SteerX: Disentangled Steering for LLM Personalization

Xiaoyan Zhao[1], Ming Yan[1], Yilun Qiu[1], Haoting Ni
Yang Zhang, Fuli Feng, Hong Cheng, Tat-Seng Chua
xzhao@se.cuhk.edu.hk,{ym689,nhting}@mail.ustc.edu.cn,{qiuyilun,zhangy}@nus.edu.sg
The Chinese University of Hong Kong, University of Science and Technology of China, National University of Singapore

Abstract

Large language models (LLMs) have shown remarkable success in recent years, enabling a wide range of applications, including intelligent assistants that support users' daily life and work. A critical factor in building such assistants is personalizing LLMs, as user preferences and needs vary widely. Activation steering, which directly leverages directions representing user preference in the LLM activation space to adjust its behavior, offers a cost-effective way to align the model's outputs with individual users. However, existing methods rely on all historical data to compute the steering vector, ignoring that not all content reflects true user preferences, which undermines the personalization signal. To address this, we propose **SteerX**, a disentangled steering method that isolates preference-driven components from preference-agnostic components. Grounded in causal inference theory, SteerX estimates token-level causal effects to identify preference-driven tokens, transforms these discrete signals into a coherent description, and then leverages them to steer personalized LLM generation. By focusing on the truly preference-driven information, SteerX produces more accurate activation steering vectors and enhances personalization. Experiments on two representative steering backbone methods across real-world datasets demonstrate that SteerX consistently enhances steering vector quality, offering a practical solution for more effective LLM personalization.

CCS Concepts

• **Computing methodologies** → **Natural language generation**;
• **Information systems** → **Personalization**.

Keywords

Large Language Model, LLM, LLM Personalization, Personalized Text Generation, Causal Inference, Activation Steering

ACM Reference Format:
Xiaoyan Zhao[1], Ming Yan[1], Yilun Qiu[1], Haoting Ni and Yang Zhang, Fuli Feng, Hong Cheng, Tat-Seng Chua. 2018. SteerX: Disentangled Steering for LLM Personalization. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Equal contribution.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

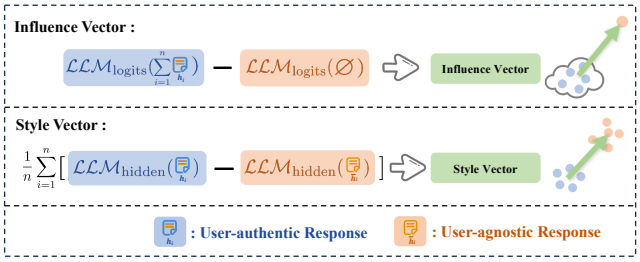

Figure 1: Two representative training-free activation steering methods. *Influence Vector* is the logit difference with vs. without history. *Style Vector* is the mean hidden-state difference between user-authentic response h_i and user-agnostic response $\bar{h}_i$.

1 Introduction

Large language models (LLMs) have achieved remarkable success over the past two years [2, 10, 12, 59], giving rise to a wide range of applications [22, 71]. Among these, leveraging LLMs for building intelligent assistants [55, 73] has emerged as one of the most promising directions to support everyone's daily life and work. Many efforts from both academia and industry have been devoted to this direction [52, 70], leading to early products such as Apple Intelligence[1] and the Meta AI app[2]. A key factor in building such assistants is personalization. LLMs are originally designed under a one-size-fits-all paradigm; however, when integrated into individuals' daily lives and work, they inevitably encounter diverse user preferences and needs [22]. Personalizing LLMs—adapting them to accommodate individual differences—is essential for delivering high-quality service to enhance user experiences, giving rise to the emerging field of LLM personalization [8, 21, 25, 57, 69].

For LLM personalization, activation steering has emerged as a promising and cost-effective solution [5, 7, 13, 64]. This method directly modifies the LLM's internal states to achieve desired outputs by adding a vector—a direction in the activation or hidden state space that can affect the characteristics of the model's responses. Based on how the steering vector is computed for LLM personalization, two main methods have emerged. The first derives a user's "style vector" [64] by comparing historical user-authentic responses (containing both content and style) with user-agnostic responses (content-only) and then uses it to steer the model's output toward the user's style. The second computes an influence vector [13] from the used historical user textual data on the model activations and amplifies it to strengthen personalization. Both approaches can effectively adjust model behavior to better align with user preferences, with promising results demonstrated across various scenarios.

[1]https://www.apple.com/apple-intelligence/
[2]https://ai.meta.com/meta-ai/

Despite notable progress made by these efforts, we argue that a common limitation is that both of their steering vector computations rely blindly on all the historical user data, overlooking the fact that not all parts of the user data reflect the user's preferences. Even though the textual data is produced by the user, each data piece often contains heterogeneous information. Different elements play different roles and can naturally show different degrees of relevance to user preferences [74]. For example, there are some boilerplate parts [62], e.g., common connecting words or sentences, which usually do not reflect user preference but only work as connecting functions. Treating all components as equally relevant risks dilutes the true preference signal and introduces noise into the steering vector, finally preventing the personalization quality.

In this work, we propose to disentangle preference-driven components from less preference-related ones, focusing on deriving the steering vector from the preference-driven components to enable more accurate personalization. This task is non-trivial, as there is no direct signal indicating which parts of the data are driven by user preferences. Inspired by prior studies [68, 74], we employ a causal analysis framework to address this challenge. As illustrated in Figure 2, the observed user data arises from the interaction between the user and the context. According to causal inference theory, determining whether a component is preference-driven requires measuring the extent to which it is influenced by the user's inherent traits. This influence corresponds to the causal effect of the user on the observed data, quantified by the difference between the factual world and a counterfactual world where a reference user (without individual preference) generates the data. By leveraging the magnitude of these causal effects, we can effectively identify the preference-driven components of each user data instance.

To address this, we propose SteerX, a disentangled steering method inspired by causal analysis. SteerX follows a disentangling–smoothing–steering paradigm for LLM personalization: it disentangles preference-driven tokens through causal effect estimation, transforms them into smooth descriptions, and then leverages them in the steering process. Specifically, we first estimate token-level causal effects: for each historical data instance, we represent the user trait using all other histories, then compare the factual prediction probabilities of the target history tokens (with the user trait) against their counterfactual predictions (without the user trait) to approximate causal effects via the LLM. Based on these effects, tokens are classified as either preference-driven or not. Since the extracted preference-driven tokens are discrete and lack smooth semantics, we reconstruct them into coherent, fluent sentences to facilitate downstream steering. Finally, these smoothed preference-driven segments replace the original history data in existing steering methods to achieve personalization.

The main contributions of this work are summarized as follows:

- To the best of our knowledge, we are the first to study the disentanglement of preference-driven and other components in user data to enable more accurate steering for LLM personalization.
- We propose a novel disentangled steering method, **SteerX**, for LLM personalization, which leverages causal effects to separate preference-driven components from user data, followed by a smoothing–steering process to achieve personalization.
- We apply SteerX to two existing representative steering methods, and extensive experiments on multiple real-world datasets demonstrate that it effectively enhances activation steering accuracy, thereby improving LLM personalization.

