Personalized LLM Decoding via Contrasting Personal Preference

Anonymous ACL submission

Abstract

As large language models (LLMs) are progressively deployed in various real-world applications, personalization of LLMs has become increasingly important. While various approaches to LLM personalization such as prompt-based and training-based methods have been actively explored, the development of effective decoding-time algorithms remains largely overlooked, despite their demonstrated potential. In this paper, we propose CoPE (**Co**ntrasting **P**ersonal Preference), a novel decoding-time approach applied after performing parameter-efficient fine-tuning (PEFT) on user-specific data. Our core idea is to leverage reward-guided decoding specifically for personalization by maximizing each user's implicit reward signal. We evaluate CoPE across five open-ended personalized text generation tasks. Our empirical results demonstrate that CoPE achieves strong performance, improving personalization by an average of 10.57% in ROUGE-L,without relying on external reward models or additional training procedures.

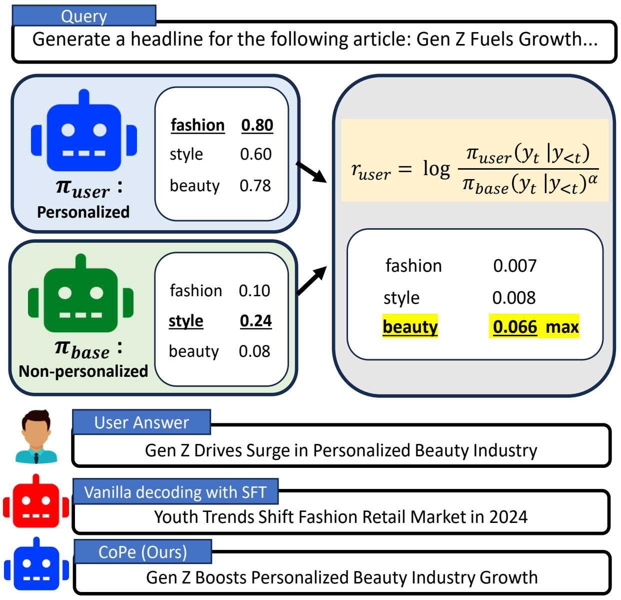


Figure 1: **Implicit reward maximization via contrastive preference**. Under an implicit reward model that leverages the interaction between a personalized and a non-personalized generic model, generated texts better align with user preferences.

1 Introduction

Personalization of large language models (LLMs) (Achiam et al., 2023; Team et al., 2023; Anthropic, 2024; Touvron et al., 2023) — the process of aligning model outputs with individual user preferences — has received growing attention as LLMs are increasingly deployed in real-world applications such as writing assistants (Mysore et al., 2024), content recommendation (Zhang et al., 2024), and review generation (Peng et al., 2024). Prompt-based personalization (Santurkar et al., 2023; Hwang et al., 2023), which augments a user query by retrieving prior interactions or constructing a summarized user profile, is arguably considered as one of the most straightforward approaches. However, its effectiveness is often limited by the absence of direct learning from user data. In contrast, training-based personalization (Zhao et al., 2024; Kim and Yang, 2025) more effectively captures user preferences by updating model parameters, but it introduces challenges such as catastrophic forgetting and increased computational costs. To mitigate these limitations, recent work such as One PEFT per User (Tan et al., 2024) has demonstrated that lightweight parameter-efficient fine-tuning (PEFT) offers an effective solution for personalizing LLMs. Unlike prior works mentioned above, we turn to a new perspective to effectively personalize LLMs.

In this work, we introduce CoPE (**Co**ntrasting **P**ersonal preference), a new paradigm for LLM personalization that operates at the decoding stage, applied after PEFT on user-specific data. At a high level, CoPE is a form of *reward-guided decoding* (Deng and Raffel, 2023; Khanov et al., 2024; Lightman et al., 2024), an approach that effectively steers LLM outputs toward desired properties (*e.g.*,

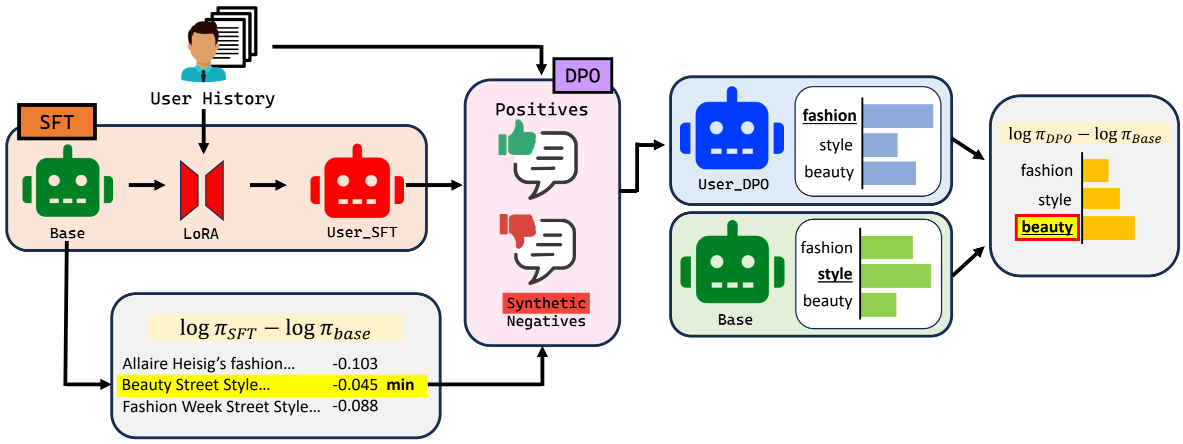


Figure 2: **Illustration of CoPe** (**Co**ntrasting Prefernece for **Pe**rsonalized LLM Decoding). The training pipeline (left) builds an expert user model via Direct Preference Optimization (DPO) with synthetic negatives. The reward-guided decoding method (right) contrasts this user model with a base model at the token level, maximizing implicit user reward during both training and decoding for improved personalization.

improved reasoning) by maximizing a reward function, adapted specifically for personalizing LLMs across varying contexts and user goals.

Unlike conventional reward-guided decoding methods, CoPe does not require an external reward model to estimate rewards. Instead, it leverages the implicit user reward signal, which can be efficiently approximated using the likelihoods from both the PEFT-tuned model and the original base model. Building on our key insight which connects this implicit reward to the objective of contrastive decoding (Li et al., 2023), the proposed CoPe can be implemented easily (see overview in Figure 2).

In addition, we further enhance PEFT for LLM personalization by encouraging the model to better capture the implicit user reward. The core idea is to contrast implicit rewards between a *positive* response (provided by the user) and a *negative* response (unlikely to be from the user, *e.g.*, from other users), using Direct Preference Optimization (DPO) (Rafailov et al., 2023). To avoid the practical and privacy challenges of relying on data from other users, we synthesize negative responses by generating outputs with low implicit rewards via Best-of-N sampling (Gui et al., 2024). This training method not only improves the effectiveness of PEFT, but also enhances the performance of our proposed reward-guided decoding by enabling more accurate modeling of the implicit user reward. An overview of full pipeline is shown in Figure 2.

We demonstrate the effectiveness of CoPe with the experiments in five different personalized open-ended text generation tasks from Language Model Personalization (LaMP) (Salemi et al., 2024) and LongLaMP (Kumar et al., 2024) benchmarks. Specifically, CoPe achieves an average relative improvement of 10.57% in ROUGE-L across all tasks, compared to the task-finetuned model. Notably, CoPe also outperforms a simply personalized model that lacks the contrastive mechanism, with an average ROUGE-L gain of 5.67% across tasks. Furthermore, the effectiveness of CoPe is well-generalized across different scales and types of state-of-the-art LLMs. Our robust experimental results show that the implicit reward maximization of CoPe further enhances alignment with individual user preferences. Together, these findings highlight CoPe as a promising approach for scalable and effective LLM personalization.

