from CMCC*. This set contains 334 pairwise samples.

- **Cross-Domain (Cr. Do.) Test Set**: This test set is constructed using documents from the *3 CMCC authors* from the validation and test splits. Crucially, we exclusively use genres that are *entirely withheld* from both the training and validation sets. These unseen genres include *blog articles, interview transcripts, and chat logs*. This challenging set consists of 439 pairwise samples.

Models. We use Qwen2.5-3B-Instruct and Qwen2.5-7B-Instrcut as base models, as prior work demonstrates that the strong fundamental reasoning abilities of Qwen series facilitate effective post-training and adaptation to reasoning-focused downstream tasks (Xie et al., 2025). Our resulting models are referred to as PersRM-R1-3B and PersRM-R1-7B, respectively. Qwen2.5-72B-Instruct (Qwen Team, 2024) is employed for both preference data construction and reasoning trace generation. [1]

Model Training. Our models are trained following the SFT and RFT procedure introduced in Sec. 3.2 and Sec. 3.2. Further training details, such as hyperparameter setups, are provided in Appendix B.

Evaluation Metric. Following previous work (Stiennon et al., 2022; Lambert et al., 2024), we evaluate RM performance using *accuracy* with respect to the ground-truth preference labels, where random guessing yields an accuracy of 50%. Further details on the evaluation methodology for different types of RMs are provided in Appendix B.

4.2 Baselines

We compare PersRM-R1 with a diverse set of well-known reward models across three baseline categories: (1) **Scalar RMs**: Internlm2-7B-Reward, RM-Mistral-7B, Skywork-Reward-Llama3.1-8B (abbreviated as SR-Llama3.1-8B), all of which are top-performing models of approximately 7B parameters on the RewardBench leaderboard (Lambert et al., 2024); (2) **Generative RMs**: Mistral-v0.3-7B-Instruct (Jiang et al., 2023), the Qwen2.5 series (ranging from 7B to 32B), Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct (MetaAI, 2024), representing mainstream open-source LLMs across a large range of model size; (3) **Reasoning RMs**: We also compare against two recently released, state-of-the-art RMs that have been specifically enhanced for reasoning capabilities: RM-R1-7B (Chen et al., 2025) and RRM-7B (Guo et al., 2025).

[1] DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI et al., 2025), another strong open-source LLM known for its reasoning capabilities, is also examined. However, in our experimental scenario, its generated reasoning traces often lack proper formatting. In contrast, Qwen2.5-72B-Instruct exhibits satisfactory performance in both reasoning quality and adherence to formatting requirements.

4.3 Experimental Results

Table 1 presents a performance comparison between our models and various baselines. We report the performance on the standard test set, with separate results for each dataset (CCAT and CMCC), as well as on the cross-domain test set (Cr. Do.).

Category	Model	CCAT	CMCC	Cr. Do.
Scalar RMs	Internlm2-7B-Reward	67.8	69.2	64.3
	RM-Mistral-7B	65.7	68.1	62.8
	SR-Llama3.1-8B	65.3	68.4	68.8
Generative RMs	Mistral-v0.3-7B-Ins.	73.9	76.2	74.8
	Qwen2.5-7B-Ins.	77.6	75.8	72.4
	Qwen2.5-14B-Ins.	82.3	82.8	83.1
	Qwen2.5-32B-Ins.	86.4	87.1	86.7
	Llama3.1-8B-Ins.	77.3	79.3	73.8
	Llama3.1-70B-Ins.	**94.3**	<u>94.3</u>	**93.7**
Reasoning RMs	RM-R1-7B	89.8	88.2	89.7
	RRM-7B	87.2	89.6	89.3
	PersRM-R1-3B (Ours)	91.8	92.2	89.7
	PersRM-R1-7B (Ours)	<u>93.8</u>	**94.6**	<u>92.3</u>

Table 1: Performance comparison of various models across different test sets, reported as reward modeling accuracy (%), where higher values indicate better performance. Best results are in bold, runner-up in <u>underline</u>.

Performance of Baselines. Our evaluation of the baseline models yields three important findings. First, existing scalar-based RMs perform poorly in scoring personal preferences, with accracies below 70.0%, indicating their inadequacy in distinguishing personalization divergence. Second, reward modeling performance for personalization exhibits a clear scaling trend, with larger models consistently achieving higher accuracy. For example, the Qwen2.5 series shows improved performance as model size increases, and Llama3.1-70B-Instruct achieves the best results across most benchmarks, underscoring the effectiveness of greater model capacity in capturing user-specific preferences.Third, reasoning-based reward models consistently outperform non-reasoning generative models of comparable size. For example, RM-R1-7B achieves nearly 10% higher accuracy than Qwen2.5-7B and Llama3.1-8B, and performs comparably to the larger Qwen2.5-32B. These results underscore the superiority and potential of reasoning reward models in personalization alignment.

Effectiveness of PersRM-R1. Our proposed model, PersRM-R1-7B, achieves 93.8% accuracy on CCAT and 94.6% on CMCC, substantially outperforming other reasoning-based models of sim-

ilar size. This demonstrates the effectiveness of our training paradigm, even with a limited amount of personalized data. Moreover, despite having significantly fewer parameters, our models approach the performance of much larger models such as Llama3.1-70B-Instruct, highlighting their efficiency and scalability. These findings support our hypothesis that existing reward models, typically trained on datasets focused on mathematical or common-sense reasoning, lack the architectural and data-centric design necessary for capturing personalized preferences. This highlights the need for specialized personalization reward models, as advanced in this study.

