

Steerable Chatbots: Exploring Personalization Control Interfaces via LLM Activation Steering

Jessica Y. Bo^{*}
University of Toronto
Toronto, Ontario, Canada
jbo@cs.toronto.edu

Tianyu Xu[†]
Google AR
Mountain View, California, USA
tyx@google.com

Ishan Chatterjee
Google AR
Seattle, Washington, USA
ishanc@google.com

Katrina Passarella-Ward
Google AR
San Francisco, California, USA
kpassarella@google.com

Achin Kulshrestha[‡]
Google AR
Toronto, Ontario, Canada
kulac@google.com

D Shin^{†‡}
Google AR
Mountain View, California, USA
deshin@google.com

Activation steering can manipulate the expression and behaviours of LLMs, but how should these controls be exposed to users for personalization?

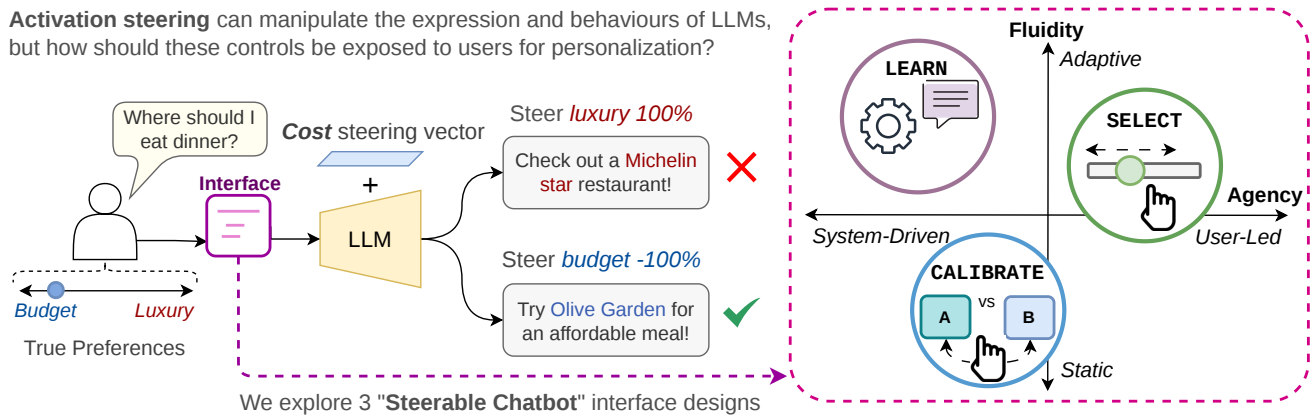


Figure 1: We explore how users can directly control the personalization of LLM-based chatbots to better align with their personal preferences, using a scalar factor to steer the model along a desired preference dimension. We evaluate three designs of 'steerable chatbots' that embody different personalization values of *fluidity* and *agency* – **SELECT**, **CALIBRATE**, and **LEARN**.

Abstract

Personalizing LLM responses typically requires users to articulate their preferences through prompting, which can be burdensome at cold start and difficult to articulate in natural language. We introduce an alternative paradigm, steerable chatbots: rather than asking users to describe what they want, let them directly manipulate it via a linear factor. We implement this through activation steering, leveraging a linear scalar to control how strongly a preference is expressed in the LLM's output. We first assess the computational viability of activation steering as a method to control granular preference expression, then we explore how the factor can be exposed

to users. We prototype three activation steering interface designs that vary on the axes of *agency* (user-led vs. system-driven) and *fluidity* (static vs. adaptive). A within-subjects user study ($n=14$) in cold-start personalization tasks shows the potential for steerable chatbots to align better with underlying user preferences than prompting alone, while revealing heterogeneous values around control, persistence, and transparency in LLM personalization.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

LLM Personalization, Activation Steering, Chatbot Interfaces

ACM Reference Format:

Jessica Y. Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D Shin. 2026. Steerable Chatbots: Exploring Personalization Control Interfaces via LLM Activation Steering. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3830398.3830533>

^{*}This work was conducted as a Student Researcher at Google.

[†]Now at Google DeepMind.

[‡]Equal supervision.



This work is licensed under a Creative Commons Attribution 4.0 International License. <https://creativecommons.org/licenses/by/4.0/>

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2856-3/2026/11

<https://doi.org/10.1145/3830398.3830533>

1 Introduction

Large language models' (LLMs) ability to generate tailored responses to their users' preferences and needs lends them to be increasingly used as personal assistants [53, 95]. In day-to-day planning tasks where preferences can be highly variable — such as picking a restaurant for an anniversary dinner — the LLM assistant needs to understand the user's underlying preferences and incorporate them in its response [17, 43]. Currently, LLMs are trained to assume the preferences of an average user [14, 34], while personalization relies on retrieving prior mentions of preferences from long-term memory [50, 63, 78], or well-specified prompts written by the user. The former is computationally intensive and does not support *cold-start* personalization [76], while the latter demands structured prompting skills that many lay users lack [64, 88, 100].

To shift this burden away from users, we explore the paradigm of *directly steerable* chatbot interfaces for LLM personalization, provisioning users with manipulatable control over aligning the LLM with their latent preferences. In our steerable chatbots paradigm, users would not need to describe their preferences through prompting, as they can achieve their desired personalization through manipulating a linear steering factor. Figure 1 displays an example of how a question “Where should I eat dinner?” can diverge significantly — steering towards *luxury* (+100) results in high-end Michelin restaurants, while steering towards *budget* (-100) produces casual chain restaurants that are matched to the user.

We leverage activation steering, a method for inducing desired behaviours in LLMs, and explore how it can be exposed through diverse user-facing controls for LLM personalization. Activation steering has emerged as a promising method for controlling LLM behaviours through injecting lightweight ‘steering vectors’ into the internals of the model [54, 85, 90]. During forward inference, the pre-computed vectors are added to the representations at the residual stream of the LLM to modulate its outputs. The strength of activation steering is set with a *linear factor*, which controls the expression of the behaviour with a direction (positive or negative) and a magnitude. This factor is persistent, prompt-independent, and mappable to user-facing controls. Activation steering provides computational benefits as it does not require any re-training of model weights like fine-tuning [32]; it offers more granular control than natural language prompting [73]; and it can be applied persistently or conditionally once the user's preferences are learned [49]. Although the technique has been primarily used in AI safety for model alignment, it creates an opportunity to expose the behavioural control of LLMs to users. Specifically, we adapt and apply the method developed by von Rütte et al.

Can activation steering be used to help users improve their experience in tasks with high variability in preferences? We perform an exploration into user-controlled direct manipulation interfaces for LLM personalization, identifying opportunities where such steerable chatbots are most useful. In Section 3, we lay out how existing activation steering methods can be used to parameterize personalization via controlling preference expressions. In Section 4, we computationally assess its viability for aligning with varied preferences in day-to-day lifestyle planning tasks. Finally, in Section 5, we evaluate how users interact with three steering interfaces that embody distinct design values (**SELECT**, **CALIBRATE**, and **LEARN**).

We contextualize our contributions within the domain of lifestyle planning, which is low-risk but has high variation in preferences across the population.

Summary of Contributions.

- (1) We introduce **steerable chatbots**, a personalization paradigm that exposes a direct manipulation LLM control parameter to users via activation steering. We iterate within the interaction design space of steerable chatbot interfaces and propose three diverse control interfaces.
- (2) We validate the effectiveness of preference-based activation steering through computational experiments across five open-source LLMs in lifestyle planning tasks.
- (3) In an exploratory user study ($n = 14$), we show that steerable chatbots have potential to improve preference alignment over an unscaffolded, prompting-only chatbot in cold-start personalization. Qualitative findings discuss how individual values shape what users prefer in steering interfaces.

2 Related Works

The related works synthesize literature on personalizing LLMs, activation steering, and the design values for personalization controls.

2.1 LLM Personalization

Research on personalization for LLMs is vast, ranging from adopting personas to customizing content [3, 17, 37, 56, 77, 87, 93]. Technical methods for personalization have focused on storing unique user preferences in data-efficient manners, such as through retrieval-augmented generation [78], model post-training [34, 52, 69, 76, 92], and blending different LLMs or learning reward models [16, 19, 34, 68, 71, 76, 81, 96]. These methods tend to require significant amounts of individual user data and would not be easily adaptable to cold start personalization scenarios. On the other hand, interface-driven methods focus on scaffolding the user interface to support better intent specification, such as through decomposing LLM outputs into sub-tasks [58], structuring the dimensions of the generation space [86], and interactively resolving ambiguities [57, 67]. These techniques tend to augment existing prompt-driven interactions, whereas we expose a direct manipulation linear steering factor for fine-grained control over preference expression.

2.2 Activation Steering

Activation steering is a method for manipulating the internal states of LLMs at inference time to control the model's output without any model post-training. This is achieved by adding a steering vector — computed by contrasting the desirable and undesirable representations of a behaviour — to the residual stream of the LLM. Steering has been commonly used in AI safety to steer away from hallucinatory or harmful responses [5, 11, 51, 73, 102] and aligning with other desired model behaviours [13, 27, 41, 44, 54, 84]. The effect of steering can be further amplified, reduced, or even inverted by controlling the steering strength, a numerical factor that is multiplied to the vector. This property of activation steering maps well to real-world preferences, which can be strong, weak, or in opposite directions. Currently, applications of activation steering for LLM personalization have not studied interactions with users, nor interface designs [31, 46, 101]. Our work focuses on exploring

the interaction design space of directly manipulatable LLMs, rather than improving the technical execution of activation steering. We adopt the existing steering method from von Rütte et al. and apply it to a new downstream context of lifestyle planning, which is highly relevant to personalization.

2.3 Manipulatable Personalization Interfaces

Interfaces for human-LLM collaboration predominantly rely on text inputs [20, 36, 42], but the linear nature of the steering factor opens an opportunity for direct manipulation of model behavior. We organize the design space around two axes relevant to personalization [79]: **agency** (how much control users have) and **fluidity** (how adaptively preferences are updated). The right-hand side of Figure 1 maps out how our interfaces fit within these dimensions, and Section 3.2 discusses the design rationale in more detail.

In the *user-led and adaptive* design, we ground the **SELECT** interface, which provides users real-time control via a dynamic slider, with the principle of direct manipulation [23, 61, 82]. This type of transparent and controllable personalization is well-supported: users value retaining explicit control over recommendations [6, 45, 55, 62, 66] and generative outputs [22, 26]. As some users may find sliders to lack interpretability [21], the *system-driven and persistent* design of **CALIBRATE** performs a calibration step to infer user preferences before the conversation begins. This method draws on the framework of active learning [75], Bayesian optimization [97], and item response theory [15]. Lastly, the *highly-adaptive and system-driven* **LEARN** interface operates without explicit user input, updating the steering factor in the background from conversational sentiment, analogous to feedback-driven approaches in LLM alignment [80] and recommender systems [40, 65]. This prioritizes an unobtrusive, temporally adaptive experience [8].

3 Steerable Chatbots Paradigm

In this section, we describe the proposed paradigm for steering-based LLM personalization, the design space for activation steering control interfaces, and real-life applications of this personalization method. We target everyday LLM-assisted tasks with *very high* variability in preferences, landing on *lifestyle planning* as the task domain.

3.1 Parameterizing Preference-Based Steering

We describe the generalized personalization objective as follows. Given a task domain T (e.g., *lifestyle planning*) that necessitates highly variable solutions for each user, there is a set of interpretable preference dimensions, $\mathbf{d} = \{d_1, d_2, \dots, d_n\}$. Each d_i represents a preference dimension's strength, where a large positive value corresponds to a strong preference towards one end of the preference spectrum (e.g. *luxury* for the dimension of **cost**) and a negative value to a preference towards the opposite end (e.g. *budget*). The preference can also be uni-directional instead of bi-directional, through truncating the steering factor range with 0 as the minimum (e.g., *neutral* $\rightarrow$ *prefers seafood*).