2 Related Works

LLM personalization. Given the diversity of user goals and preferences, various approaches to personalization of LLM have been explored. One common strategy is prompt-based personalization, wherein techniques such as retrieval-augmented generation (RAG) (Lewis et al., 2021) and prompt-augmented generation (PAG) (Richardson et al., 2023) dynamically inject user-specific context into each prompt at inference. However, these methods lack parametric memory and rely entirely on prompt construction, making them vulnerable to context length limitations and insufficient grounding. On the other hand, training-based personalization methods, which fine-tune the model on user-specific data, have demonstrated superior performance in capturing user preferences compared to prompting-based approaches (Zhao et al., 2024; Zhuang et al., 2024). Nevertheless, even these methods face several limitations. Firstly, these methods are computationally intensive, as they involve modifying model parameters. In fact, in the worst case, frequent retraining may be necessary to reflect evolving user preferences (Madotto et al., 2021). Moreover, these methods are susceptible to catastrophic forgetting—a phenomenon where adapting to new user data can lead the model to forget previously learned preferences or general knowledge (McCloskey and Cohen, 1989; de Masson d'Autume et al., 2019).

A recent and practical method to address these limitations is the utilization of lightweight parameter-efficient fine-tuning (PEFT), which offers an effective and scalable approach to personalizing LLMs (Zhang et al., 2024, 2025). Meanwhile, personalization at the decoding stage remains largely unexplored in existing methods. Motivated by this gap, we aim to address the aforementioned limitations through a decoding-based approach to personalization.

LLM decoding. Various decoding strategies have been explored and applied in LLMs to boost their performance. For instance, contrastive decoding has demonstrated strong effectiveness not only in open-ended text generation (Li et al., 2023), but also in reasoning (O'Brien and Lewis, 2023), retrieval-augmented generation (RAG) (Shi et al., 2023), and even multi-modal generation (Leng et al., 2023). On the other hand, reward-guided decoding has emerged as another promising approach, aiming to improve alignment and reasoning capabilities directly at the decoding stage, without additional model training. To further explain, reward-guided decoding guides the generation process using reward signals, offering a lightweight yet effective alternative for steering outputs toward desired behaviors (Deng and Raffel, 2023; Lightman et al., 2024). In fact, adaptive reward shaping, as proposed by Khanov et al. (2024), has also been shown to improve sample efficiency during decoding. Despite the growing interest in both decoding strategies and personalization, there is no prior work that effectively leverages decoding methods for personalization due to the challenge of modeling separate rewards for each user. In this aspect, we propose the first guided decoding approach for personalization that does not require any external reward models. Specifically, our method can be easily implemented using contrastive decoding, thereby enabling more practical and scalable deployment in real-world settings.

Preference learning. Preference learning is an approach that ensures alignment with human or task-specific preferences by leveraging relative feedback between outputs, rather than relying on absolute labels. One traditional approach to preference learning is Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), which involves fitting a reward model based on human-labeled comparisons and optimizing model policies through reinforcement learning. However, RLHF often requires complex and costly training procedures. To address this limitation, recent methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) simplify the process by directly fine-tuning models through binary classification between preferred and dispreferred outputs.

Building on these advances, we propose a personalized fine-tuning method that integrates preference learning by treating user profile responses as positive examples and non personalized outputs as negative examples. This training formulation supports contrastive decoding, due to the fact that maximization of implicit user reward is plausible both in the training and decoding section. In other words, this conceptual alignment between preference learning and contrastive decoding ensures consistency between training and inference, enabling more effective personalization without external reward models or additional training procedures.

3 CoPE: Contrasting Preference for Personalized LLM Decoding

In this section, we present our new decoding framework for LLM personalization by **Co**ntrasting **P**ersonal preference (CoPE). Our key idea is incorporating *implicit reward signals* for user preference to guide both training and inference. We first present our problem setup in Section 3.1. Next, we present the proposed decoding scheme, CoPE, in Section 3.2. Lastly, in Section 3.3, we present our training scheme to further improve PEFT for the personalization, by explicitly maximizing user reward based on the synthetic negative response.

3.1 Preliminary

Let us first assume that we have the historical interaction data $H_{\texttt{user}} = \{(x^i, y^i)\}_{i=1}^N$ for a target user. Then, for a given input query x, the goal of LLM personalization is to generate a personalized output y from LLM π that aligns with the user's preferences and behaviors exhibited in $H_{\texttt{user}}$. A representative approach for LLM personalization is to adapt a generic pre-trained LLM $\pi_{\texttt{base}}$ using parameter-efficient fine-tuning (PEFT) techniques, such as LoRA (Hu et al., 2021).

Formally, let $\Delta_{\texttt{user}}$ denote the user-specific PEFT module.[1] The personalized model is then defined as $\pi_{\texttt{user}} = \pi_{\texttt{base}} + \Delta_{\texttt{user}}$, such that only $\Delta_{\texttt{user}}$ is optimized using the user's data $H_{\texttt{user}}$. For example, Tan et al. (2024) optimizes $\Delta_{\texttt{user}}$ on $H_{\texttt{user}}$ via conventional supervised fine-tuning (SFT) that minimizes cross-entropy between the output of $\pi_{\texttt{user}}(x^i)$ and ground-truth label y^i. After optimizing $\Delta_{\texttt{user}}$, $\pi_{\texttt{user}}$ is expected to generate the responses that align with the user's preferences.

3.2 Optimizing personal preference via contrastive decoding with PEFT

Assume that we have access to a generic base model $\pi_{\texttt{base}}$ and a personalized model $\pi_{\texttt{user}}$. Then, to generate response y that better align with user's preferences for a given test query x, CoPE adopts a reward-guided decoding strategy that contrasts the token-level likelihoods under these two models.

Let $y_{<t} = (y_1, \dots, y_{t-1})$ denote the partial output sequence at decoding step t. Then, following Li et al. (2023), we first define a plausibility-constrained candidate set of next tokens as:

$$\mathcal{V}_{\text{head}}^t = \{y_t \in \mathcal{V} \,|\, \pi_{\texttt{user}}(y_t \mid y_{<t}) \geq \tau_t\}, \quad (1)$$

where $\tau_t := \tau \cdot \max_{w \in \mathcal{V}} \pi_{\texttt{user}}(w \mid y_{<t})$ is an adaptive threshold determined by a hyperparameter $\tau \in [0, 1]$ and $\mathcal{V}$ denotes the vocabulary for $\pi_{\texttt{user}}$. For each candidate token $y_t \in \mathcal{V}_{\text{head}}^t$, we compute an implicit user reward by contrasting its likelihoods under the personalized and base models:

$$r_{\texttt{user}}(y_t) = \log \frac{\pi_{\texttt{user}}(y_t \mid y_{<t})}{\pi_{\texttt{base}}(y_t \mid y_{<t})^\alpha}, \quad (2)$$

where $\alpha \geq 0$ is a contrastive weight hyperparameter. This reward encourages the selection of tokens that are strongly preferred by the personalized model while being penalized under the base model, yields the outputs that are both user-aligned and distinctive. Finally, the next token y_t^* is selected which maximizes the implicit user reward:

$$y_t^* = \arg\max_{y_t \in \mathcal{V}_{\text{head}}^t} r_{\texttt{user}}(y_t). \quad (3)$$

Rationale behind implicit user reward. Here, we present the theoretical intuition behind our proposed implicit user reward $r_{\texttt{user}}$ (Eq. 2). To this end, we revisit the concept of *implicit reward* introduced in DPO (Rafailov et al., 2023), which has been widely adopted in the LLM alignment literature (Chen et al., 2025; Kim et al., 2025; Cui et al., 2025). Specifically, Rafailov et al. (2023) show that the reward function r, which captures human preferences, can be approximated under RLHF framework (Ouyang et al., 2022) as the log-likelihood ratio between the optimal (aligned) LLM policy π_r and a reference policy $\pi_{\texttt{ref}}$:

$$r(y) \approx \beta \cdot \log \frac{\pi_r(y)}{\pi_{\texttt{ref}}(y)}, \quad (4)$$

where β is a hyperparameter controlling the strength of KL regularization in RLHF.[2] This derivation of implicit reward enables reward modeling without an explicit reward model using only the relative likelihoods under two LLM policies, and yields much efficient preference learning algorithm, called DPO (see details in Appendix F).