Cross-Domain Generalizability. The cross-domain test setup is challenging, especially for RMs with relatively small size. PersRM-R1-7B achieves an accuracy of 92.3%, while the smaller PersRM-R1-3B attains 89.7%, exhibiting strong cross-domain generalizability to genres not included in the training data. This strong performance on unseen genres highlights our method's ability to learn the underlying principles of personal preference rather than merely overfitting to the topics in the training data. This capability is crucial for building RMs that are practical and reliable in real-world applications.

Generalizability to Additional Exemplars. We evaluate the generalizability to additional exemplars in inference time, as in realistic scenarios multiple exemplars are available for a target user.We observe from experimental results in Table 2 that incorporating additional personal exemplars leads to performance improvements across all models. This effect is particularly pronounced for Qwen2.5-7B-Instruct, which starts with relatively low performance when using only one exemplar. Notably, PersRM-R1-7B demonstrates even greater improvement, despite already achieving the best results in the single-exemplar setting, when compared to Llama3.1-70B-Instruct. This highlights the strong generalizability of PersRM-R1-7B with respect to the number of exemplars provided, even tuned on preference data with only one exemplar. Additional experimental results regarding different training paradigms and additional exemplars are provided in Appendix F.3.

4.4 Tweet Style Test and User Study

To further assess the robustness and practical applicability of our model, we conducted two additional tests: a cross-domain test on the LaMP-Tweet dataset (Salemi et al., 2023)—whose semantic structure differs from our writing-style data—and a user study to measure alignment with human perception, detailed in Appendix H and I respectively.

Model	# Ex.	CCAT	CMCC	Cr. Do.
Qwen2.5-7B-Ins.	1	77.6	75.8	72.4
	3	78.3 (+0.7)	79.2 (+3.4)	76.2 (+3.8)
Llama3.1-70B-Ins.	1	94.3	94.3	93.7
	3	**94.3** (+0.0)	94.6 (+0.3)	93.9 (+0.2)
PersRM-R1-7B	1	93.8	94.6	92.3
	3	94.1 (+0.3)	**95.1** (+0.5)	**92.6** (+0.3)

Table 2: Reward modeling accuracy (%) with varying numbers of exemplars (# Ex.). Values in parentheses indicate improvement over the one-exemplar setting.

5 Analysis

In this section, we present a series of analytical experiments to provide deeper insights into our proposed training methodology. The primary goal is to investigate the underlying reasons for its effectiveness and to characterize the dynamic changes that occur throughout the training process.

5.1 Effectiveness of Different Paradigms

We explore the effectiveness of different training paradigms: 1) SFT only, 2) RFT only, and 3) SFT followed by RFT (SFT+RFT). Our experiments are conducted based on the Qwen2.5-7B-Instruct model and previously constructed datasets. As illustrated in Table 3, both standalone SFT and RFT are effective in enhancing the model's capabilities in measuring personality similarity. However, the paradigm of SFT followed by RFT, embodied in our PersRM-R1-7B model, delivers the most significant performance boost, substantially outperforming all other approaches. This highlights the value of integrating both supervised and reinforcement fine-tuning strategies, and suggests that the strong capabilities of PersRM-R1 are not solely due to knowledge distillation through SFT from the off-the-shelf LLM used in data construction, but also stem from the reinforcement learning procedure, which enables the model explore and generalize to novel skills for solving the task.

To gain deeper insights into the value of the SFT stage, we track the average sequence length and reward scores during the RFT stage for the RFT Only and SFT+RFT training paradigms, as shown in Fig. 4. The results demonstrate that performing SFT prior to RFT yields a model capa-

Training Paradigm	CCAT	CMCC	Cr. Do.
Base Model	77.6	75.8	72.4
SFT Only	86.1	87.7	82.2
RFT Only	83.7	84.2	80.2
SFT + RFT	**93.8**	**94.6**	**92.3**

Table 3: Reward modeling accuracy (%) across different training paradigms.

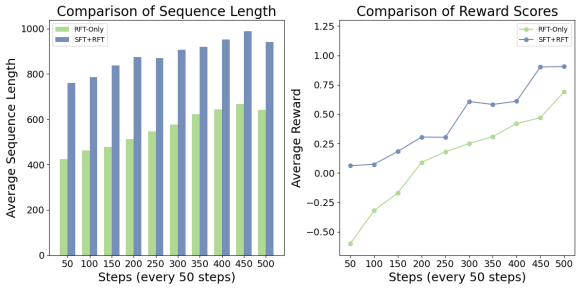

Figure 4: Comparison between the RFT stage for the RFT only and SFT followed by RFT training paradigms.

ble of generating longer sequences and achieving higher average reward scores compared to the base model, Qwen2.5-7B-Instruct. RFT applied to this improved starting point consistently outperforms RFT applied directly to the base model in both sequence length and average reward. This underscores the importance of an initial SFT phase for establishing a foundational reasoning capabilities, without which the model struggles to learn effectively. For more in-depth insights into the RFT process, we provide the training curves for PersRM-R1-3B and PersRM-R1-7B in Appendix F.2.

5.2 Task-specific Behaviors Emerge

Task-specific behaviors beneficial for personality trait analysis emerge from RFT's objective of maximizing cumulative rewards. Identified emergent abilities include 1) *the discovery of novel and case-specific stylistic criteria* that are not present in the SFT dataset; and 2) *dynamic and context-aware prioritization for various criteria*.

