Preference Heads in Large Language Models: A Mechanistic Framework for Interpretable Personalization

Weixu Zhang[1,2], **Ye Yuan**[1,2], **Changjiang Han**[3], **Yuxing Tian**[4], **Zipeng Sun**[1,2], **Linfeng Du**[1,2], **Jikun Kang**[5], **Hong Kang**[1,2], **Xue Liu**[1,2,3], **Haolun Wu**[1,2*]
[1]McGill University [2]Mila - Quebec AI Institute
[3]MBZUAI [4]University of Montreal [5]Salesforce
{weixu.zhang, haolun.wu}@mail.mcgill.ca

Abstract

Large Language Models (LLMs) exhibit strong implicit personalization ability, yet most existing approaches treat this behavior as a black box, relying on prompt engineering or fine tuning on user data. In this work, we adopt a mechanistic interpretability perspective and hypothesize the existence of a sparse set of **Preference Heads**, attention heads that encode user specific stylistic and topical preferences and exert a causal influence on generation. We introduce **Differential Preference Steering (DPS)**, a training free framework that (1) identifies Preference Heads through causal masking analysis and (2) leverages them for controllable and interpretable personalization at inference time. DPS computes a **Preference Contribution Score (PCS)** for each attention head, directly measuring its causal impact on user aligned outputs. During decoding, we contrast model predictions with and without Preference Heads, amplifying the difference between personalized and generic logits to selectively strengthen preference aligned continuations. Experiments on widely used personalization benchmarks across multiple LLMs demonstrate consistent gains in personalization fidelity while preserving content coherence and low computational overhead. Beyond empirical improvements, DPS provides a mechanistic explanation of where and how personalization emerges within transformer architectures. Our implementation is fully open-sourced.[1]

1 Introduction

Large Language Models (LLMs) exhibit strong implicit personalization ability. They often adapt to a user-specific tone, writing style, and topical interests with minimal conditioning. This capability is critical for applications such as personalized writing assistance, recommendation generation, and conversational agents, where users expect outputs to reflect individual preferences. Recent surveys highlight rapid progress in personalized LLMs via prompting, retrieval, and fine-tuning-based approaches (Liu et al., 2025). However, most existing methods treat LLMs as black-box systems. They rely on prompt engineering, retrieval-based user profiles, or fine-tuning on user data (Salemi et al., 2024; Kumar et al., 2024; Chen et al., 2025a), without understanding how user preferences are internally represented or propagated. As a result, these approaches often improve surface-level consistency but lack interpretability, controllability, and scalability across users.

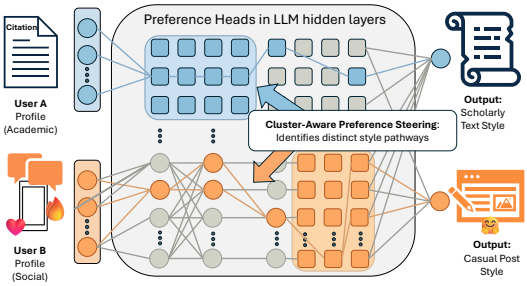

Figure 1: Overview of preference-based personalization in LLMs. Distinct user profiles activate different subsets of *Preference Heads*, forming sparse internal pathways that steer generation toward user-aligned styles. Cluster-aware preference steering further captures shared structure across users with similar preferences.

Recent advances in mechanistic interpretability suggest that many transformer behaviors can be traced to small, specialized internal components. Prior work has identified attention heads responsible for sequence copying, long-range retrieval, and abstract task representations, showing that high-level capabilities may emerge from localized circuits rather than diffuse parameter interactions (Wu et al., 2025; Zhang et al., 2025b; Yin and Steinhardt, 2025). Related studies further demonstrate that attention heads can act as efficient zero-shot re-rankers or encode structured interaction patterns

*Corresponding author.
[1]https://github.com/weixuzhang/DPS

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47742–47754
July 2-7, 2026

(Chen et al., 2025b; Qu et al., 2025). These findings raise a natural but underexplored question: *where does personalization arise within large language models?*

In this work, we argue that personalization is mediated by a sparse subset of attention heads that encode user-specific stylistic and topical signals and exert a direct influence on generation. We refer to these components as **Preference Heads**. As illustrated in Figure 1, different users activate distinct internal pathways through these heads, leading to systematically different generation behaviors (e.g., scholarly versus conversational styles). Unlike general semantic or retrieval-oriented heads, Preference Heads selectively amplify tokens that are consistent with a user profile. While prior work has explored preference learning from user edits or feedback (Tucker et al., 2024; Gao et al., 2024; Pang et al., 2024), it does not examine how preferences are represented at the level of internal model components.

To identify Preference Heads, we introduce a causal head discovery procedure based on targeted masking and contribution analysis. We define a **Preference Contribution Score (PCS)** that measures the causal impact of each attention head on user-aligned generation. This approach builds on recent mechanistic analyses of attention heads for retrieval and factuality (Wu et al., 2025; Gema et al., 2024), but focuses specifically on personalization.

Building on this analysis, we propose **Differential Preference Steering (DPS)**, a decoding-time framework that enables interpretable and controllable personalization without modifying model parameters. DPS contrasts model predictions with and without Preference Heads and amplifies their difference during decoding, strengthening preference-aligned continuations while suppressing generic ones. To account for user heterogeneity, we further extend DPS with cluster-aware preference steering, allowing users with similar profiles to share partially overlapping preference circuits. While related to contrastive decoding and activation steering methods (Gema et al., 2024; Zhang et al., 2025a; Wang et al., 2025), DPS targets causally identified personalization circuits rather than externally defined behaviors.

We evaluate DPS on widely used personalization benchmarks covering both short-form and long-form user-conditioned generation (Salemi et al., 2024; Kumar et al., 2024). Across multiple open-source LLMs, DPS consistently improves personalization fidelity and stylistic alignment while preserving content coherence and computational efficiency. Further analyses show that Preference Heads are causally meaningful and interpretable: masking them disrupts personalized behavior, and their activation patterns correspond to structured stylistic and topical dimensions.