An individual user u 's preference profile can be defined as $\mathbf{d}^u = \{d_1^u, d_2^u, \dots, d_n^u\}$ and the preferred LLM response of user u can be represented in a simplified manner as: $o = M(x, \text{steer}(\mathbf{d}^u))$. The generalized function $M(\cdot)$ represents the LLM's inference process,

$\text{steer}(\cdot)$ represents the idealized activation steering method that uses $\mathbf{d}^u$ to modify the activations during inference, and x represents the user's input to the LLM. The steering process described is $h_{\text{steered}} \leftarrow h + \mathbf{d}^u \cdot \mathbf{v}$, where h represents the intermediate representations at the residual stream of the LLM, $\mathbf{v} = \{v_1, v_2, \dots, v_n\}$ are the steering vectors associated with different preference dimension. Here, steering vectors capture desirable behaviours for the LLM to align with. By adding them into the residual stream during inference, alongside the *strength* of steering represented by $\mathbf{d}^u$, the LLM's outputs are steered in the direction of the behaviour [73].

While we present a model of multi-preference activation steering here, we focus on steering one dimension at a time in our experiments. This decision to limit the complexity of the problem space decreases model instability from excessive steering, which would negatively impact the user studies. Under idealized conditions, however, multiple preferences reflecting a richer user profile could be steered simultaneously. Degeneracy and preliminary multi-dimensional results are discussed further in Appendix B.

3.2 Design Space of Steering Interfaces

With preference-based steering defined, we explore the design space around how the linear factor can be exposed to users. We investigate how activation steering enables different personalization interfaces across the axes of **agency**, referencing how preferences are set; and **fluidity**, referencing when user profiles are learned and updated. Agency determines if the user has the ability to manipulate the parameters themselves, or if the personalization algorithm drives the learning in the background [38, 39, 59]. This narrows in on the tension between designs favouring higher user control [6, 45] versus those favouring higher automation [38, 59]. Fluidity determines if a user's preferences are pre-computed at the start of the interaction, or adaptively learned throughout the ongoing conversation [10, 24], which reflects how easily changing preferences can be accommodated.

We developed three differentiated interface designs across the two axes in order to broadly explore the design space. While there exist other axes of personalization, such as interpretability and granularity, we focus on these designs as preliminary exploration. These chosen designs also resemble existing interaction paradigms that users are generally familiar with, which reduces the learning curve of the experiment. The designs are visualized on the **agency-fluidity** axes in Figure 1.

- **SELECT** [*High Agency, Medium Fluidity*]: Users directly select and update the steering factor at-will in conversation with the chatbot. The factor can also be statically set and treated as a part of the permanent user profile. This interface is similar to volume and temperature control sliders, which are adjusted by the user.
- **CALIBRATE** [*Medium Agency, Low Fluidity*]: Prior to entering the conversation, the user is prompted to iteratively select their preferred output produced by sampling different steering factors, and the factor is updated until their preference converges. This interface is similar to the gaze calibration step executed prior to using AR/VR headsets, where the individual parameters are pre-calibrated.

- **LEARN** [*Low Agency, High Fluidity*]: The steering parameter is updated in the backend of the chatbot based on feedback inferred in the user’s response. To the user, the experience is similar to the typical interaction pattern with an AI chatbot. However, instead of personalizing by updating the memory bank with the user’s written preference, the system makes an implicit inference based on the sentiment of the user’s responses to update the steering factor.

3.3 Application of Steerable Chatbots

Use Case. Where would steerable chatbots be the most useful? We envision that they have the highest utility when users’ preferences are persistent but difficult to articulate. In cold-start scenarios, users cannot yet rely on conversational history to calibrate the LLM, and crafting prompts requires metacognitive effort that many users lack or find burdensome. Activation steering addresses this by operating beneath the prompt layer: rather than asking users to specify what they want, it shifts the LLM’s outputs along pre-defined preference dimensions. Steering can be applied continuously and uniformly across all outputs, functioning as a *soft constraint* on the LLM’s behavior that contextualizes every response without overriding the user’s in-conversation input.

Contextualizing Interactions with Prompting. Activation steering can be applied *in tandem* with prompting. Steering and prompting serve fundamentally distinct goals in personalization — the former aligns LLM outputs with users’ latent preferences, while the latter allows users to input additional requirements. Consider a low-income user who expresses, “*I want to treat myself!*” in her request for restaurant recommendations. She may not realize that the LLM is interpreting her preference as luxury-focused and suggests a Michelin-starred restaurant out of her budget range. By applying a *budget* steering vector, her prompted request would be better contextualized based on her latent cost preference, and her prompt simply shifts the chatbot’s output to the higher end of her personal range.

Complex Preference Representation. We propose that user profiles can be represented through d^n , a set of steering factors across pre-computed, interpretable preference dimensions. In the real world, personal preferences can be intricate and evolving, requiring more complex representations. For example, instead of grounding user profiles using the same dimensions, personalized steering vectors can be constructed from a user’s own preference data. In long-term usage, the user can build up substantial interactions with the chatbot and provide positive and negative ratings for its responses (similar to [76]). With high-quality data, it is possible to capture the nuances of an individual directly and dynamically update their personalized steering vector as new samples become available.

Adapting to New Contexts. Enabling steering in new task and preference domains is flexible. Given an open-source model where the internals are accessible, the steering vector of a desired concept can be computed from datasets that capture the desirable and undesirable LLM behaviours. These datasets can be sampled from real data, synthesized from LLMs, or even manually crafted by the

user. Once computed, the vector is saved and cached, and then can be used with the user-defined strength factor during the inference of new outputs. The same steering dataset can be used for all downstream users, eliminating the need of user-specific data.

4 Computation Experiments

Before studying interactions with users, we establish two technical preconditions for steerable chatbots: that steering reliably shifts expressed preferences (E1,) and that it composes with user prompting (E2). We also conduct experiments exploring multi-preference steering (E3 in Appendix B.3). In this section, we describe task and preference dimensions, evaluation, technical implementation, and the results of the computational experiments.

4.1 Implementation and Evaluation of Steering

To compute steering vectors, we follow von Rütte et al.’s [91] method of training layer-wise linear probes that can accurately separate the representations of the positive and negative data samples. The coefficients of the logistic regression probe are used as the steering vector v_i , which can then be applied with the steering strength d_i to amplify, dampen, or negate the preference expression in the LLM. More details about the technical implementation, including the parameter selection process, are provided in Appendix A.1. For the dataset of positive and negative behavioural examples, we constructed the samples by prompting GPT-4o to output responses that would satisfy users of both the positive and negative sides of the preference. See Appendix A.2 for examples.

We validate the steering procedure with five open-source LLMs available on Huggingface, ordered in increasing size: *stablelm-21-6b-chat* (1.6B) [9], *gemma-2-2b-it* (2B) [89], *Mistral-7B-Instruct-v0.3* (7B) [35], *Qwen2.5-7B-Instruct* (7B) [70], and *gemma-2-9b-it* (9B) [89]. As the model’s intermediate activations must be directly accessible, larger and closed-source models could not be used. All experiments are run on an NVIDIA L4 GPU. The hyperparameters are fixed with *temperature* = 0.7, *top_k* = 50, and *top_p* = 0.95 to encourage some diversity through sampling. We constrain *max_new_tokens* = 100 to restrict the length of the output.

4.2 Evaluation of Preference Expression

To evaluate steered LLM outputs against diverse, granular representations of human preferences in *lifestyle planning* tasks, we use human-written Yelp reviews as a reference. We collect reviews representing five common, interpretable preferences extractable from Yelp reviews metadata¹: **cost**, **ambiance**, **age**, **time**, and **cuisine**, outlined in Table 1. These dimensions are selected to show that activation steering can generalize across preferences, but are not meant to be comprehensive. Each dimension is instantiated with bi-directional traits (e.g. *luxury* and *budget* for the dimension of **cost**), which correspond with the positive and negative directions of the steering vector. A limitation in this implementation is that nuanced and multi-dimensional preferences, such as **cuisine**, have to be minimized into bi-directional representations for simplicity.

For both the computational and user experiments, we measure the *effect of expressed preference* as the relative cosine similarity between the LLM output’s embeddings and each of the bidirectional

¹<https://business.yelp.com/data/resources/open-dataset/>

Table 1: Bidirectional preference dimensions used in the experiments. The effect of the expressed preference is computed based on cosine similarity to a corpus of exemplary reviews from the Yelp Dataset, the processing details of which are provided.

Preference	Negative	Positive	Yelp Dataset Processing Approach
<i>Cost</i>	Budget	Luxury	Price point of 1 (budget) vs 4 (luxury).
<i>Ambiance</i>	Touristy	Hipster	Ambiance of <i>touristy</i> vs <i>hipster</i> .
<i>Age</i>	Kids	Adults	Attribute of <i>kids-friendly</i> being True vs False.
<i>Time</i>	Evening	Morning	Business hours in the evenings vs morning.
<i>Cuisine</i>	Asian	American	Keywords related to Asian vs American food.

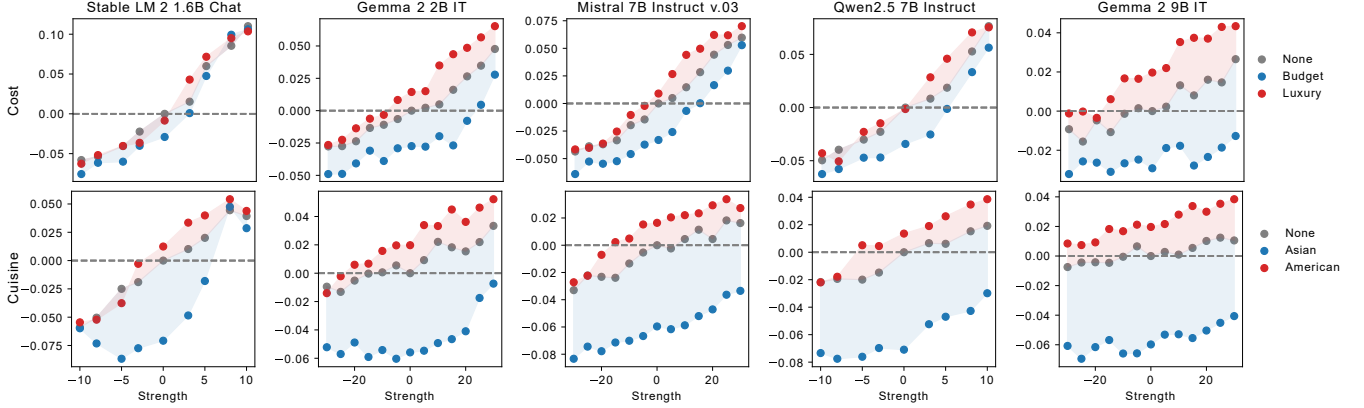


Figure 2: Interaction between prompting and steering for two preferences: cost (top) and cuisine (bottom) for E2.

traits’ reference Yelp reviews — for example, the reference datasets for **cost** are sets of 700-900 positive reviews for establishments with a low and high price points, respectively. See Appendix A.4 for examples. $Effect = \cos(e_o, \bar{e}_+) - \cos(e_o, \bar{e}_-)$, where e_o is the BERT embeddings of the LLM’s output, $\bar{e}_+$ is the mean BERT embeddings of the positive trait’s reference dataset, and $\bar{e}_-$ is that of the negative trait’s reference dataset. All embeddings are computed using a pretrained Sentence-BERT model [72] (*sentence-transformers/stsb-roberta-base-v2* on HuggingFace). We use cosine similarity as the measure following prior works that evaluate subjective constructs against a reference corpus [4, 48]. Additionally, similarity scores, rather than hard classification, allow the evaluation to capture more granular intensities of preference expression.

We evaluate preference-based activation steering on a dataset of $n = 30$ real user queries sampled from the OASST2 dataset [47], which encompasses topics like travel planning, restaurant recommendations, recipe selection, and gift shopping. We filtered questions that appeared at the start of an interaction and were open-ended and not grounded in particular preferences or specifications, like “What are the best restaurants in San Francisco?” See examples of the queries used in Appendix A.3.

4.3 Computational Experiment Results

4.3.1 (E1) Effect of Activation Steering on Content and Quality.

How does steering strength affect the preferences expressed? The objective of **E1** to establish a stable and functional steering range, where model outputs express the desired preference while maintaining response quality. We apply steering for all models and

preference dimensions with strength factors between $-30 < d < 30$, measuring the mean preference effect across the LLM’s responses to the OASST2 query dataset. In addition to measuring the preference expression as described previously, we also compute the perplexity-normalized effect (PNE), which normalizes the expressed preference with the model’s perplexity (a proxy for degeneration). If perplexity increases disproportionately more than the preference effect, then the PNE will be low, reflecting reduced quality.