In our setting, however, the personalized model $\pi_{\texttt{user}}$ is not trained with explicit KL regularization, as in standard RLHF. Nevertheless, we argue that the PEFT used for training $\pi_{\texttt{user}}$ implicitly imposes a similar constraint. For example, in LoRA (Hu et al., 2021), only the newly introduced low-rank matrices are updated, while the original model parameters remain fixed. This architectural constraint implicitly regularizes the updated model, preventing it from deviating significantly from the base model. As a result, the personalized model $\pi_{\texttt{user}}$ trained via PEFT remains close to the base model $\pi_{\texttt{base}}$, and the log-likelihood ratio between them can serve as a valid proxy for an implicit reward signal—namely, $r_{\texttt{user}}$.

Interestingly, we note that this formulation, based on the ratio of log-likelihoods between two models, also appears in contrastive decoding (Li et al., 2023). In this sense, our insight reveals

[1] In this work, we only consider LoRA.

[2] While y is generated for input x, we omit this in Eq. 4 for the simplicity.

a novel connection between two popular decoding paradigms, contrastive decoding and reward-guided decoding. Following Li et al. (2023), we additionally introduce a hyperparameter α to control the strength of contrastive adjustment during decoding and further enhance personalization.

3.3 Aligning PEFT to user preference via DPO with synthetic negative response

While CoPE effectively maximizes the implicit user reward during decoding with the personalized model $\pi_{\texttt{user}}$, its performance can be further improved by explicitly aligning $\pi_{\texttt{user}}$ with the user's preferences during training. One natural approach is to apply preference learning algorithms such as RLHF or DPO. However, a key practical challenge is a lack of negative examples ((*i.e.*, responses unlikely to come from the user) in the user dataset $H_{\texttt{user}}$. To address this, we propose a simple yet effective approach that synthesizes negative examples leveraging the implicit user reward $r_{\texttt{user}}$. Specifically, for each train query $x^i \in H_{\texttt{user}}$, we sample K candidate responses $\{\widetilde{y}^{i,1}, \ldots, \widetilde{y}^{i,K}\}$ from the generic base model $\pi_{\texttt{base}}$. Among these, we select the response with the lowest implicit user reward, *i.e.*, the one most unlikely from the user:

$$\widetilde{y}^{i,*} = \arg \min_{y \in \{\widetilde{y}^{i,1}, \ldots, \widetilde{y}^{i,K}\}} \sum_t r_{\texttt{user}}(y_t), \quad (5)$$

where the contrastive weight α is set to 1. Then, we construct a preference dataset $\mathcal{D}_{\text{pref}} := \{(x^i, y^i_{\text{pos}}, y^i_{\text{neg}})\}_{i=1}^N$ where (x^i, y^i_{pos}) from H_u, *i.e.*, $y^i_{\text{pos}} = y^i$, and $y^i_{\text{neg}} = \widetilde{y}^{i,*}$.

Using this preference dataset $\mathcal{D}_{\text{pref}}$, we further fine-tune $\pi_{\texttt{user}}$ with the following DPO loss:

$$\mathcal{L}_{\text{dpo}} = \mathop{\mathbb{E}}_{(x, y^{\text{pos}}, y^{\text{neg}}) \in \mathcal{D}_{\text{pref}}} \left[-\log \sigma \left(\beta \cdot r_{\text{dpo}} \right) \right], \quad (6)$$

where $r_{\text{dpo}} = r_{\texttt{user}}(y^{\text{pos}}) - r_{\texttt{user}}(y^{\text{neg}})$, and $\sigma(\cdot)$ denotes the sigmoid function. Optimizing this loss encourages the personalized model $\pi_{\texttt{user}}$ to assign higher reward to user-aligned responses compared to generic ones. This better modeling of implicit user reward further improves the effectiveness of reward-guided decoding through CoPE.

4 Experiments

In this section, we design our experiments to investigate the following questions:

- Does CoPE yield better personalization than existing baselines? (Table 1)
- Is CoPE applicable to models of varying architectures and parameter scales? (Table 2)
- How different components in CoPE contribute to personalization performance? (Table 3)
- How sensitive is the performance of CoPE to different configuration settings? (Figure 3)

4.1 Setups

Datasets and metrics. We evaluate the effectiveness of CoPE primarily on personalized text generation tasks from the Large Language Model Personalization (LaMP) (Salemi et al., 2024) and LongLaMP (Kumar et al., 2024) benchmarks. Specifically, we focus on the following five tasks: news headline generation (LaMP 4), scholarly title generation (LaMP 5), abstract generation (LongLaMP 2), review writing (LongLaMP 3), and topic writing (LongLaMP 4). Throughout our framework, we follow the setup of an earlier work (Tan et al., 2024): we use 100 users with the longest activity histories as the test set, and the remaining users to train the task-adapted base model.

For evaluation, we mainly report ROUGE-1 and ROUGE-L scores across all tasks, which serve as standard metrics to measure the content relevance and structural similarity between the generated and ground-truth texts.

Baselines. We compare CoPE against several baselines to generate personalized responses from LLMs as follows: (1) *Base* – generation using a vanilla model without any supervised fine-tuning; (2) *RAG* (Lewis et al., 2021) – a retrieval-augmented generation method that directly injects user-related histories into the prompt without additional training; (3) *PAG* (Richardson et al., 2023) – a prompt-augmented generation approach that additionally incorporates user profiles to the prompt; (4) *TAM* (Tan et al., 2024) – generation with a task-adapted model trained on data from users excluding the test user, allowing familiarity with the task but lacking personalization; (5) *OPPU* (Tan et al., 2024) – generation with a personalized model equipped with user-specific adapters trained via simple supervised fine-tuning on user data.

Implementation details. Under the methods including training step (TAM, OPPU, CoPE), all models are trained using AdamW (Loshchilov and Hutter, 2019) with a weight decay of 0.01. Linear learning rate decay was used with a warm-up ratio of 0.1. The batch size for the initial training of the

Table 1: **Main Results.** ROUGE-1 and ROUGE-L scores are reported for five tasks: Abstract Generation, Review Writing, and Topic Writing from LongLaMP; News Headline and Scholarly Title from LaMP. All experiments are conducted using `Mistral-7B-Instruct-v0.3`.

Methods	**Abstract Generation**		**Review Writing**		**Topic Writing**		**News Headline**		**Scholarly Title**	
	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L
Base	0.341	0.186	0.287	0.126	0.246	0.105	0.119	0.105	0.409	0.324
RAG	0.347	0.205	0.272	0.128	0.243	0.115	0.141	0.124	0.425	0.347
PAG	0.344	0.186	0.256	0.125	0.262	0.107	0.118	0.102	0.372	0.289
TAM	0.357	0.204	0.289	0.122	0.253	0.107	0.200	0.179	0.514	0.456
OPPU	0.378	0.218	0.319	0.134	0.278	0.112	0.203	0.182	0.510	0.454
CoPE (Ours)	**0.392**	**0.239**	**0.335**	**0.146**	**0.281**	**0.120**	**0.205**	**0.184**	**0.519**	**0.461**

task-adapted model is set to 8, while subsequent training stages use 4 to better capture the style of each user. Supervised training is conducted for 2 epochs with a learning rate of 1e-4 for LongLamP and 1e-5 for Lamp. Subsequently, DPO training uses a 5e-6 learning rate for 1 epoch on LongLaMP and 2 epochs on LaMP. Also, we note that OPPU is continuously applied after TAM, following Tan et al. (2024). Similar to this, the proposed DPO step (Eq. 6) is applied after OPPU (see Figure 1).