Our key contributions are summarized as follows:

- We present the first mechanistic analysis of personalization in LLMs by identifying **Preference Heads**, a sparse set of attention heads that causally encode user-specific preferences.
- We introduce **Preference Contribution Score** and **Differential Preference Steering**, a training-free framework for interpretable and controllable personalization at decoding time.
- Through extensive experiments and analyses, we demonstrate that targeted intervention on Preference Heads yields consistent personalization gains across model architectures, offering a principled alternative to black-box personalization methods.

2 Related Work

2.1 Personalization in Large Language Models

Personalization in large language models aims to adapt generation behavior to user specific preferences such as writing style, topical interests, or decision patterns. Prior work explores fine tuning, prompting, retrieval, and decoding time methods to achieve personalization, with benchmarks such as LaMP and LongLaMP enabling systematic evaluation (Salemi et al., 2024; Kumar et al., 2024). Several approaches model preferences explicitly, including personalized decoding (Chen et al., 2025a), difference aware user modeling (Qiu et al., 2025), and latent preference learning from user edits or feedback (Tucker et al., 2024; Gao et al., 2024). While effective, these methods treat personalization as an external conditioning problem and do not examine how user preferences are internally represented within transformer components.

2.2 Mechanistic Interpretability and Attention Heads

Mechanistic interpretability studies show that transformer behavior can often be attributed to specialized internal components. Induction heads support

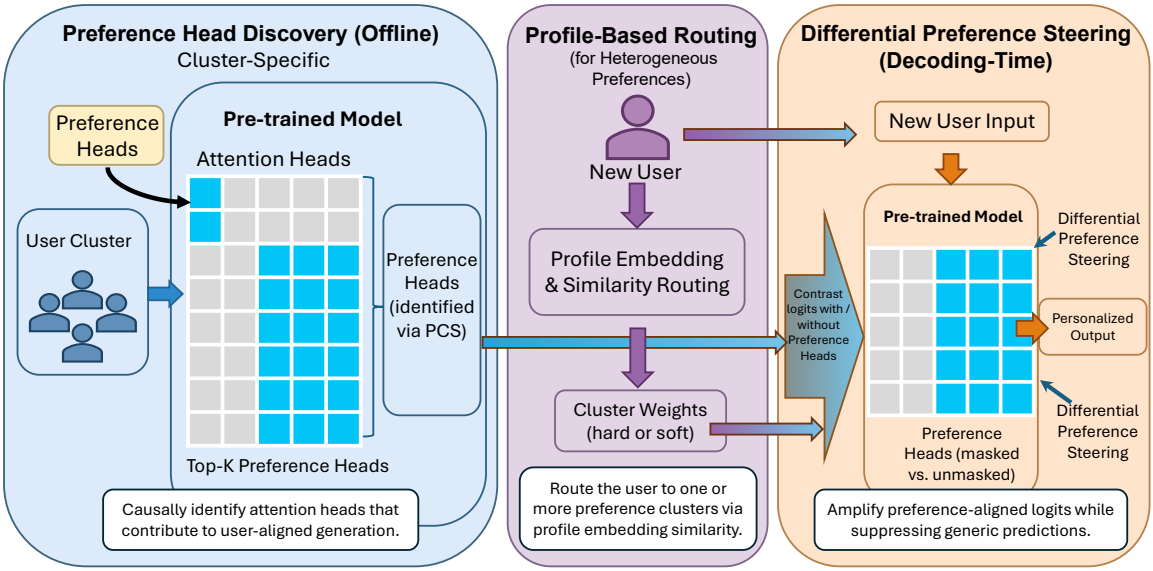

Figure 2: Overview of the proposed framework. We first perform offline, causal discovery of Preference Heads by measuring the effect of head ablation on user-aligned generation. For heterogeneous users, profiles can optionally be clustered and routed via embedding similarity. At decoding time, Differential Preference Steering contrasts model logits with and without Preference Heads and amplifies preference-aligned differences to produce personalized outputs without parameter updates.

pattern matching in context (Crosbie and Shutova, 2025), retrieval heads enable long context factual recall (Wu et al., 2025; Zhang et al., 2025b), and other attention heads act as efficient rerankers or encode structured interactions (Chen et al., 2025b; Qu et al., 2025). Related work also examines which attention heads matter for specific tasks and shows that a sparse subset of internal units can dominate model behavior (Yin and Steinhardt, 2025; Jiao et al., 2024). However, existing analyses primarily focus on reasoning, factuality, or in context learning rather than personalization.

2.3 Decoding Time Control and Contrastive Methods

Decoding time control provides a model agnostic way to steer language model behavior without retraining. Contrastive approaches guide generation by amplifying differences between model variants, including layer level contrast in DoLa (Chuang et al., 2024) and head level contrast in DeCoRe (Gema et al., 2024). Related methods apply activation steering or contrastive decoding to influence generation style or behavior (Zhang et al., 2025a; Wang et al., 2025). Although some work applies decoding time control to personalization (Chen et al., 2025a), they do not identify the internal components responsible for personalization, whereas our approach grounds decoding time intervention in a mechanistic analysis of attention heads.

3 Preference Head Discovery

In this section, we formalize the notion of *Preference Heads* and introduce a causal procedure for identifying them. Our goal is to determine which attention heads inside a pretrained transformer are directly responsible for injecting user specific preferences into the generation process.

3.1 Problem Setup and Intuition

We consider a pretrained transformer based language model parameterized by θ. Given a task input x and a user profile u, the model generates an output sequence $y = (y_1, \ldots, y_T)$ autoregressively. At decoding step t, the model produces a distribution over the vocabulary

$$p_\theta(y_t \mid x, u, y_{<t}) = \mathrm{Softmax}(\ell_t), \qquad (1)$$

where $\ell_t \in \mathbb{R}^{|\mathcal{V}|}$ denotes the output logits.

The logits ℓ_t reflect a mixture of generic linguistic knowledge and user specific preference signals. While existing personalization methods influence this mixture externally through prompting or fine tuning, we seek to identify where preference information is internally encoded. Motivated by prior mechanistic analyses, we hypothesize that personalization is mediated by a sparse subset of attention heads whose activations causally affect user aligned generation. We refer to these components as *Preference Heads*.