Discovery of Novel and Case-specific Criteria. A key capability unlocked by the RFT phase is the discovery of evaluation criteria that are more granular and insightful than the general rubrics provided during SFT. The SFT-only model is confined to applying these pre-defined rubrics from its training data. In contrast, the RL-trained model demonstrates an exploratory capacity, identifying the latent stylistic principles that make an author's voice unique. For instance, when evaluating responses to the prompt, *"Do you feel that gender discrimination is still an issue in the present-day United States?"*, the reference author constructs their entire argument around a detailed personal anecdote. Our SFT+RFT model moves beyond the SFT-only model's predefined criteria by identifying an emergent one: *"Argumentation Grounded in Personal Narrative."* It learned that the author's core stylistic signature isn't just mentioning an experience, but using it as the foundational structure for the entire essay. Therefore, it correctly favors responses mimicking the author's narrative-driven argumentative style over generic essays on the same topic.

Dynamic and Context-aware Rubric Prioritization. In contrast to the SFT-only model's fixed-order application of rubrics, the SFT+RFT model dynamically prioritizes evaluation criteria based on the specific stylistic context of the reference and responses. This flexible, context-aware reasoning leads to more accurate and nuanced preference judgments. For example, its "<eval>" block might state: *"Response A is decisively superior as it mirrors the reference author's core technique of 'Argumentation Grounded in Personal Narrative.' The entire essay is built around a central story, which is the defining characteristic of the target style. Response B, while addressing the topic, adopts an abstract and impersonal tone, completely missing this crucial stylistic signature."* In this example, a specific criterion is prioritized in the analysis.

A full example of a PersRM-R1 generation is in Appendix E.

5.3 Cognitive Behaviors Emerge

During RFT, we observe the emergence of four cognitive behaviors (Gandhi et al., 2025). These behaviors (e.g., Verification, which involves reasoning like "Let me double-check..."; detailed in Appendix J) directly shape the reasoning flow and improve prediction accuracy.

6 Conclusion

We introduce PersRM-R1, the first reasoning-based reward model for personalization alignment. Trained via a novel pipeline of guided data augmentation, supervised and reinforcement fine-tuning, it learns to perform fine-grained, personality-centric reasoning. Experimentally, PersRM-R1 achieves state-of-the-art, parameter-efficient performance and generalizes robustly to unseen domains and new user exemplars by capturing transferable personal preference principles.

7 Limitations

The superior capabilities of PersRM-R1 in personality trait analysis and similarity measurement open up several promising avenues for future exploration. In the following, we discuss the scope of our current work, framing these boundaries as key opportunities that build upon the foundation we have established. First, the evaluation of personalized policy tuning was not conducted, as this subsequent step is a significantly more resource-intensive task. Our empirical validation, which was deliberately focused on the reward model itself, provides sufficient and robust evidence for its efficacy through extensive experiments. Second, the training and evaluation of PersRM-R1 are conducted using datasets particularly focused on writing tasks, due to the limited availability of open-source user-specific data in other domains. Crucially, this data limitation underscores the value of our approach, as the data construction pipeline and training paradigm for PersRM-R1 were designed to be highly adaptable and are not confined to these specific domains. Scaling and evaluating the approach cross additional domains remains an important direction for future research.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv: 2204.05862*. ArXiv: 2204.05862 [cs.CL].

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022b. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv: 2212.08073*. ArXiv: 2212.08073 [cs.CL].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165. ArXiv: 2005.14165 tex.bibsource: dblp computer science bibliography, https://dblp.org tex.biburl: https://dblp.org/rec/journals/corr/abs-2005-14165.bib tex.timestamp: Wed, 03 Jun 2020 11:36:54 +0200.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization. *arXiv preprint*. ArXiv:2406.00045 [cs].

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research (TMLR)*. ArXiv: 2307.15217 [cs.AI].

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. 2025a. PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. 2025b. RM-R1: Reward Modeling as Reasoning. *arXiv preprint*. ArXiv:2505.02387 [cs].

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. 2025. RM-R1: Reward Modeling as Reasoning. *arXiv e-prints*, arXiv:2505.02387.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. ArXiv: 1706.03741 [stat.ML].

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. 2025. Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs. *arXiv e-prints*, arXiv:2503.01307.

Jade Goldstein-Stewart, Kerri Goodwin, Roberta Sabin, and Ransom Winder. 2008. Creating and Using a

Correlated Corpus to Glean Communicative Commonalities. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2025. Reward Reasoning Model. *arXiv preprint*. ArXiv:2505.14674 [cs].

Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2025. Reward Reasoning Model. *arXiv e-prints*, arXiv:2505.14674.

Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Weikai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. 2025. OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework. *arXiv preprint*. ArXiv:2405.11143 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv e-prints*, arXiv:2310.06825.

W. Bradley Knox and Peter Stone. 2013. Learning non-myopically from human-generated reward. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, IUI '13, pages 191–202, New York, NY, USA. Association for Computing Machinery.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style Vectors for Steering Generative Large Language Model. *arXiv preprint*. ArXiv:2402.01618 [cs].

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. RewardBench: Evaluating reward models for language modeling. ArXiv: 2403.13787 [cs.LG].

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv preprint*. ArXiv:2309.00267 [cs].

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *J. Mach. Learn. Res.*, 5:361–397.

Mengdi Li, Jiaye Lin, Xufeng Zhao, Wenhao Lu, Peilin Zhao, Stefan Wermter, and Di Wang. 2025. Curriculum-RLAIF: Curriculum Alignment with Reinforcement Learning from AI Feedback. *arXiv preprint*. ArXiv:2505.20075 [cs].

Mengdi Li, Xufeng Zhao, Jae Hee Lee, Cornelius Weber, and Stefan Wermter. 2023. Internally rewarded reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of machine learning research*, pages 20556–20574. PMLR.

Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Liu Leqi. 2024. Personalized Language Modeling from Personalized Human Feedback. *arXiv preprint*. ArXiv:2402.05133 [cs].

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-Time Scaling for Generalist Reward Modeling. *arXiv preprint*. ArXiv:2504.02495 [cs].

MetaAI. 2024. Introducing llama 3.1: Our most capable models to date. `https://ai.meta.com/blog/meta-llama-3-1/`.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024. OpenAI o1 System Card. *arXiv preprint*. ArXiv:2412.16720 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*. Tex.copyright: arXiv.org perpetual, non-exclusive license.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. ArXiv: 2304.03442 [cs.HC].

Qwen Team. 2024. Qwen2. 5: A party of foundation models. *Qwen (Sept. 2024). url: https://qwenlm. github. io/blog/qwen2*, 5.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. ArXiv: 2304.11406 [cs.CL].

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

Wolfram Schultz. 2015. Neuronal Reward and Decision Signals: From Theories to Data. *Physiological Reviews*, 95(3):853–951.

Omar Shaikh, Michelle Lam, Joey Hejna, Yijia Shao, Michael Bernstein, and Diyi Yang. 2024. Show, Don't Tell: Aligning Language Models with Demonstrated Feedback. *arXiv preprint*. ArXiv:2406.00888.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. *arXiv preprint*. ArXiv:2310.10158.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint*. ArXiv:2402.03300 [cs].

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. HybridFlow: A Flexible and Efficient RLHF Framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297. ArXiv:2409.19256 [cs].

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. ArXiv: 2009.01325 [cs.CL].

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning. *arXiv preprint*. ArXiv:2502.14768 [cs].

Diji Yang, Linda Zeng, Kezhen Chen, and Yi Zhang. 2024a. Reinforcing Thinking through Reasoning-Enhanced Reward Models. *arXiv preprint*. ArXiv:2501.01457 [cs].

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2024b. RLCD: Reinforcement Learning from Contrastive Distillation for Language Model Alignment. *arXiv preprint*. ArXiv:2307.12950 [cs].

Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2024. PersLLM: a personified training approach for large language models. ArXiv: 2407.12393 [cs.CL].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint*. ArXiv:2306.05685 [cs].

Appendix

A Preliminary

A.1 Reward Modeling

Canonical reward models are typically formulated as a Bradley-Terry (BT) models (Knox and Stone, 2013; Christiano et al., 2017) and are trained in supervised learning. The model estimates the probability that one response is preferred over another, and the optimization objective is formulated with loss function:

$$\mathcal{L}_{\text{BT}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\left[\log\sigma\left(r_\theta(x,y^+) - r_\theta(x,y^-)\right)\right], \tag{2}$$

where $\mathcal{D}$ represents a pairwise preference database containing triplet entries (x, y^+, y^-), x is the query, y^+ and y^- are the preferred or rejected responses, respectively. $r_\theta(x, y)$ denotes the predicted reward score for response y given input x by a reward model that is parameterized by θ. $\sigma(\cdot)$ is the sigmoid function.

As causal LLMs exhibit great generalizability across diverse domains (Radford et al., 2019; Brown et al., 2020), reward modeling has been formulated as a generative task recently (Liu et al., 2025; Guo et al., 2025), where a causal LLM works as the reward model to predict the reward score in its generated output. The training objective is identical to the next-token prediction in pre-training:

$$\mathcal{L}_{\text{Causal}}(\theta) = -\mathbb{E}_{x'\sim\mathcal{D}}\left[\sum_{t=1}^{T-1}\log p_\theta(x_{t+1} \mid x_1,\ldots,x_t)\right], \tag{3}$$

where x' is a piece of text constructed with (x, y^+, y^-), $p_\theta(x_{t+1} \mid x_1, \ldots, x_t)$ represents the predicted probability of the next token given previous ones. The training sample x' used in this formulation is flexible where intermediate reasoning process can be included before the production of the final reward score. We adopt the causal formulation to perform supervised fine-tuning (SFT) in this paper, as it aligns with our goal to incorporate reasoning process into reward modeling.

A.2 Reinforcement Fine-tuning

Proximal policy optimization (PPO) (Schulman et al., 2017) and its variant group relative policy optimization (GRPO) (Shao et al., 2024) are two mainstreaming RL methods for RFT to enhance reasoning capabilities of LLMs. In this work, we utilize GRPO for RFT following existing work in enhancing reasoning capabilities for LLMs (DeepSeek-AI et al., 2025; Chen et al., 2025b; Guo et al., 2025).

GRPO is an improved version of PPO designed to boost memory and computational efficiency Traditionally, PPO uses a value model to estimate the state value of generated responses for advantage estimation (Schulman et al., 2015). In the domain of RFT for LLMs, this value model is typically initialized with a pretrained LLM similar in size to the policy model and is optimized along with it in supervised learning. In contrast, GRPO removes the need for a value model by estimating advantages through a Monte Carlo approach. It calculates advantages based on the rewards from a set of randomly sampled outputs, which greatly reduces both memory and computational cost while ensuring effective policy optimization.