2 Related Work

2.1 Personalized Text Generation

With the rapid development of LLMs, these models have shown exceptional generative performance across a wide range of tasks, revolutionizing many areas of artificial intelligence [1, 9, 17, 38, 49, 72]. However, despite their impressive capabilities, LLMs still struggle to adapt to individual preferences and user needs. As users increasingly demand personalized interactions, this limitation has sparked growing interest in LLM personalization [8, 21, 25, 57, 69], with researchers focusing on how to tailor these models to provide more relevant, context-aware, and user-specific responses.

Recently, LLM personalization has been extensively explored across various domains, including image generation [41, 56, 58], conversational agents [19, 20, 28], recommendation systems [24, 51, 61], and multimedia applications [6, 32, 53]. Among these, personalized text generation remains a fundamental and central topic within the field of LLM personalization [18, 35, 36, 40], as it directly impacts the quality and relevance of user interactions. Existing methods for personalized text generation can be mainly categorized into two groups: 1) retrieval-based methods [43, 76], where ROPG [39] proposes a retrieval selection model that dynamically chooses the best retriever to capture valuable personalized information, while DPL [36] focuses on extracting inter-user differences to improve LLM personalization; and 2) fine-tuning-based methods [46, 65], where OPPU [47] learns a set of parameter-efficient adapters for each user, while PPlug [26] encodes the user history as a soft prompt.

2.2 Causal Inference in LLMs

Causal inference provides a principled framework to reason about cause-effect relationships, and its integration with LLMs has recently attracted increasing attention [11, 27, 75]. Unlike purely correlational approaches, causal methods aim to disentangle spurious correlations from genuine causal effects, which is crucial for a broader range of NLP tasks [16, 44, 63]. For instance, in recommender systems, causal techniques have been applied to better model user-item interactions, mitigate biases, and improve personalization [54, 67, 68]. NextQuill [74] is the first work to systematically apply causal inference to personalized text generation. It proposes a unified causal framework that models user preference effects from both the model side and the data side, demonstrating the importance and effectiveness of causal inference for enhancing LLM personalization.

2.3 Activation Steering

Activation steering focuses on controlling the behavior of LLMs by directly injecting a direction vector into their internal activations during inference, rather than modifying model weights or input prompts [42, 45, 77]. This research is motivated by recent findings that LLMs can control specific semantics or behaviors via linear directions in their activation space [29]. Building upon this, several

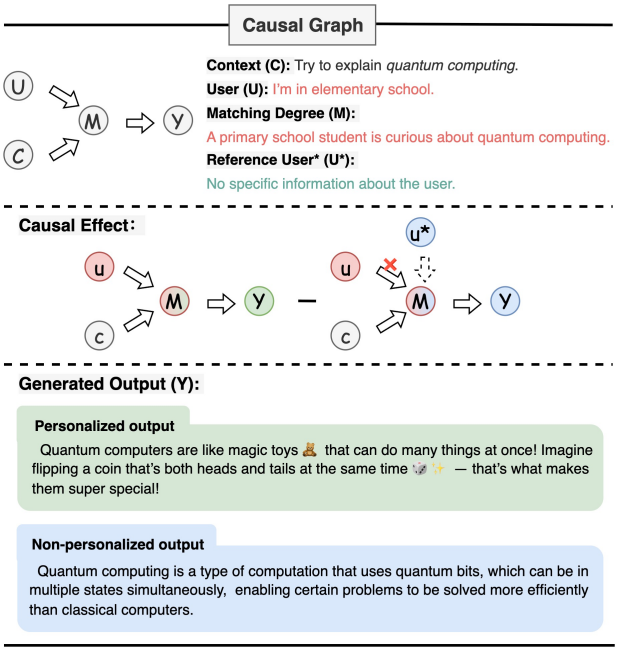

Figure 2: Illustration of our causal inference framework for personalized text generation.

methods have been proposed: ActAdd [48] performs activation addition using contrastive-derived steering vectors for controlling sentiment and topic, while CAA [37] enhances steering precision by leveraging mass-mean activation differences.

Recently, some works [5, 7] have begun exploring activation steering for LLM personalization, focusing on two main implementations: influence vector [13] and style vector [64], which show promising potential for adapting model behavior to individual users. However, these steering-based methods overlook the fact that user data is not entirely produced by the user and that each data piece often contains heterogeneous information, leading to suboptimal performance in LLM personalization. To the best of our ability, our proposed SteerX addresses this limitation.

3 Preliminary: Activation Steering

Let $x \in \mathcal{X}$ denote the input (e.g., a user-specific query) and $y \in \mathcal{Y}$ denote the target output. P is the **task prompt** describing the desired task. Each historical interaction consists of an input x_i and its corresponding *authentic* user-specific response h_i. The user's total history is

$$H = \{(x_1, h_1), (x_2, h_2), \ldots, (x_n, h_n)\} \tag{1}$$

$\mathcal{LLM}_{\text{logits}}(y \mid \cdot)$ denotes the output logits of the language model for y given the specified input condition. $\mathcal{LLM}_{\text{hidden}}(\cdot)$ is the hidden state representation of the LLM at a chosen layer ℓ.

Activation steering modifies intermediate hidden states during inference to incorporate user-specific preferences. Current methods can be divided into two categories: *influence vector* and *style vector*. Both compute a steering direction $a \in \mathbb{R}^d$ in the activation space and add it to the model's hidden states during inference, but differ in how a is derived.

3.1 Influence Vector

The **influence vector** (IV) measures the change in the model's logits caused by the user's histories, a training-free steering mechanism. The user-specific preferences can be obtained by the logits difference between the output of LLM with the user histories H (*i.e.,* inputting both H and prompt P) and the output of LLM without the user history (*i.e.,* inputting P only). A representative activation steering method of this approach is Context Steering (CoS) [13], which applies influence vectors to control personalization strength directly from user histories.

Formally, let $\mathcal{LLM}_{\text{logits}}(x, H; P)$ and $\mathcal{LLM}_{\text{logits}}(x, \varnothing; P)$ denote the generation logits given the input x with and without context, the activation steering direction is:

$$a_{IV} = \mathcal{LLM}_{\text{logits}}(x, H; P) - \mathcal{LLM}_{\text{logits}}(x, \varnothing; P). \tag{2}$$

Scaling this direction by λ can modulate the personalization strength during LLM generation:

$$\text{Steer}_{IV}(x, H; P) = \mathcal{LLM}_{\text{logits}}(x, H; P) + \lambda \cdot a_{IV}. \tag{3}$$

3.2 Style Vector

The **style vector** (SV) encodes user-specific preferences derived from the user's historical responses, also as a training-free steering mechanism. It is derived by contrasting *user-authentic* responses against *user-agnostic* responses generated by a general-purpose LLM, isolating the stylistic component independent of task content. A representative method is the Contrastive Activation Steering (CAS) [64] method, which gets the style direction of each instance and then aggregates them to obtain an overall style representation for the user.