Algorithm 1 Differential Preference Steering with Heterogeneous Preferences

Input: User-conditioned dataset $\mathcal{D} = \{(x_i, u_i, y_i^*)\}_{i=1}^N$, context x, user profile u, pretrained model M_θ, personalization strength γ

Output: Personalized output sequence $y = (y_1, \ldots, y_T)$

/* Phase 1: Preference Head Discovery */
for each attention head $h_{l,k}$ in M_θ **do**
    Construct ablated model $M_{\theta \setminus h_{l,k}}$
    Compute $\mathrm{PCS}(h_{l,k}) \leftarrow \mathbb{E}_{(x,u,y^*)\sim\mathcal{D}}\big[\mathcal{L}(M_{\theta\setminus h_{l,k}}, x, u, y^*) - \mathcal{L}(M_\theta, x, u, y^*)\big]$
Select top K heads with highest PCS values
$\mathcal{H}_{\text{pref}} \leftarrow \mathrm{TopK}(\mathrm{PCS})$

/* Phase 2: Clustering and Routing */
Compute profile embeddings for all users
Cluster users into groups $\{c\}$ using embedding similarity
Assign user u to cluster(s) with weight(s) $\{w_c\}$ ▷ hard or soft routing
for each cluster c **do**
    Obtain cluster-specific Preference Heads $\mathcal{H}_{\text{pref}}^{(c)}$

/* Phase 3: Differential Preference Steering */
Initialize decoding context $z \leftarrow (x, u)$
for $t = 1$ to T **do**
    Compute personalized logits $\ell_t^{\text{pref}} \leftarrow f_{M_\theta}(z)$
    Compute depersonalized logits with weighted head suppression $\ell_t^{\text{gen}} \leftarrow f_{M_{\theta \setminus \sum_c w_c \mathcal{H}_{\text{pref}}^{(c)}}}(z)$
    Combine logits $\tilde{\ell}_t \leftarrow (1+\gamma)\,\ell_t^{\text{pref}} - \gamma\,\ell_t^{\text{gen}}$
    Sample or select next token $y_t \sim \mathrm{Softmax}(\tilde{\ell}_t)$
    Update context $z \leftarrow (z, y_t)$
return y

3.2 Causal Identification via Head Ablation

Let $h_{l,k}$ denote the kth attention head in layer l. To assess whether a head contributes to personalization, we perform targeted ablation by masking its output during inference. Let M_θ denote the original model and $M_{\theta\setminus h_{l,k}}$ denote the model with head $h_{l,k}$ masked. Given a dataset of user conditioned examples:

$$\mathcal{D} = \{(x_i, u_i, y_i^*)\}_{i=1}^N, \tag{2}$$

we measure how ablating $h_{l,k}$ affects the likelihood of the reference output y_i^*.

For a single example, we define the negative log likelihood

$$\mathcal{L}(M, x, u, y^*) = -\frac{1}{T}\sum_{t=1}^{T} \log p_M(y_t^* \mid x, u, y_{<t}^*). \tag{3}$$

If masking a head increases this loss, it indicates that the head contributes causally to user aligned generation behavior.

3.3 Preference Contribution Score

Based on this intuition, we define the *Preference Contribution Score* (PCS) for each attention head. For head $h_{l,k}$, PCS is computed as the expected loss difference between the original model and the ablated model:

$$\mathrm{PCS}(h_{l,k}) = \mathbb{E}_{(x,u,y^*)\sim\mathcal{D}}\big[\mathcal{L}(M_{\theta\setminus h_{l,k}}, x, u, y^*) - \mathcal{L}(M_\theta, x, u, y^*)\big]. \tag{4}$$

A positive PCS indicates that removing the head degrades user aligned generation, suggesting a causal contribution to personalization. Because PCS is computed through direct intervention, it measures causal effect rather than correlation.

3.4 Selecting Preference Heads

We compute PCS for all attention heads and select a small subset with the highest scores as Preference Heads. Formally, let $\mathcal{H}$ denote the set of all attention heads and define

$$\mathcal{H}_{\text{pref}} = \mathrm{TopK}_{h\in\mathcal{H}}\,\mathrm{PCS}(h), \tag{5}$$

Model	Method	News Headline Generation			Scholarly Title Generation			Tweet Paraphrasing		
		R-1 ↑	R-L ↑	METEOR ↑	R-1 ↑	R-L ↑	METEOR ↑	R-1 ↑	R-L ↑	METEOR ↑
LLaMA-3-8B	CAD	0.1681	0.1498	0.1568	0.3530	0.3068	0.3925	0.3368	0.2893	0.2813
	DeCoRe	0.1768	0.1572	0.1626	**0.4010**	**0.3527**	0.4004	0.3231	0.2764	0.2729
	DoLa	0.1694	0.1508	0.1592	0.3636	0.3117	**0.4079**	0.3365	0.2877	0.2795
	DPS (ours)	**0.1787**	**0.1596**	**0.1650**	0.3243	0.2787	0.3826	**0.3389**	**0.2898**	**0.2884**
Qwen2-7B	CAD	0.1580	0.1392	0.1255	0.4197	**0.3780**	0.4381	**0.3590**	**0.3106**	**0.3384**
	DeCoRe	0.1581	0.1305	0.1232	**0.4311**	0.3729	0.4565	0.3470	0.3065	0.3173
	DoLa	0.1642	0.1473	0.1272	0.4277	0.3746	**0.4596**	0.3524	0.3046	0.3246
	DPS (ours)	**0.1627**	**0.1450**	**0.1318**	0.4071	0.3421	0.4230	0.3533	0.2981	0.3269
Mistral-7B	CAD	0.1361	0.1132	0.0980	**0.4375**	0.3712	0.4561	0.3342	0.2916	**0.3097**
	DeCoRe	0.1299	0.1085	0.0908	0.4135	0.3648	0.4419	0.3407	0.2927	0.2978
	DoLa	0.1362	0.1136	0.0962	0.4364	**0.3733**	**0.4605**	0.3291	0.2852	0.3026
	DPS (ours)	**0.1536**	**0.1366**	**0.1399**	0.3983	0.3350	0.4162	**0.3441**	**0.2998**	0.2990

Table 1: Results on LaMP generation tasks. ↑ indicates higher is better. Best results per model are in bold.

where K is chosen such that $\mathcal{H}_{\text{pref}}$ remains sparse relative to the total number of heads.

