A Concise Agent is Less Expert: Revealing Side Effects of Using Style Features on Conversational Agents

Anonymous ACL submission

Abstract

Style features such as friendly, helpful, or concise are widely used in prompts to steer the behavior of Large Language Model (LLM) conversational agents, yet their unintended side effects remain poorly understood. In this work, we present the first systematic study of cross-feature stylistic side effects. We conduct a comprehensive survey of 127 conversational agent papers from ACL Anthology and identify 12 frequently used style features. Using controlled, synthetic dialogues across task-oriented and open-domain settings, we quantify how prompting for one style feature causally affects others via a pairwise LLM-as-a-Judge evaluation framework. Our results reveal consistent and structured side effects, such as prompting for conciseness significantly reduces perceived expertise. They demonstrate that style features are deeply entangled rather than orthogonal. To support future research, we introduce CASSE (Conversational Agent Stylistic Side Effects), a dataset capturing these complex interactions. We further evaluate prompt-based and activation steering–based mitigation strategies and find that while they can partially restore suppressed traits, they often degrade the primary intended style. These findings challenge the assumption of faithful style control in LLMs and highlight the need for multi-objective and more principled approaches to safe, targeted stylistic steering in conversational agents.

1 Introduction

The use of natural language prompts to steer the persona and output style of LLMs is now a ubiquitous practice in conversational agent design. Developers and researchers routinely employ system instructions, such as "be empathetic," "be professional," or "be concise", to tailor model behaviors for specific applications ranging from mental health support to customer service (Feng et al., 2025; Zhao et al., 2025a; Rachidi et al., 2025). These prompt-based controls often ignore a fundamental yet untested assumption: that styles function are entangled, correlated features rather than independently controllable dimensions. In practice, prompting for a particular style implicitly activates a constellation of associated traits, shaped by training data, social conventions, and latent representations within the model.

In this work, we interrogate this entanglement and identify a pervasive phenomenon we term *stylistic side effects*: unintended and systematic behavioral shifts in styles distinct from the prompted control feature. While prior work examines prompt side effects and behavioral changes from using persona (i.e. an Asian persona) (Luz de Araujo and Roth, 2025; Gupta et al., 2024), a comprehensive statistical investigation into how stylistic controls interfere with one another remains absent.

We address this gap by presenting the first systematic study of cross-feature stylistic side effects in LLMs. To ground our analysis in real-world usage, we first conduct a comprehensive survey of 127 conversational agent papers from the ACL Anthology (2023–2025), identifying 12 distinct, frequently used style features such as *helpful*, *concise*, and *expert*. We then implement a rigorous causal evaluation framework, generating synthetic dialogues across task-oriented and open-domain settings using models like `Llama3`, `Qwen3-8B`, and `GPT-5-mini`. By employing an LLM-as-a-Judge pairwise comparison protocol with win rates, we quantify how prompting for a Main Feature causally impacts the expression of unrelated Side Features. To facilitate reproducible research, we introduce CASSE (Conversational Agent Stylistic Side Effects), a dataset annotated with these side-feature interactions.

Our results reveal that style features in high-dimensional space are deeply entangled. We demonstrate consistent and structured side effect patterns, such as prompting for *Concise* significantly reduces perceived *Expertise*, while prompt-

ing for *Efficient* leads to a drop in *Helpfulness*. Furthermore, we evaluate two mitigation strategies, Prompt Intervention and Steering Intervention, and find that attempting to restore suppressed traits often degrades the primary intended style. These findings suggest that simple prompt concatenation and activation editing are insufficient to disentangle these opposing effects, highlighting the need for more principled, multi-objective approaches to safe stylistic steering.

- **We conduct a comprehensive survey** of style feature usage across 127 ACL Anthology papers and identify frequently used style features.
- **We present CASSE**, a dataset that includes 12,200 synthetically generated messages annotated with 12 style features.
- **We reveal Stylistic Side Effects** from causal relationship between style features and find that style controls are deeply entangled.
- **We evaluate Side Effect Mitigation methods** and find that attempting to restore suppressed effects often degrades the primary intended style.

The paper is organized as follows. Section 2 reviews related work in the field of style transfer and model steering. Section 3 details the survey methodology used to define and extract style features. Section 4 outlines the experimental setup for prompt-based message generation. Section 5 presents our framework for quantifying correlations and identifying style features' side effects. Section 6 presents prompting algorithm and steering vector algorithm for Side Effect Mitigation, Section 7 discusses the findings, and Section 8 concludes the paper.

2 Related Work

Recent research highlights the critical role of stylistic controls in LLMs, especially for interactive tasks like "role-playing" and "personalization" (Tseng et al., 2024; Chevi et al., 2025). However, the significance of style extends beyond user satisfaction to fundamental model safety and robustness. Studies reveal that "superficial style alignment" can inadvertently bypass safety guardrails by prioritizing format over harmlessness (Xiao et al., 2025), while style biases in LLM-based evaluation often prioritize pleasing outputs over factual substance (Feuer et al., 2025). To address these risks, new benchmarks like PersonalLLM (Zollo et al., 2025), PERSONAMEM (Jiang et al., 2025), and Crab (He et al., 2025b) rigorously test adaptive capabilities. Furthermore, findings that model performance is brittle to stylistic perturbations (Truong et al., 2025) underscore the urgent need for style-aware optimization techniques to ensure consistent effectiveness across contexts (Pu et al., 2024). These studies show that LLM stylistic control is a critical research direction, affecting both user experience with LLM and LLM safety guarantee.

Research on LLM stylistic steering relies heavily on natural language prompts like "be empathetic" to modulate behavior (Feng et al., 2025; Zhao et al., 2025a; Rachidi et al., 2025). However, recent audits suggest these controls are often superficial or counterproductive. Zheng et al. (2024) show that expert personas can degrade performance on objective tasks, while other studies reveal that style prompts frequently function as retrieval cues for demographic caricatures, triggering latent stereotypes rather than precise constraints (Lutz et al., 2025; Malik et al., 2024).

To address prompting limitations, activation steering techniques like Contrastive Activation Addition (CAA) have emerged to enforce traits without fine-tuning (Panickssery et al., 2024; Zhang et al., 2025). Recent work extends this to granular stylistic control: Bo et al. (2025) align chatbots with personality traits via preference-based steering, while Zhao et al. (2025b) and Liu (2025) propose "SteerX" and "StyleVector" to isolate style components for parameter-efficient transfer. These methods establish activation steering as an effective approach for behavioral modification.

However, these methods often fail to recognize the intricate web of stylistic side effects introduced by these methods. They fail to recognize that style features in high-dimensional space are deeply entangled; our work identifies that amplifying a social style feature (e.g., friendliness) systematically and unintentionally suppresses orthogonal functional style features (e.g., efficiency), a critical form of cross-feature side effects that prior research has neglected.