Given each history pair (x_i, h_i), a style-neutral response $\bar{h}_i$ is first generated with task prompt P:

$$\bar{h}_i = \mathcal{LLM}(x_i; P) \tag{4}$$

The style direction is computed as the mean hidden-state difference:

$$a_{SV} = \frac{1}{n} \sum_{i=1}^{n} \left[\mathcal{LLM}_{\text{hidden}}(h_i) - \mathcal{LLM}_{\text{hidden}}(\bar{h}_i) \right]. \tag{5}$$

During inference, steering is applied to obtain the personalized generation:

$$Steer_{SV}(x, H; P) = \mathcal{LLM}_{\text{hidden}}(x, H; P) + \gamma \cdot a_{SV}, \tag{6}$$

where γ controls personalized strength.

Comparison. As illustrated in Figure 1, the IV is computed by contrasting model outputs with and without the entire historical context for a given (H, P) pair, capturing the global effect of history. In contrast, the SV is computed at the instance level by contrasting $(h_i, \bar{h}_i)$ pairs, then averaged across all histories to form a global style representation.

4 Methodology

To overcome the indiscriminate use of full user histories in existing activation steering methods, we propose **SteerX**, a three-stage (*"disentangling–smoothing–steering"*) disentangled steering framework

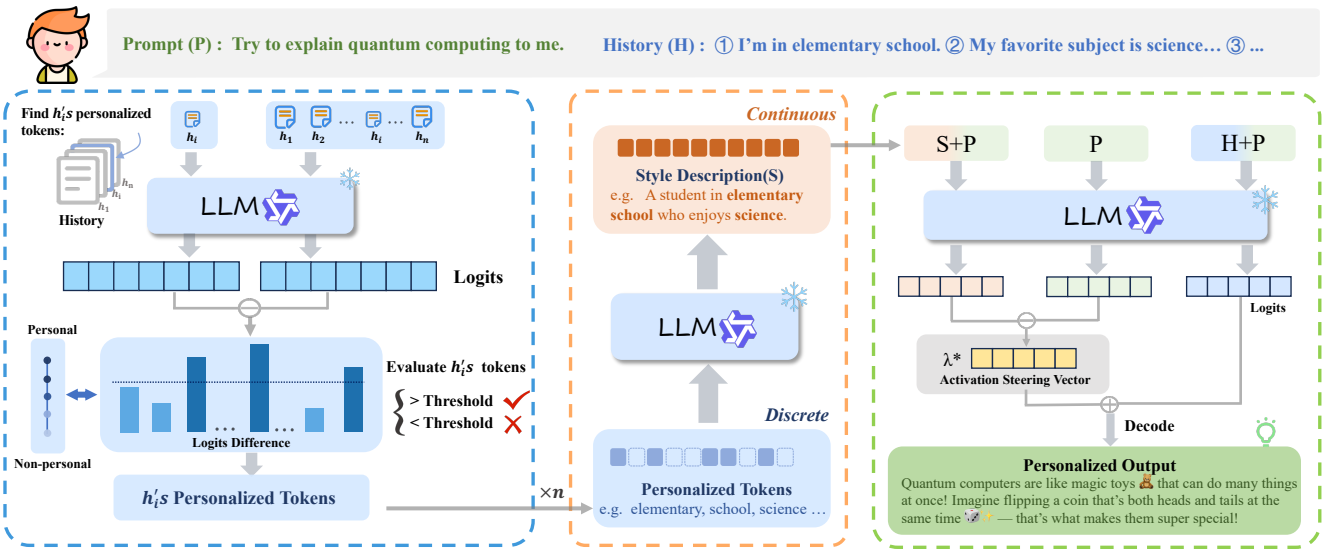

Figure 3: Overview of the proposed SteerX framework for personalized text generation. (Left) *Personalized anchor token identification*: For each history h_i, we apply a leave-one-history-out strategy and mark tokens with causal influence scores exceeding a predefined threshold as *personalized tokens*. (Middle) *Style profile generation*: All personalized tokens are transformed from a discrete token set into a continuous, coherent *preference description* S. (Right) *Personalization steering*: S is injected into the generation process. A personalization scaling factor λ controls the strength of steering based on the style vector.

for enhancing personalized text generation. As shown in Figure 3, grounded in causal inference theory, SteerX estimates token-level causal effects to identify preference-driven tokens (*disentangling*), transforms these discrete signals into a coherent natural-language description (*smoothing*), and leverages this representation to steer personalized LLM generation (*steering*). Next, we first present the causal foundation of the work, followed by the detailed description of the three steps of SteerX.

4.1 Task Formulation as Causal Graph

Personalized text generation involves multiple intertwined factors, including the specific task context faced by the user, the user's inherent preferences, and the interaction between the two. Therefore, directly quantifying personalized components from observed behavior is challenging, as it is entangled by other non-preference-related influences. To explicitly disentangle the preference-driven components from these contextual effects, we formalize the process using a causal graph.

Formulation. We define the following variables:

- **Context (C):** the task scenario and objectives (or items) that the user need to response.
- **User (U):** the underlying preferences or interests of user.
- **Matching (M):** the interaction (matching) between the current context C and the user U.
- **Generated Output (Y):** the content produced by the user in response to the given context.

Factual Process. As shown in Figure 2, the factual generation process can be expressed as:

$$H \rightarrow M \leftarrow C, \quad M \rightarrow Y \tag{7}$$

Here, U and C jointly determine the matching variable M, which mediates the personalization effect on the generated output Y.

Counterfactual Process. To isolate the contribution of personalization, we need to define a counterfactual by intervening U to a reference user u^*, as shown in Figure 2:

$$do(U = u^*) \tag{8}$$

Here, u^* corresponds to a generic user without individual preference. This intervention severs the causal link from the original $U = u$ to M, while keeping C fixed. By comparing the output Y under the factual and counterfactual settings, we can obtain the personalization effect:

$$\begin{aligned} \Delta(Y) &= P(Y \mid do(U = u), C) - P(Y \mid do(U = u^*), C) \\ &= P(Y \mid U = u, C) - P(Y \mid U = u^*, C) \end{aligned} \tag{9}$$

The causal effect $\Delta(Y)$ can be attributed to the personalization pathway $U \rightarrow M \rightarrow Y$.

4.2 Causal Analysis

In the context of LLM personalization, we approximate the *user* U by the *user history* H, while the *context* C corresponds to the task instructions and any accompanying situational information (including the item/objective the user needs to response to). Our goal is to isolate the *preference-driven components* within a user's data instance—i.e., the specific parts of the content that are causally attributable to the personalization pathway $H \rightarrow M \rightarrow Y$.

To achieve this, we require a quantitative measure of personalization influence at the *token level*. By comparing the generation behavior under factual and counterfactual conditions, we can determine how strongly each token reflects the user's preference,

enabling the identification of tokens that carry user-preference-related signals.

Token-Level Causal Effect. After representing each user with their history, the token-level causal effect for t-th token Y^t in Y can be defined as follows, based on Equation (9):

$$\Delta_t = \log p(Y^t \mid H, C) - \log p(Y^t \mid H^*, C) \tag{10}$$

where H^* denotes the corresponding reference that can be set to $\varnothing$ to indicate no preference. A larger Δ_t indicates a stronger influence from the personalization pathway $H \longrightarrow M \longrightarrow Y$, which is obvious.