4 Differential Preference Steering

In this section, we introduce *Differential Preference Steering* (DPS), a decoding time method that leverages discovered Preference Heads to enable controllable and interpretable personalization. DPS modifies the generation process by isolating and amplifying the internal preference signal contributed by Preference Heads, without changing model parameters or requiring additional training.

4.1 Decomposing Generic and Preference Conditioned Behavior

At each decoding step t, the output logits ℓ_t reflect the combined influence of multiple internal components. Based on the analysis in Section 3, we view these logits as consisting of a generic component that captures task relevant and linguistic information, and a preference conditioned component that reflects user specific stylistic and topical tendencies.

Let M_θ denote the original model and let $M_{\theta \setminus \mathcal{H}_{\text{pref}}}$ denote the model with Preference Heads masked. For a given context $(x, u, y_{<t})$, we define

$$\ell_t^{\text{pref}} = f_{M_\theta}(x, u, y_{<t}), \tag{6}$$

$$\ell_t^{\text{gen}} = f_{M_{\theta \setminus \mathcal{H}_{\text{pref}}}}(x, u, y_{<t}), \tag{7}$$

where $f_M(\cdot)$ denotes the logit function of M.

4.2 Differential Logit Combination

DPS constructs modified logits by amplifying the contrast between preference conditioned and generic predictions:

$$\tilde{\ell}_t = (1 + \gamma)\, \ell_t^{\text{pref}} - \gamma\, \ell_t^{\text{gen}}, \tag{8}$$

where $\gamma \geq 0$ controls personalization strength.

The next token distribution is given by

$$p(y_t \mid x, u, y_{<t}) = \text{Softmax}(\tilde{\ell}_t), \tag{9}$$

and decoding proceeds autoregressively using standard decoding strategies.

5 Extending to Heterogeneous Preferences

Preference Heads are not universal. Different users may rely on different internal circuits depending on their stylistic and topical tendencies. To account for this heterogeneity, we extend our framework with cluster aware Preference Head discovery and weighted DPS.

5.1 Profile Embedding and Clustering

We represent each user profile as a dense embedding computed from historical text using a pretrained sentence encoder. Users are clustered based on embedding similarity using k means, with the number of clusters chosen to ensure sufficient samples per group. This clustering captures coarse preference types such as formal versus conversational style or domain specific interests.

5.2 Cluster Specific Preference Heads

For each cluster, we repeat the PCS based head discovery procedure using only samples from that cluster. This yields cluster specific Preference Head sets and corresponding importance weights. Heads that are important for one cluster may be less important for others, reflecting diversity in personalization mechanisms.

Model	Method	Citation Identification		Movie Tagging		Product Rating	
		Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	MAE ↓	RMSE ↓
LLaMA-3-8B	CAD	0.6240	0.6070	0.4552	0.3839	0.4426	0.9300
	DeCoRe	0.6232	0.6200	**0.4639**	**0.4034**	0.4442	0.9458
	DoLa	0.6156	0.5961	0.2800	0.1643	**0.4200**	**0.8718**
	DPS (ours)	**0.6356**	**0.6288**	0.4610	0.3910	0.4236	0.9278
Qwen2-7B	CAD	0.6230	0.6250	0.1850	0.1181	**0.3180**	**0.6240**
	DeCoRe	0.5400	0.5891	0.2320	0.1217	0.3250	0.6325
	DoLa	0.6790	0.6795	0.2412	0.0958	0.3200	0.6300
	DPS (ours)	**0.6932**	**0.7078**	**0.3902**	**0.3202**	0.3276	0.6719

Table 2: Results on LaMP classification and regression tasks. ↑ indicates higher is better and ↓ indicates lower is better. Best results per model are in bold.

5.3 Weighted Differential Preference Steering

At inference time, we route each user to clusters based on profile similarity, using either hard assignment to the nearest cluster or soft assignment that distributes weight across multiple clusters. We aggregate cluster specific head weights into a single weighted importance map, which determines how strongly each head is suppressed in the depersonalized pass. Weighted DPS then applies contrastive decoding using these soft head suppressions, enabling smooth personalization control across heterogeneous users.

Algorithm 1 provides a complete description of Differential Preference Steering with support for heterogeneous user preferences.

6 Experiments

6.1 Experimental Setup

Models We evaluate Differential Preference Steering (DPS) on multiple instruction tuned open source large language models: LLaMA-3-8B-Instruct, Mistral-7B-Instruct, and Qwen2-7B-Instruct. These models represent different architectural and training design choices while maintaining comparable parameter scales.

Datasets We conduct experiments on the **LaMP** benchmark (Salemi et al., 2024), which evaluates personalized language modeling across both generation and prediction tasks. LaMP includes user conditioned text generation tasks (e.g., headline generation, scholarly title generation, tweet paraphrasing) as well as classification and regression tasks (e.g., citation identification, movie tagging, product rating). We follow the official data splits and evaluation protocols.

Metrics We report task appropriate automatic metrics. For generation tasks, we use ROUGE-1, ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). For classification tasks, we report Accuracy and F1. For regression tasks, we report Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Baselines We compare DPS against strong decoding time intervention baselines: **Context Aware Decoding (CAD)** (Shi et al., 2024), **Decoding by Contrasting Layers (DoLa)** (Chuang et al., 2024), and **Decoding by Contrasting Retrieval Heads (DeCoRe)** (Gema et al., 2024). All baselines are implemented using their publicly released code or faithful reimplementations, with hyperparameters tuned following the original papers.

6.2 Results

Tables 1 and 2 report results on LaMP generation, classification, and regression tasks. Across models and tasks, DPS consistently achieves strong or best performance compared to existing decoding-time baselines.