3 Style Feature Extraction - A Survey

To ground our study in contemporary practice, we perform a systematic survey of conversational-

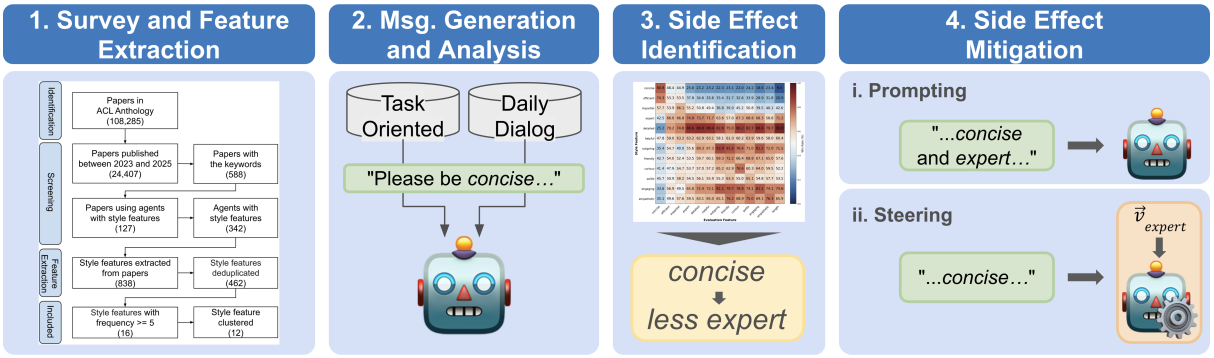


Figure 1: Overall pipeline of our study. We first collect popular style features from our survey, and generate feature-guided messages in both task-oriented and daily dialog domain. Then we identify side effects of using style features, and finally mitigate these side effects via prompting and steering.

agent papers published in the *ACL Anthology* between 2023 and 2025, and extracted commonly used style features from the selected papers. The overview of our pipeline is visualized in Figure 2. The distribution of extracted features is shown in Figure D. Additionally, the summary of our survey is presented in Table 3.

Starting with all papers from ACL Anthology from January 2023 to June 2025, we first select papers that have keyword 'conversational agent', 'dialogue system', 'dialog system' and 'chatbot' in their titles or abstracts. Then we extract all style features used in agent prompts from the papers. Next, we transform these features to adjectives, to get a list of unique features. Then we identify candidate features by filtering for those with a frequency ≥ 5, which results in 16 features. Figure 10 shows that features with similar meaning are subsequently merged via hierarchical clustering based on cosine similarity thresholds greater than 0.5, using embeddings generated by OpenAI's `text-embedding-3-small` model (OpenAI, 2024). As the result, we extracted 12 distinct style features: *concise*, *expert*, *helpful*, *empathetic*, *friendly*, *detailed*, *engaging*, *curious*, *polite*, *impartial*, *outgoing*, and *efficient*, that typify how recent papers employ prompt-based message control.

Out of these style features, *helpful*, *empathetic*, *friendly* and *concise* are most frequently used. Right skewed distribution shows style feature usages are concentrated into a few terms, with a significant long tail where 75% of the 462 unique features appear only once in our list of papers. We further analyze the number of features employed per agent. Among the 342 agents studied, the median number of features is 1, with 177 agents (52%) utilizing only a single feature. This finding suggests

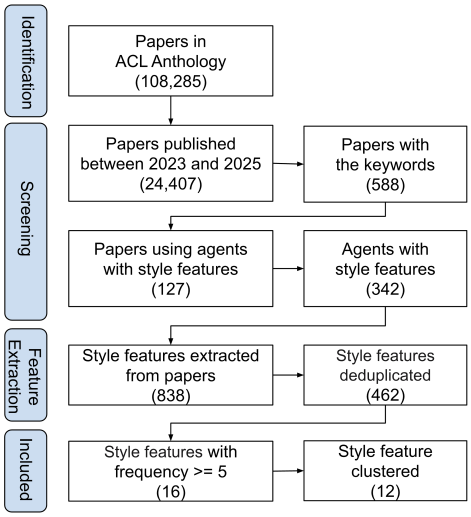


Figure 2: Data collection pipeline for papers and style features selection of our survey.

that contemporary research and applications predominantly rely on isolated keywords rather than more sophisticated style feature usage. A more detailed discussion on style features' utilization patterns from the survey is shown in Appendix L.

4 Message Generation and Analysis

In this section, we introduce a controlled generation and evaluation pipeline to measure the causal effect of style feature usage on an agent's style shifts. To cover diverse, real-world domains for message generation, we select two dialogue datasets: LMSYS-Chat-1M (Task, Zheng et al. 2023) for task-oriented inquiries and DailyDialog (Daily, Li et al. 2017) for open-domain, daily-life interactions. In each domain, we select 10 topics and sample 10 conversations per topic.[1] Example of topics in each

[1] See Appendix B for topic selection details.

domain is shown in Table 4.

Next, we employ agents powered by `GPT-5-mini` (OpenAI, 2025), `Qwen3-8B` (Bai et al., 2023), or `Llama-3-8B-Instruct` (`Llama3`) (Grattafiori et al., 2024) models to generate responses. To apply a style feature, an agent is queried with: *"You are having a conversation about {topic}. Please be {style_feature} in your response."* For the control group (*Neutral*), we utilized a system prompt devoid of stylistic instruction: *"You are having a conversation about {topic}."* Only the first message of each conversation is used to initiate the interaction, and we record the agent's first response.

For each message and style feature, we use high temperatures[2] to repeatly sample five responses, and one additional *Neutral* response to serve as a reference for future evaluation. This procedure yields a comprehensive corpus of 12,200 generated utterances, allows for significant statistical power when analyzing the consistency and variance of stylistic expression.

To evaluate stylistic shifts, we employ a pairwise comparison protocol that analyzes the relationship between a **Main Feature**, i.e., the style feature explicitly specified in the prompt, and a **Side Feature**, i.e., the style feature evaluated in the agent's response. Each comparison pair consists of a *Styled* response generated with the Main Feature and its corresponding *Neutral* response generated without the Main Feature. We then use `Qwen3-8B` as a judge to determine which response exhibits the Side Feature more strongly.[3]

This procedure is applied to all conversations across 12 Main Features (treated as causes), with each evaluated against 12 Side Features (treated as effects), resulting in a total of 144 win-rate measurements. We use **win rate** as the primary metric to quantify both the strength and direction of the causal effect, capturing how the use of a given style feature influences model behavior across all stylistic dimensions. Covering the generated messages and Side Feature's win rate annotations, we release CASSE dataset for further research endeavor.