The correlation between steering strength and preference expression is shown in Figure B.1 in Appendix B.1. To summarize the findings, all models’ preference expression is well correlated with the input steering strength, reflecting both the magnitude and direction of the steering factor. However, the PNE is much lower outside the model’s specific *functional steering range* — the approximate range of steering strengths for which steering does not significantly impact the LLM’s output quality. We observe near-linear relationships between steering strength and perplexity-normalized preference effects. For all ensuing experiments, we only steer within the functional range of the LLM to preserve quality. The finding of **E1** is that **steering, in moderation, can effectively guide the preference expressed by LLMs.**

4.3.2 (E2) Effect of Prompting on Activation Steering. How does prompting interact with steering?

In our paradigm, steering is used to align with the user’s latent preference profile, and the prompting adds contextualized variability — e.g., “I want to treat myself!” should yield different responses for luxury-seeking vs budget-oriented users when the LLM is correctly steered. We repeat

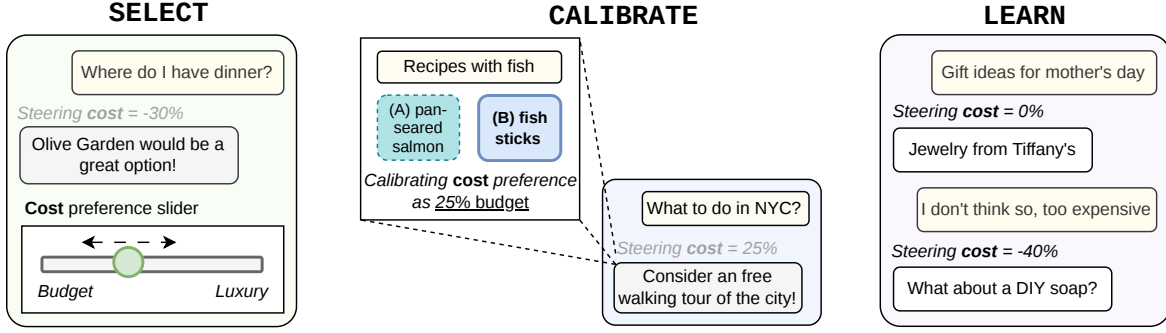


Figure 3: Graphical representations of the three steerable chatbot designs. See Figure C.1 for screenshots of the real interfaces.

E1, but with the addition of preference-based prompts to each question in the query dataset. For each preference dimension, we augment the queries with prompting for both the positive and negative traits, such as “*I am an early riser*” for the *morning* trait and “*I am a night owl*” for the *night* trait. The objective is to understand how steering and prompting interact, especially what happens to the expressed preferences if the directions contradict or reinforce each other.

Figure 2 depicts how adding preference-based prompts induces an *offset* on the preferences expressed by the LLMs. Red represents prompting towards the positive trait, inducing a mostly positive offset; and blue represents the negative trait, inducing mostly a negative offset. We limit the main results to two dimensions (**cost** and **cuisine**) for brevity, but the other dimensions can be found in Figure B.2, as well as specific examples in Table B.1, both in Appendix B.2. Using the two-sample Kolmogorov-Smirnov test for significance ($p < 0.05$), we find that for **cost**, only *gemma-2-9b-it* is significantly affected by prompting, while other models are not. For this reason, we use *gemma-2-9b-it* as the base model of the user study due to its sensitivity to prompting, which allows the model to respond to user prompts in addition to steering.

The significance of these results, while variable across models, indicates that **prompting can intensify or reduce the effect of preference expressions on top of activation steering**. This is a practical trait for a steerable chatbot — steering to the user’s underlying preference *contextualizes* the overall conversation, and adding prompting introduces variability within the context.

5 User Study

With preference-based steering validated, we investigate how real users interact with and perceive steerable chatbots for personalization in an exploratory user study. We evaluate if activation steering aligns chatbot outputs with users’ ground-truth preferences (U1), and how users value the affordances of each steering interface design (U2). We compare the three different steering control interfaces with a prompting-only baseline with no added steering, **PROMPT**. The user study simulates cold-start conversations where the chatbot is not provided with any additional information about the user or scaffolding strategies specific to personalization.

5.1 Implementation of Interfaces

We implement all four chatbot interfaces using Gradio, an open-source Python package for prototyping visual interfaces of machine learning models [1]. Participants were provided with a link hosting the chatbot interface, while the LLM (*gemma-2-9b-it*) is hosted in the backend. We standardize the preference strength range from -100 to 100 to make it interpretable for participants, but map it to *gemma-2-9b-it*’s effective steering range. See Figure 3 for a graphical representation of each interface, and Figure C.1 for screenshots of the Gradio prototypes.

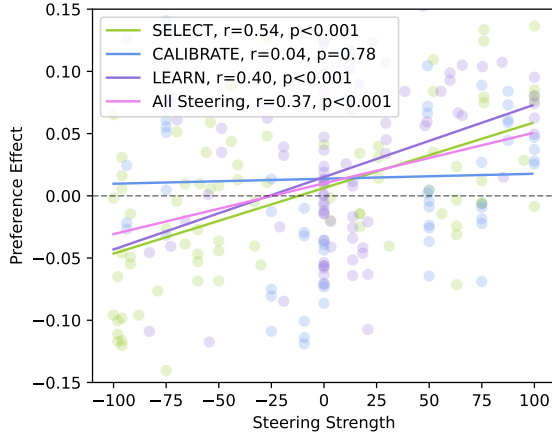
PROMPT: The baseline interface has no added steering. Participants were encouraged to specify any personalization that they wish to add directly in the prompt.

SELECT: The steering parameter is controlled with a slider positioned at the bottom of the chat interface, with a range of -100 to 100. Participants are briefed that the selected slider value is instantaneously applied to the chatbot, and they can dynamically adjust it throughout the conversation.

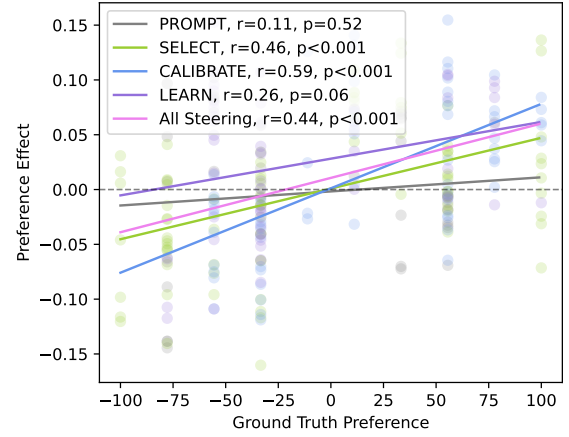
CALIBRATE: Prior to starting the conversation, users perform a calibration step to initialize the steering strength that is applied to the chatbot. Pairwise outputs (Response A and B) for are generated by sampling two different steering strengths. Based on the user’s preferred response, the steering strengths are updated to be closer to the preferred steering strength. For example, updating $d_A \leftarrow (d_A + d_B)/2$ if the preference is *Slightly B*. This calibration step is repeated 2-3 times until the outputs A and B are roughly equivalent, with the final d taken as the average of d_A and d_B .

LEARN. Similar to **PROMPT**, the user can only control the personalization through prompting. To align the chatbot to the latent preferences of the user, we implement a simple learning algorithm that modifies the magnitude and direction of the steering factor based on the sentiment and preference direction of the user’s latest message. For example, “*I don’t like these gifts, they’re too expensive*” would update the **cost** steering negatively towards *budget*. The learned steering strength is displayed as a percentage (e.g. 40% *budget*) in the interface for added transparency. For a message at timestep t , the update equation is given by Eq. (1), with implementation details in Appendix C.3.

$$d_{t+1}^{u*} \leftarrow d_t^{u*} + p(\text{dissatisfaction}(x_t) \cdot \text{direction}(x_t)) \quad (1)$$



(a) Correlation between the applied steering strength and the preference expressed in the chatbots' outputs.



(b) Correlation between the participant's ground truth preference and the preference expressed in the chatbots' output.

Figure 4: Correlation analyses that demonstrate the steering produces preference-aligned outputs in the user study (U1).

5.2 User Study Procedure

We conduct a within-subjects study where participants used the prompting-only baseline and all three steering interfaces in randomized order to complete four lifestyle planning tasks. Each task corresponds to a preference dimension: gift shopping (**cost**), vacation planning (**ambiance**), restaurant recommendations (**cuisine**), and choosing meal prep recipes (**age**). Prior to starting the tasks, we collect participants' self-rated preferences on a scale of 1 to 10, where 1 mapped to the maximum negative preference and 10 mapped to the maximum positive preference. In each task, participants were instructed to engage in a conversation with the chatbot for at least several turns until they reached satisfactory responses, or until they exceeded 5 minutes in the task. Lastly, participants rated each chatbot on 7-pt Likert scale questions:

- **Likelihood to Use:** *How likely would you use the chatbot again for personalization tasks?*
- **Satisfaction:** *How satisfied were you by the personalization?*
- **Perceived Control:** *How in control did you feel of the personalization done by the chatbot?*
- **Perceived Persistency:** *How well do you expect your preferences to be remembered by the chatbot for future use?*

The study concluded with a semi-structured interview where participants could openly discuss the interfaces that they liked and disliked, as well as what features and designs they considered to be important for personalized chatbots. To reduce biases, the order of conditions was randomized, the assignment of tasks to interfaces was counter-balanced, participants were not given the names of the chatbots, and participants were encouraged to honestly rate and discuss each interface.

5.3 User Study Results

5.3.1 Participants and Conversation Statistics. We recruit 14 participants from the USA and Canada (women=7, men=4, non-binary=1) with varied LLM use (daily=7, weekly=5, and monthly/occasionally=2). The study lasted up to 60 minutes, for which participants

were compensated with 25 USD gift vouchers. Figure C.2 in Appendix C.2 shows the participants' self-reported preferences for each of the four tasks and preference dimensions. For **ambiance** and **cost**, the preferences were well-distributed; for **cuisine** and **age**, the preferences were biased towards the negative and positive steering dimensions, respectively.

The number of turns of conversation for **PROMPT** ranged from 2-10 ($M = 5.25, SD = 1.96$), **SELECT** from 4-10 ($M = 6, SD = 1.46$), **CALIBRATE** from 1-9 ($M = 4.57, SD = 1.76$), and **LEARN** from 3-9 ($M = 5.93, SD = 1.98$). Using the Friedman test for differences between three or more non-parametric groups, we find that the lengths are not significantly different ($\chi^2(3) = 5.75, p = 0.12$). **CALIBRATE** conversations tend to be shorter because the participants spent more time at the calibration stage. The number of characters per message for **PROMPT** ranged from 15-188 ($M = 58.52, SD = 27.40$), **SELECT** from 10-118 ($M = 49.40, SD = 23.68$), **CALIBRATE** from 9-119 ($M = 54.34, SD = 24.46$), and **LEARN** from 15-188 ($M = 58.41, SD = 27.10$). **SELECT** has shorter messages because participants did not specify their preferences in-prompt. Friedman test reveals a slightly significant difference ($\chi^2(3) = 8.83, p = 0.03$), but further post-hoc Wilcoxon signed-rank tests with a Holm-Bonferroni correction do not show significant differences between any pairs of conditions.

5.3.2 (U1) Aligning to Users' Latent Preferences. Does activation steering work in noisy, real user conversations? We examine if the method can a) **control the chatbot's expressed preferences**, and b) **produce preference expressions that align with users' latent preferences**. We measure the preference expression using the same method from Section 4.2 and perform analyses with Pearson Correlation. The correlations between steering strength and the expressed preference for all steered chatbots are shown in Figure 4a. The correlation between participants' ground truth preferences and the chatbot's expressed preferences is shown in 4b. For this latter analysis, we dropped the first two turns of conversation for

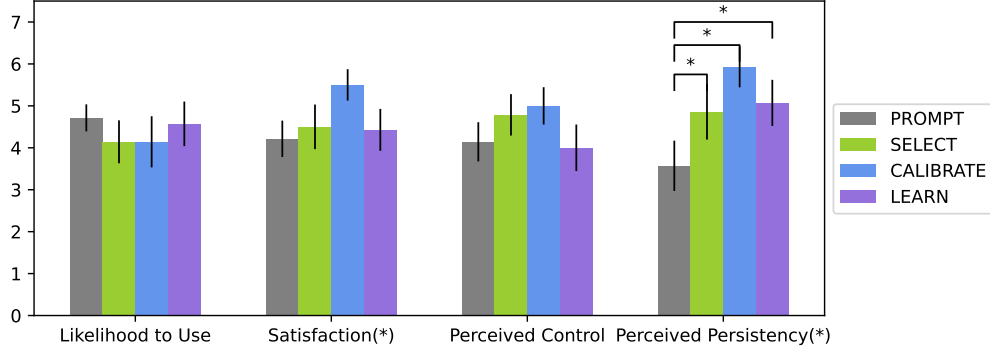


Figure 5: Post-experiment perceptions Likert ratings for all four interfaces. Significant contrasts between the steering conditions and PROMPT are indicated by * in the figure. Perception categories that were rated significantly higher in participants’ favourite interface are further indicated with (*) in the x-label.