In GRPO, a group of outputs for a given prompt x is generated by sampling from the policy π_θ:

$$\{o_1, o_2, \ldots, o_G\} \sim \pi_\theta(\cdot \mid x), \tag{4}$$

where, G is a hyperparameter that specifies how many outputs are sampled per prompt x. To estimate the advantage for each sampled output, GRPO uses the normalized reward within the group:

$$A_i = \frac{r_i - \bar{\mathbf{r}}}{\sigma(\mathbf{r})}, i \in \{1, 2, \ldots, G\}, \tag{5}$$

where $\mathbf{r} = \{r_1, r_2, \ldots, r_G\}$ are the rewards assigned to each output in the group, calculated by a rule-based reward function or a learned reward model. $\bar{\mathbf{r}}$ and $\sigma(\mathbf{r})$ denote the mean and standard deviation of the rewards in the group, respectively. With these advantage estimates, GRPO updates the policy by maximizing the following objective, closely following the PPO framework:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big\{\min\Big[\lambda_t(\theta)A_{i,t}, \text{clip}\Big(\lambda_t(\theta), 1-\varepsilon, 1+\varepsilon\Big)A_{i,t}\Big] - \beta\mathbb{D}_{KL}[\pi_\theta\|\pi_{\text{ref}}]\Big\},$$

where $\lambda_t(\theta) = \frac{\pi_\theta(o_{i,t}|x,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|x,o_{i,<t})}$ measures the change in predicted probability of token $o_{i,t}$ between the current and previous policy models. The advantage $A_{i,t}$ is set to $= A_i$ for all tokens in output o_i, and $|o_i|$ denotes the number of tokens in output o_i. The hyperparameter ε controls the clipping range for stable updates, while β determines the weight of the KL-divergence constraint with respect to a reference policy π_{ref}. For a deeper dive

into the theoretical foundations and optimization strategies, we refer readers to the original PPO and GRPO papers (Schulman et al., 2017; Shao et al., 2024).

B Additional Training and Evaluation Details

SFT. SFT Stage is carried out using the SFTTrainer module from OpenRLHF (Hu et al., 2025), with a batch size of 64 and a single training epoch. To improve GPU memory efficiency, we enable gradient checkpointing, leverage FlashAttention, and apply optimizer offloading. The model is optimized using the Adam optimizer with a learning rate of 5e-6.

RFT. We adopt the verl framework (Sheng et al., 2025) for all GRPO training. The training is performed using Fully Sharded Data Parallel (FSDP) with parameter, gradient, and optimizer offloading enabled to improve memory efficiency. The training batch size is set to 32. Gradient checkpointing is enabled to further reduce memory usage. Rollout generation is handled by the vLLM backend, configured with tensor parallelism size of 1 and GPU memory utilization capped at 0.6. Sampling uses default parameters, with temperature = 1.0 and top-p = 1.0. Each prompt is decoded using 8 candidate responses. KL regularization is applied with a coefficient of 1e-3 and uses the low-variance KL approximation. The learning rate is set to 3e-7 for all model sizes, with no scheduler applied. We train the 3B and 7B models on 1 node equipped with 8 A100 (80GB) GPUs.

Reward Model Evaluation. The evaluation methodology for RMs is fundamentally dependent on their architecture and task formulation (cf. Sec. A.1), which dictates the format of the input data and the method to extract reward scores to calculate reward modeling accuracy. We define distinct input structures for scalar-based and generative models.

- **Scalar RMs** are trained to predict a quality score for a single response in isolation. The training instances for these models are structured as a `{query, response}` entity, where the model learns to output a scalar value corresponding to an absolute quality rating. To evaluate on pairwise data `{query, response_a, response_b}`, the RM performs inference separately on each response in the triplet to obtain individual reward scores.
- **Generative RMs** are trained on pairwise preference data to directly learn a comparative function. The input for these models is a triplet formatted as `{query, chosen_response, rejected_response}`. During training, the model is optimized to assign a higher score to the `chosen_response` over the `rejected_response` for the same query. For evaluation on pairwise data `{query, response_a, response_b}`, the RM processes an input constructed with the triplet and predicts reward scores for both responses simultaneously.

C Prompt Examples for Response Sample Generation

Prompt Example 1 : Style Mimicking

You are an expert writer trained to imitate human-written responses.
Your task is to write a continuation based on the sentence inside <Problem>, while closely following the tone, structure, style, and content of the example inside <Context>. The example provided does not need to be copied.
Instead, you should carefully mimic its language patterns, coherence, detail level, and writing flow to generate a similar-quality output.
<Problem> {problem} </Problem>
<Context> {context} </Context>
Now, please generate a continuation that matches the writing style and quality of the example above, based on the problem.
Your response should read as if it could have been written by the same author who wrote the example in <Context>.

Prompt Example 2: Minor Replacement

You must perform only minimal rewriting of the paragraph below. Strictly replace exactly 5 to 6 individual words only, preferably adjectives, adverbs, or verbs, with close synonyms.
Do not change sentence structure, punctuation, or the order of words.
Do not alter nouns, names, numbers, or proper terms.
Do not insert or remove any words — the total

word count should remain nearly identical.
The meaning, tone, and style of the paragraph must remain unchanged.
Do not explain your changes — only output the rewritten paragraph.
Here is the paragraph to rewrite: {paragraph}

Prompt Example 3 : Random Style

You are a creative writer.
Your task is to randomly continue the given sentence in any style, tone, or direction.
Each time you see the same input, you must write something completely different from before.
You are encouraged to be unpredictable, humorous, absurd, poetic, or even surreal.
Avoid logical consistency or factual accuracy.
Focus on creativity, stylistic variation, and divergence from expected content.
The sentence you must continue is shown below between <Problem> tags. Your response should be a free-form continuation, based solely on the content within <Problem>.
<Problem> {problem} </Problem>
Now, please generate the continuation.