4.1 Result

We present the average causal effect of style prompting across all domains and LLMs in Figure 3. By aggregating these win rates, we construct

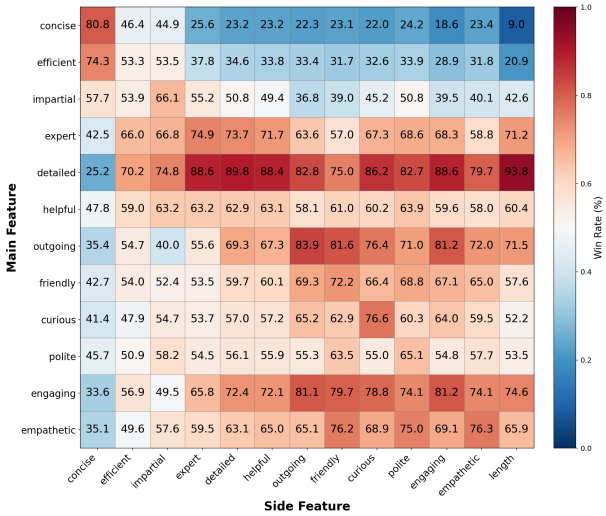


Figure 3: Style Feature Win Rate Matrix Across Three Models. The y-axis shows Main Features, and the x-axis shows Side Features for comparison. Each cell represents the average win rate of the styled response against the neutral response across three models. Red indicates a win rate $> 50\%$ (positive alignment), while blue indicates a win rate $< 50\%$ (negative impact). Average results per Domain are presented in Figure 7.

a causal matrix, where the rows are Main Features (Cause) and the columns are Side Features (Effect). This matrix quantifies the degree to which a given style feature (e.g., *Concise*) induces bias in an agent toward specific style dimensions (e.g., *Expert* or *Helpful*). Length is also added as a column to present causal impact on the generated length in word counts. The color scale encodes the average win rate: values above 50% (red) indicate that the styled response presents stronger tendency to a Side Feature, while values below 50% (blue) indicate a negative impact.

In general, the prominent red diagonal shows that style prompting steers the model toward the intended feature effects. Beyond this primary effect, we observe that features naturally cluster based on their impact on response length. Features such as *Detailed*, *Outgoing*, and *Expert* consistently show high win rates in the *Length* column (typically $> 70\%$), while *Concise* and *Efficient* have significantly lower *Length* win rates ($< 21\%$).

However, distinct stylistic signatures exist even within these length-based groups. For instance, while both *Concise* and *Efficient* reduce output length, *Concise* imposes a much stronger penalty on the *Helpful* and *Expert* dimensions (23.2% and 25.6%) compared to *Efficient* (33.8% and 37.8%), suggesting nuanced differences in semantically

[2] $temperature = 0.7$ is used for `Qwen3-8B` and `Llama-3-8B-Instruct`, while `GPT-5-mini`is fixed as 1.0.

[3] See Appendix A.1 for the evaluation prompt.

Prompted Feature	Side Effect	Model	Domain
Concise	Less Expert	Llama3	Task, Daily
Efficient	Less Helpful	Llama3	Task, Daily
Curious	Less Empathetic	Llama3	Daily
Engaging	Less Impartial	Qwen3	Daily
Polite	Less Efficient	Qwen3	Task, Daily

Table 1: Critical side effect pairs identified for mitigation. These specific pairs represent instances where the prompted style significantly and unexpectedly degrades a secondary style feature. For a complete analysis of side effects across all models and domains, see the full heatmaps in Appendix 6.

similar features. Crucially, the matrix reveals significant "side effects"—unintended behavioral shifts in style features distinct from the prompted feature. For example, prompting for *Concise* inadvertently but significantly reduce perceived expertise (25.6% win rate on *Expert*), even though a concise response can be recognized as a professional response from an expert.

5 Side Effect Identification

From the causal analysis, we observe that style features, when used, can cause unexpected stylistic changes in the agent's responses. We define this phenomenon as **side effect**, an unintended and unexpected behavioral shift in a style distinct from the prompted control feature. While "unintended" is inherently subjective, we ground our selection in the empirical win rate matrices across two domains and two open-sourced models (`Llama3` and `Qwen3-8B`).

Figure 6 displays the comprehensive set of win rate matrices across all models and domains, which serves as the empirical basis for our selection. We analyze these matrices to identify side effects pairs (Main Feature, Side Feature) where the prompt causes a statistically significant degradation in a secondary feature that ideally should remain unaffected or increase. We systematically screen for such counter-intuitive negative correlations across the datasets, and narrow our focus to five representative pairs that exhibit strong, statistically significant side effects to validate our mitigation experiments. We list our chosen side effect examples in Table 1.

6 Side Effect Mitigation

Unintended agent behaviors are undesirable when applying stylistic conditioning. The goal of Side Effect Mitigation (Mitigation) is to unbias the agent's responses toward the positive aspects of a Side Feature while preserving the strength of the Main Feature. To address this trade-off, we evaluate two mitigation strategies: *Prompt Intervention* and *Steering Intervention*. We analyze their effectiveness by aiming to restore compromised feature effects without sacrificing the intended stylistic objective. We partition our full constructed dataset into training, validation, and test splits with a 3:1:1 ratio. Stratified sampling is applied to ensure consistent coverage of domains and topics across all splits.

6.1 Prompt Intervention

As a simple mitigation strategy, we modify the system instructions to explicitly request both the Main Feature and the Side Feature: *"Please be {main_feature} and {side_feature} in your response."* We adopt this approach to evaluate whether explicit natural language instructions are sufficient to override the latent trade-offs triggered by single-feature prompting. By conditioning the generation on both the desired style feature and the suppressed style feature, we aim to force the model to navigate the tension between these features (e.g., maintaining brevity without sacrificing expertise) solely through in-context learning, without requiring parameter updates or activation steering.

6.2 Steering Intervention

We further explore **activation steering** to mitigate side effects at the level of a model's linear-layer activations. A detailed algorithm is presented in Appendix H.

Vector Extraction (Training) We extract steering vectors using Contrastive Activation Addition (CAA, Panickssery et al., 2024). CAA identifies direction vectors in a model's activation space by contrasting activations elicited by paired prompts that differ only in a target attribute, isolating representations associated with that attribute [4]. These contrastive directions can then be added to intermediate activations at inference time to reliably steer model behavior without modifying model weights. Following this procedure, we construct contrastive prompt pairs for each style feature and compute layer-wise steering vectors by averaging the activation differences between styled and original responses at a fixed token position, yielding raw CAA steering directions for all candidate layers.

[4] See Appendix C for contrastive pair examples.