PROMPT and LEARN to allow them time to personalize to the user (as they both start off with no preset preferences). See Figure C.3 in Appendix C.3 for how the steering strengths were progressively learned over the conversations.

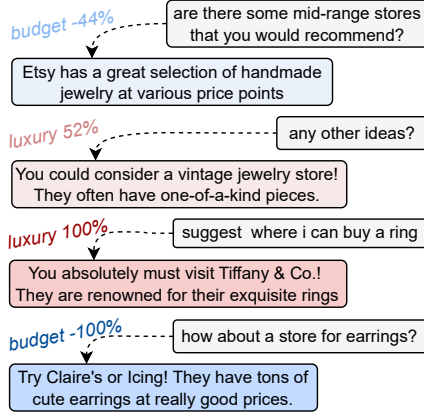
We expect the correlation between steering strength and preference expression in real conversations to be noisier than the computational experiments. Our results in Figure 4a show that SELECT ($r = 0.54, p < 0.001$) and LEARN ($r = 0.40, p < 0.001$) both have moderate positive relationships between the applied steering strength and the expressed preference, echoing the success of the computational findings. However, CALIBRATE was not correlated ($r = 0.04, p = 0.78$), likely due to the coarseness of the calibration step in capturing latent preferences accurately. To illustrate this failure: P2’s calibration result mistakenly converged at $cost = -75$, which reflects a *budget-oriented* preference that they were ultimately dissatisfied with. To correct course, they repeatedly prompted for more *luxury* options in the conversation like, “how about reputable or designer options”. Combining together all variants of the steerable chatbots, they reach a weakly significant correlation ($r = 0.37, p < 0.001$). These results have two implications. One, *steering is effective at controlling the expressed preference* when the user is satisfied with the personalization. Two, when the user is not satisfied, *prompting opposing preferences is effective at retaining control* over the conversation (as it was for CALIBRATE).

Next, Figure 4b compares the chatbot’s steered outputs and the users’ ground truth preferences. In this analysis, SELECT ($r = 0.46, p < 0.001$) and CALIBRATE ($r = 0.59, p < 0.001$) reach moderate, significant correlations, indicating that the chatbots were well-matched with their users’ preferences. LEARN ($r = 0.26, p = 0.06$) fell short of a significant correlation, and PROMPT was the worst ($r = 0.11, p = 0.52$). The sentiment-based update algorithm of LEARN may not have been robust enough to capture diverse intentions from users’ conversational feedback. Overall, all variants of steerable chatbots achieve moderate correlation with latent user preferences ($r = 0.44, p < 0.001$), outperforming PROMPT. We further analyze the error between the steering strengths and the user’s rated preferences in Figure C.4 in Appendix C.4, where the findings show

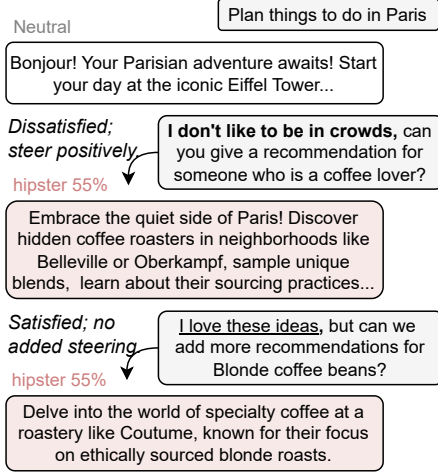
that SELECT reaches the closest alignment, followed by CALIBRATE, then LEARN. While the correlation is moderate to small and the sample size ($n = 14$) is limited, the findings cautiously suggest that *steerable chatbots provide advantages over prompting-only chatbots in aligning with user preferences in cold-start personalization*. We recommend validating these results in larger, more ecologically valid studies.

5.3.3 (U2) Subjective Perceptions of Interfaces. As each interface embeds different design values, we triangulated between participant’s Likert responses with their richer, descriptive interview data to understand the affordances provided by each steerable chatbot. The post-experiment Likert ratings are plotted in Figure 5. For every perception category, we test planned contrasts between the baseline PROMPT with each steerable chatbot condition, using the one-sided Wilcoxon signed-rank test. We apply a Holm-Bonferroni correction to the p -value to account for the multi-hypothesis testing. Strikingly, the first immediate observation is that there is no significant difference between interfaces in the categories of *likelihood to use*, *satisfaction*, and *perceived control*. For *perceived persistency* of the personalization, SELECT ($W = 37.5, p = 0.04$), CALIBRATE ($W = 55.0, p = 0.002$), and LEARN ($W = 36.0, p = 0.006$) were all rated significantly higher than the baseline, indicating that steerable chatbots are perceived to retain their personalized settings across future sessions.

Examining the ratings more closely, we observe that participants often gave strong ratings towards *one* steering interface, but not the others. As an added analysis, we operationalize each participant’s *favourite* interface as the one which received the highest rating in the *likelihood to use* category. This results in a count of 6 votes for SELECT, 5 for LEARN, and 3 for CALIBRATE — where the highest rated interface is a steered chatbot for all participants, never the baseline. With a one-sided paired sign test, we find that the *favourite* interface is rated significantly higher in *satisfaction* ($S = 0.77, p = 0.04$) and *perceived persistency* ($S = 1.0, p = 0.001$), but just short of significance for *likelihood to use* ($S = 0.75, p = 0.07$) and *perceived control* ($S = 0.8, p = 0.05$). The discrepant ratings



(a) P7's conversation for the *Gift Shopping* task, where they changed the steering strength for cost using the **SELECT** interface.



(b) P10's conversation for the *Travel Planning*, where **LEARN** updated the ambiance steering strength based on the sentiment of the user's messages (**bold** for negative, underline for positive).

Figure 6: Two sample user conversations for **SELECT** and **LEARN**

across the steerable chatbots suggest that what users want from personalization is not uniform.

For nuanced understanding, we turn to the interview data, which reveals that participants' reactions to each interface were shaped by underlying values that map directly onto the axes of our design space — **how much agency they wanted over the personalization**, and **how fluidly they expected it to adapt**. We thematically analyze the post-task interview using an inductive coding approach [25], shaped by themes related to the design values of the interfaces. Two researchers coded quotes extracted from the interview transcripts and refined the codes into agreed upon themes. We uncover marked heterogeneity in people's preferences about personalization and disseminate them through the following findings.

Theme 1: Controlled vs. Deferential Personalization. We first identify contrasting values in how much control users desire in personalization. Participants who value control prefer **SELECT**, which allowed them to directly manipulate the chatbot. "I was more in control of the kind of answers I wanted to see" said P2, and P4 agreed with "whether you felt like you were in control was extremely important". Users liked being able to use the slider to nudge the alignment of the chatbot's preference expression while communicating more granular requirements through prompting. P4 observes, "I can control this dimension [via the slider] and then I can tune the other parts that I want via conversation", and P6 echoes that it was like "drawing a box' around the personalization factors that I want to explore". Users could also be fully alleviated from the mental load of prompting, like P7 who exploited the slider on **SELECT** to nudge the chatbot's response towards *luxury* without having to specify their preferences in-prompt (see the conversation in Figure 6a).

On the other end of the spectrum, some users like deferring to the system, relieving them of making the personalization choices themselves. P12 believed **LEARN** was more straightforward and expressive, with wanting to "let me just dump my thoughts and let the LLM handle what's going on", and P14 agrees with saying they prefer to let the chatbot "adjust my preference through multi-turn prompting". In terms of the user experience, P8 said "learning from conversations feels a little bit more flowy and natural", and P9 echoes with "it was more seamless and I felt like I could best express what I was looking for in more detail". P10 expressed being pleasantly surprised by how well **LEARN**'s steering strength for *hipster* adapted to the their preference to *not be in crowds* (see Figure 6b for the conversation). However, participants who did not like **LEARN** pointed out the lack of control over what the system learns. P4 expressed feeling judged, especially "if [the chatbot] thinks that I have a certain preference but that's not what I intended". P11 wants the learned steering factor to be adjustable and P13 agrees that "what a chatbot knows about me" should be editable.

This summarizes the trade-off on the **agency** axis, where many prefer to retain control with the *user-led* **SELECT**, but others acknowledge that the *system-led* design of **LEARN** offers seamless personalization via conversation. In either case, users indicate that they want to understand the personalization applied and be able to modify the information when necessary, highlighting the importance of transparency.

Theme 2: Situational Personalization Persistence. Overall, participants generally found value in personalization that persists over sessions. This is reflected in the significance of ratings in the *perceived persistency* category in Figure 5. In the interview, P3 said the baseline **PROMPT** interface felt like having "a random conversation with a stranger", expressing dissatisfaction in having to describe their preferences. However, people held different views on how stable personalization should be. Some, like P9, valued the ability to update their preference through **SELECT**, while others, like P3, believed their preference "is probably set for the entire conversation" and does not need to be changed. In general, we find disparate orientations towards the permanence of preferences. P14 believes that "people's thoughts can change all of a sudden", even throughout the same conversation. P6 acknowledges that "preferences tend to change over time", hesitantly saying that "I don't know if I actually

want my chatbots to remember my preferences for future use”. P13 notes that they only want “personalization for certain things, but not others” — referring to a separation between their work and personal profiles, where the preferences would diverge. These examples showcase how a seemingly desirable trait, like persistent personalization, should be approached with contextual nuance.

Our personalization interfaces support different levels of **fluidity**. **LEARN** and **SELECT** allows lightweight and dynamic preference updates, while **CALIBRATE** preferences are fixed after the calibration step. Users who liked **CALIBRATE** generally believed that their preferences are more static, and the calibration step has the advantages of saving time and effort. P3 reports enjoying the calibration steps, saying “I’ll be less frustrated while [using the chatbot], because I’ve already personalized it up to some extent with my liking”. P8 also liked “having the [personalization] set beforehand... that way I don’t have to think about it and have a back-and-forth conversation”. The focused, interactive calibration questions also invoked a perception of precision, with P10 saying “it kind of started to narrow down... so I feel like that was understanding very detailed preferences” and P7 comparing it to an eye doctor appointment, “is it option A or option B?... eventually they get to your prescription”. This is empirically reflected in the correlation analysis in Figure 4b, showing that **CALIBRATE** matched ground truth user preferences the best.

We find that for **fluidity**, higher *adaptivity* is particularly valued if the inference for user preferences is erroneous, such as in the system-driven interfaces. P9 asserts that “if [a chatbot] was making a false inference that was going to lead to more inaccurate outputs, I would want to be able to stop that in its tracks”. This sentiment highlights what many felt — if the personalization is not helpful or accurate, there should be a way to change it. However, if preferences can be accurately inferred with overhead steps, some users appreciate persistence in personalization.

Theme 3: Interpretability of steering. Outside of the design space dimensions, participants also brought up general feedback on the interpretability of preference-based steering. Participants were split over the transparency feature of **LEARN**, where the learned steering strength is exposed in the conversation. P12 found the transparency helpful, “I liked being able to see how the model was adjusting”, and P14 said it helped them understand the “chatbot’s mental model of the user’s preferences”. However, P11 said the preference percentage “was overloading me... honestly I was ignoring it”, and P4 said “it felt judgy”. For **SELECT**, P5 found the numerical steering strength of the slider to lack interpretability, critiquing that “it gets into the nuance of how [the chatbot] understands those [preferences] and how you understand them”. P11 believed that “people might not know how to [use the slider] correctly for more complex dimensions”, reflecting how interpretability challenges will scale up. The strength of **CALIBRATE** is clear here, with P5 saying, “instead of just picking a word that you and the AI are defining differently, it can ask you questions”. This shows that the interpretability of preference strength should be improved, such as by providing more in-context examples and calibrating users to the correct mental model ahead of the conversation.