All of the experiments are conducted using `Mistral-7B-Instruct-v0.3`,[3] except for those reported in Table 2. Greedy decoding is used to eliminate randomness, except for negative sample generation. In this case, we use `vLLM` (Kwon et al., 2023) with a temperature of 1.0 for faster decoding, generate $K = 3$ candidates using the task-adapted model, and select the final negative using the reward function (Eq. 5). For DPO training (Rafailov et al., 2023), we set coefficient for KL regularization $\beta = 3.0$ for LaMP tasks and $\beta = 0.05$ for LongLaMP tasks. At this point, we treat the task-adapted model as the base model $\pi_{\texttt{base}}$ and the DPO-trained model as the user model $\pi_{\texttt{user}}$ in Eq. 2. To implement the proposed reward-guided decoding (Eq. 3), we adopt the contrastive decoding (Li et al., 2023), with a plausibility threshold of $\tau = 0.1$ for both LaMP and LongLaMP tasks. The contrastive weight α is set to 0.3 for LaMP and 0.1 for LongLaMP tasks. We apply a repetition penalty of 1.0 for LaMP and 7.0 for LongLaMP, after observing that these values offered acceptable control over repetition in preliminary experiments.

4.2 Main results

Table 1 summarizes the experimental results on five personalized open-ended text generation tasks. First, it is observed that the effectiveness of prompting-based methods is indeed limited. In particular, RAG and PAG exhibit limited improvement compared to training-based approaches, and even they are sometimes worse than Base method, which does not apply any personalization technique. This observation validates the necessity for developing a training-based method like the proposed framework. Next, the experimental results in Table 1 also demonstrate that CoPE consistently outperforms all baseline methods across all tasks and metrics. For instance, CoPE achieves an average relative improvement of 10.57% in ROUGE-L compared to the task-adapted model, TAM. Notably, CoPE even outperforms a personalized model OPPU that relies solely on explicit user-specific fine-tuning, with average relative improvement of 5.67% in ROUGE-L. These results highlight the effectiveness of our framework, which maximizes implicit reward signals to better align with user preferences.

We further observe a task-specific trend across benchmarks. While RAG shows some effectiveness in LaMP tasks, its performance declines in the LongLaMP setting. For instance, RAG scores 5.23% lower than Base in review writing (ROUGE-1) and 1.22% lower in topic writing (ROUGE-1). This highlights the increased difficulty of LongLaMP tasks, where simple retrieval of user history is no longer sufficient. In contrast, CoPE remains effective even in this more demanding setting. In fact, CoPE demonstrates a significantly higher relative improvement in the more challenging LongLaMP setting—achieving a 16.33% gain in ROUGE-L over the task-adapted model, compared to just 3.89% in LaMP. This suggests that LongLaMP tasks may offer greater room for personalization gains when properly modeled. We also note that the tasks in LongLaMP tend to involve more subjective or user-specific expression, mak-

[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Table 2: **Compatibility of CoPE.** ROUGE-L scores on the Abstract Generation task across different LLMs.

Methods	LLaMA 3.1-8B	Gemma 3-4B	Qwen 2.5-1.5B
Base	0.172	0.135	0.130
RAG	0.183	0.170	0.128
PAG	0.183	0.169	0.130
TAM	0.198	0.181	0.150
OPPU	0.202	0.194	0.163
CoPE (Ours)	**0.261**	**0.237**	**0.233**

Table 3: **Ablation study.** The effects of contrastive decoding (CD) and direct preference optimization (DPO).

	CD	DPO	Abstract Generation		News Headline	
			ROUGE-1	ROUGE-L	ROUGE-1	ROUGE-L
OPPU	✗	✗	0.378	0.218	0.203	0.181
	✓	✗	0.385	0.232	0.204	0.183
	✗	✓	0.386	0.230	0.203	0.182
CoPE (Ours)	✓	✓	0.392	0.239	0.205	0.184

ing them especially well-suited for personalized generation when guided by an effective framework like CoPE.

4.3 Additional analyses

Here, we provide additional analyses of CoPE with the experiments on abstract generation from LongLaMP and news headline generation from LaMP. More analyses are in Appendix E.

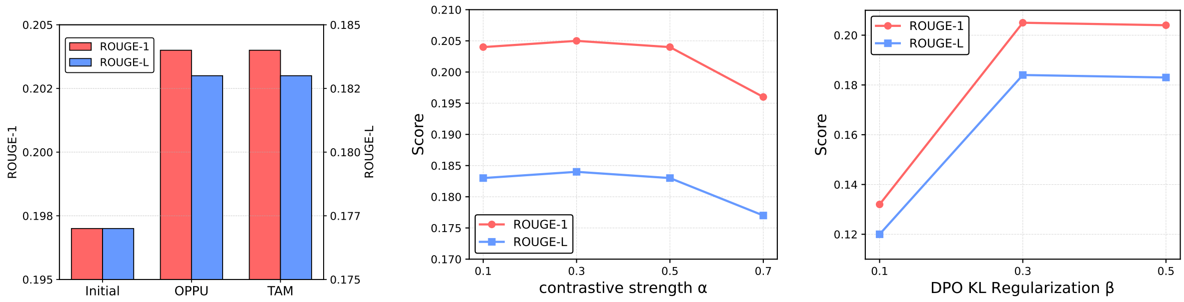
Generalization to various LLMs. In this section, we explore the applicability of CoPE to various LLMs and sizes. Results are presented in Table 2. The experimental results validate that CoPE generalizes well across a diverse range of LLMs, including LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024), Gemma-3-4B-it (Team et al., 2025), and Qwen2.5-1.5B-Instruct (Qwen et al., 2025). Compared to TAM, CoPE significantly improves ROUGE-L by 31.8% on LLaMA-3.1-8B , 30.9% on Gemma-3-4B-it, and 55.3% on Qwen2.5-1.5B. Similarly, compared to OPPU, CoPE achieves a relative improvement of 29.2% on LLaMA-3.1-8B, 22.2% on Gemma-3-4B-it, and 42.9% on Qwen2.5-1.5B. These consistent improvements suggest that CoPE does not simply rely on a specific environment setting. Instead, our framework is generalizable and flexible with respect to model architecture and parameter scale. This makes CoPE a broadly applicable framework for deployment across diverse LLMs .

Ablation study. We now proceed to validate the individual components of CoPE. To assess the contribution of each component to the overall performance of CoPE, we perform a detailed ablation study. For this analysis, we primarily conducted experiments on abstract generation and news headline tasks, serving as representative tasks for LongLaMP and LaMP, respectively. The results are presented in Table 3. Here, it is observed that adding each component progressively improves the performance. Comparing with the OPPU baseline, applying only contrastive decoding increases the scores in both tasks, as it encourages the model to generate outputs that are more distinguishable from less preferred candidates. Meanwhile, in the training side, introducing only preference-aligned training also improves the performance of the model, as it guides the model to internalize user preferences by learning to favor higher-quality responses over inferior ones during fine-tuning.

Finally, when combining these components to formulate an implicit reward maximization objective both during training and decoding, we observe the highest performance. These results indicate that each component independently contributes to performance improvements, and their integration yields the most substantial gains across tasks. This is because both components work synergistically to align model outputs with implicit user preferences: training encourages the model to internalize preference signals through comparisons between better and worse responses, while decoding promotes outputs that more closely reflect these learned preferences at inference time. Together, they implicitly guide the model to maximize a user-aligned reward signal, even in the absence of explicit supervision from external model.