	Model	Llama3			Qwen3	
	Features / Method	Concise Expert	Efficient Helpful	Curious Empathetic	Engaging Impartial	Polite Efficient
Win Rate of Main Feature	Only Main	0.812*	0.532*	0.822*	0.982*	0.697*
	Only Side	0.479	0.587*	0.730*	0.248*	0.221*
	Prompting	0.482	0.394*	0.934*	0.990*	0.983*
	Steering	0.367*	0.315*	0.940*	0.948*	0.959*
Win Rate of Side Feature	Only Main	0.281*	0.291*	0.438*	0.304*	0.440*
	Only Side	0.709*	0.599*	0.838*	0.788*	0.522
	Prompting	0.709*	0.945*	0.878*	0.484	0.641*
	Steering	0.660*	0.941*	0.820*	0.474	0.459*

Table 2: The Side Effect Mitigation Experimental Results. Concise-Expert and Efficient-Helpful are conducted in the task and daily domains. Curious-Empathetic is conducted in the daily domain. Engaging-Impartial is conducted in the daily domain. Polite-Efficient is conducted in the task and daily domains. We use a two-sided binomial test ($p <= 0.05$) to calculate statistical significance.

Layer Selection (Validation) To determine the optimal injection layer, we utilize the validation split to generate responses with steering vectors applied at various layers. We compare these against Neutral responses to identify the specific layer that maximizes the restoration of the side effect while maintaining generation quality.

Inference (Testing) During testing, we prompt the model with the single-style prompt (*"Please be {style_feature} in your response."*) but intervene during the forward pass by injecting the steering vector derived from the best performing layer. Since the identified side effects represent an unexpected loss of a style feature (e.g., a loss of expertise when being concise), we add the Side Feature steering vector to shift the model's internal state toward the compromised feature, effectively counteracting the suppression caused by the prompt.

We subsequently calculate the win rates of prompting and steering mitigations against Neutral responses to assess their effectiveness in mitigating side effects, measured by performance on Side Features and Main Features. While the mitigation aims to improve win rates on Side Features, an effective approach should also preserve performance on the Main Features.

6.3 Results

We report the efficacy of our Mitigation in Table 2 and Figure 4. The results reveal a complex landscape of trade-offs between maintaining stylistic adherence and mitigating unintended behavioral degradation. Mitigation result across all 12 style features is reported in Figure 8.

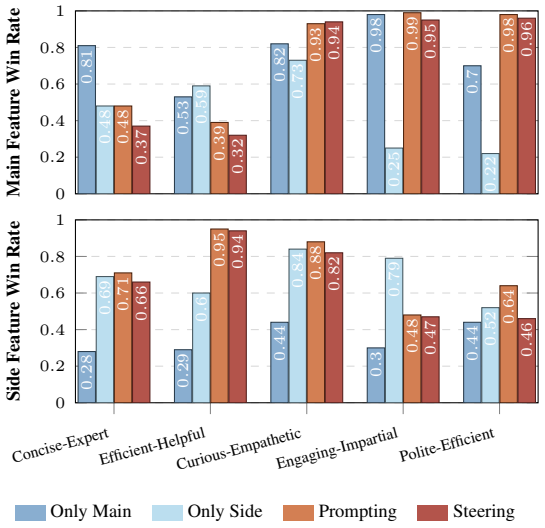


Figure 4: Performance comparison of prompting vs. steering Interventions. Top: Main Feature win rates. Bottom: Side Feature win rates.

Efficacy of Side Effect Restoration Both *Prompt Intervention* and *Steering Intervention* successfully restore the compromised Side Features across nearly all experimental settings. As shown in Figure 4, the Mitigation methods consistently outperform the non-intervened styled responses. For example, in the `Llama3` Task setting, the *Efficiency* prompt originally suppresses *Helpfulness* to a win rate of 0.29. Prompt Intervention restores this to 0.95, and Steering Intervention to 0.94, effectively neutralizing the negative side effect. Similarly, for the `Llama3`, Open Domain, *Curious* prompt, the suppressed *Empathy* score (0.44) is robustly restored by both Prompt (0.88) and Steering (0.82) Interventions.

The Alignment Tax: Main Feature Degradation While side effects are successfully mitigated, the result shows mixed outcome on preserving the strength on the Main Feature. Figure 4 illustrates that some Main Features' win rates drop significantly, while others maintain or even strengthen the Main Feature.

High Degradation: For `Llama3`, enforcing *Conciseness* while trying to restore *Expertise* causes the *Concise* win rate to drop from 0.81 (Only Main) to 0.48 (prompt) and 0.37 (steering). This suggests that "being expert" requires a minimum length that fundamentally conflicts with "being concise."

Synergistic Pairs: Conversely, some features exhibit synergy. Mitigating the side effects of *Curious*

(Llama3) and *Polite* (Qwen3) actually *improves* or maintains the Main Feature win rates. For instance, adding *Efficiency* restoration to the *Polite* prompt raises the *Polite* win rate from 0.70 to 0.98, suggesting that a polite agent that is also efficient is perceived as *more* polite than one that is merely polite but inefficient.

Mitigation Comparison: Prompting vs. Steering Comparing the two mitigation strategies, *Prompt Intervention* generally offers a more stable balance. *Steering Intervention*, while still powerful at restoring side effects (e.g., restoring *Helpfulness* to 0.94 in the *Efficient* setting), tends to degrade the Main Feature more aggressively than prompting. Specifically, in the *Concise* task, steering reduces the Main Feature adherence to 0.37, whereas Prompt Intervention maintains it at 0.48. This suggests that while steering vectors can forcefully inject a trait, they may disrupt the delicate activation patterns required for the primary style more than natural language instructions.

7 Discussion

7.1 The Influence of Response Length on Feature Strength

We observe a substantial correlation between the length of a generated response and its perceived strength across various style features. As illustrated in Figure 3 and 6, the win rate on average word count (represented by the "length" column) is strongly correlated with the specific style feature used in the system prompt. Specifically, prompting for *Efficient* consistently decreases response length (win rate 20.9%), whereas social prompts like *Outgoing* (71.5%) and *Engaging* (74.6%) significantly increase verbosity.

This variation in length positively correlates with the strength of style features in social contexts; longer responses generally achieve higher win rates on dimensions such as *Friendly* and *Empathetic*. However, this correlation does not uniformly app to all features. For instance, when models are prompted to be *Impartial*, they achieve a high win rate on the *Impartial* evaluation feature (57.7%) while maintaining a response length (42.6%) that is relatively close to the baseline. This suggests that while some features to manifest, certain features can be steered without significant deviations in output length.