6 Discussion

6.1 Key Findings

In this work, we propose a paradigm for LLM personalization using activation steering, where relevant preference dimensions for a task can be amplified or dampened with pre-computed steering vectors moderated by linear strength factors. Our computational experiments validate the generalizability of steering across five different models and preference dimensions (**E1**), and its effectiveness in tandem with prompting (**E2**). Our exploratory user study further suggests that steering-based chatbots can align better with latent user preferences than a prompting-only chatbot in cold-start personalization (**U1**). We explore how the steering factor can be exposed to users via three interface designs, guided by a design space of agency and fluidity. Participants report highly divergent preferences shaped by personal values, highlighting trade-offs among control, persistence, and transparency — for which our steerable chatbots design space offers structured exploration (**U2**).

6.2 Implications and Future Work

Interfaces for Interacting with Chatbots. The key difference between steerable chatbots and prompting-based chatbots lies the interaction affordances enabled through the linear steering factor. Methods for interacting with chatbots outside of natural language prompting are rising in popularity, including graphical interfaces that can promote structured exploration across diverse spaces [86], improve prompt structuring [58], manipulate and refine ideas [60, 61], and sketch relationships between components [74, 99]. Enabling different modalities of inputs, control, and sensemaking can be highly transformative in supporting complex LLM workflows.

Personalizing the Personalization Experience. The variance in participant feedback on the steerable chatbot interfaces indicates highly heterogeneous preferences towards how personalization is implemented. We found that **CALIBRATE** suits users with stable preferences; **LEARN** is liked by users whose preferences shift mid-conversation; and **SELECT** fits users who want explicit, ongoing control. Among the main themes that participants expressed in the interviews, values related to *agency* and *fluidity* where echoed in explanations about which chatbots participants liked most. In particular, the need to understand how personalization is performed and retain control over what is learned and stored stands out, and this is reflected in the findings of prior research [6, 45, 62]. Ultimately, meta-preferences within personalization may be determined through a combination of individual and task factors. Blanket designs (including the unscaffolded, prompt-only chatbots) are not always suitable — pointing towards malleable interfaces that can be customized to suit personalized experiences [12].

Viability of Activation Steering as LLM Control. Activation steering, as a technical method of LLM control, offers both benefits and limitations. A strength of steering is that it requires only the inference-time application of a pre-computed vector; it is incredibly resource-efficient in comparison to methods that require updating model weights or storing extensive user data [73]. Since the steering vector and strengths contain no personally-identifiable data (unlike personalized finetuning weights or explicit user profiles),

it is also more sensitive to user privacy concerns [2, 33]. However, the question of preserving generation quality and robustness is important to address [91, 94], and more stability is needed of the method to adapt it to steering complex personalization setups.

Generalization to Long-Term Personalization. In our focus on cold-start personalization, we chose to parameterize steering via a top-down approach, relying on pre-defined preference dimensions where every user can be modelled as a simple vector $\mathbf{d}^u$, rather than requiring prior interaction data. In the long-term, steering can be enhanced via integration with other personalization strategies like storing user details in RAG systems [78]. Alternatively, steering can also be used in a bottom-up, data-driven approach, where the user’s own data (such as likes and dislikes) is used to create the personalization signal. This is viable given comprehensive user preference data collected from diverse scenarios over prolonged interactions. This approach is more generalizable to complex preferences, but would require substantial individual data.

6.3 Limitations

We conduct the user study with low-fidelity prototypes, where the calibration and learning algorithms were meant to convey proof-of-concept capabilities. For **LEARN**, this limited the accuracy of the learned preferences, which impacted user satisfaction in the steered outputs. For **CALIBRATE**, the ambiguity of the underlying preference sometimes resulted in misalignment with users’ interpretations. Complex preferences had to be reduced down to singular dimensions (although we do perform preliminary multi-dimensional steering in Appendix B.3), limiting the ecological validity of how preferences are represented.

Steering was implemented using a simple method [91] with small, local models, which do not reflect the personalization potential of state-of-the-art chatbots. As such, we asked the participants to concentrate on the interaction modality of the interfaces in their post-task interviews, rather than the personalization quality. The prompting-only baseline was implemented without sophisticated prompting interventions, such as providing structured templates for expressing preferences, as it was meant to approximate a naive, unscaffolded chatbot. Accordingly, our results showing that steering improves personalization to latent user preferences should therefore be taken in the context of a relatively weak baseline.

Our low-stakes task domain does not represent all personalization use cases, particularly more critical domains like *value alignment* [83]. Future explorations can include tasks such as creative writing [28], communications [98], and persona adoption [17, 30]. We were also constrained to a set of five preferences based on their relevance to the task and their presence within the Yelp Dataset, which represents a relatively narrow preference subspace as the dataset is dominated by restaurant reviews. Our evaluation setup therefore may not apply to tasks outside of the lifestyle planning domain. With the advancement of *LLM-as-judge* evaluation capabilities [29], future evaluation techniques may not be constrained by existing human preference datasets [18].

7 Conclusion

We demonstrate how activation steering can be applied towards personalizing LLMs in preference-driven tasks, with both computational and user studies. We develop three interaction interfaces for *steerable chatbots* based on design values relevant to personalization, and compare their results in aligning chatbot outputs with latent user preferences and subjective perceptions from end users. Our findings reframe LLM personalization as an interaction design problem and showcase activation steering as a resource-efficient mechanism for driving directly manipulatable chatbots.

References

- [1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569* (2019).
- [2] Elizabeth Aguirre, Anne L Roggeveen, Dhruv Grewal, and Martin Wetzels. 2016. The personalization-privacy paradox: implications for new media. *Journal of consumer marketing* 33, 2 (2016), 98–110.
- [3] Rumi A Allbert and James K Wiles. 2024. Identifying and manipulating personality traits in LLMs through activation engineering. *arXiv [cs.CL]* (Dec. 2024).
- [4] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*. 2450–2461.
- [5] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems* 37 (2024), 136037–136083.
- [6] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, René Witte, Greg Butler, and Adrian Tsang. 2013. An approach to controlling user models and personalization effects in recommender systems. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. 49–56.
- [7] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the association for computational linguistics: EMNLP 2020*. 1644–1650.
- [8] Christine Bauer, Chandni Bagchi, Olusanmi A Hundogan, and Karin van Es. 2024. Where are the values? a systematic literature review on news recommender systems. *ACM Transactions on Recommender Systems* 2, 3 (2024), 1–40.
- [9] Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshith Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. 2024. Stable lm 2.1.6 b technical report. *arXiv preprint arXiv:2402.17834* (2024).
- [10] Peter Brusilovsky and Eva Millán. 2007. User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web: methods and strategies of web personalization*. Springer, 3–53.
- [11] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net. <https://openreview.net/forum?id=ETKGuby0hcs>
- [12] Yining Cao, Peiling Jiang, and Haijun Xia. 2025. Generative and Malleable User Interfaces with Generative and Evolving Task-Driven Data Model. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [13] 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. *Advances in Neural Information Processing Systems* 37 (2024), 49519–49551.
- [14] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krashennnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Trans. Mach. Learn. Res.* 2023 (2023). <https://openreview.net/forum?id=bx24KpJ4Eb>
- [15] Chih-Ming Chen, Hahn-Ming Lee, and Ya-Hui Chen. 2005. Personalized e-learning system using item response theory. *Computers & Education* 44, 3 (2005), 237–255.
- [16] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakavi Vinayak. [n. d.]. PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment. In *The Thirteenth International Conference on Learning Representations*.
- [17] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-Playing Language Agents. *arXiv [cs.CL]* (April 2024).
- [18] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509* (2025).
- [19] Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2025. Pad: Personalized alignment of llms at decoding-time. In *International Conference on Learning Representations, Vol. 2025*. 9014–9039.
- [20] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. Creativeconnect: Supporting reference recombination for graphic design ideation with generative ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–25.
- [21] John Joon Young Chung and Eytan Adar. 2023. Artinter: AI-powered Boundary Objects for Commissioning Visual Arts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 1997–2018.
- [22] John Joon Young Chung and Eytan Adar. 2023. Promptpaint: Steering text-to-image generation through paint medium-like interactions. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–17.
- [23] John Joon Young Chung, Woosook Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [24] Tuck Siong Chung, Michel Wedel, and Roland T Rust. 2016. Adaptive personalization using social networks. *Journal of the Academy of Marketing Science* 44, 1 (2016), 66–87.
- [25] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.
- [26] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. Ganslider: How users control generative models for images using multiple sliders with and without feedforward information. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [27] Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 11275–11288.
- [28] Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 14504–14528.
- [29] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2026. A survey on llm-as-a-judge. *The Innovation* 7, 6 (2026).
- [30] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding how people customize, interact, and experience personas in large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–24.
- [31] Zhiyang He, Sashrika Pandey, Mariah Schrum, and Anca Dragan. 2025. Context steering: Controllable personalization at inference time. In *International Conference on Learning Representations, Vol. 2025*. 87863–87895.
- [32] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net. <https://openreview.net/forum?id=nZvKeeFYf9>
- [33] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda Van Noort, and Edith Smit. 2019. Privacy concerns in chatbot interactions. In *International workshop on chatbot research and design*. Springer, 34–48.
- [34] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv [cs.CL]* (Oct. 2023).
- [35] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv:arXiv:2310.06825*
- [36] Peiling Jiang, Jude Rayan, Steven P Dow, and Haijun Xia. 2023. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*. 1–20.
- [37] Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2025. A Survey on Human Preference Learning for Aligning Large Language Models. *Comput. Surveys* 58, 6 (2025), 1–39.
- [38] Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Scholkopf, Rada Mihalcea, and Mrinmaya Sachan. 2024. Implicit personalization in language models: A systematic study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 12309–12325.
- [39] Maurits Kaptein, Panos Markopoulos, Boris De Ruyter, and Emile Aarts. 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies* 77 (2015), 38–51.
- [40] Ikram Karabila, Nossayba Darraz, Anas EL-Ansari, Nabil Alami, and Mostafa EL Mallahi. 2024. BERT-enhanced sentiment analysis for personalized e-commerce recommendations. *Multimedia Tools and Applications* 83, 19 (2024), 56463–56488.
- [41] Sheer Karny, Anthony Baez, and Pat Pataranutaporn. 2026. Neural transparency: Mechanistic interpretability interfaces for anticipating model behaviors for personalized ai. In *Proceedings of the 31st International Conference on Intelligent*