Sensitivity of CoPE. Figure 3 presents a sensitivity analysis of key components in the proposed framework. In this section, we conduct experiments on the news headline generation task, chosen for its shorter runtime, to explore the behavior of CoPE under different settings.

We begin by examining the choice of base model for contrastive decoding (*i.e.*, $\pi_{\texttt{base}}$ to calculate likelihood for the denominator in Eq. 2). We first note that TAM is originally used as the base model in CoPE, as it yields better understanding of the target task. To investigate this, we performed experiments by varying the base models from TAM to init (*i.e.*, initial mistral model) and OPPU (*i.e.*, after adaption to user and before DPO). The results are presented in Figure 3(a), and one can verify that the current design choice is the best and using



(a) Choice of base model π_{base} (b) Contrasting strength α (c) DPO KL regularization β

Figure 3: **Different hyperparameters.** (a) Performance variation by base model choice. (b) Effect of contrastive strength α. (c) Effect of KL regularization β in DPO. ROUGE-1 and ROUGE-L scores are reported.

News Headline	
Query	`Generate a headline for the following article: When we first saw Michael H. Rohde's photography series \"Below The Floor\" on Design-Milk.com, we were floored. (Pun intended`
User Answer	`Michael H. Rohde, German Photographer, Shoots Breathtaking Series 'Below The Floor' (PHOTOS)`
TAM	`Craft Of The Day: Create A Floating Photo Gallery With This DIY`
OPPU	`'Below The Floor' Photography Series Reveals The Hidden Beauty In The Places We Walk On`
CoPe (ours)	`'Below The Floor' Photography Series By Michael H. Rohde Is A Whole New Perspective On The World (PHOTOS)`

Figure 4: **A qualitative example of CoPE on the News Headline task (LaMP 4).** the output of CoPE contains more words that align with the user gold response compared to TAM and OPPU. Words overlapping with the User Answer are highlighted. Additional qualitative examples from other tasks are provided in Appendix G.

init is the worst. The findings suggest that using either OPPU or TAM as the base model yields the best performance. We hypothesize that these models help isolate and downweigh non-personalized features, allowing user-specific characteristics to be more prominently reflected.

Next, we analyze the sensitivity of CoPE to two key hyperparameters: the contrastive strength (α) and the KL regularization coefficient (β) in preference-aligned training. These two hyperparameters are crucial in the decoding and training components of our framework, respectively. Figure 3(b) shows the effect of varying the contrastive strength α under fixed $\beta = 3.0$. We observe that CoPE performs reliably across a range of α values, with a slight peak around $\alpha = 0.3$. While stronger contrastive signals may lead to marginal decreases in output quality, the overall performance remains stable, demonstrating the robustness of CoPE to decoding-time variations.

Figure 3(c) illustrates the impact of varying the KL regularization coefficient β during training. As β increases from 0.1 to 0.3, both ROUGE-1 and ROUGE-L scores improve, after which performance growth starts to hinder. This suggests that CoPE benefits from moderate regularization while remaining resilient to further increases. These results indicate that CoPE performs consistently well across a range of configurations, underscoring its robustness and reliability without signs of overfitting to specific hyperparameter values.

5 Conclusion

In this work, we propose CoPE, the first decoding-based framework for personalizing LLMs. Specifically, CoPE is a reward-guided decoding approach that maximizes implicit rewards of each user, thereby enhancing personalization without requiring external reward models. Our comprehensive experiments show that CoPE consistently outperforms various baselines across multiple tasks and also is well-generalized to various types and scales of LLMs. Consequently, these results demonstrate that it is not only effective but also a practical framework for decoding-time personalization.

Limitations

While the proposed CoPE shows consistent improvements in personalized generation, it applies a fixed set of hyperparameters (*e.g.*, learning rate, batch size, LoRA rank) across all users, regardless of the size or characteristics of each user's dataset. This uniform setting may not be optimal, especially when user data varies widely in volume or domain. Future work should investigate adaptive strategies that dynamically adjust hyperparameters based on user-specific data profiles. In addition, we only consider LoRA as PEFT for the experiments, but different PEFT approaches (Li and Liang, 2021; Liu et al., 2022) are also considerable. Nevertheless, we expect that CoPE is also easily deployed for these approaches, as our method does not explicitly depend on them and PEFT methods commonly assume architectural constraint similar to LoRA.

Ethics Statement

Our work investigates LLM adaptation to specific user, using PEFT methods such as LoRA. To ensure user privacy, our approach does not store or expose raw user data, and only updates a small number of task- and user-specific parameters. In addition, all negative samples used for preference optimization are synthetically generated from a base model, rather than extracted from real user outputs. While we do not explicitly evaluate membership inference risks, the structure of our approach, especially the use of PEFT and synthetic negatives, may offer improved privacy protection compared to full-model fine-tuning. All datasets and models used in this study are publicly available and used in accordance with their intended purposes. We also used an AI assistant (ChatGPT) to refine the writing during manuscript preparation.

References

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

Anthropic. 2024. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Changyu Chen, Zichen Liu, Chao Du, Tianyu Pang, Qian Liu, Arunesh Sinha, Pradeep Varakantham, and Min Lin. 2025. Bootstrapping language models with dpo implicit rewards. In *International Conference on Learning Representations (ICLR)*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, and et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org. Accessed 14 April 2023.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Preprint*, arXiv:1906.01076.

Haikang Deng and Colin Raffel. 2023. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. *arXiv preprint arXiv:2310.09520*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lin Gui, Cristina Gârbacea, and Victor Veitch. 2024. Bonbon alignment for large language models and the sweetness of best-of-n sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. 2024. Args: Alignment as reward-guided search. In *International Conference on Learning Representations (ICLR)*.

Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2025. Spread preference annotation: Direct preference judgment for efficient llm alignment. In *International Conference on Learning Representations (ICLR)*.

Jaehyung Kim and Yiming Yang. 2025. Few-shot personalization of llms with mis-aligned responses. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *Preprint*, arXiv:2311.09180.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *Preprint*, arXiv:2309.09117.

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. *Preprint*, arXiv:2203.02155.

Qiyao Peng, Hongtao Liu, Hongyan Xu, Qing Yang, Minglai Shao, and Wenjun Wang. 2024. Review-llm: Harnessing large language models for personalized review generation. *Preprint*, arXiv:2407.07487.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*.

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

Stephen Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241. Springer.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the International Conference on Machine Learning (ICML)*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *Preprint*, arXiv:2305.14739.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Chiyu Zhang, Yifei Sun, Jun Chen, Jie Lei, Muhammad Abdul-Mageed, Sinong Wang, Rong Jin, Sem Park, Ning Yao, and Bo Long. 2024. Spar: Personalized content-based recommendation via long engagement attention. *Preprint*, arXiv:2402.10555.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2025. Personalization of large language models: A survey. *Preprint*, arXiv:2411.00027.

Siyan Zhao, John Dang, and Aditya Grover. 2024. Group preference optimization: Few-shot alignment of large language models. In *International Conference on Learning Representations (ICLR)*.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. Hydra: Model factorization framework for black-box llm personalization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Datasets

For the experiments, we focus mainly on the text generation tasks provided in the LaMP (Salemi et al., 2024) and LongLaMP (Kumar et al., 2024) benchmarks. Following these benchmarks, we use ROUGE-1 and ROUGE-L as metrics for evaluation. Detailed descriptions of each task are as follows.

LaMP 4: News Headline. This task evaluates the ability of a model to generate headlines for news articles, conditioned on an author profile containing historical article-title pairs, thereby capturing distinctive stylistic patterns in journalistic writing.