On generation tasks, DPS improves ROUGE and METEOR scores on News Headline Generation and Tweet Paraphrasing across all evaluated models. While some baselines achieve strong performance on individual tasks, DPS demonstrates more consistent gains across task types, indicating improved personalization fidelity without sacrificing content quality.

On classification and regression tasks, DPS achieves the best or near-best performance in most settings. In particular, DPS substantially improves performance on Movie Tagging for Qwen2-7B and achieves the lowest error on Product Rating tasks

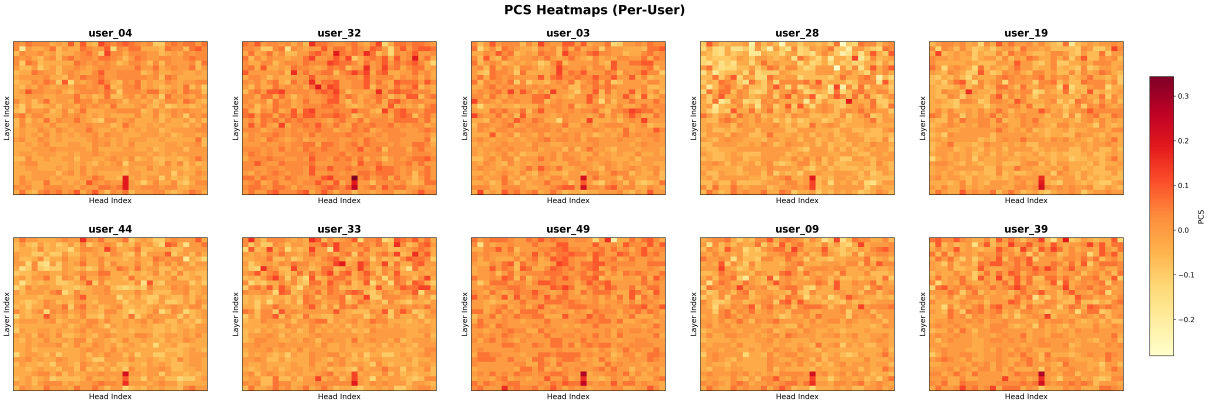

Figure 3: Per-user PCS heatmaps across layers and attention heads. Users are selected randomly. Preference Heads are sparse and causally significant within users.

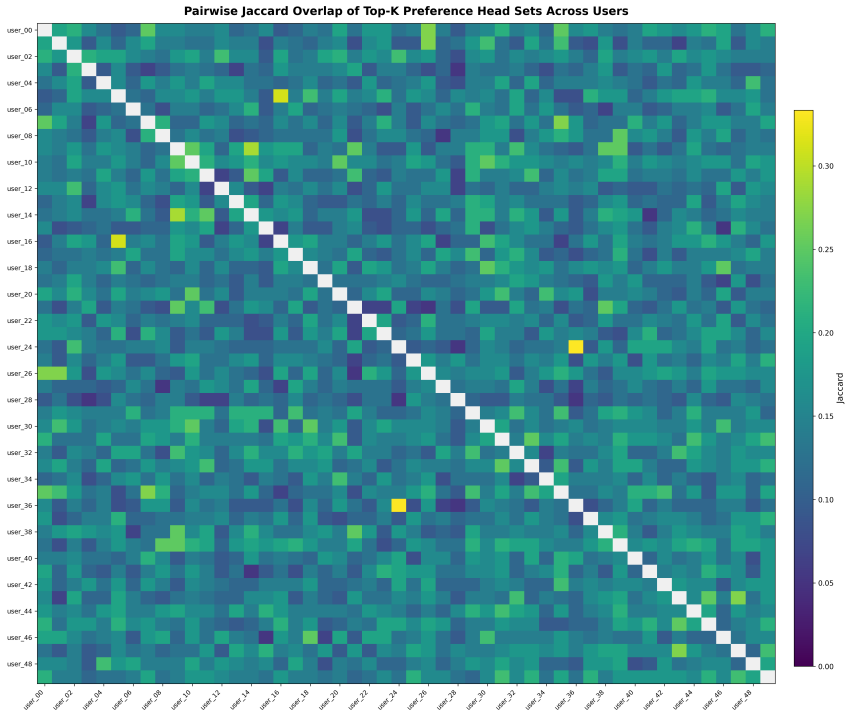

Figure 4: Pairwise Jaccard overlap of top-K Preference Head sets across users. Preference Heads exhibit limited overlap across users, motivating cluster-aware head discovery.

compared to contrastive decoding baselines. These results suggest that amplifying preference-specific internal signals benefits both generative and predictive personalization tasks.

Overall, DPS delivers consistent gains across model families and task formats, suggesting that preference head-based decoding is a robust and broadly applicable personalization mechanism rather than a model- or task-specific effect.

6.3 Head Discovery Analysis

We analyze the structure and consistency of Preference Heads identified by the Preference Contribution Score (PCS).

Sparse and user-specific Preference Heads. Per-user PCS heatmaps (Figure 3) reveal that high-PCS heads are sparse and unevenly distributed across layers. While a small number of heads consistently exhibit strong causal influence for each user, the locations of these heads vary substantially across users. This indicates that personalization is driven by localized internal components rather than uniformly distributed attention behavior.

Limited cross-user overlap. To quantify consistency across users, we compute pairwise Jaccard overlap between the top-K Preference Head sets identified independently for each user. As shown in Figure 4, overlap between different users remains