Anchor Preference Token Identification. We can then define the set of *anchor preference tokens* as:

$$\text{AnchorTokens}(y) = \{\, Y^t \mid \Delta_t \geq \tau \,\} \tag{11}$$

where τ is a fixed threshold or is chosen adaptively so that the top-$k\%$ tokens with the highest Δ_t values are selected, treating them as preference-driven tokens.

4.3 Disentangling: Personalized Anchor Token Identification

Building on the causal effect analysis in Section 4.2, which identifies *preference-driven tokens* in the generated output, we now turn to the *input side* to locate the corresponding preference-driven components within the user history. The goal is to extract the parts of the history that contribute most to the personalization signal, thereby enabling the steering vector to be derived from truly preference-relevant content rather than from the entire history.

Leave-One-History-Out Strategy. Let $H = \{h_1, \ldots, h_n\}$ denote a user's history, omitting the input part x_i of each data point here for brevity. To identify preference tokens for a given history $h \in H$, we treat h as the outcome variable Y in Equation (10), and represent the user with the remaining histories $H^- = H \setminus \{h\}$. The causal effect of a token h^t in h is then computed as, based on Equation (10):

$$\Delta_t = \log p(h^t \mid H^-, C) - \log p(h^t \mid \varnothing, C), \tag{12}$$

where $p(\cdot)$ is estimated using LLMs.

Tokens with larger Δ_t are considered more preference-driven. Specifically, tokens whose attribution scores exceed a predefined threshold τ are collected to form the *anchor preference token set* for history h_i:

$$\mathcal{T}_{\text{anchor}} = \{\, h^t \in h \mid \Delta_t \geq \tau \,\} \tag{13}$$

This procedure disentangles the preference-driven components of the user history from those less related to preferences.

4.4 Smoothing: Style Description Generation

In Section 4.3, we disentangled the preference-driven parts of the user history from the less preference-related ones. Although these tokens offer fine-grained attribution, they are discrete and fragmented, lacking global coherence and contextual continuity necessary for effective personalization.

To steer LLM personalization more effectively, we transform the identified personalized anchor tokens into a compact, interpretable, and coherent representation. Specifically, we apply a coherence transformation to $\mathcal{T}_{\text{anchor}}$ to obtain the more personalized style profile:

$$s = \text{Coherence}(\mathcal{T}_{\text{anchor}}) \tag{14}$$

where $\text{Coherence}(\cdot)$ employs an LLM to generate a fluent, context-aware description that summarizes the key content/style reflected in the anchor tokens. Applying this process to each historical data point's anchor tokens yields the complete preference-driven portion of the all user history for steering, denoted by $S = \{s_1, \ldots, s_n\}$, where s_i denotes the result for the i-th history.

4.5 Steering: Personalized Generation

Following the *disentangling–smoothing–steering* paradigm, the previous two stages isolated preference-driven components from the user history (*disentangling*) and transformed them into a compact, coherent style description S (*smoothing*). In this final *steering* stage, we build on the activation steering paradigms—*influence vector* and *style vector*—to incorporate S as an explicit and interpretable personalization signal for guiding LLM generation at inference time. Unlike prior methods [13, 64] that directly use raw user history or unstructured style representations, we replace the original history with the preference-driven parts S, deriving the steering vector from preference-driven components to achieve more accurate personalization.

Inference Vector with Style Description. For the influence vector setting, given a prompt P and treating S as the context, the steering direction is

$$F_{S,P}(x_i) = \mathcal{LLM}_{\text{logits}}(x_i, S; P) - \mathcal{LLM}_{\text{logits}}(x_i, \varnothing; P) \tag{15}$$

and the steered logits become

$$\text{Steer}_{IV}(x, H; P) = \mathcal{LLM}_{\text{logits}}(x_i, H; P) + \lambda \cdot F_{S,P}(x_i) \tag{16}$$

where λ controls the strength of S's influence.

Style Vector with Style Description. For the style vector setting, we enhance CAS [64] by incorporating the style description S into the hidden-state difference computation. Given each history pair (x_i, h_i), we first generate a style-neutral response $\bar{h}_i = \mathcal{LLM}(x_i; P)$. The refined style direction is then computed by averaging three hidden-state differences: the user-authentic response vs. style-neutral, the style description vs. style-neutral, and (optionally) the original CAS term. Formally, at layer ℓ:

$$\begin{aligned} a_{SV} = \frac{1}{n}\Bigg[\sum_{i=1}^{n} \gamma_1 \cdot \big(\mathcal{LLM}_{\text{hidden}}(h_i) - \mathcal{LLM}_{\text{hidden}}(\bar{h}_i)\big) \\ + (1-\gamma_1) \cdot \big(\mathcal{LLM}_{\text{hidden}}(s_i) - \mathcal{LLM}_{\text{hidden}}(\bar{h}_i)\big) \Bigg] \end{aligned} \tag{17}$$

During inference, steering is applied to the hidden states as:

$$\text{Steer}_{SV}(x, H; P) = \mathcal{LLM}_{\text{hidden}}(x, H; P) + \gamma_2 \cdot a_{SV} \tag{18}$$

where γ_1, γ_2 controls the steering strength. Here, S can also be used to filter or select history entries most aligned with the target style, further improving the relevance and precision of the steering signal.

By introducing the structured style profile S into activation steering, we enable fine-grained personalization control without retraining the base model, directly linking more preference-driven parts of history content to output personalization, and condensing personalization cues into a single steering representation for flexible control during generation.

5 Experiments

In this section, we conduct extensive experiments in real-world datasets to answer the following research questions:

- **RQ1:** How does our proposed SteerX perform on the personalized text generation tasks compared to existing baseline methods?
- **RQ2:** What is the contribution of each individual design of SteerX to its overall performance?
- **RQ3:** How do the specific hyperparameters and settings in SteerX affect its overall performance?
- **RQ4:** In what ways does SteerX illustrate its personalized generation capabilities through qualitative case studies?

5.1 Experimental Setup

Datasets. Building upon prior work, our experiments focus on the review generation[3] [14] and topic writing[4] [50], which serve as two representative personalized text generation tasks. Following the standard settings of OPPU [47], we form the test dataset by selecting the 100 most active users with the longest historical records. For review generation, we use the *Movies & TV* category from Amazon, and for topic writing, we use *Reddit* posts.

Baselines. We conduct a comprehensive comparison between SteerX and a set of existing baseline methods:

- **Non-Perso:** This represents the non-personalized setting, where all user-specific information is excluded from the model input.
- **RAG [40]:** This method employs a retrieval approach to extract key user-specific information, which is then used as instructional context for personalization.
- **COS (Context Steering) [13]:** This method amplifies the influence of context by comparing LLM outputs with and without the context and linearly scales this effect to control the degree of personalization at inference time.
- **CAS (Contrastive Activation Steering) [64]:** This method generates style-agnostic responses for a user's history with a base LLM, derives style vectors by contrasting hidden activations between real and neutral responses, and steers generation at inference via linear activation interventions using these vectors.

Notably, COS and CAS are the most recent state-of-the-art steering methods, representing the influence vector and style vector approaches, respectively. We apply our method to both methods and compare them with their original versions to verify the superiority of our disentangled steering over different LLM backbones.