D Prompt Examples for Reasoning Trace Generation

Prompt Example 4: Reasoning Trace Generation

You are a skilled assistant for scoring and comparing responses.
You need to evaluate two given responses based on how well they match the context's personal style, tone, preferences, and writing style.
Your overall scores must reflect a single, comprehensive judgment per response that integrates all criteria.
Personal-Align-Specific Criteria (you may choose from or add to these as appropriate):
1. Personal Style Adherence:
2. Tone and Voice Consistency:
3. Language Fluency and Coherence:
4. Relevance to Personal Preferences and Experiences:
You may also add other relevant criteria if they help evaluate the two responses.
Your task is to use the most suitable criteria for the current context to judge the responses holistically and then output one overall integer score per response.
Conversation Context: { context }
Responses to be Scored:

Response A: { response a}
Response B: { response b}
Output Format Requirements Produce your answer as exactly three sections:
<criteria> Specific Criteria: Clearly list the criteria you decided to use for this context (you can select from the ones provided or add new ones if appropriate). Do not include any scores in this section.</criteria>
<eval> Analysis: Provide a very detailed, thorough, and step-by-step reasoning process. Break your reasoning into multiple points as you go through each chosen criterion one by one. Explain how Response A and Response B satisfy or fail that criterion, and clearly state the impact of each observation on your final judgment. End this section with a concluding sentence that ties all observations together and justifies the overall scores you will give. Do not include any numeric scores in this section.</eval>
<scores> Scores: [[x,y]]</scores>

E Case Study

In this section, we present a concrete example generated by PersRM-R1, accompanied by explanatory comments (highlighted in gray) to clarify its behavior. These annotations are added for illustration purposes and are not part of the original model output.

Response Generated by PersRM-R1

<criteria>
Specific Criteria:

1. Personal Style Adherence
2. Tone and Voice Consistency
3. Language Fluency and Coherence
4. Argumentation Grounded in Personal Narrative # Note: PersRM-R1 discovers novel sample-specific criteria that are be-

yond those included in synthetic data generation for SFT.

</criteria>

<eval>
Analysis:
First, let's establish a clear evaluation plan: I will assess the four standard criteria sequentially, then focus on the more nuanced, discovered criterion, and finally, trace back through all points to confirm the final scores. # Note: Cognitive Behavior: Subgoal Setting

1. **Personal Style Adherence:**
   - **Response A:** This response closely mirrors the user's personal style. It effectively integrates casual language and humor, making it feel like a natural continuation of the user's thoughts.
   - **Response B:** This response also adheres well, but is slightly more formal in some parts, representing a minor deviation from the user's typically more relaxed style.
2. **Tone and Voice Consistency:**
   - **Response A:** The tone is fully aligned. It maintains the same level of emotional intimacy and casualness as the user's original message, feeling natural and consistent.
   - **Response B:** The tone is mostly aligned. However, the slight formality can make it feel a bit less personal and more distant compared to the user's voice.
3. **Language Fluency and Coherence:**
   - **Response A:** The language is very fluent and smooth, with no awkward phrasing.
   - **Response B:** Mostly fluent, but the intermittent formality creates a slightly disjointed reading experience compared to the user's seamless style.
4. **Argumentation Grounded in Personal Narrative:** This is the crucial differentiator that goes beyond surface-level style matching. It's not just about *including* anecdotes (covered by 'Relevance'), but about using a personal story as the foundational structure for the entire argument. This insight represents the Discovery of Emergent and Naunced Criteria. In this context, this criterion is weighted heavily, demonstrating Dynamic and Context-Aware Rubric Prioritization.
   - **Response A:** Excels here. It correctly identifies and replicates the user's method of building the entire response around a central personal story.
   - Initially, I thought Response B's use of examples was sufficient. However, upon reconsideration, that route doesn't hold. Response B only sprinkles in anecdotes as supporting evidence for a list of points, failing to grasp that the narrative *is* the argument. This is a fundamental stylistic misunderstanding. # Note: Cognitive Behavior: Backtracking

Concluding Summary: Let me double-check the overall logic before scoring. (Cognitive Behavior: Verification) While both responses perform adequately on surface-level criteria like tone and fluency, Response A demonstrates a profoundly deeper understanding of the author's style by correctly identifying and replicating the core **Argumentation Grounded in Personal Narrative** technique. Response B's failure on this single, prioritized criterion creates a significant quality gap. Therefore, if Response A is superior, it must have fulfilled this key requirement, which it has. Tracing back, this single point justifies a much higher score. (Cognitive Behavior: Backward Chaining) This clear, verifiable distinction based on a critical discovered criterion perfectly exemplifies Enhanced Interpretability and Faithful Adherence and is the basis for the final scores.
</eval>

<scores> [[9,7]] **</scores>**

F Additional Experimental Results

F.1 Evaluation of Constructed Preference Dataset

In this part, we evaluate the quality of the preference dataset constructed after the first stage of our data curation pipeline (see Fig. 2, left panel). Following previous work (Zheng et al., 2023; Shaikh et al., 2024; Yang et al., 2024b), we use GPT-4o as a proxy for human judges to compare the stylish similarity between responses and personal exemplars. GPT-4o is prompted to produce a similarity score given a pair of responses over 300 randomly selected samples for each of the 5 methods, for a total of 1,500 samples as our evaluation metric, where a higher similarity score indicates closer personality traits exhibited in the given pair of responses. The prompt used for this evaluation is provided below. Specifically, GPT-4o is prompted to assign a similarity score to each response and exemplars pair on 1,500 samples. A higher similarity score reflects a greater alignment in the exhibited personality traits between the response and the personal exemplars. The prompt used for this evaluation is presented below.