LaMP 5: Scholarly Title. This task assesses the capacity of a model to generate appropriate titles for scholarly article abstracts conditioned on an author profile of historical article-title pairs, reflecting distinct academic writing style.

LongLaMP 2: Abstract Generation. This task focuses on evaluating the proficiency of a model in generating scientific abstracts given paper titles and keywords by leveraging an author profile of previous publications to emulate characteristic academic writing style and domain-specific terminology

LongLaMP 3: Review Writing. This task tests the ability of a model to generate comprehensive product reviews based on product specifications and user experiences, conditioned on a user profile of review history to reflect distinctive evaluative style and subjective perspective.

LongLaMP 4: Topic Writing. This task evaluates the capability of a model to generate Reddit post content based on post summaries while maintaining the unique writing style of individual users, requiring the generation of content from a given summary conditioned on a user profile containing their previous posts.

B Baselines Details

Detailed explanations for each baseline are provided below. Black boxes indicate vanilla models and prompt-base baselines (*i.e.*, training-free), while white boxes represent training-base ones.

- ■ **Base model** refers to the generation with the original, unmodified LLM without any task-specific fine-tuning or additional conditioning. It represents the vanilla, pre-trained model as released.
- ■ **RAG: Retrieval-Augmented Generation** (Lewis et al., 2021) is a method that retrieves user-related history records and directly incorporates them into the prompt. Following the setup in LaMP (Salemi et al., 2024), we retrieve the top-k history records for each user. In our experiments, we set $k = 3$, meaning the three most relevant records are selected using BM25 (Robertson and Walker, 1994)—a standard keyword-based retrieval method. We implement BM25 using the `rank_bm25` library with `BM25Okapi`.
- ■ **PAG: Profile-Augmented Generation** (Richardson et al., 2023) is a technique for personalizing LLM outputs by conditioning on structured user profiles. Following the prior work (Tan et al., 2024), we generate user profiles using the `vicuna-7B` model (Chiang et al., 2023), based on the past responses of a typical user. Each profile captures key stylistic characteristics, such as tone, lexical choices, and recurring templates. The model then uses these profiles as a guide to generate output that aligns closely with the user style.
- ❑ **TAM: Task Adapted Model** (Tan et al., 2024) is trained on data from users other than the selected 100 test users. The objective of this model is to adapt the base model to the task in a general manner via LoRA (Low-Rank Adaptation) (Hu et al., 2021), enabling it to understand the task setup without being exposed to the specific styles of the target users.
- ❑ **OPPU: One PEFT Per User Model** (Tan et al., 2024) is a baseline that fine-tunes the LoRA adapter from the TAM model on individual users. Specifically, the historical data of each user is used to fine-tune the LoRA adapter from the TAM model, resulting in 100 separate personalized adapters. Intuitively, each LoRA adapter is specialized to learn the unique style of a specific user.

C Prompts

Below are prompts used in our experiments. Note that the text in `{BRACES}` is a placeholder for user- and query-specific input.

News Headline

You are a news headline generator.
Generate a headline for the following article.

Table 4: **Dataset statistics.** Base LLM training corresponds to *TAM*, and Personal PEFT training to OPPU.

Task	Base LLM Training (TAM)			Personal PEFT Training (OPPU)		
	#Train	L_{in}	L_{out}	#Profile	L_{in}	L_{out}
Abstract Generation	31,808	70.4 ± 13.3	233.1 ± 117.5	$1{,}296.7 \pm 446.4$	604.4 ± 142.7	210.5 ± 92.8
Review Writing	19,649	185.1 ± 109.0	407.2 ± 299.5	759.3 ± 324.2	$1{,}143.0 \pm 343.3$	511.8 ± 294.2
Topic Writing	21,119	56.6 ± 54.8	358.3 ± 316.9	260.6 ± 314.0	759.8 ± 321.8	358.3 ± 255.4
News Headline Generation	7,275	53.6 ± 19.0	15.5 ± 6.0	270.1 ± 182.1	92.2 ± 11.3	18.6 ± 5.2
Scholarly Title Generation	16,076	230.6 ± 97.9	17.9 ± 6.1	444.0 ± 121.6	266.4 ± 85.9	16.4 ± 5.8

article: {ARTICLE}
headline:

Scholarly Title
You are a scholarly title generator.
Generate a title for the following abstract of a paper.
abstract: {ABSTRACT}
title:

Abstract Generation
You are an abstract writer.
Generate the review text written by a reviewer who has a given an overall rating of "{RATING}" for a product with description "{PRODUCT}". The summary of the review text is "{SUMMARY}".
Review:

Review Writing
You are a review writer.
Generate an abstract for the title “{TITLE}”.
Abstract:

Topic Writing
You are a creative content generator for Reddit posts.
Generate the content for a reddit post.
post: {POST}
content:

D Chat Templates

In this section, we provide the chat templates we applied for experiments. We also include the chat templates of other LLMs used to test the generalization of COPE.

Mistral-7B-Instruct-v0.3

```
MISTRAL_CHAT_TEMPLATE = """
{% if messages[0]['role'] == 'system' %}
{% set loop_messages = messages[1:] %}
{% set system_message = messages[0]['content'].
    strip() + '\n' %}
{% else %}
{% set loop_messages = messages %}
{% set system_message = '' %}
{% endif %}
{% for message in loop_messages %}
   {% if loop.index0 == 0 %}
      {% set content = system_message + message
          ['content'] %}
   {% else %}
      {% set content = message['content'] %}
   {% endif %}
   {% if message['role'] == 'user' %}
      {{ '[INST] ' + content.strip() + ' [/INST
          ]' }}
   {% elif message['role'] == 'assistant' %}
      {{ ' ' + content.strip() + ' ' +
          eos_token }}
   {% endif %}
{% endfor %}
"""
```

LLaMA-3.1-8B-Instruct

```
LLAMA_CHAT_TEMPLATE = """
 {{- bos_token }}
 {%- if messages[0]['role'] == 'system' %}
    {%- set system_message = messages[0]['
        content'].strip() %}
    {%- set loop_messages = messages[1:] %}
    {{- '<|start_header_id|>system<|
        end_header_id|>\\n\\n' + system_message
         + '<|eot_id|>' }}
 {%- else %}
    {%- set loop_messages = messages %}
 {%- endif %}
 {%- for message in loop_messages %}
    {%- if message['role'] == 'user' %}
       {{- '<|start_header_id|>user<|
           end_header_id|>\\n\\n' + message['
           content'].strip() + '<|eot_id|>' }}
    {%- elif message['role'] == 'assistant' %}
       {{- '<|start_header_id|>assistant<|
           end_header_id|>\\n\\n' + message['
           content'].strip() + '<|eot_id|>' }}
    {%- endif %}
 {%- endfor %}
 {%- if add_generation_prompt %}
    {{- '<|start_header_id|>assistant<|
        end_header_id|>\\n\\n' }}
 {%- endif %}"""
```

GEMMA-3-4B-it

```
GEMMA_CHAT_TEMPLATE = """
"{% set bos_token = '<bos>' %}
{% set eos_token = '<eos>' %}

{{ bos_token }}
{% if messages[0]['role'] == 'system' %}
  {{ 'System: ' + messages[0]['content'].strip()
       + '\n' }}
```

```
  {% set loop_messages = messages[1:] %}
{% else %}
  {% set loop_messages = messages %}
{% endif %}

{% for message in loop_messages %}
  {% if message['role'] == 'user' %}
    {{ 'User: ' + message['content'].strip() + '\
        n' }}
  {% elif message['role'] == 'assistant' %}
    {{ 'Assistant: ' + message['content'].strip()
        + eos_token + '\n' }}
  {% endif %}
{% endfor %}
{{ 'Assistant:' }}"
"""
```