Evaluation metrics. To provide a rigorous and comprehensive evaluation of our method, we compute multiple metrics that quantify the correspondence between the generated outputs and the ground-truth data. Specifically, we use standard metrics commonly adopted in personalized text generation tasks [18, 30, 40], including ROUGE-1 [23], METEOR [3], BLEU[5] [34], and BERTScore[6] [66]. Among these, we take METEOR as the primary evaluation metric owing to its extensive adoption in personalized text generation.

Implementation Details. We conduct experiments on two open source LLMs of different sizes [59]: `Qwen3-8B`[7] and `Qwen3-14B`[8]. We use the `Contriever` [15] model to perform a retrieval-based ranking of the current input prompt against the user's historical samples. The top two ranked samples are selected as historical context for all included methods. Subsequently, these two historical samples are combined with the current context and embedded using the `Qwen2.5-1.5B`[9] [60]. We then apply the method described in Section 4.2 to obtain personalized anchor tokens. After generating anchor tokens from historical samples, we concatenate them into a single sequence and feed it into the `Qwen3-14B` [59] model to produce a fluent, context-aware style description that captures the user's stylistic and preference tendencies. Finally, we generate the personalized output based on the method described in Section 4.5. As for the hyperparameters, for the review generation task, we set $(\gamma_1, \gamma_2, l, \lambda, \tau)$ to $(0.4, 0.4, 20, 0.7, 1.0)$ for the `Qwen3-8B` model and $(0.4, 0.1, 20, 0.7, 1.0)$ for the `Qwen3-14B` model, while for the topic writing task, we set $(\gamma_1, \gamma_2, l, \lambda, \tau)$ to $(0.1, 0.1, 32, 0.75, 0.7)$ for the `Qwen3-8B` model and $(0.1, 0.1, 20, 0.3, 0.7)$ for the `Qwen3-14B` model. For inference, we disable sampling and always select the token with the highest probability, with the maximum generation length set to 2048 tokens. Experiments are run on NVIDIA A100 and A800 GPUs with 80GB of GPU memory.

5.2 Main Results (RQ1)

We begin by evaluating the overall performance of all compared methods. The main experimental results across the two personalized text generation tasks are presented in Table 1, from which we can draw the following observations:

- **Leveraging user history boosts the model's ability to generate personalized text.** For both LLM sizes, Non-Person demonstrates the weakest personalization performance across two tasks, reflecting its inability to utilize user-specific contextual information to guide personalized generation. In contrast, RAG achieves notable gains by retrieving and incorporating user-specific data, producing more tailored and relevant output.
- **Activation steering approaches enhance personalization by directly modulating model behavior.** Methods including CAS and COS achieve general achievements in both personalized tasks compared to RAG. By injecting user-specific directional signals into the model's internal activations, these approaches enable the LLM to produce outputs that more accurately reflect individual user preferences, highlighting the promise of steering-based personalization.
- **SteerX achieves further improvements on LLMs of different sizes across tasks and metrics.** By disentangling the preference-driven component from user data for steering, our method enhances the quality of the steering vector, thereby improving personalization and generally outperforming the original steering method. Moreover, these improvements persist across diverse LLM backbones and steering methods, demonstrating both the effectiveness and the broad applicability of our approach.

[3] Processed dataset: https://huggingface.co/datasets/SnowCharmQ/DPL-main [36].
[4] Processed dataset: https://huggingface.co/datasets/LongLaMP/LongLaMP [18].
[5] We utilize the standard `SacreBLEU` [33] library to calculate the BLEU score here: https://github.com/mjpost/sacrebleu.
[6] We adopt the `led-base-16384` [4] model to obtain embeddings for calculating the BERTScore: https://huggingface.co/allenai/led-base-16384.
[7] https://huggingface.co/Qwen/Qwen3-8B
[8] https://huggingface.co/Qwen/Qwen3-14B
[9] https://huggingface.co/Qwen/Qwen2.5-1.5B

Table 1: Performance comparison between baseline methods and the proposed SteerX across two personalized text generation tasks using LLMs of two different sizes. For each LLM, the best result is shown in bold. Higher values indicate better performance across all metrics. *Impr.* denotes the relative average improvement across four metrics for each method compared with the Non-Person baseline.

Datasets (→)	Review Generation					Topic Writing				
Methods (↓)	ROUGE-1	METEOR	BLEU	BERTScore	*Impr.* ↑	ROUGE-1	METEOR	BLEU	BERTScore	*Impr.* ↑
Qwen3-8B (Non-Perso)	0.2903	0.1911	1.9381	0.4646	—	0.2380	0.1649	1.1104	0.4448	—
+ RAG	0.3061	0.2090	2.8455	0.4810	16.29%	0.2527	0.1913	2.5582	0.4587	38.92%
+ CAS	0.3119	0.2099	2.8429	0.4880	17.25%	0.2554	0.1896	2.6874	0.4581	41.83%
+ SteerX (CAS)	0.3118	0.2140	3.1307	**0.4898**	21.59%	0.2624	0.1940	2.6344	0.4600	42.14%
+ COS	0.3149	0.2168	3.1092	0.4839	21.63%	0.2625	0.2086	3.0324	0.4604	53.35%
+ SteerX (COS)	**0.3156**	**0.2238**	**3.5912**	0.4848	**28.87%**	**0.2658**	**0.2089**	**3.2097**	**0.4617**	**57.81%**
Qwen3-14B (Non-Perso)	0.2835	0.1884	1.8485	0.4641	—	0.2648	0.1952	1.8238	0.4509	—
+ RAG	0.3143	0.2184	2.8731	0.4845	21.65%	0.2893	0.2221	2.4741	0.4701	15.74%
+ CAS	0.3205	0.2199	2.9782	0.4847	23.83%	0.2918	0.2148	2.5982	0.4711	15.48%
+ SteerX (CAS)	**0.3209**	0.2191	3.0231	**0.4861**	24.44%	0.2937	0.2154	2.5961	0.4728	17.12%
+ COS	0.3146	0.2184	3.1719	0.4838	25.68%	0.2861	0.2237	2.0095	0.4674	9.12%
+ SteerX (COS)	0.3201	**0.2243**	**3.5787**	0.4855	**32.54%**	**0.3063**	**0.2329**	**3.6407**	**0.4749**	**34.98%**

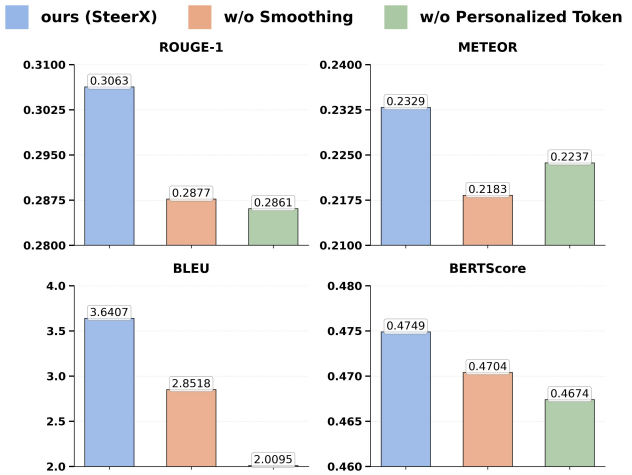

Figure 4: Ablation results on the topic writing task, comparing the full SteerX framework with two reduced variants: (1) w/o Smoothing and (2) w/o Personalized Token. Performance is reported across ROUGE-1, METEOR, BLEU, and BERTScore.