- User Interfaces*. 868–884.
- [42] Tae Soo Kim, Yoonjoo Lee, Minsuk Chang, and Juho Kim. 2023. Cells, generators, and lenses: Design framework for object-oriented interaction with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.
 - [43] Yoonsu Kim, Jueon Lee, Seoyoung Kim, Jaehyuk Park, and Juho Kim. 2024. Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 385–404.
 - [44] Hannah Rose Kirk, Henry Davidson, Ed Saunders, Lennart Luettgau, Bertie Vidgen, Scott A Hale, and Christopher Summerfield. 2025. Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships. *arXiv preprint arXiv:2512.01991* (2025).
 - [45] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*. 43–50.
 - [46] Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*. 782–802.
 - [47] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richard Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems* 36 (2024).
 - [48] Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84, 5 (2019), 905–949.
 - [49] Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehl, Pierre Dognin, Manish Nagreddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *International conference on learning representations*, Vol. 2025. 90960–90985.
 - [50] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
 - [51] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems* 36 (2023), 41451–41530.
 - [52] Xinyu Li, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized Human Feedback. *arXiv [cs.CL]* (Feb. 2024).
 - [53] 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).
 - [54] Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. 2024. In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering. In *Forty-first International Conference on Machine Learning, ICLR 2024, Vienna, Austria, July 21–27, 2024 (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR / OpenReview.net, 32287–32307. <https://proceedings.mlr.press/v235/liu24bx.html>
 - [55] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
 - [56] Zhenyi Lu, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Danyang Chen, and Jixiong Chen. 2023. Miracle: Towards Personalized Dialogue Generation with Latent-Space Multiple Personal Attribute Control. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023 (Findings of ACL, Vol. EMNLP 2023)*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 5933–5957. doi:10.18653/V1/2023.FINDINGS-EMNLP.395
 - [57] Jiaju Ma, Lei Shi, Kenneth Aleksander Robertsen, and Peggy Chi. 2025. AmbigChat: Interactive Hierarchical Clarification for Ambiguous Open-Domain Question Answering. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–18.
 - [58] Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Ying Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2024. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.
 - [59] Marcos Maroto-Gómez, María Malfaz, José Carlos Castillo, Álvaro Castro-González, and Miguel Ángel Salichs. 2024. Personalizing activity selection in assistive social robots from explicit and implicit user feedback. *International Journal of Social Robotics* (2024), 1–19.
 - [60] Damien Masson, Young-Ho Kim, and Fanny Chevalier. 2025. Textshop: Interactions Inspired by Drawing Software to Facilitate Text Editing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–14.
 - [61] Damien Masson, Sylvain Malacria, Géry Casiez, and Daniel Vogel. 2024. Directgpt: A direct manipulation interface to interact with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.
 - [62] Sean M McNee, Shyong K Lam, Joseph A Konstan, and John Riedl. 2003. Interfaces for eliciting new user preferences in recommender systems. In *International Conference on User Modeling*. Springer, 178–187.
 - [63] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*. 198–219.
 - [64] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. 2024. Prompting AI art: An investigation into the creative skill of prompt engineering. *International Journal of Human-Computer Interaction* (2024), 1–23.
 - [65] NA Osman, SA Mohd Noah, and M Darwich. 2019. Contextual sentiment based recommender system to provide recommendation in the electronic products domain. *International Journal of Machine Learning and Computing* 9, 4 (2019), 425–431.
 - [66] Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67.
 - [67] Yingzhe Peng, Xiaoting Qin, Zhiyang Zhang, Jue Zhang, Qingwei Lin, Xu Yang, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. 2025. Navigating the unknown: a chat-based collaborative interface for personalized exploratory tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1048–1063.
 - [68] Silviu Pitit, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordani. 2024. Improving context-aware preference modeling for language models. *Advances in Neural Information Processing Systems* 37 (2024), 70793–70827.
 - [69] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems* 37 (2024), 52516–52544.
 - [70] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuhong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. arXiv:arXiv:2412.15115
 - [71] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems* 36 (2023), 71095–71134.
 - [72] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. <http://arxiv.org/abs/1908.10084>
 - [73] Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt 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)*, ACL 2024, Bangkok, Thailand, August 11–16, 2024, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 15504–15522. doi:10.18653/V1/2024.ACL-LONG.828
 - [74] Karl Toby Rosenberg, Rubaiat Habib Kazi, Li-Yi Wei, Haijun Xia, and Ken Perlin. 2024. Drawtalking: Building interactive worlds by sketching and speaking. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–25.
 - [75] Neil Rubens, Mehdi Elahi, Masashi Sugiyama, and Dain Kaplan. 2015. Active learning in recommender systems. *Recommender systems handbook* (2015), 809–846.
 - [76] Michael J Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Barr 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)*. 8045–8078.
 - [77] 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, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 7370–7392. doi:10.18653/V1/2024.ACL-LONG.399

- [78] Alireza Salemi and Hamed Zamani. 2025. Comparing Retrieval-Augmentation and Parameter-Efficient Fine-Tuning for Privacy-Preserving Personalization of Large Language Models. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025*, Hamed Zamani, Laura Dietz, Benjamin Piwowarski, and Sebastian Bruch (Eds.). ACM, 286–296. doi:10.1145/3731120.3744595
- [79] Claire M Segijn, Joanna Strycharz, Amy Riegelman, and Cody Hennesy. 2021. A literature review of personalization transparency and control: Introducing the transparency-awareness-control framework. *Media and Communication* 9, 4 (2021), 120–133.
- [80] Omar Shaikh, Michelle S. Lam, Joey Hejna, Yijia Shao, Hyundong Justin Cho, Michael S. Bernstein, and Diyi Yang. 2025. Aligning Language Models with Demonstrated Feedback. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net. <https://openreview.net/forum?id=1qGkuxi9UX>
- [81] Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hanna Hajishirzi, Noah A. Smith, and Simon S. Du. 2024. Decoding-Time Language Model Alignment with Multiple Objectives. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/57c89126d60c209f48d0e6395c766bb3-Abstract-Conference.html
- [82] Ben Shneiderman. 2010. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India.
- [83] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mirehghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A Roadmap to Pluralistic Alignment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024 (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR / OpenReview.net, 46280–46302. <https://proceedings.mlr.press/v235/sorensen24a.html>
- [84] Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. Improving Instruction-Following in Language Models through Activation Steering. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net. <https://openreview.net/forum?id=w0zhdnRCtw>
- [85] Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting Latent Steering Vectors from Pretrained Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22–27, 2022 (Findings of ACL, Vol. ACL 2022)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 566–581. doi:10.18653/V1/2022.FINDINGS-ACL48
- [86] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11–16, 2024*, Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, Corina Sas, Max L. Wilson, Phoebe O. Toups Dugas, and Irina Shklovski (Eds.). ACM, 644:1–644:26. doi:10.1145/3613904.3642400
- [87] Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv [cs.CL]* (Dec. 2023).
- [88] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.
- [89] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).
- [90] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv [cs.CL]* (Aug. 2023).
- [91] Dimitri von Rütte, Sotiris Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2024. A Language Model's Guide Through Latent Space. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024 (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR / OpenReview.net, 49655–49687. <https://proceedings.mlr.press/v235/von-rutte24a.html>
- [92] Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andrea Michi, Marco Gelmi, Yunxuan Li, Raghav Gupta, Kumar Avinava Dubey, et al. 2024. Conditional language policy: A general framework for steerable multi-objective finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 2153–2186.
- [93] Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Ai persona: Towards life-long personalization of llms. *arXiv preprint arXiv:2412.13103* (2024).
- [94] Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13–19, 2025 (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR / OpenReview.net. <https://proceedings.mlr.press/v267/wu25a.html>
- [95] 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.
- [96] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-Context: Multi-objective Alignment of Foundation Models with Dynamic Preference Adjustment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21–27, 2024 (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR / OpenReview.net, 56276–56297. <https://proceedings.mlr.press/v235/yang24q.html>
- [97] Xiwang Yang, Yang Guo, and Yong Liu. 2012. Bayesian-inference-based recommendation in online social networks. *IEEE Transactions on Parallel and Distributed Systems* 24, 4 (2012), 642–651.
- [98] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–35.
- [99] Ryan Yen, Jian Zhao, and Daniel Vogel. 2025. Code Shaping: Iterative Code Editing with Free-form AI-Interpreted Sketching. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [100] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [101] 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)*. 7128–7141.
- [102] Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering Knowledge Selection Behaviours in LLMs via SAE-Based Representation Engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, 5117–5136. doi:10.18653/V1/2025.NAACL-LONG.264

A Implementation of Preference-Based Steering

This appendix describes the technical steering parameters and datasets that were used in the computational experiments in Section 4.

A.1 Steering Parameters

We followed von Rütte et al.’s procedure for selecting the probe type (in addition to the logistic regressor, we also tested difference-in-means and principal component analysis) and the k hyperparameter for the *top-k* layers, where steering is selectively applied to the *top-k* of the layers with the highest detection accuracy. Table A.1 shows the best performing parameters that were used for each model. We also experimentally determined an approximate functional steering range where the model is not affected significantly by degeneracy.

Table A.1: Steering parameters selected for each model.

Model	Top K Layers	Functional Steering Range	Probe
<i>stablelm-21-6b-chat</i>	16	(-10, 10)	Logistic
<i>gemma-2-2b-it</i>	16	(-30, 30)	Logistic
<i>Mistral-7B-Instruct-v0.3</i>	24	(-30, 30)	Logistic
<i>Qwen2.5-7B-Instruct</i>	24	(-10, 10)	Logistic
<i>gemma-2-9b-it</i>	32	(-30, 30)	Logistic

A.2 Steering Dataset

Due to a lack of LLM output datasets on preferences, all steering datasets are generated by GPT-4o to emulate what the gold standard of embodying this preference should look like – e.g. “Generate a list of responses that a chatbot might reply to a user who is looking for ***kid-friendly*** suggestions across user queries in lifestyle planning tasks such as travel planning, restaurant recommendations, recipe selections...” See Listing A.1 for sample instances from the positive and negative steering dataset for **age**. Across all steering datasets in the five preference dimensions, we generated 50-80 samples per trait. We explored alternative datasets such as a list of synonymous adjectives (e.g., *cheap* suggestions, *affordable*, *low-cost*... for **cost**) and the Yelp reviews, but the former did not have an appropriate format to guide LLM outputs and the latter is too noisy to learn meaningful concepts from.

Listing A.1: Sample instances from the steering dataset for age

```

1 [Positive steering dataset for adults-oriented]
2 Visit the rooftop bar downtown for stunning views and a curated selection of craft cocktails.
3 Plan a day of wine tasting at the region’s top vineyards, complete with private tours.
4 Book a spa day with massages, saunas, and relaxation areas exclusively for adults.
5 Attend a live jazz performance at the speakeasy-style venue known for its intimate ambiance.
6 Join a mixology workshop to learn the secrets behind making perfect cocktails at home.
7 Take a sunset cruise offering drinks and live music in a serene, adults-only setting.
8 Visit the cigar lounge downtown, featuring an extensive selection and a cozy atmosphere.
9 Book a brewery tour with tastings of seasonal beers and behind-the-scenes insights from brewmasters.
10
11 [Negative steering dataset for kids-oriented]
12 Visit the interactive science museum with hands-on exhibits perfect for kids of all ages.
13 Plan a day at the local zoo, featuring kid-friendly animal encounters and feeding sessions.
14 Explore the city’s largest playground with climbing structures, slides, and picnic areas for families.
15 Spend time at the local aquarium, known for its touch tanks and playful sea otters.
16 Visit the city’s amusement park with rides designed specifically for younger kids and toddlers.
17 Head to the local library for storytime sessions and engaging activities for kids.
18 Visit the butterfly garden – it’s colorful, educational, and a hit with younger children.
19 Take a scenic train ride that offers kid-friendly entertainment and stunning views along the way.
20 Visit the local zookeeper for a behind-the-scenes tour – kids love learning about animals up close.

```

A.3 Evaluation Queries Dataset

Listing A.2 shows a sample of real user queries selected from the OASST2 dataset in the domain of *lifestyle planning* that were used throughout the computational experiments. These queries were manually identified and filtered by the research team.

Listing A.2: Sample task queries used in the computational experiments.

```

1 Give a recipe idea for a vegetarian meal which has tofu?
2 what are some popular souvenirs people take home when they visit the usa from europe?
3 What are the best restaurants in San Francisco?
4 Hi, I am trying to plan a camping holiday with my family. What supplies do you suggest I purchase or pack?
5 I want to buy a gift for my gir friend for the valentines day
6 What are some things I should do on a 5-day vacation in Thailand?
7 What is the best way to cook a tomato?
8 ...

```

A.4 Reference Yelp Reviews for Preference Effect Computation

Listing A.3 shows samples of the reference Yelp reviews datasets for *hipster* and *touristy* of the **ambiance** preference dimension. The samples are first collected based on the criteria listed in Table 1, then filtered down for relevancy to the trait through a combination of manual human evaluation and automated LLM evaluation with GPT-4o-mini as the judge.

Listing A.3: Sample Yelp reviews of the traits *touristy* and *hipster* for the ambiance dimension (some are abbreviated for length).

```

1 [Yelp reviews for hipster]
2 There are restaurants that you go to because they are unique enough to stand out from anything in their category
3 I would highly recommend if you are looking for somewhere different to eat that does not have the usual bar fare.
4 I wish I wasn't reviewing this place because I want it to remain a hidden gem.
5 This is a great find just outside of the hustle and bustle of downtown...
6 The Mercy Lounge hosts great bands and events, including local burlesque acts, and sports pin-up gals on the bar.
7 Amazingly friendly staff. Building is just bursting with personality.
8 Simply lovely. Every dish had an unusual but delicious mix of flavors and textures.
9
10 [Yelp reviews for touristy]
11 The signature iced tea was fabulous! There was a long wait but worth every minute of it!
12 Reading Terminal Market has everything that you need everything is fresh good or good for you
13 The highlight of the cruse was the tour guide/narrator Charles and his eloquent oration.
14 We were so excited to visit here from Las Vegas. What a Christmas experience for us!!!
15 It is very bustling so if you're one who doesn't like waiting in lines, you'll have to suck it up!
16 If touristy is what you want, they serve great cheesesteaks!
17 Genos vibe was very touristy (we were tourists as well haha) but crowded.

```

B Additional Computational Experiment Results

This appendix includes additional computational evaluation results for E1 and E2, as well as an additional experiment E3 for multi-preference steering.