Prompt Example for Style Similarity Evaluation

Evaluate the similarity between the following two sentences in terms of sentence structure, grammar, style, tone, and word choice.
Give a comprehensive similarity score ranging from 0 to 10, where a higher score indicates greater similarity.
Only return a single number. Do not provide any explanation.

Table 4 shows that different generation strategies yield clearly distinguishable similarity scores, confirming the effectiveness and stylistic separability of our preference data. The high scores for target-author writings and low scores for randomized styles suggest that the dataset can provide reliable training signals for style alignment.

F.2 RFT Training Curves

Fig. 5 illustrates training curves of RFT during the developing of PersRM-R1-3B and PersRM-R1-7B. The average reward obtained by each model (cf. Eq. 1) steadily increases until reaching a plateau. Notably, PersRM-R1-7B achieves a higher average reward than PersRM-R1-3B at both the initial stage and at convergence, indicating its superior final performance after both SFT and RFT. Interestingly, the smaller model surpasses the larger model in terms of reward during steps [100, 250] and [300, 400], though the larger model eventually overtakes it. These curves clearly illustrate the effectiveness of our RFT training in enhancing the model's capability to analyze personality traits and produce reward scores that reflect personality similarity.

Category	GPT-4o-eval score
Other writings by the target author	9.41
Minor Replacement	9.39
Style Mimicking	5.89
Writings by other authors	3.86
Style Randomization	2.41

Table 4: Style similarity scores evaluated by GPT-4o

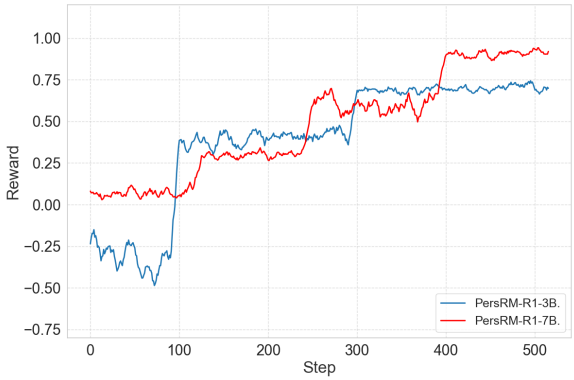

Figure 5: Reward curves during RFT for PersRM-R1-3B and PersRM-R1-7B.

F.3 Generalizability to Additional Exemplars of Different Training Paradigms

We evaluate the generalizability to additional exemplars in inference time for models that are trained in different training paradigms (cf. Sec. 5.1) using Qwen2.5-7B-Instruct as the base model. Experimental results are presented in Table 5. This demonstrates that the model trained through SFT followed by RFT, i.e., PersRM-R1-7B, exhibits stronger generalizability to additional exemplars.

G Dataset Details

Our dataset is constructed from two corpora: **CCAT** (news) and **CMCC** (multi-genre). We adopt a strictly author-disjoint split to ensure that the evaluation measures the model's generalizability. Details of dataset splits are provided in Table 6.

Training Paradigm	# Ex.	CCAT	CMCC	Cross-Domain
SFT Only	1	86.1	87.7	82.2
	3	86.3 (+0.2)	87.7 (+0)	82.4 (+0.2)
RFT Only	1	83.7	84.2	80.2
	3	83.7 (+0)	84.2 (+0)	80.5 (+0.3)
SFT + RFT	1	93.8	94.6	92.3
	3	**94.1** (+0.2)	**95.1** (+0.5)	**92.6** (+0.4)

Table 5: Reward modeling accuracy (%) with varying number of exemplars (# Ex.). Values in parentheses indicate the improvement achieved relative to using one exemplar. The best-performing results in the three-exemplar setup are highlighted in **bold**, while the runner-up results are underlined.

Split	# Authors (CCAT / CMCC)	# Pairwise Samples
Training Set	45 / 18	17.2k
Validation Set	2 / 1	200
Test Set (In-domain)	3 / 2	334
Test Set (Cross-Domain)	0 / 3	439

Table 6: Author and sample counts of each dataset split.

H Cross-Domain Generalization on the LaMP-Tweet Dataset

To rigorously assess the generalization capabilities of PersRM-R1 beyond its training distribution, we conduct a challenging cross-domain evaluation using the LaMP dataset (Salemi et al., 2023).

Motivation. Our primary motivation is to test the model's robustness against a significant domain shift. While our training data (CMCC, CCAT50) is composed of long-form, structured content like emails and essays, the chosen LaMP task features text from the social media domain. Specifically, we use the **LaMP7: Personalized Tweet Paraphrasing task**, where tweets are characterized by their brevity, informality, and strong, idiosyncratic stylistic personalization. This stark contrast makes LaMP7 an ideal out-of-distribution testbed to evaluate if PersRM-R1 has learned transferable principles of personal preference, rather than merely overfitting to the stylistic patterns of formal writing.

Dataset and Protocol. For this evaluation, we designed a rigorous protocol to create a stylistic coherence identification task. We selected two disjoint sets of users from LaMP7: 120 target users for generating evaluation instances and 120 distractor users for negative sampling. The process for each of the 120 target users is as follows:

- **User Profile**: A stylistic profile is created from 5 historical tweets of the target user.
- **Positive Sample**: An authentic, unseen tweet from the same target user, representing perfect stylistic alignment.
- **Negative Sample**: A tweet randomly drawn from the distractor user pool.

This strict separation between target and distractor pools ensures a clean and robust evaluation, tasking the model to identify the stylistically coherent tweet (the positive sample) based solely on the user's profile.