5.3 Ablation Studies (RQ2)

To assess the contribution of each component in our *disentangling–smoothing–steering* paradigm, we compare three variants on topic writing dataset: (1) **w/o Personalized Token**: the strong personalized baseline without the token-level disentangling step; (2) **w/o Smoothing**: using only the disentangled personalized anchor tokens for steering, without coherence transformation; (3) **Ours (SteerX)**: our proposed model with preference smoothing via the preference description S. The results are presented in Figure4.

Effectiveness of Personalized Tokens. Compared with **w/o Personalized Token**, the **w/o Smoothing** variant already yields notable improvements across all metrics, demonstrating that the identified anchor tokens indeed capture meaningful user preferences. This validates the effectiveness of our causal disentangling step in isolating preference-driven signals from noisy history.

Benefit of Smoothing. When replacing the discrete tokens with the smoothed preference description S in **Ours (SteerX)**, we observe further gains in all metrics. This indicates that the continuous, coherent representation better preserves contextual semantics and provides a stronger, more stable steering signal during generation. The improvement is consistent across LLM sizes, highlighting that the smoothing step not only enhances personalization quality but also improves robustness.

5.4 In-Depth Analysis (RQ3)

5.4.1 Token-Level Causal Influence Analysis. To examine the behavior of our disentangling stage, we analyze the distribution of token-level causal effects, computed as the logit difference between the *factual* setting (with full user history) and the *counterfactual* setting (with the history removed). This difference, as defined in Section 4.2, quantifies how strongly each token is influenced by personalization.

Figure 6 visualizes the distribution of causal effect scores across all generated tokens. We observe that a large subset of tokens exhibits low causal influence, while a smaller subset forms a long-tailed distribution with significantly higher scores. These high-influence tokens correspond to the *preference-driven components* identified by our method, which we term *personalized anchor tokens*. This highlights the necessity of our causal analysis for isolating truly personalization-relevant content from noisy background information.

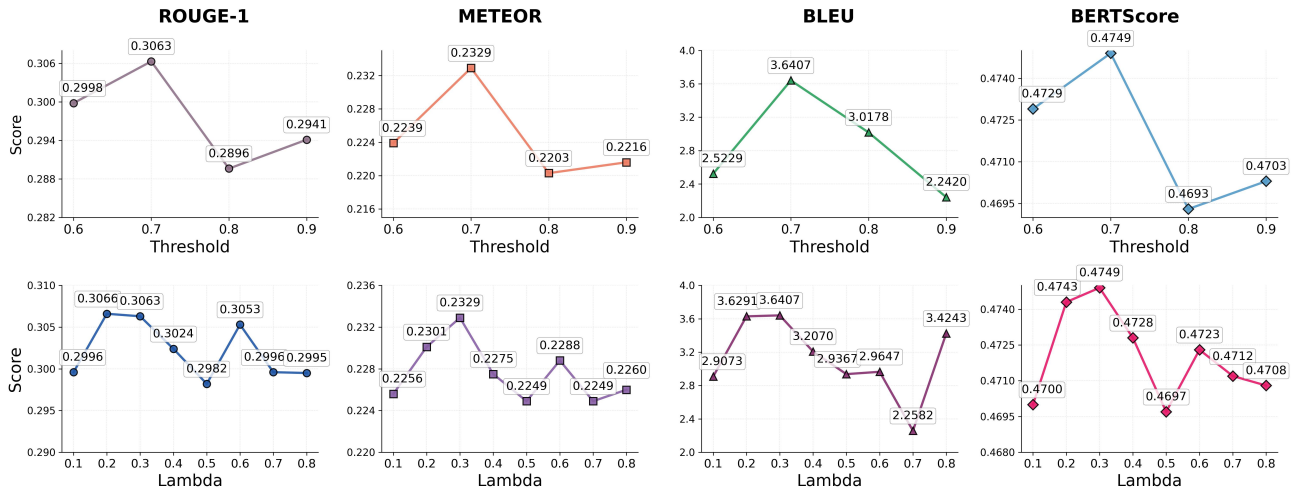

Figure 5: Effects of different hyperparameters on SteerX (COS) for the topic writing task: (1) upper panel — varying the causal effect threshold τ; (2) lower panel — varying the steering coefficient λ. Performance is evaluated using four metrics: ROUGE-1, METEOR, BLEU, and BERTScore.

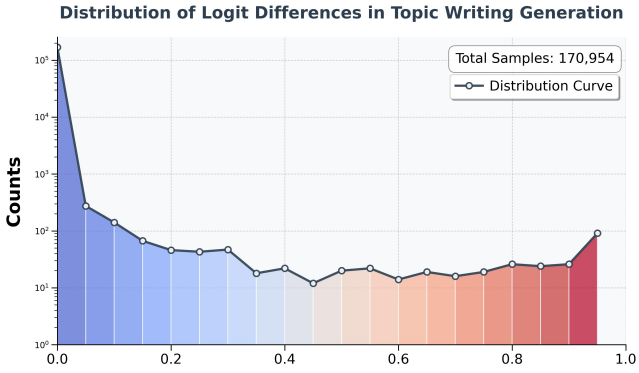

Figure 6: Distribution of token-level causal influence scores, quantified as logit differences between factual (with history) and counterfactual (without history) settings on the topic writing task. A subset of high scores corresponds to personalization-relevant tokens identified by SteerX.

5.4.2 Effect of Threshold τ on Personalized Token Selection. We further investigate the sensitivity of SteerX to the causal influence threshold τ, which determines which tokens are retained as personalized anchor tokens. The upper part of Figure 5 reports performance (ROUGE-1, METEOR, BLEU, BERTScore) on the development set for $\tau \in \{0.6, 0.7, 0.8, 0.9\}$.

Across all metrics, $\tau = 0.7$ consistently yields the best or near-best results, indicating a balanced trade-off between retaining sufficient personalization signals and excluding noisy tokens. Lower thresholds (e.g., $\tau = 0.6$) include more tokens but also admit irrelevant content, diluting the personalization signal. Conversely, higher thresholds ($\tau = 0.8, 0.9$) result in fewer selected tokens, improving precision but risking the omission of useful preference cues, which leads to performance drops in BLEU and ROUGE.

Overall, the performance curve remains relatively smooth, suggesting that SteerX is robust to moderate variations in τ. These findings highlight that a well-chosen threshold is essential for isolating preference-driven components while preserving sufficient coverage for coherent personalization.