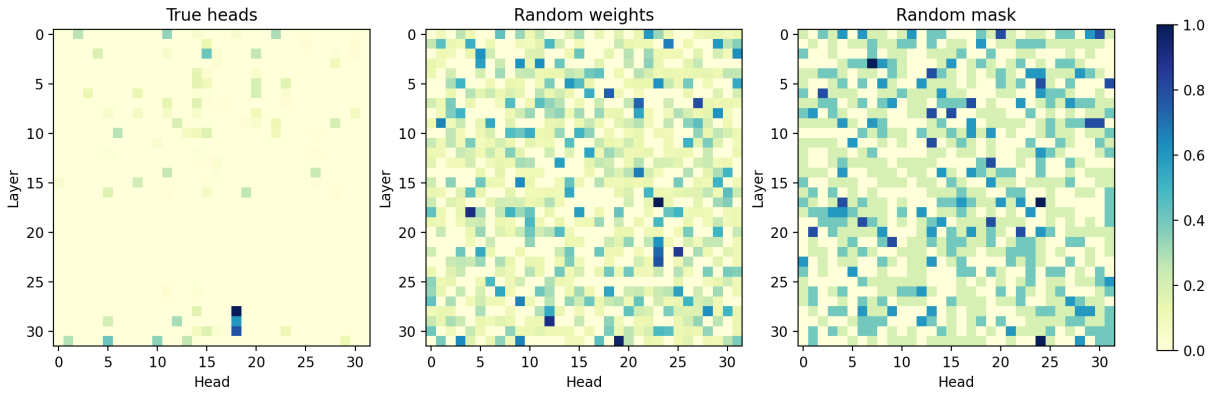

Figure 5: Ablation analysis of attention head selection. Discovered Preference Heads exhibit sparse and structured importance (left), while random weights and random masks (middle, right) produce diffuse patterns and fail to recover comparable personalization behavior.

Prompt Length	Baseline	DPS	Overhead
512	6.57	6.96	1.06×
1024	13.04	13.43	1.03×
2048	26.80	27.21	1.02×

Table 3: Estimated inference FLOPs (TFlop) for LLaMA-3-8B with $T = 32$ generated tokens. DPS shares prompt prefill cost with standard decoding and adds a second decoding pass.

low, with most values clustered near zero. This suggests that Preference Heads are not shared globally, and that directly aggregating per-user head sets would dilute preference-specific signals.

Implications. Together, these results show that Preference Heads are causally meaningful but heterogeneous across users. This motivates the cluster-aware extension introduced in Section 5, which stabilizes head discovery by sharing preference signals among users with similar profiles.

6.4 Ablation Studies

We conduct ablation studies to validate that the gains of DPS arise from causally identified Preference Heads rather than arbitrary head manipulation.

True heads vs random controls. We compare DPS using discovered Preference Heads against two control variants: (i) random head selection with the same cardinality, and (ii) random masking patterns with matched sparsity. Figure 5 visualizes the resulting head importance maps. In contrast to the sparse and localized structure of true Preference Heads, random controls produce diffuse and unstructured patterns.

Quantitatively, replacing Preference Heads with random heads consistently degrades performance across tasks, while random masking yields unstable and task-dependent behavior. These results confirm that DPS relies on semantically meaningful internal components rather than generic sparsification effects.

Effect of head selection strategy. We further evaluate performance when progressively increasing the number of selected heads. As shown in Figure 6, performance improves when including more high-PCS heads, but saturates beyond a moderate head count. This indicates that personalization signals are concentrated in a limited subset of heads, and that including too many heads introduces noise rather than additional benefit.

6.5 Hyperparameter Sensitivity

We analyze the sensitivity of DPS to key hyperparameters, including the number of Preference Heads K, cluster size, personalization strength γ, and routing strategy. Additional analysis is provided in Appendix B.

Number of Preference Heads. Figure 6 shows performance as a function of the number of selected heads. DPS remains stable across a wide range of K, with peak performance achieved using a moderate number of heads. This robustness suggests that precise tuning of K is not critical, as long as the selection focuses on high-PCS heads.

Head set stability across scales. To assess how head sets evolve as K increases, we compute pairwise Jaccard overlap between head sets selected at different values of K. As shown in Figure 7, head sets become increasingly stable at larger K,

indicating that top-ranked heads are consistently preserved while additional heads contribute diminishing returns.

Routing strategy. Finally, we compare hard routing and soft routing for cluster-aware DPS. Figure 8 in Appendix B shows that hard routing performs slightly better on classification tasks, while soft routing provides more robust gains on generation tasks. This trade-off suggests that soft routing offers smoother personalization control, whereas hard routing emphasizes stronger specialization.

6.6 Efficiency Analysis

We analyze the inference-time computational cost of Differential Preference Steering (DPS). DPS is training-free and introduces no additional parameters or prompt modifications.

Inference Cost At decoding time, DPS performs two forward passes per step: one using the original model and one with Preference Heads masked. This cost is comparable to existing contrastive decoding methods such as CAD and DeCoRe. Unlike CAD, DPS does not require re-encoding modified prompts and therefore incurs no additional prefill overhead. Since prompt prefill typically dominates the inference cost for realistic context lengths, the relative overhead of our DPS decreases as prompt length increases.

Computation Cost Table 3 reports estimated inference FLOPs for LLaMA-3-8B with $T = 32$ generated tokens. Despite requiring two decoding passes, DPS introduces only a modest end-to-end overhead due to shared prompt computation.

Preference Head discovery is performed offline and amortized across users and inference runs, and therefore does not affect deployment-time latency. Overall, DPS achieves interpretable personalization with inference efficiency comparable to prior contrastive decoding methods.

6.7 Human Evaluation

Automatic metrics do not fully capture alignment with user-specific preferences. We therefore complement them with both human and LLM-based evaluation on LaMP-4, a personalized news headline generation task.

Human Annotation. We sample 100 test instances from LaMP-4. For each instance, annotators are shown the user profile, the input article, and two anonymized candidate headlines produced by CAD and DPS in randomized order. They are asked to choose the output that better matches the user profile while remaining relevant, fluent, and faithful to the input, and final decisions are obtained by majority vote.

Evaluator	Metric	CAD	DPS
Human	Alignment Win (%)	34	**40**
GPT-5.2	Relevance	3.97	**4.45**
	Fluency	4.51	**4.83**
	Style	3.62	**3.91**
	Alignment	3.63	**3.93**
	Factuality	4.08	**4.21**

Table 4: Human and LLM-based evaluation results on LaMP-4. Human evaluation reports profile-alignment win rates. LLM-based evaluation uses GPT-5.2 and reports average per-method ratings on a 1–5 scale.

LLM-based Evaluation. We additionally use **GPT-5.2** as an automatic judge. Using the same blinded pairwise setup, the judge compares two anonymized outputs and assigns a 1–5 score to each output on five dimensions: relevance, fluency, style, preference alignment, and factuality. We report the average per-method ratings across all evaluated instances.