B.1 Additional Results for E1: Perplexity-Normalized Preference Effect

As described in the main text, we evaluate how steering strength modulates the expressed preferences of the model, including examining how model degeneracy impacts outputs with higher amounts of steering. To evaluate this, we compute the perplexity-normalized effect (PNE), which incorporates a measure of how the model’s perplexity, a proxy for degeneration, changes with steering compared to a non-steered baseline. This equation is given by von Rütte et al. as: $(Effect_{steer=d} - Effect_{steer=0}) / (PPL_{steer=d} / PPL_{steer=0})$, where PPL is perplexity and d is the steering strength. If perplexity increases disproportionately more than the preference effect, then the PNE will be low, reflecting reduced quality. Figure B.1 shows the standard (*top*) and perplexity-normalized (*bottom*) preference effects for all models and preferences. The values are standardized to center around $d = 0$ as the neutral point with $effect=0$. The PNE shows the approximate stable ranges of steering for each model, which was mapped to the functional steering range in Table A.1.

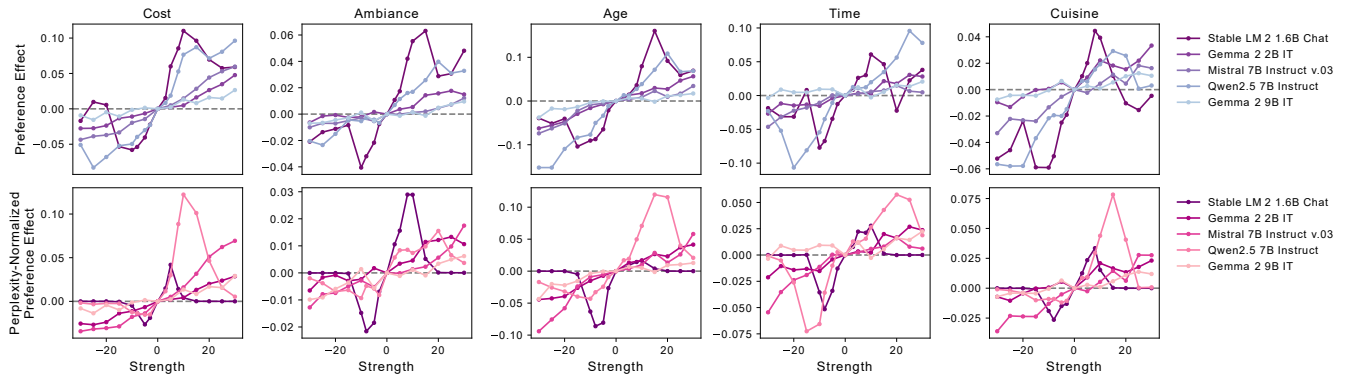


Figure B.1: Effect of preferences (*top*) and perplexity-normalized effect (*bottom*) for all preferences and models for (E1).

B.2 Additional Results for E2: Interactions Between Steering and Prompting

The results for the effect of prompting in addition to steering for other preference dimensions in E2 are in Figure B.2. In particular, **time** is very sensitive to prompting, as evidenced by the large positive and negative offsets due to prompting towards *morning* and *night*. For the other dimensions, we observe that the default behaviour of the model for **ambiance** close aligns with *touristy*, and **age** with *adult*. This demonstrates what the default representations of the models are biased towards. Qualitative examples of how prompting impacts a **cost**-steered model are provided in Table B.1. For example, when the *budget* user asks for more *luxurious* options, then model outputs affordable suggestions that are reasonable for the user.

Table B.1: Samples of steered outputs with preference-based prompting for cost (E2).

Cost steering	No prompting	Prompt towards <i>luxury</i>	Prompt towards <i>budget</i>
“Average User” ($d=0$)	Here’s a list of essential items you might want to consider for your camping trip. 1) Tent: Depending on the size of your family, you’ll need a tent that can comfortably accommodate everyone...	Here’s a comprehensive list of camping supplies for your family camping holiday. 1) Tent: A spacious and high-quality tent is essential. Consider a tent with multiple rooms to provide privacy and separate areas for each family member...	Here’s a list of essential camping supplies to help you get started while keeping it budget-friendly. 1) Tent: A 4-person dome tent is a popular and affordable choice...
“Wealthy User” ($d=20$)	To ensure your comfort and convenience, I have prepared a comprehensive list of essentials. 1) Tent: A four-season, safari-style tent with a large canopy...	To ensure that your stay is nothing less than perfection, I have prepared a meticulously curated list of the most exquisite and indulgent camping supplies...	To ensure a comfortable and enjoyable adventure, I have compiled a list of essential items that will cater to your needs without exceeding your budget. 1) Tent: A spacious 4-person dome tent with a waterproof flysheet and a mesh canopy for stargazing...
“Budget User” ($d=-20$)	Tent: Depending on your camping style, you may need a tent. There are various types of tents available, such as dome tents, backpacking tents, and family tents...	Here are some essential items to consider when planning a camping trip: 1) Tent: There are various types of tents available, from simple pop-up tents to more complex backpacking tents...	Here are some budget-friendly supplies for a camping holiday: 1) Tent: Consider buying used tents or borrowing from friends...

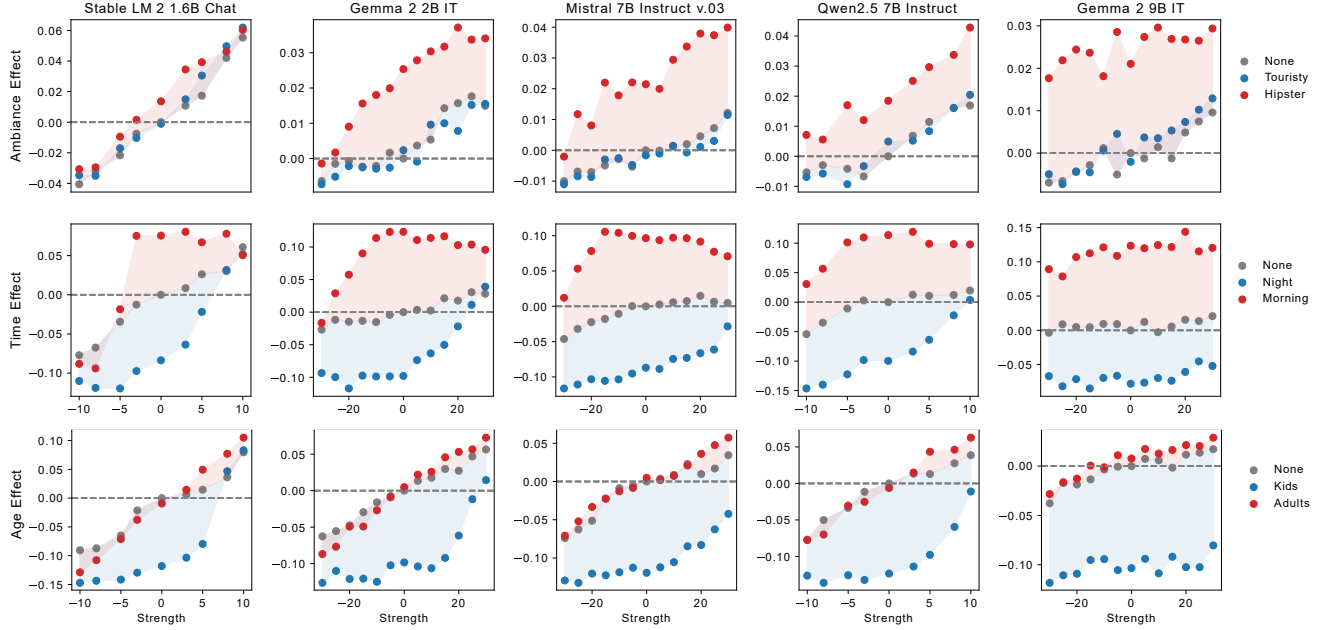


Figure B.2: Effect of prompting on steering for preferences ambiance (top row), time (middle row), and age (bottom row) in E2.

B.3 Additional Results for E3: Multi-Preference Steering

In this additional experiment on the preliminary implementation of multi-preference steering, we focus on two dimensions instead of all five at once. In future context, this should be extended to incorporate more preferences, ideally ones that are orthogonal to each other. Since our preferences are not inherently orthogonal, we select two dimensions with the least collinearity, which are **cuisine** and **age**. We vary the steering strength for each dimension uniformly and measure the effect of each preference separately. The steering vector is applied as a weighed combination of the individual preference vectors. With compounded steering, the main concern is that the effects will deteriorate. Figure B.3 shows 3D plots of the variation of the preference effect for **cuisine** and **age**. While the range of effects is slightly degraded at large values of complementary steering, the surfaces follow the expected trend. Qualitative samples of multi-steered texts are shown in Table B.2, which highlights phrases that exemplify each of the target traits. The results of E3 suggest that *multiple preferences can be steered additively*, which is promising for scaling up steering as a personalization strategy.

Table B.2: Samples from multi-preference steering with preferences age and cuisine (E3).

	Mistral 7B ($d_{cuisine} = 30$, steer $\rightarrow$ American)	Mistral 7B ($d_{cuisine} = -30$, steer $\rightarrow$ Asian)
Mistral 7B ($d_{age} = 30$, steer $\rightarrow$ Adults)	San Francisco is known for its vibrant dining scene, offering a wide range of culinary experiences . Here are some of the top-rated restaurants in the city. 1) The Slanted Door - Overlooking the Bay, this restaurant offers breathtaking views alongside a menu of contemporary American dishes ...	San Francisco, being a vibrant city with a rich culinary scene , offers a wide array of exceptional restaurants that cater to various gastronomic preferences. 1) Izakaya Roku: Known for its authentic and vibrant izakaya atmosphere, this restaurant offers a comprehensive selection of japanese small dishes and sake ...
Mistral 7B ($d_{age} = -30$, steer $\rightarrow$ Kids)	San Francisco offers a variety of family-friendly restaurants that are popular among tourists and locals. Here are some of the best restaurants in the San Francisco Bay Area. 1) The Original Pancake House : A roadside attraction featuring giant dinosaur statues and a restaurant serving breakfast foods ...	Here are some popular restaurants that you can consider visiting. 1) Golden Gate Fortune Cookie Factory : This is a fun and educational experience where you can make your own fortune cookies . It's a great way to introduce kids to different shapes and designs...

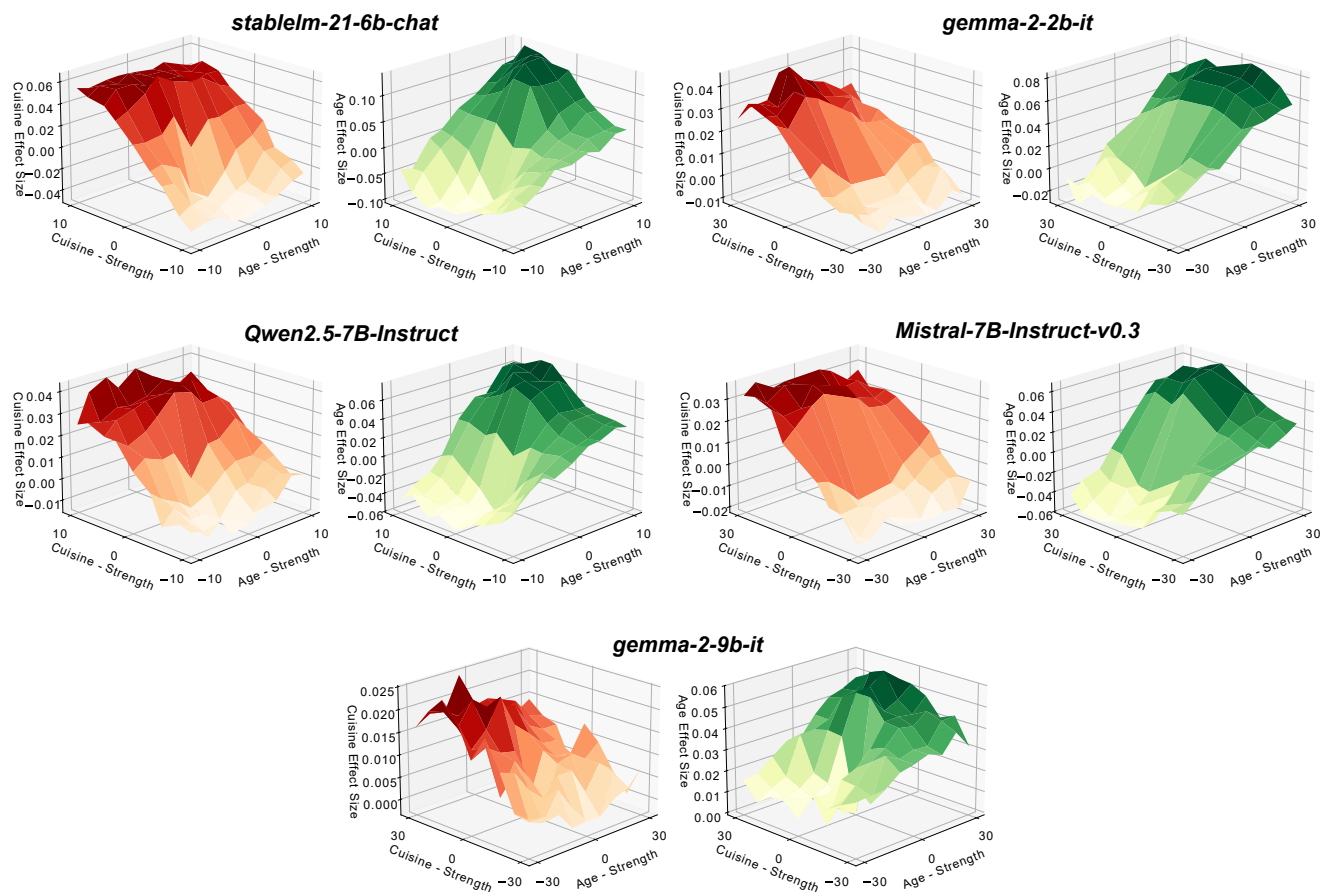


Figure B.3: Additional multi-preference steering surface plots for E3.