Qwen2.5-1.5B-Instruct

```
QWEN_CHAT_TEMPLATE = ''' {%- if messages[0]['
     role'] == 'system' %}
     {{- '<|im_start|>system\\n' + messages[0]['
          content'].strip() + '<|im_end|>\\n' }}
     {%- set loop_messages = messages[1:] %}
 {%- else %}
     {%- set loop_messages = messages %}
 {%- endif %}
 {%- for message in loop_messages %}
     {%- if message['role'] == 'user' %}
         {{- '<|im_start|>user\\n' + message['
              content'].strip() + '<|im_end|>\\n'
              }}
     {%- elif message['role'] == 'assistant' %}
         {{- '<|im_start|>assistant\\n' + message
              ['content'].strip() + '<|im_end|>\\n
              ' }}
     {%- endif %}
 {%- endfor %}
 {%- if add_generation_prompt %}
     {{- '<|im_start|>assistant\\n' }}
 {%- endif %}
'''
```

E More Quantitative Results

In this section, we provide more quantative results. First, in Table 5, we present the results under various LLMs on Abstract Generation using Rouge-1, instead of Rouge-L in Table 2. One can verify that CoPE significantly improve Rouge-1 as well. Next, in Tables 7 and 6, we present the results on News Headline Generation using Rouge-L and Rouge-1, respectively. Here, it is observed that the proposed CoPE is continuously effective to improve the performance.

Table 5: **Compatibility of CoPE.** ROUGE-1 scores on the Abstract Generation task across different LLMs.

Methods	LLaMA 3.1-8B	Gemma 3-4B	Qwen 2.5-1.5B
Base	0.340	0.270	0.278
RAG	0.330	0.295	0.240
PAG	0.333	0.292	0.241
TAM	0.355	0.326	0.298
OPPU	0.363	0.347	0.304
CoPE (Ours)	**0.417**	**0.393**	**0.384**

Table 6: **Compatibility of CoPE.** ROUGE-1 scores on the News Headline Generation task across different LLMs.

Methods	LLaMA 3.1-8B	Gemma 3-4B	Qwen 2.5-1.5B
Base	0.127	0.070	0.117
RAG	0.146	0.098	0.136
PAG	0.129	0.099	0.128
TAM	0.188	0.161	0.142
OPPU	0.191	0.164	0.143
CoPE (Ours)	**0.211**	**0.168**	**0.147**

Table 7: **Compatibility of CoPE.** ROUGE-L scores on the News Headline Generation task across different LLMs.

Methods	LLaMA 3.1-8B	Gemma 3-4B	Qwen 2.5-1.5B
Base	0.110	0.063	0.104
RAG	0.129	0.089	0.121
PAG	0.112	0.089	0.114
TAM	0.169	0.144	0.127
OPPU	0.171	0.147	0.127
CoPE (Ours)	**0.190**	**0.151**	**0.131**

F Background for RLHF and DPO

Let us denote LLM as π_θ, which generates an output sequence (*e.g.*, response) y for a given input sequence (*e.g.*, prompt) x, *i.e.*, $y \sim \pi_\theta(\cdot|x)$. Then, the goal of LLM alignment is to make π_θ provide human-aligned responses to various input prompts. To this end, let assume that the preference dataset $\mathcal{D} = \{(x, y_l, y_w)\}$ is available which consists of the triplets of input prompt x, preferred response y_w, and dispreferred response y_l. Here, the preference labels were annotated by a ground truth annotator, that is usually a human expert.

Reward modeling and RL fine-tuning. Since a pairwise preference between y_w and y_l is hard to model directly, one of the common practices is introducing reward function $r(x, y)$ and modeling the preference based on this using the Bradley-Terry model (Bradley and Terry, 1952):

$$p(y_w \succ y_l \mid x) = \frac{\exp\left(r(x, y_w)\right)}{\exp\left(r(x, y_w)\right) + \exp\left(r(x, y_l)\right)}.$$

From this, one can introduce a parametrized reward model $r_\phi(x, y)$ by estimating its parameters with the maximum-likelihood objective:

$$\mathcal{L}_r = \mathop{\mathbb{E}}_{(x,y_w,y_l)\sim\mathcal{D}} \left[-\log \sigma\left(r_\phi(x, y_w) - r_\phi(x, y_l)\right)\right],$$

where σ is a sigmoid function. After this reward modeling procedure, one could improve the alignment of LLM π_θ by optimizing it to maximize the reward from r_ϕ. Here, KL-distance from the reference model π_{ref} is incorporated as a regularization to prevent the reward over-optimization of π_θ, with a hyper-parameter $\beta > 0$ (Ouyang et al., 2022):[4]

$$\begin{aligned}\mathcal{L}_{\text{RLHF}} = &-\mathbb{E}_{y\sim\pi_\theta,\, x\sim\rho}\left[r_\phi(x, y)\right] \\ &+ \beta D_{\text{KL}}\left(\pi_\theta(y|x) \| \pi_{\text{ref}}(y|x)\right).\end{aligned}$$

Direct preference optimization. Rafailov et al. (2023) propose an alternative approach to align LLM π_θ with the preference dataset $\mathcal{D}$, which is called Direct Preference Optimization (DPO). DPO integrates a two-step alignment procedure with reward modeling and RL fine-tuning into a single unified fine-tuning procedure. Specifically, the optimal reward function is derived from the RLHF objective (Eq. **??**), with the target LLM π_θ and the reference model π_{ref}, which is often called implicit reward:

$$r(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x),$$

where $Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)$. Then, the preference between two responses could be measured using this reward derivation, and π_θ is optimized to maximize this preference of y_w over y_l using the preference dataset $\mathcal{D}$.

$$\begin{aligned}p_\theta(y_w \succ y_l \mid x) = \sigma\Bigg(&\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} \\ &- \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\Bigg),\end{aligned}$$

$$\mathcal{L}_{\text{DPO}} = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[-\log p_\theta(y_w \succ y_l | x)\right].$$

G More Qualitative Examples

In this section, we present the additional qualitative examples similar to Figure 4. Figures 5, 6, 7, and 8 clearly show the advantages of CoPE, compared to the baseline methods.

[4] π_{ref} is usually initialized with supervised fine-tuned (SFT) LLM. Also, π_θ is initialized with π_{ref}.

Scholarly Title	
Query	Generate a title for the following abstract of a paper: Despite advances in multicore smartphone technologies, battery consumption still remains one of customer's least satisfying features. This is because existing energy saving techniques do not consider the electrochemical characteristics of batteries, which causes battery consumption to vary unpredictably, both within and across applications. Additionally, these techniques provide application specific fixed performance degradation in order to reduce energy consumption. Having a performance penalty, even when a battery is fully charged, adds to customer dissatisfaction. We propose a control-based approach for runtime power management of multicore smartphones, which scales the frequency of processing cores in response to the battery consumption, taking into account the electrochemical characteristics of a battery. The objective is to enable graceful performance modulation, which adapts with application and battery availability in a predictable manner, improving quality-of-user-experience. Our control approach is practically demonstrated on embedded Linux running on Cortex A15-based smartphone development platform from nvidia. A thorough validation with mobile and Java workloads demonstrate 2.9\u00d7 improvement in battery availability compared to state-of-the-art approaches.
User Answer	Graceful Performance Adaption through Hardware-Software Interaction for Autonomous Battery Management of Multicore Smartphones
TAM	Predictable Quality-of-User-Experience Management for Multicore Smartphones.
OPPU	Predictable performance management for multicore smartphones
CoPe (ours)	Predictable Performance Management for Battery-Powered Multicore Smartphones