Results. Table 4 shows that human annotators prefer DPS over CAD in terms of profile alignment on LaMP-4. The LLM-based evaluation shows a similar trend, with DPS achieving higher scores on style and preference alignment while maintaining strong relevance, fluency, and factuality.

7 Conclusion

We present Differential Preference Steering (DPS), a training-free personalization framework grounded in mechanistic interpretability. By causally identifying Preference Heads and amplifying their contribution at decoding time, DPS enables controllable and interpretable personalization without modifying model parameters. Experiments show that DPS consistently improves personalization fidelity while preserving content quality and efficiency. Beyond performance gains, our results provide evidence that user preferences are encoded in sparse, specialized attention heads, offering a mechanistic view of personalization in transformer-based language models. These findings suggest that personalization can be understood through localized internal components rather than only through black-box adaptation.

Limitations

While Differential Preference Steering improves personalization fidelity in large language model outputs, it has several limitations. DPS requires access to internal model components, including attention heads and intermediate activations, which makes it unsuitable for black-box API models. In addition, DPS introduces additional computation during decoding by requiring a second forward pass with masked Preference Heads. Although this overhead is modest in practice, it may be undesirable in latency-critical settings. Future work could explore approximating Preference Head effects in black-box models through lightweight logit biasing or prompt-based steering, as well as further optimizing inference efficiency.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2025a. PAD: personalized alignment of llms at decoding-time. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Shijie Chen, Bernal Jimenez Gutierrez, and Yu Su. 2025b. Attention in large language models yields efficient zero-shot re-rankers. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Joy Crosbie and Ekaterina Shutova. 2025. Induction heads as an essential mechanism for pattern matching in in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5034–5096. Association for Computational Linguistics.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning LLM agents by learning latent preference from user edits. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2024. Decore: Decoding by contrasting retrieval heads to mitigate hallucinations. *CoRR*, abs/2410.18860.

Difan Jiao, Yilun Liu, Zhenwei Tang, Daniel Matter, Jürgen Pfeffer, and Ashton Anderson. 2024. SPIN: sparsifying and integrating internal neurons in large language models for text classification. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 4666–4682. Association for Computational Linguistics.

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

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

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Jieming Zhu, Minda Hu, Menglin Yang, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *CoRR*, abs/2502.11528.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025. Measuring what makes you unique: Difference-aware user modeling for enhancing LLM personalization. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 21258–21277. Association for Computational Linguistics.

Xiaoye Qu, Zengqi Yu, Dongrui Liu, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. 2025. Cooperative or competitive? understanding the interaction between attention heads from A game theory perspective. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14079–14099. Association for Computational Linguistics.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7370–7392. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 783–791. Association for Computational Linguistics.

Aaron David Tucker, Kianté Brantley, Adam Cahall, and Thorsten Joachims. 2024. Coactive learning for large language models using implicit user feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Mengru Wang, Ziwen Xu, Shengyu Mao, Shumin Deng, Zhaopeng Tu, Huajun Chen, and Ningyu Zhang. 2025. Beyond prompt engineering: Robust behavior control in llms via steering target atoms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23381–23399. Association for Computational Linguistics.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Kayo Yin and Jacob Steinhardt. 2025. Which attention heads matter for in-context learning? In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 7128–7141. Association for Computational Linguistics.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025b. Query-focused retrieval heads improve long-context reasoning and re-ranking. *CoRR*, abs/2506.09944.

A Additional Experimental Details

A.1 Dataset Details

We provide additional details of the LaMP benchmark used in our experiments. LaMP evaluates personalized language modeling by conditioning generation or prediction on user profiles constructed from historical interactions. Tasks span multiple output formats and supervision types, including free form generation, classification, and regression.

Table 5 summarizes the statistics of the datasets used in this work.

Task	Type	# Users	# Instances
News Headline Gen.	Gen.	∼1.5K	∼18K
Scholarly Title Gen.	Gen.	∼1.2K	∼14K
Tweet Paraphrasing	Gen.	∼1.0K	∼12K
Citation Identification	Cls.	∼2.0K	∼20K
Movie Tagging	Cls.	∼1.8K	∼16K
Product Rating	Reg.	∼2.5K	∼25K

Table 5: Dataset statistics for LaMP tasks used in our experiments. Gen. denotes generation, Cls. denotes classification, and Reg. denotes regression. Exact counts follow the official LaMP benchmark release.

A.2 Baseline Methods

Context Aware Decoding (CAD) CAD improves generation by contrasting logits conditioned on context with a weakened contextual signal. It has been shown effective for resolving knowledge conflicts and improving contextual faithfulness.

Decoding by Contrasting Layers (DoLa) DoLa contrasts logits produced by different transformer layers to amplify factual knowledge encoded in later layers. It is a representative contrastive decoding method operating over internal model representations.

Decoding by Contrasting Retrieval Heads (DeCoRe) DeCoRe identifies retrieval oriented attention heads and applies contrastive decoding to mitigate hallucinations. Unlike DPS, DeCoRe focuses on factual grounding rather than user specific personalization.

B Additional Hyperparameter Analysis

This section reports additional sensitivity analyses for the number of Preference Heads K. We evaluate how varying K affects personalization performance and the stability of the selected head sets across tasks and models. The results show that DPS is robust over a wide range of settings.