7.2 Inconsistencies Between Semantic Similarity and Behavioral Outcomes

Our results highlight a critical distinction between the semantic intent of a prompt and its evaluation in the feature space. We find that distinct prompts can converge to similar evaluation profiles, while semantically related features can behave divergently.

Prompt–Evaluation Non-Equivalence While *Concise* and *Efficient* produce similar effects when used as prompts—both depressing social feature ratings and response length—they are evaluated differently as output features. For example, prompting with *Expert* leads to a significant increase in *Efficient* ratings (66.0%) but a decrease in *Concise* ratings (42.5%). This demonstrates that a response perceived as efficient by the judge is not necessarily concise; the model can demonstrate efficiency through density of information rather than mere brevity.

Evaluation-Space Convergence Conversely, prompts with disparate meanings can converge on specific evaluation metrics. The *Efficient* and *Detailed* prompts exhibit opposing effects across most of the evaluation spectrum (e.g., *Efficient* lowers length, *Detailed* raises it). However, the *Helpful* prompt leads to statistically significant increases in both *Efficient* (59.0%) and *Detailed* (62.9%) ratings. This convergence suggests that the abstract quality of "helpfulness" effectively bridges contradictory traits, optimizing for both information density and completeness.

7.3 Asymmetry of Causal Relationship

We identify a distinct directionality in the side effect patterns between style features, where the causal link between two features is often asymmetric. This phenomenon manifests in three key pairs:

Impartiality and Conciseness: Instructing a model to be *Impartial* increases its *Concise* rating (57.7%), likely because neutrality discourages conversational filler. Conversely, instructing a model to be *Concise* makes it less *Impartial* (44.9%), as extreme brevity may necessitate omitting nuanced perspectives.

Expertise and Efficiency: Prompting for *Expert* makes a model more *Efficient* (66.0%), but prompting for *Efficient* makes it significantly less *Expert* (34.6%). This suggests that expertise naturally en-

compasses efficiency, whereas forced efficiency often sacrifices the depth required for expertise.

Impartiality and Empathy: While *Impartial* prompts predictably decrease *Empathetic* ratings (40.1%), *Empathetic* prompts unexpectedly increase *Impartial* ratings (57.6%). This may indicate that the judge perceives the validation and active listening typical of empathy as a form of unbiased engagement.

7.4 Domain and Model Differences

Task-oriented and open-domain conversations are the two most widely studied scenarios for conversational agents (Yi et al., 2024). Figure 6 shows that the effects of style features are highly domain-dependent. In particular, tendencies toward Side Features are generally more pronounced in the Daily domain than in the Task domain. This suggests that although causal relationships among style features already exist in constrained settings such as task-oriented conversations, the increased freedom of open-domain interactions amplifies stylistic biases.

We also observe that certain Style Features exhibit qualitatively different effects across domains. For example, applying *Impartial* to `Llama3` generally increases bias toward all Side Features in the Task domain, whereas in the Daily domain it leads to a significant decrease in win rates across most Side Features.

In addition to domain dependence, the causal relationships among style features are also model-dependent. While different models share broadly similar behavioral trends, they differ substantially in their detailed responses. For instance, `GPT-5-mini` exhibits only weak impacts across style features in task-oriented scenarios, in contrast to `Qwen3-8B`, which shows strong effects on Side Features in both Task and Daily domains.

Together, these findings suggest that when a particular style feature is applied, users should carefully consider its interactions with other stylistic dimensions. Such interactions can vary substantially depending on both the conversational domain and the underlying model.

7.5 Side Effect Mitigation Experiments

Our experiments highlight significant limitations in current mitigation strategies, revealing that lightweight "counter prompts" and steering vectors struggle to decouple conflicting stylistic traits. When attempting to neutralize trade-offs, adding a Side Feature steering vector often actively degrades the primary intended effect rather than balancing the output. For instance, in the *Concise-Expert* pair, the steering vector reduces the *Concise* win rate from 81% to 37%—a loss of nearly half the original gain. Similarly, prompting for *Efficient-Helpful* drops the *Efficient* score from 53% to 39%, suggesting that style features are deeply entangled with each others in causal relationship and cannot be cleanly separated by simple concatenation or activation editing.

To verify that these side effects are not merely artifacts of prompt positioning, we conduct an ablation study reversing the feature order in Prompt Intervention (Figure 11). The results show that Side Feature effects persist regardless of position, disproving the hypothesis that recency bias drives these trade-offs and pointing instead to intrinsic, high-dimensional correlations between these features. This persistence indicates that using style features with a specific order is insufficient for signaling the primary and secondary focus of styles. Achieving a robust balance between Main and Side Feature without collateral drift will likely require more principled approaches, such as iterative reinforcement learning, targeted fine-tuning, or multi-objective optimization.

8 Conclusion

This work demonstrates that commonly used style prompts in conversational agents introduce systematic and non-trivial side effects, revealing that stylistic controls are deeply entangled rather than independent. Through a large-scale survey and controlled causal experiments, we show that prompting with a style feature often degrades other important qualities, such as expertise, helpfulness, or efficiency. It challenges the assumption that style prompts are faithful and fairly isolated controls. Our mitigation experiments further indicate that both prompt-based and activation steering approaches struggle to resolve these trade-offs without sacrificing the primary intended effect. Together, these findings highlight the need to rethink style control as a multi-objective problem and motivate more principled methods for achieving reliable, safe, and targeted stylistic steering in LLM-based conversational agents. Our released CASSE dataset, with its Side Feature annotations, will help future researchers in this research direction.

Limitation

A key limitation of our study is ecological validity: all findings stem from short, synthetic conversational agent responses, rather than from longer, human-to-agent dialogues assessed by real users. This design offers scale and control but risks over-estimating side effects that might be attenuated or that may manifest differently when humans adapt their wording, challenge inconsistencies, or engage in multi-topic conversations. Moreover, we do not evaluate on the largest language models, due to limitations in compute resources; less common cues and other languages could yield different side effect patterns. Finally, our mitigation test uses simple prompt concatenation and activation steering, so the negative results do not rule out more sophisticated techniques such as iterative re-prompting or fine-tuning, which remain for future work.

References

Yelaman Abdullin, Diego Molla, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2023. Synthetic dialogue dataset generation using LLM agents. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 181–191, Singapore. Association for Computational Linguistics.

Anum Afzal, Alexander Kowsik, Rajna Fani, and Florian Matthes. 2024. Towards optimizing and evaluating a retrieval augmented QA chatbot using LLMs with human-in-the-loop. In *Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024)*, pages 4–16, Mexico City, Mexico. Association for Computational Linguistics.