C Additional User Study Details

This appendix provides the details of implementing the user study as well as additional results not covered by the main text.

C.1 Interfaces Screenshots

As a correspondence to the interface diagrams in Figure 3, we also provide screenshots of what the participants actually saw in Figure C.1. The interfaces were implemented using Gradio [1].

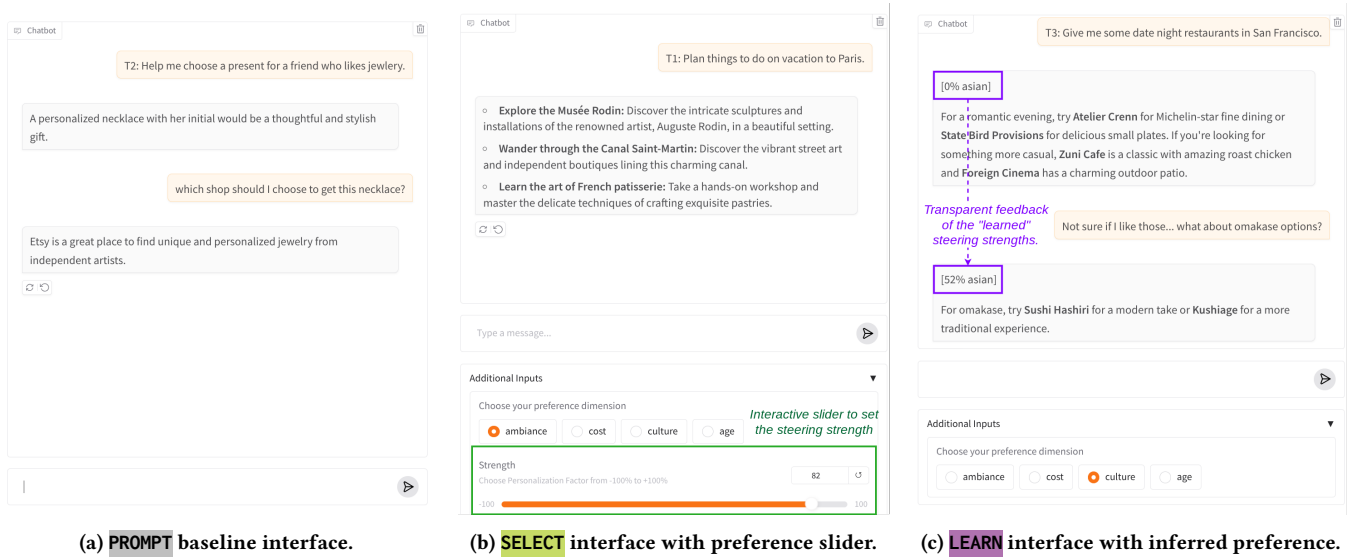
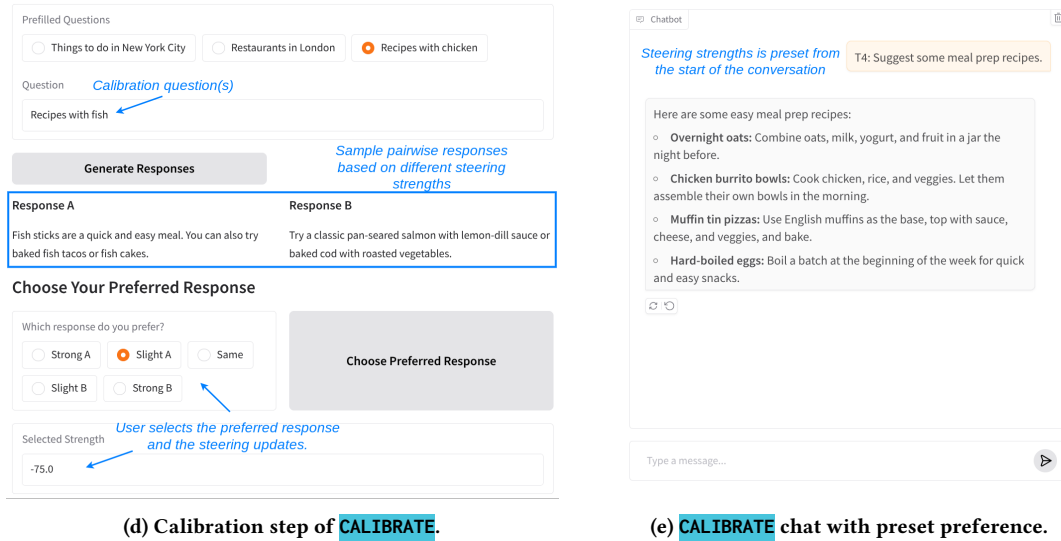


Figure C.1: Screen captures of all four chatbot interfaces that participants used.

C.2 Participants' Ground Truth Preferences

Figure C.2 shows the distribution of true preferences across each of the four tasks/preferences pairings as they were self-reported by the user study participants. The preferences of **ambiance** and **cost** are well-distributed across both positive and negative traits, while **cuisine** leaned towards *Asian* and **age** leaned towards *adults*.

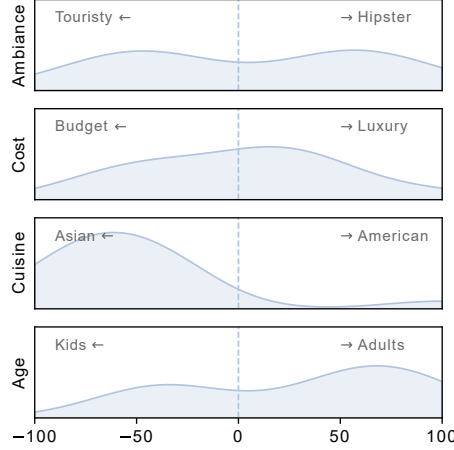


Figure C.2: Ground truth preference distributions.

C.3 LEARN Implementation and Experimental Outcomes

We provide more detail on how the **LEARN** algorithm is implemented. As presented in the main text, the steering strength is updated based on the content and sentiment of the user's message:

$$d_{t+1}^{u*} \leftarrow d_t^{u*} + p(\text{dissatisfaction}(x_t) \cdot \text{direction}(x_t)) \quad (2)$$

Here, x_t is the user's message at round t of the conversation, $\text{dissatisfaction}(\cdot)$ measures the granular sentiment of dissatisfaction in the user's feedback, $\text{direction}(\cdot)$ is a binary classification that determines whether the user wants the model's outputs to shift positively or negatively, and $p(\cdot)$ is a simple linear transformation. While there are many possible choices for carrying out these computations, we demonstrate the aptitude of the high-level methodology using straightforward choices:

- $\text{dissatisfaction}(\cdot)$ is taken as the weighted sum of the prediction probabilities of negative and neutral sentiment detected with the pretrained classifier TweetEval [7]².
- $\text{direction}(\cdot)$ is determined based on the greater cosine similarity to Sentence-BERT [72] embeddings of reference phrases (e.g., "I want more luxury" and "I want lower cost" for the preference dimension of **cost**).
- $p(\cdot)$ is a remapping of the estimated preference value, for example, to the functional steering range of the model.

In the user study itself, Figure C.3 shows the progression of the learned steering strengths across all 14 participant conversations, with red indicating more positive strengths and blue indicating more negative.

C.4 Error in Steering Strengths

For **U1**, we also examine if the applied steering strengths match the ground truth preferences of the participants. For **SELECT**, we take the steering strength set by the participant in the first step; for **CALIBRATE**, we use the post-calibration steering strength; and for **LEARN**, we use the learned strength in the last turn of the conversation. Figure C.4 compares the chatbots across two measures — do the steering strengths directionally agree with ground truth preferences (e.g. is the steering value *positive* if the user indicates they prefer *luxury* for **cost**?); and what is the absolute error between the steering value and ground truth for the given task and preference?

Averaged across all participants, **SELECT** had the lowest mean absolute error ($M = 27.8, SD = 31.9$) and highest proportion of agreement at 86% — although it is surprisingly not 100%, as participants chose it themselves. **CALIBRATE** matched the same agreement rate, but had higher error ($M = 41.4, SD = 34.6$). **LEARN** was less robust and achieved slightly above random rate of agreement at 57%, with the highest error ($M = 50.1, SD = 40.9$), again reflecting the potential weakness in the learning algorithm. In summary, we observe varying levels of accuracy in the steerable chatbots' alignment with ground-truth preferences, but we contend that future advances in calibration and learning algorithms can substantially improve this performance.

²We use a weighted computation of $0.75 * p_{\text{Negative}} + 0.25 * p_{\text{Neutral}}$ because it offers greater discrimination than just the negative probability alone.

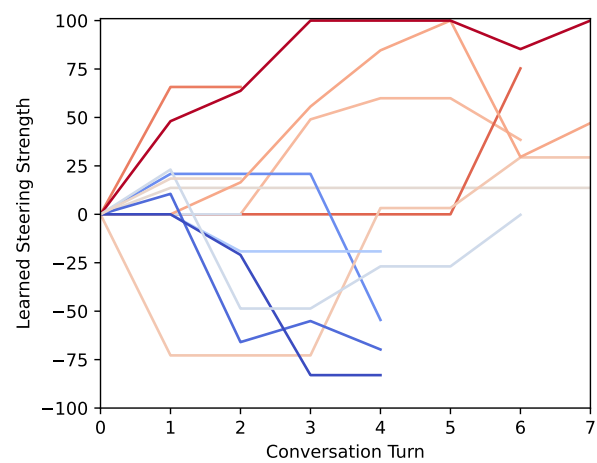


Figure C.3: Progression of strengths learned by **LEARN** across all conversations.

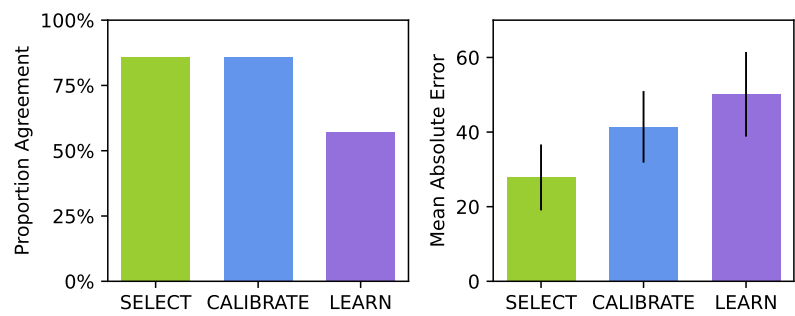


Figure C.4: Proportion of agreement between the applied steering direction with the participants' true preference (left) and mean absolute error of the steering strengths (right).