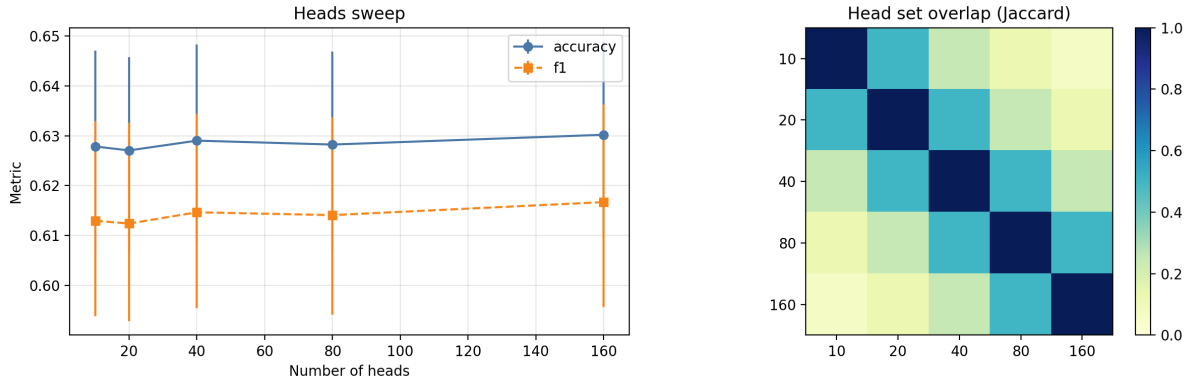

Figure 6: Performance as a function of the number of selected Preference Heads K. Accuracy and F1 scores remain stable across a wide range of K, with performance saturating at moderate values, indicating that personalization signals are concentrated in a limited subset of heads.

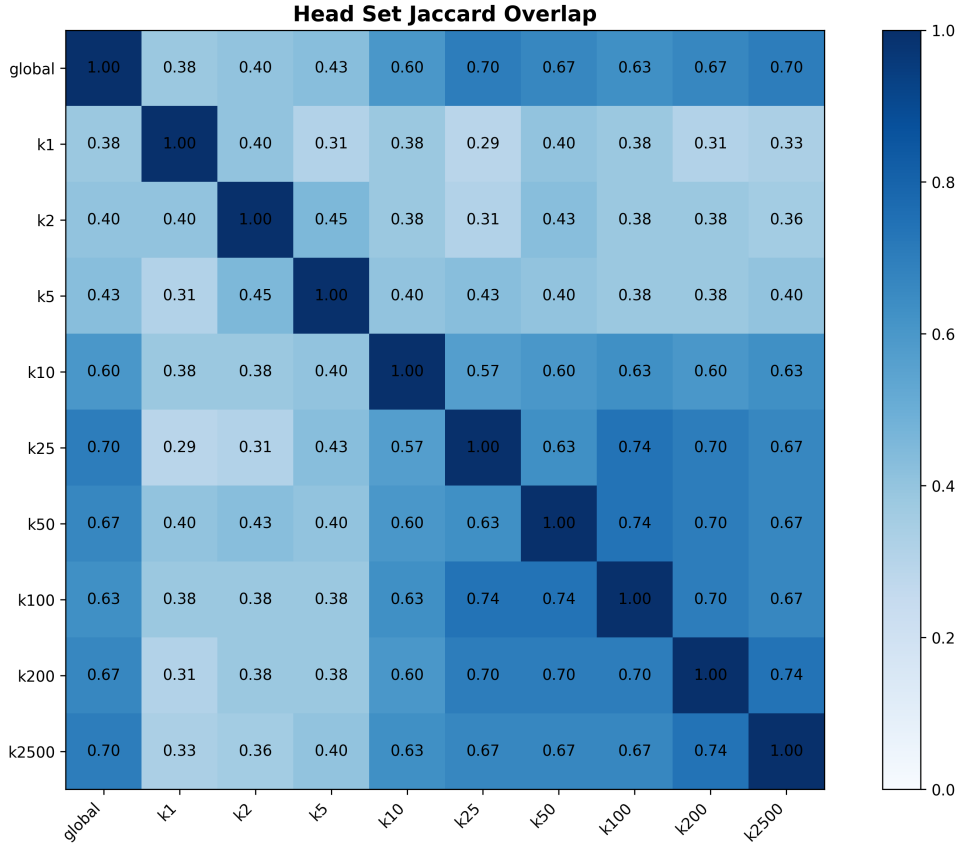

Figure 7: Jaccard overlap between Preference Head sets selected at different values of K. Overlap increases smoothly as K grows, indicating that top-ranked Preference Heads are consistently preserved while additional heads contribute diminishing variation.

In particular, performance improves rapidly as K increases from small values and saturates at moderate K, indicating that personalization is driven by a compact set of heads. We also observe that the selected head sets remain increasingly stable as K grows, suggesting that top-ranked Preference Heads are consistently preserved while additional heads contribute diminishing returns.

C Use of AI Assistants

We used AI assistants to support limited paraphrasing and language refinement during the writing process. Specifically, AI tools were used to improve clarity, conciseness, and grammatical correctness of draft text written by the authors. All technical content, methodological design, experimental setup, analysis, and conclusions were conceived,

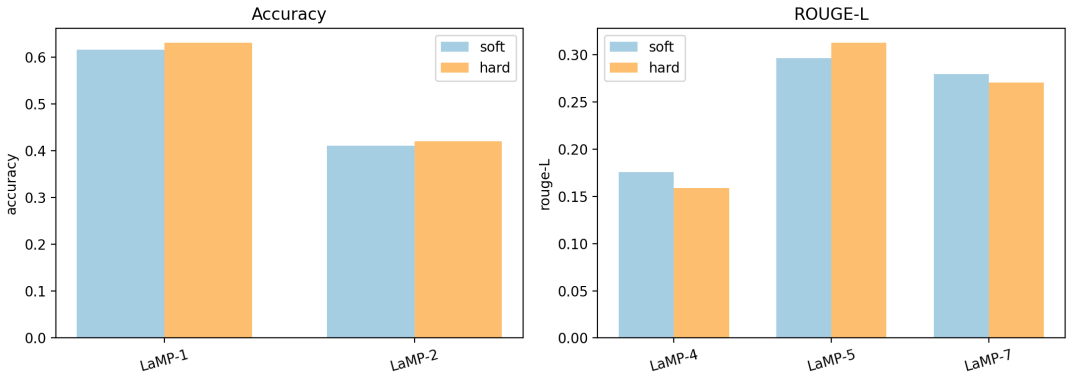

Figure 8: Comparison of hard and soft routing strategies for cluster-aware DPS across LaMP tasks. Hard routing favors classification tasks, while soft routing yields more consistent improvements on generation tasks.

implemented, and verified by the authors.