Stuti Agrawal, Pranav Pillai, Nishi Uppuluri, Revanth Gangi Reddy, Sha Li, Gokhan Tur, Dilek Hakkani-Tur, and Heng Ji. 2024. Dialog flow induction for constrainable LLM-based chatbots. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 66–77, Kyoto, Japan. Association for Computational Linguistics.

Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic Arabic medical dialogues using advanced multi-agent LLM techniques. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 11–26, Bangkok, Thailand. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

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

Andrew Brown, Jiading Zhu, Mohamed Abdelwahab, Alec Dong, Cindy Wang, and Jonathan Rose. 2024. Generation, distillation and evaluation of motivational interviewing-style reflections with a foundational language model. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1241–1252, St. Julian's, Malta. Association for Computational Linguistics.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.

Sijia Cheng, Wen Yu Chang, and Yun-Nung Chen. 2025. Exploring personality-aware interactions in salesperson dialogue agents. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 60–71, Bilbao, Spain. Association for Computational Linguistics.

Rendi Chevi, Kentaro Inui, Thamar Solorio, and Alham Fikri Aji. 2025. How individual traits and language styles shape preferences in open-ended user-LLM interaction: A preliminary study. *arXiv preprint arXiv:2504.17083*.

Yuntian Deng, Wenting Zhao, Jack Hessel, Xiang Ren, Claire Cardie, and Yejin Choi. 2024. WildVis: Open source visualizer for million-scale chat logs in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 497–506, Miami, Florida, USA. Association for Computational Linguistics.

Chengfeng Dou, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2023. PlugMed: Improving specificity in patient-centered medical dialogue generation using in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5050–5066, Singapore. Association for Computational Linguistics.

Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, Chenxi Yuan, Qi Long, Ying Ding, Tianlong Chen, and Kaidi Xu. 2025. GuideLLM: Exploring LLM-guided conversation with applications in autobiography interviewing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics:*

Human Language Technologies (Volume 1: Long Papers), pages 5558–5588, Albuquerque, New Mexico. Association for Computational Linguistics.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An open-source avatar multimodal empathetic chatbot. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–71, Bangkok, Thailand. Association for Computational Linguistics.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. 2025. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6218–6240, Albuquerque, New Mexico. Association for Computational Linguistics.

Yujie Feng, Bo Liu, Xiaoyu Dong, Zexin Lu, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2024. Continual dialogue state tracking via reason-of-select distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7075–7087, Bangkok, Thailand. Association for Computational Linguistics.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P. Dickerson. 2025. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sarah E. Finch and Jinho D. Choi. 2025. Leveraging explicit reasoning for inference integration in commonsense-augmented dialogue models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2222–2235, Abu Dhabi, UAE. Association for Computational Linguistics.

Sarah E. Finch, Ellie S. Paek, and Jinho D. Choi. 2023. Leveraging large language models for automated dialogue analysis. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 202–215, Prague, Czechia. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *Preprint*, arXiv:2311.04892.

Junqing He, Liang Zhu, Rui Wang, Xi Wang, Gholamreza Haffari, and Jiaxing Zhang. 2025a. MADial-bench: Towards real-world evaluation of memory-augmented dialogue generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9902–9921, Albuquerque, New Mexico. Association for Computational Linguistics.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. 2025b. Crab: A novel configurable role-playing LLM with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Pegah Jandaghi, Xianghai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2024. Faithful persona-based conversational dataset generation with large language models. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 114–139, Bangkok, Thailand. Association for Computational Linguistics.

Jihyoung Jang, Taeyoung Kim, and Hyounghun Kim. 2024. Mixed-session conversation with egocentric memory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11786–11815, Miami, Florida, USA. Association for Computational Linguistics.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Proceedings of the 2025 Conference on Language Modeling (COLM)*.

Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. 2023. SuperDialseg: A large-scale dataset for supervised dialogue segmentation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4086–4101, Singapore. Association for Computational Linguistics.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, and Monica Lam. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. In *Proceedings of the 2024 Conference on Empirical Methods in*

Natural Language Processing, pages 9917–9955, Miami, Florida, USA. Association for Computational Linguistics.

Frederic Kirstein, Terry Ruas, Robert Kratel, and Bela Gipp. 2024. Tell me what I need to know: Exploring LLM-based (personalized) abstractive multi-source meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 920–939, Miami, Florida, US. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, and 1 others. 2023. Openassistant conversations – democratizing large language model alignment. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Fabian Lechner, Allison Lahnala, Charles Welch, and Lucie Flek. 2023. Challenges of GPT-3-based conversational agents for healthcare. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 619–630, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted LLMs as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554, Toronto, Canada. Association for Computational Linguistics.

Jeehyun Lee, Seung-Moo Yang, and Won Ik Cho. 2025a. AMAN: Agent for mentoring and assisting newbies in MMORPG. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 522–532, Abu Dhabi, UAE. Association for Computational Linguistics.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025b. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large scale language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yu Li, Shang Qu, Jili Shen, Shangchao Min, and Zhou Yu. 2024a. Curriculum-driven edubot: A framework for developing language learning chatbots through synthesizing conversational data. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 400–419, Kyoto, Japan. Association for Computational Linguistics.

Yuetai Li, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Dinuka Sahabandu, Bhaskar Ramasubramanian, and Radha Poovendran. 2024b. CleanGen: Mitigating backdoor attacks for generation tasks in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9101–9118, Miami, Florida, USA. Association for Computational Linguistics.

et al. Liu. 2025. Personalized text generation with contrastive activation steering. *arXiv preprint arXiv:2503.05213*.

June M. Liu, He Cao, Renliang Sun, Rui Wang, Yu Li, and Jiaxing Zhang. 2025. CAPE: A Chinese dataset for appraisal-based emotional generation in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6291–6309, Albuquerque, New Mexico. Association for Computational Linguistics.

Junhua Liu, Tan Yong Keat, Bin Fu, and Kwan Hui Lim. 2024. LARA: Linguistic-adaptive retrieval-augmentation for multi-turn intent classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1096–1106, Miami, Florida, US. Association for Computational Linguistics.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. *Preprint*, arXiv:2507.16076.

Pedro Henrique Luz de Araujo and Benjamin Roth. 2025. Helpful assistant or fruitful facilitator? investigating how personas affect language model behavior. *PLOS One*, 20(6):e0325664.

Navid Madani, Anusha Bagalkotkar, Supriya Anand, Gabriel Arnson, Rohini K. Srihari, and Kenneth

Joseph. 2025. A recipe for building a compliant real estate chatbot. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 213–235, Abu Dhabi, UAE. Association for Computational Linguistics.

Manuj Malik, Jing Jiang, and Kian Ming A. Chai. 2024. An empirical analysis of the writing styles of persona-assigned LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19369–19388, Miami, Florida, USA. Association for Computational Linguistics.

Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tur. 2023. Using in-context learning to improve dialogue safety. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11882–11910, Singapore. Association for Computational Linguistics.

Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023. e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967, Singapore. Association for Computational Linguistics.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062, Miami, Florida, USA. Association for Computational Linguistics.

Lidiya Murakhovs'ka, Philippe Laban, Tian Xie, Caiming Xiong, and Chien-Sheng Wu. 2023. Salespeople vs SalesBot: Exploring the role of educational value in conversational recommender systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9823–9838, Singapore. Association for Computational Linguistics.

Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, and Fabrice Lefèvre. 2025. Enabling trait-based personality simulation in conversational LLM agents: Case study of customer assistance in French. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 299–308, Bilbao, Spain. Association for Computational Linguistics.

Shinnosuke Nozue, Yuto Nakano, Shoji Moriya, Tomoki Ariyama, Kazuma Kokuta, Suchun Xie, Kai Sato, Shusaku Sone, Ryohei Kamei, Reina Akama, Yuichiroh Matsubayashi, and Keisuke Sakaguchi. 2024. A multimodal dialogue system to lead consensus building with emotion-displaying. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 669–673, Kyoto, Japan. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o mini model. `https://platform.openai.com/docs/models/gpt-4o-mini`. Accessed: 2026-01-05; a fast, affordable small AI model with large context support and multimodal capabilities.

OpenAI. 2024. Text embedding 3 small. `https://platform.openai.com/docs/guides/embeddings`. Accessed: 2025-07-29.

OpenAI. 2025. GPT-5-mini model. `https://platform.openai.com/docs/models`. Accessed: 2026-01-05.

Ryosuke Oshima, Seitaro Shinagawa, and Shigeo Morishima. 2024. The gap in the strategy of recovering task failure between GPT-4V and humans in a visual dialogue. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 728–745, Kyoto, Japan. Association for Computational Linguistics.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. DialogBench: Evaluating LLMs as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170, Mexico City, Mexico. Association for Computational Linguistics.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. *Preprint*, arXiv:2312.06681.

Dominic Petrak, Thy Thy Tran, and Iryna Gurevych. 2024. Learning from implicit user feedback, emotions and demographic information in task-oriented and document-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4573–4603, Miami, Florida, USA. Association for Computational Linguistics.

Arpan Phukan, Manish Gupta, and Asif Ekbal. 2024. ECIS-VQG: Generation of entity-centric information-seeking questions from videos. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14411–14436, Miami, Florida, USA. Association for Computational Linguistics.

Xiao Pu, Tianxing He, and Xiaojun Wan. 2024. Style-compress: An LLM-based prompt compression framework considering task-specific styles. *arXiv preprint arXiv:2410.14042*.

Inass Rachidi, Anas Ezzakri, Jaime Bellver-Soler, and Luis Fernando D'Haro. 2025. Design, generation and evaluation of a synthetic dialogue dataset for contextually aware chatbots in art museums. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 20–28, Bilbao, Spain. Association for Computational Linguistics.

Ravi Shanker Raju, Swayambhoo Jain, Bo Li, Jonathan Lingjie Li, and Urmish Thakker. 2024. Constructing domain-specific evaluation sets for LLM-as-a-judge. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 167–181, Miami, Florida, USA. Association for Computational Linguistics.

Mansi Rana, Kadri Hacioglu, Sindhuja Gopalan, and Maragathamani Boothalingam. 2025. Zero-shot slot filling in the age of LLMs for dialogue systems. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 697–706, Abu Dhabi, UAE. Association for Computational Linguistics.

Gonçalo Raposo, Luisa Coheur, and Bruno Martins. 2023. Prompting, retrieval, training: An exploration of different approaches for task-oriented dialogue generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 400–412, Prague, Czechia. Association for Computational Linguistics.

Traian Rebedea, Makesh Sreedhar, Shaona Ghosh, Jiaqi Zeng, and Christopher Parisien. 2024. CantTalkAboutThis: Aligning language models to stay on topic in dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12232–12252, Miami, Florida, USA. Association for Computational Linguistics.

Revanth Reddy, Hao Bai, Wentao Yao, Sharath Chandra Etagi Suresh, Heng Ji, and ChengXiang Zhai. 2023. Social commonsense-guided search query generation for open-domain knowledge-powered conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 873–885, Singapore. Association for Computational Linguistics.

Cristina Reguera-Gómez, Denis Paperno, and Maaike H. T. de Boer. 2025. Empathy vs neutrality: Designing and evaluating a natural chatbot for the healthcare domain. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 508–517, Tallinn, Estonia. University of Tartu Library.

Sougata Saha and Rohini Srihari. 2024. Integrating argumentation and hate-speech-based techniques for countering misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11109–11124, Miami, Florida, USA. Association for Computational Linguistics.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam . 2024. MediTOD: An English dialogue dataset for medical history taking with comprehensive annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.

Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2387–2413, Singapore. Association for Computational Linguistics.

Seungyeon Seo and Gary Geunbae Lee. 2024. DiagESC: Dialogue synthesis for integrating depression diagnosis into emotional support conversation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 686–698, Kyoto, Japan. Association for Computational Linguistics.

Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Context does matter: Implications for crowdsourced evaluation labels in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1258–1273, Mexico City, Mexico. Association for Computational Linguistics.

Li Siyan, Teresa Shao, Julia Hirschberg, and Zhou Yu. 2024a. Using adaptive empathetic responses for teaching English. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 34–53, Mexico City, Mexico. Association for Computational Linguistics.

Li Siyan, Teresa Shao, Zhou Yu, and Julia Hirschberg. 2024b. EDEN: Empathetic dialogues for English learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3492–3511, Miami, Florida, USA. Association for Computational Linguistics.

Dominic Sobhani, Ruiqi Zhong, Edison Marrese-Taylor, Keisuke Sakaguchi, and Yutaka Matsuo. 2025. Language models can categorize system inputs for performance analysis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6241–6257, Albuquerque, New Mexico. Association for Computational Linguistics.

Rickard Stureborg, Sanxing Chen, Roy Xie, Aayushi Patel, Christopher Li, Chloe Zhu, Tingnan Hu, Jun Yang, and Bhuwan Dhingra. 2024. Tailoring vaccine messaging with common-ground opinions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2553–2575, Mexico City, Mexico. Association for Computational Linguistics.

Albert Sun, Varun Nair, Elliot Schumacher, and Anitha Kannan. 2024. CONSCENDI: A contrastive and scenario-guided distillation approach to guardrail models for virtual assistants. In *Proceedings of the*

2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4009–4030, Mexico City, Mexico. Association for Computational Linguistics.