5.4.3 Effect of the Steering Strength λ. We analyze the sensitivity of SteerX to the scaling factor λ that controls the magnitude of activation steering in the hidden-state space. The lower part of Figure 5 reports ROUGE-1, METEOR, BLEU and BERTScore on the development set for $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Performance peaks in the light-to-moderate regime ($\lambda \approx 0.2-0.3$; e.g., ROUGE-1 and BLEU reach their maxima near $\lambda = 0.2-0.3$), then gradually declines as λ increases. This pattern indicates a trade-off: small λ under-utilizes the preference signal, whereas large λ over-amplifies it and perturbs content fidelity. The curves are relatively smooth, suggesting robustness to moderate shifts in λ. Overall, these results support that steering in the activation space has a meaningful "sweet spot," where the learned direction remains valid while the scaling preserves a good balance between personalization strength and task performance.

5.4.4 Static LIWC Lexicon vs. Adaptive Causal Token Identification. To further validate the effectiveness of our personalized anchor token identification, we compare SteerX with a LIWC-based approach. LIWC [31] is an expert-crafted static lexicon widely used in psycholinguistics, containing predefined categories for personalization. These categories are designed by domain experts to map linguistic features to psychological traits and are applied uniformly across all text without adaptation to individual users. In our comparison, the LIWC-based method flags tokens in the generated text that match entries in its personalized dictionary, treating them as stylistic tokens.

Figure 7 shows that SteerX consistently outperforms LIWC across all four metrics (ROUGE-1, METEOR, BLEU, and BERTScore) with absolute gains. This performance gap highlights two critical

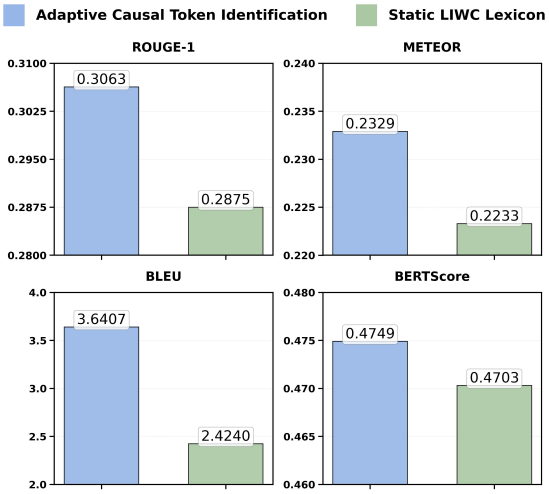

Figure 7: Performance comparison between our proposed adaptive causal token identification (SteerX) and the static LIWC lexicon across four evaluation metrics: ROUGE-1, METEOR, BLEU, and BERTScore.

differences: (1) LIWC's static nature limits its ability to capture domain-specific or idiosyncratic user preferences that fall outside its fixed vocabulary. (2) Our causal-effect-based identification dynamically adapts to each user's history, detecting tokens with empirically validated influence on the model's generation.

Besides, LIWC yields a sparse set of matched tokens, often missing subtle stylistic or topical cues. SteerX produces a richer and more contextually aligned token set, enabling finer-grained and more accurate personalization. This adaptivity is crucial in open-domain generation, where user style often diverges from standardized lexicons. This analysis highlights the advantage of SteerX in learning personalized token representations directly from causal influence signals, rather than relying on fixed, expert-defined resources.

5.5 Case Study (RQ4)

To illustrate the effectiveness of SteerX in personalized review generation, we present a representative example from the review generation task. Following our *disentangling–smoothing–steering* paradigm, SteerX first **disentangles** the user's history by identifying *personalized anchor tokens* through causal-effect analysis. These tokens are then **smoothed** into a coherent *style description*, which serves as an explicit personalization signal. Finally, this description is **used to steer** the generation process at inference time.

In the example shown in Figure 8, the style description highlights the user's tendency to blend emotional engagement with factual evaluation—combining expressive sentiment (e.g., "real gem") with concrete assessments (e.g., "five-star movie"). The generated review by SteerX aligns closely with the user's authentic review, not only reproducing the evaluative tone and stylistic markers but also capturing the rhythm and phrasing patterns characteristic of the user's past writing. Besides, it preserves lexical overlap with the user's expressions, integrates sentiment intensities, and provides informative detail. This case demonstrates SteerX's ability to go

Figure 8: Case study of SteerX's personalized review generation under the disentangling–smoothing–steering paradigm. SteerX preserves lexical overlap, sentiment intensity, and informative detail, aligning closely with the user's preferences.

beyond topical matching, generating text that faithfully reflects both the user's preferences and their distinctive communicative style—enhancing authenticity, engagement, and user alignment in personalized content.

6 Conclusion

In this work, we introduced SteerX, a disentangled steering framework for LLM personalization that addresses key limitations of existing activation steering methods. Rather than constructing steering vectors from an entire user history—risking the inclusion of preference-irrelevant content—SteerX adopts a causal inference perspective to explicitly isolate *preference-driven* components. By estimating token-level causal effects between factual and counterfactual histories, our approach pinpoints the tokens most influenced by intrinsic user traits. These high-impact tokens are then smoothed into coherent semantic descriptions, ensuring interpretability, reducing noise amplification, and enhancing compatibility with downstream steering. Extensive evaluations on multiple real-world personalization datasets and two representative steering backbones confirm that SteerX delivers consistent improvements on ROUGE, METEOR, BLEU, and BERTScore. In-depth analyses highlight that the causal disentanglement stage effectively filters out irrelevant background signals, while the smoothing stage preserves fluency and semantic consistency—together enabling more precise and robust steering vector construction.

Future Potential. SteerX offers a generalizable paradigm for causally grounded personalization, where steering is informed by the true drivers of user-specific behavior rather than surface correlations. This opens new opportunities for integrating fine-grained causal analysis into large-scale LLM personalization pipelines, enabling adaptive, interpretable, and privacy-conscious personalization strategies. Future work may extend this framework to multimodal personalization and real-time interactive systems, further advancing the alignment of LLM outputs with diverse user needs and preferences.

Ethical Consideration

Our method relies on the collection and analysis of users' historical interaction data, which necessitates the application of strict security measures and adherence to robust ethical standards. The creation of user profiles for personalization purposes carries inherent risks, such as unauthorized surveillance or targeted manipulation. To mitigate these risks, the approach should integrate comprehensive safeguards, including secure data storage, effective anonymization, and clear mechanisms for obtaining informed user consent. The handling of personal history data requires careful oversight to ensure it is not misused and remains protected throughout its lifecycle. All experiments in this work utilize open-source datasets. The research upholds principles of data integrity, privacy protection, and ethical responsibility, ensuring the fair and transparent use of open-source resources.

References

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 8–17.

[2] 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* (2023).

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

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

[5] Jessica Y Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D Shin. 2025. Steerable chatbots: Personalizing llms with preference-based activation steering. *arXiv preprint arXiv:2505.04260* (2025).

[6] Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. 2025. Large language models empowered personalized web agents. In *Proceedings of the ACM on Web Conference 2025*. 198–215.

[7] Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

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

[9] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Forty-first International Conference on Machine Learning, ICML 2024*. OpenReview.net.

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[11] Norman E. Fenton, Martin Neil, and Anthony C. Constantinou. 2020. The Book of Why: The New Science of Cause and Effect, Judea Pearl, Dana Mackenzie. Basic Books (2018). *Artif. Intell.* 284 (2020), 103286.