Results and Analysis. As shown in Table 7, our models demonstrate highly competitive performance in this challenging cross-domain setting. PersRM-R1-7B achieves an accuracy of 84.8%, performing on par with the significantly larger Llama3.1-70B-Ins. (86.7%). Moreover, our more parameter-efficient PersRM-R1-3B model attains a robust accuracy of 82.1%, outperforming several larger models. While the absolute accuracies are expectedly lower than on our in-domain datasets due to the profound domain shift, the results are particularly noteworthy. The challenge lies in inferring a consistent authorial voice from noisy and fragmented data. For instance, a user's profile might contain a mix of colloquialisms, abbreviations, and direct replies, such as: "got an email saying they've posted my sims 3. it'll take 3-5 days wtff." and "@KahunaLaguna YOU BETTER NOT JUST SIT AT HOME." The success of PersRM-R1 in this setting proves that it has internalized transferable principles of personalization that extend beyond formal writing structures. It validates our model's ability to capture the nuanced essence of a user's informal tone and style, underscoring the robustness of our reasoning-based approach.

I User Study on Perceived Personalization Quality

To validate whether PersRM-R1's scoring mechanism aligns with human perception of personalization, we conducted a dedicated user study. This study directly addresses the model's ability to capture the subjective quality of personalized responses as judged by humans.

Experimental Setup. We randomly selected 50 test instances, yielding a total of 50 positive and

Category	Model	Cr. Do. LaMP-Tweet
Scalar RMs	Internlm2-7B-reward	42.3
	RM-Mistral-7B	48.1
	SR-Llama-3.1-8B	39.12
Generative RMs	Mistral-v0.3-7B-Ins.	69.3
	Qwen2.5-7B-Ins.	64.6
	Qwen2.5-14B-Ins.	73.2
	Qwen2.5-32B-Instruct	81.6
	Llama3.1-8B-Ins.	62.1
	Llama3.1-70B-Ins.	**86.7**
Reasoning RMs	RM-R1 7B	79.3
	RRM-7B	77.2
Our Models	PerRM-R1-3B (ours)	82.1
	PerRM-R1-7B (ours)	<u>84.8</u>

Table 7: Performance comparison of various models on the Cr.Do.LaMP-Tweet test set, reported as reward modeling accuracy (%), where higher values indicate better performance.Best results are in bold, runner-up in <u>underline</u>

50 negative responses for evaluation. To ensure a fair and consistent evaluation, 15 human annotators were instructed to score the personalization quality of each response on a 0-10 scale, following the identical scoring rubric established during the training phase. Concurrently, we used our PersRM-R1-7B to score the same set of 100 responses.

Results and Analysis. The score distributions from both human evaluators and PersRM-R1 are presented in Figure 6. A key observation is that both distributions exhibit a clear bimodal pattern, as visualized by the histograms and the Kernel Density Estimation (KDE) curves. For human evaluators, the scores tend to cluster at the lower end (around 1-2) and the higher end (around 7-8), suggesting that humans can distinctly categorize responses into "low-quality" and "high-quality" personalization. Crucially, PersRM-R1's scoring distribution mirrors this bimodal trend, with its own peaks around scores of 3 and 8. The strong alignment of the higher-score peak (around 8) indicates that our model's criteria for what constitutes a well-personalized response are highly consistent with those of human judges. While there is a slight shift in the low-score peak, the overall structural similarity of the distributions strongly suggests that PersRM-R1 has successfully internalized nuanced, human-centric principles of personal preference. This result provides compelling evidence that our reasoning-based reward model effectively captures perceived personalization quality, bridging the gap between automated metrics and genuine human experience.

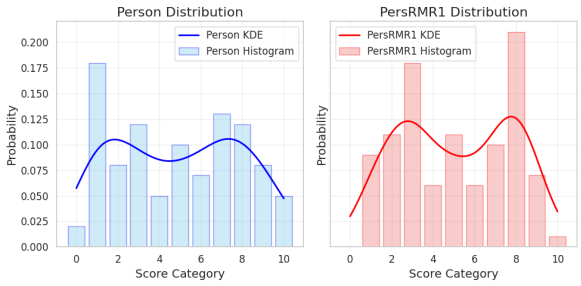

Figure 6: Comparison of score distributions between human evaluators (left, blue) and our PersRM-R1 model (right, red) on 100 test responses. Both the histograms and their corresponding Kernel Density Estimation (KDE) plots show a distinct bimodal distribution.

J Cognitive Behaviors

As mentioned in the main text, RFT elicits four cognitive behaviors that enhance reasoning. In contrast, with SFT alone, none of these behaviors except for subgoal setting are observable. This suggests that the base LLM, i.e., Qwen2.5-7B-Instruct, does not inherently demonstrate these human-like reasoning skills when analyzing personality traits. Instead, these abilities are explored and reinforced during RFT with our constructed preference data, a process analogous to the emergence of the "Aha moment" observed in the RFT of DeepSeek-R1 for solving math and coding problems (DeepSeek-AI et al., 2025).We provide illustrative examples of these behaviors below:

Examples of Identified Cognitive Behaviors

Verification: "Let me double-check whether Response A's tone truly aligns with the user's style ..."

Backtracking: "Initially, I thought Response B matched better, but upon reconsideration, that route doesn't hold due to its formal tone ..."

Subgoal Setting: "First, let's evaluate tone consistency, then move on to language fluency and finally check personal relevance ..."

Backward Chaining: "If Response A is indeed superior, it must fulfill all the criteria—let me trace back through each evaluation point to confirm ..."