Figure 5: **Qualitative example for Scholarly Title Generation**

Abstract Generation	
Query	Generate an abstract for the title \"A resolved CFD-DEM coupling model for modeling two-phase fluids interaction with irregularly shaped particles\" using the following items: 1. Resolved coupling model\n2. Multi-sphere clumps\n3. Benchmark cases\n4. Dambreak wave impact\n5. Coastal engineering
User Answer	\u2022A resolved coupling model to directly simulate interaction between two-phase fluids and irregularly shaped particles.\u2022The irregularly shaped particles are modeled by multi-sphere clumps in DEM.\u2022A number of benchmark cases are conducted and compared with experimental works.\u2022A case study of dambreak wave impact on a rock pile demonstrates the model's potential to coastal engineering.
TAM	\u2022Resolved coupling model is developed to simulate two-phase fluid interactions with multi-sphere clumps.\u2022The model can be applied in a wide range of particle sizes and concentrations.\u2022Benchmark cases show good agreement between simulation results and experimental data.\u2022Dambreak wave impacts on single sphere, cluster of spheres and cylinder are investigated.\u2022Applications include coastal engineering, sediment transport and environmental studies.
OPPU	\u2022Resolved coupling model is developed to simulate two-phase fluid interactions with irregularly shaped particles.\u2022Multi-sphere clumps are used as a surrogate of real particles in simulations.\u2022Benchmark cases show that the proposed method can capture the particle motion and sediment transport accurately.\u2022Dambreak wave impacts on multi-particle bed, and dune formation process are investigated by the present model.\u2022The results provide useful insights into coastal engineering problems involving complex geometries.
CoPe (ours)	\u2022A resolved coupling model is proposed to simulate two-phase fluid interactions with irregularly shaped particles.\u2022The multi-sphere clumps method is used to represent the irregularly shaped particles.\u2022Benchmark cases are conducted to verify the accuracy of the proposed model.\u2022Dambreak wave impact on a rigid cylinder and a rough surface is simulated by the proposed model.\u2022The results show that the proposed model can be applied in coastal engineering.

Figure 6: **Qualitative example for Abstract Generation**

Review Writing	
Query	Generate the review text written by a reviewer who has a given an overall rating of \"4.0\" for a product with description \"In the taut thriller, when Nancy (Blake Lively) is surfing on a secluded beach, she finds herself on the feeding ground of a great white shark. Though she is stranded only 200 yards from shore, survival proves to be the ultimate test of wills, requiring all of Nancys ingenuity, resourcefulness, and fortitude.\". The summary of the review text is \"I found myself holding my breath in certain spots and actually willing her to survive. That is the sign of a good movie.\".
User Answer	Im gonna get off this rock. Im gonna fight. Nancy (Lively) has just lost her mom and travels to her favorite place in the world in order to hopefully reconnect with her one last time. What begins as a relaxing beach getaway takes a terrifying turn when she swims right into the feeding ground of a Great White Shark. Now, with a severely bitten leg and a tiny rock just above the water line Nancy must survive long enough for someone to find her, but she isn't the only one fighting for survival. This is a very tense movie that is scary in the fact that this is something that could actually happen. Looking back on the movie I realized there wasn't a whole lot of dialog but I didn't notice it when I was watching. You really feel for her and live and die with every scene. I found myself holding my breath in certain spots and actually willing her to survive. That is the sign of a good movie. The movie defiantly has the feel of Open Water and if you like movies like that you will enjoy this as well. Overall, the way the movie makes you feel makes this more of an experience than a movie. I liked this a lot more than I expected to. I give this a B+.
TAM	I'm not sure what it says about me that this film had me so engrossed. It was like watching a train wreck unfold before your eyes. A young woman goes out into the ocean alone while her boyfriend surfs. She gets attacked by a shark and then spends hours trying to get back to land. There isn't much dialogue or action but there were moments where I literally held my breath as she tried to swim away from the shark. This is one of those films you watch because you know something bad is going to happen and you want to see how it plays out. I don't think I would have watched it if I didn't already own it. But once I started I couldn't stop until the end. Blake Lively does a very good job playing someone who is terrified yet determined to live through this experience. If you enjoy suspenseful movies, give this one a try
OPPU	"Ive been bitten before but never like this. After being attacked by a Great White Shark while swimming alone at sea Nancy (Lively) struggles to stay alive until help arrives. This was not what I expected. Going into it I thought that there would be lots of blood and gore and a lot of suspense. While there were some tense moments the movie was more about how one woman copes with the situation rather then just trying to get away. There really wasn't much gore or anything like that so if you are looking for a horror type movie this may disappoint you. On the other hand though this was very intense and kept me interested throughout. I did find myself holding my breath in certain spots and actually willing her to survive. That is the sign of a good movie. Overall, a very interesting take on a shark attack story. I give this a B+."
CoPe (ours)	"Ive been bitten by a shark! Nancy (Lively) is out surfing alone when she is attacked by a shark. She manages to get back to shore but not before being injured. When she realizes that help is not coming she must find a way to stay alive until someone comes along. This is a very tense movie. It is hard to make a movie about one person trying to survive against something else without it getting boring or repetitive. This one does a pretty good job of keeping you interested and involved in what is happening. There are some parts that drag a little but for the most part this is a really good movie. Blake Lively does a great job in this and makes you feel everything she is feeling. Overall, a very tense and exciting movie that keeps you guessing and wondering if she will make it. I give this a B+.

Figure 7: **Qualitative example for Review Writing**

Topic Writing	
Query	Generate the content for a reddit post Indoor cat doesn't like car ride or outdoors, tries to kill me, cat poo in car preferable.
User Answer	You got off lucky. When I was in college, one of my roommates had a cat that would spend all day on the windowsill, staring out the window (it was an indoor cat). I decided one day it would be nice to drive the cat to the park to let it experience the outdoors for its first time. In the car, it tried to shred my shoulder, and then camped out under the brake pedal, almost causing me to wreck. At the park, it just cowered and wouldn't move. On the ride home, it decided the best place to be was on top of my head, using its claws for purchase.
TAM	I have an indoor/outdoor cat that is about 10 years old now and she has always been very skittish around people (she was found as a stray when she was younger). She will not let anyone pick her up unless they know how to handle her properly. If you try to grab her by the scruff of the neck she will bite your hand off. So we just pet her on the head and rub her belly until she calms down enough to be picked up. We also keep her nails trimmed so she can't do any real damage if she does decide to attack. \n The problem is that my wife wants to take our cat with us when we go camping this weekend but I am worried because she hates being in cars and hates going outside. When we first got her she would hide under the bed whenever someone came over and wouldn't come out until everyone left. Now she comes out from time to time but still only likes to stay inside. \n My question is what should I do? Should I put her in a carrier and hope for the best? Or should I leave her at home? Also, if it helps, she has never had diarrhea before and she is litter trained.
OPPU	I have an indoor/outdoor cat that hates being outside and hates riding in cars. He is also very smart (and has been known to open doors). One day we were going on vacation and he was not allowed to come with us because of his issues. We put him in a carrier and took it into our bedroom so he wouldn't be able to see us leave. When we got back from vacation, he had somehow gotten into the house while we were gone, found the carrier, opened it up, climbed inside, and then proceeded to shit all over himself and the box. Then he waited until we came home, walked around the corner when we entered the room, and jumped at my face as hard as he could. It knocked me off balance and I fell onto the bed. The little bastard ran away before I could get my hands on him.
CoPe (ours)	I was driving my mom's car with her cat in it. The cat had been an indoor cat all of its life and didn't like being in the car. It kept trying to get out through the window (it wasn't rolled down). So I put on some gloves and held onto him so he couldn't escape. He then proceeded to try to claw my eyes out. \n 2) I was driving home from work one day when I saw a cat sitting in the middle of the road. I stopped and got out to move it off the road. As soon as I touched it, it started spraying pee everywhere. I quickly dropped it back into the street and drove away.

Figure 8: **Qualitative example for Topic Writing**