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

[13] Jerry Zhi-Yang He, Sashrika Pandey, Mariah L. Schrum, and Anca D. Dragan. 2025. Context Steering: Controllable Personalization at Inference Time. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net.

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

[15] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Trans. Mach. Learn. Res.* 2022 (2022).

[16] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can Large Language Models Infer Causation from Correlation?. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.

[17] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew M. Bean, Katerina Margatina, Rafael Mosquera Gómez, Juan Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

[18] 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* (2024).

[19] Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. Hello Again! LLM-powered Personalized Agent for Long-term Dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 5259–5276.

[20] Xintong Li, Jalend Bantupalli, Ria Dharmani, Yuwei Zhang, and Jingbo Shang. 2025. Toward Multi-Session Personalized Conversation: A Large-Scale Dataset and Hierarchical Tree Framework for Implicit Reasoning. *arXiv preprint arXiv:2503.07018* (2025).

[21] Xiaopeng Li, Pengyue Jia, Derong Xu, Yi Wen, Yingyi Zhang, Wenlin Zhang, Wanyu Wang, Yichao Wang, Zhaocheng Du, Xiangyang Li, et al. 2025. A Survey of Personalization: From RAG to Agent. *arXiv preprint arXiv:2504.10147* (2025).

[22] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459* (2024).

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

[24] Zijie Lin, Yang Zhang, Xiaoyan Zhao, Fengbin Zhu, Fuli Feng, and Tat-Seng Chua. 2025. IGD: Token Decisiveness Modeling via Information Gain in LLMs for Personalized Recommendation. *arXiv preprint arXiv:2506.13229* (2025).

[25] 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. *arXiv preprint arXiv:2502.11528* (2025).

[26] Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025. LLMs + Persona-Plug = Personalized LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 9373–9385.

[27] Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2025. Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics, 7668–7684.

[28] Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. 2025. Towards Lifelong Dialogue Agents via Timeline-based Memory Management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, 8631–8661.

[29] Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *Forty-first International Conference on Machine Learning, ICML 2024*. OpenReview.net.

[30] Qiyao Peng, Hongtao Liu, Hongyan Xu, Qing Yang, Minglai Shao, and Wenjun Wang. 2024. LLM: Harnessing Large Language Models for Personalized Review Generation. *arXiv preprint arXiv:2407.07487* (2024).

[31] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. (2015).

[32] Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros, and Yannis Panagakis. 2024. Investigating personalization methods in text to music generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1081–1085.

[33] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. 186–191.

[34] Matt Post. 2018. A call for clarity in reporting BLEU scores. *arXiv preprint arXiv:1804.08771* (2018).

[35] Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. 2025. Latent Inter-User Difference Modeling for LLM Personalization. *arXiv preprint arXiv:2507.20849* (2025).

[36] 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*. Association for Computational Linguistics, 21258–21277.

[37] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 15504–15522.

[38] Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. 2025. SynthesizeMe! Inducing Persona-Guided Prompts for Personalized Reward Models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 8045–8078.

[39] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization Methods for Personalizing Large Language Models through Retrieval Augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*. ACM, 752–762.

[40] 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)*. Association for Computational Linguistics, 7370–7392.

[41] Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM Web Conference 2024*. 3833–3843.

[42] Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. 2025. AlphaSteer: Learning Refusal Steering with Principled Null-Space Constraint. *arXiv preprint arXiv:2506.07022* (2025).

[43] Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. 2025. Retrieval Augmented Generation with Collaborative Filtering for Personalized Text Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025*. ACM, 1294–1304.

[44] Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. 2023. A Causal Framework to Quantify the Robustness of Mathematical Reasoning with Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 545–561.

[45] Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting Latent Steering Vectors from Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 566–581.

[46] Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6459–6475.

[47] 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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6476–6491.

[48] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv e-prints* (2023), arXiv–2308.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. 5998–6008.

[50] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*. 59–63.

[51] Chengbing Wang, Yang Zhang, Fengbin Zhu, Jizhi Zhang, Tianhao Shi, and Fuli Feng. 2025. Leveraging Memory Retrieval to Enhance LLM-based Generative Recommendation. In *Companion Proceedings of the ACM on Web Conference 2025*. 1346–1350.

[52] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.

[53] Yanan Wang, Yan Pei, Zerui Ma, and Jianqiang Li. 2024. A user-guided generation framework for personalized music synthesis using interactive evolutionary computation. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 1762–1769.

[54] Huizi Wu, Cong Geng, and Hui Fang. 2023. Causality and correlation graph modeling for effective and explainable session-based recommendation. *ACM Transactions on the Web* 18, 1 (2023), 1–25.

[55] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68, 2 (2025), 121101.

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

[57] Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. Personalized Generation In Large Model Era: A Survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 24607–24649.

[58] Yiyan Xu, Wuqiang Zheng, Wenjie Wang, Fengbin Zhu, Xinting Hu, Yang Zhang, Fuli Feng, and Tat-Seng Chua. 2025. DRC: Enhancing Personalized Image Generation via Disentangled Representation Composition. *arXiv preprint arXiv:2504.17349* (2025).

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

[60] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

[61] Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. 2025. Harnessing multimodal large language models for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 13069–13077.

[62] Ziang Ye, Zhenru Zhang, Yang Zhang, Jianxin Ma, Junyang Lin, and Fuli Feng. 2025. Disentangling Reasoning Tokens and Boilerplate Tokens For Language Model Fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria.

[63] Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, Zijun Yao, Jing Zhang, Lei Hou, and Juanzi Li. 2024. A Cause-Effect Look at Alleviating Hallucination of Knowledge-grounded Dialogue Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, 102–112.

[64] Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025. Personalized Text Generation with Contrastive Activation Steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 7128–7141.

[65] Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. 2025. PROPER: A Progressive Learning Framework for Personalized Large Language Models with Group-Level Adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 16399–16411.

[66] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net.

[67] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 11–20.

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

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

[70] Xiaoyan Zhao, Yang Deng, Wenjie Wang, Hong Cheng, Rui Zhang, See-Kiong Ng, Tat-Seng Chua, et al. 2025. Exploring the Impact of Personality Traits on Conversational Recommender Systems: A Simulation with Large Language Models. *arXiv preprint arXiv:2504.12313* (2025).

[71] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *Comput. Surveys* 56, 11 (2024), 1–39.

[72] Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 12564–12573.

[73] Xiaoyan Zhao, Ming Yan, Yang Zhang, Yang Deng, Jian Wang, Fengbin Zhu, Yilun Qiu, Hong Cheng, and Tat-Seng Chua. 2025. Reinforced Strategy Optimization for Conversational Recommender Systems via Network-of-Experts. *arXiv e-prints* (2025), arXiv–2509.

[74] Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2025. NextQuill: Causal Preference Modeling for Enhancing LLM Personalization. *arXiv preprint arXiv:2506.02368* (2025).

[75] Yaochen Zhu, Yinhan He, Jing Ma, Mengxuan Hu, Sheng Li, and Jundong Li. 2024. Causal inference with latent variables: Recent advances and future prospectives. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6677–6687.

[76] 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 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

[77] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405* (2023).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009