Ekaterina Svikhnushina and Pearl Pu. 2023. Approximating online human evaluation of social chatbots with prompting. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 268–281, Prague, Czechia. Association for Computational Linguistics.

Maksym Taranukhin, Sahithya Ravi, Gabor Lukacs, Evangelos Milios, and Vered Shwartz. 2024. Empowering air travelers: A chatbot for Canadian air passenger rights. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 326–335, Miami, FL, USA. Association for Computational Linguistics.

Kimberly Le Truong, Riccardo Fogliato, Hoda Heidari, and Zhiwei Steven Wu. 2025. Persona-augmented benchmarking: Evaluating LLMs across diverse writing styles. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *Preprint*, arXiv:2406.01171.

Yuka Tsubota and Yoshinobu Kano. 2024. Text generation indistinguishable from target person by prompting few examples using LLM. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 13–20, Tokyo, Japan. Association for Computational Linguistics.

Dirk Väth, Lindsey Vanderlyn, and Ngoc Thang Vu. 2024. Towards a zero-data, controllable, adaptive dialog system. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16433–16449, Torino, Italia. ELRA and ICCL.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023a. Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064, Singapore. Association for Computational Linguistics.

Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023b. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1132–1143, Singapore. Association for Computational Linguistics.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.

Lanrui Wang, Jiangnan Li, Chenxu Yang, Zheng Lin, Hongyin Tang, Huan Liu, Yanan Cao, Jingang Wang, and Weiping Wang. 2025. Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 123–140, Abu Dhabi, UAE. Association for Computational Linguistics.

Xingguang Wang, Xuxin Cheng, Juntong Song, Tong Zhang, and Cheng Niu. 2024b. Enhancing dialogue state tracking models through LLM-backed user-agents simulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8724–8741, Bangkok, Thailand. Association for Computational Linguistics.

Milan Wevelsiep, Nicholas Thomas Walker, Nicolas Wagner, and Stefan Ultes. 2025. A voice-controlled dialogue system for NPC interaction using large language models. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 29–38, Bilbao, Spain. Association for Computational Linguistics.

Yuxin Xiao, Sana Tonekaboni, Walter Gerych, Vinith Suriyakumar, and Marzyeh Ghassemi. 2025. When style breaks safety: Defending LLMs against superficial style alignment. *arXiv preprint arXiv:2506.07452*.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024a. KIWI: A dataset of knowledge-intensive writing instructions for answering research questions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12969–12990, Bangkok, Thailand. Association for Computational Linguistics.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024b. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605, Bangkok, Thailand. Association for Computational Linguistics.

Chenghao Yang and Allyson Ettinger. 2023. Can you follow me? testing situational understanding for ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6385–6398, Singapore. Association for Computational Linguistics.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, and Ee-Peng Lim. 2024. Speaker verification in agent-generated conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5655–5676, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*.

Jiong Yu, Sixing Wu, Jiahao Chen, and Wei Zhou. 2024. LLMs as collaborator: Demands-guided collaborative retrieval-augmented generation for commonsense knowledge-grounded open-domain dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13586–13612, Miami, Florida, USA. Association for Computational Linguistics.

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)*, pages 7128–7141, Vienna, Austria. Association for Computational Linguistics.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024a. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, Lingshuai Wang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024b. STICKERCONV: Generating multimodal empathetic responses from scratch. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7707–7733, Bangkok, Thailand. Association for Computational Linguistics.

Weiliang Zhao, Daniel Ben-Levi, Wei Hao, Junfeng Yang, and Chengzhi Mao. 2025a. Diversity helps jailbreak large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4647–4680, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaoyan Zhao and 1 others. 2025b. SteerX: Disentangled steering for LLM personalization. *arXiv preprint arXiv:2510.22256*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. *Preprint*, arXiv:2311.10054.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2024. Think before you speak: Cultivating communication skills of large language models via inner monologue. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3925–3951, Mexico City, Mexico. Association for Computational Linguistics.

Zhouan Zhu, Shangfei Wang, Yuxin Wang, and Jiaqiang Wu. 2025. Integrating visual modalities with large language models for mental health support. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8939–8954, Abu Dhabi, UAE. Association for Computational Linguistics.

Thomas P. Zollo, Ang Li, Andrew Wei Tung Siah, Hongseok Namkoong, and Naimeng Ye. 2025. PersonalLLM: Tailoring LLMs to individual preferences. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

A Prompt Templates

In this section, we present the key prompt templates used in this project.

A.1 LLM-as-a-Judge Prompt Template

System: *You are an expert evaluator of conversation quality.*
User: *Compare the following two assistant responses and determine which one is more {feature}. Response A: {response_a} Response B: {response_b} You must choose one. Which response is more {feature}? Answer with ONLY "A" or "B" (no ties allowed). Answer:*

where the "{feature}" refers to one of the 12 features.

A.2 System Prompt Template for Message Generation With and Without Style Features

Without Style Feature: *"This is a conversation about {topic}."*
With Style Feature: *"This is a conversation about {topic}. Please be {style_features} in your response."*

A.3 Side Effect Mitigation Prompt Template - Normal and Reversed Order

Normal: *Please be {main_feature} but {side_feature} in your response.*

Reversed: *"Please be {side_feature} but {main_feature} in your response."*

A.4 Side Effect Mitigation Steeirng Prompt Template

Since we use steering vector to change model behaviors, we don't include the Side Feature in the prompt to models: *"Please be {main_feature} in your response."*

B Details in Selection of Topics

To represent a wide range of domains and topics in message generation, we select LMSYS-Chat-1M for task-oriented inquiries and DailyDialog for cosial, daily-life interactions. For each domain, we each pick 10 topics. Details of the selected topics, and examples are shown in Table 4

LMSYS-Chat-1M presents top 20 topics (clusters) in the paper, but the released dataset does not have them annotated. Instead, we first merge similar topics from the initial 20 and curate down to 10 representative topics, and used OpenAI's gpt-4o-mini model (OpenAI, 2024) to annotate full dataset until each topic has 10 conversations. On DailyDialog, we directly use the 10 topics suggested by the paper. These curation result in a total of 200 unique conversational contexts, balancing functional assistance with casual chatter.

C Constrative Prompt Pair Example for Steering Vector Extraction

Contrastive Prompt Pair Example: Expert
Based on the Contrastive Activation Addition (CAA) methodology, the extraction process uses prompts that force the model to choose between a "styled" (expert) response and an "neutral" (no style feature) response.
Positive Prompt (Eliciting Choice A): *"You are having a conversation about quantum computing. Please choose the option that shows that you are EXPERT. User message: How do superconducting qubits work? Choice A: Superconducting qubits leverage the Josephson effect to create a nonlinear oscillator, typically operating in the transmon regime to suppress charge noise. Choice B: Superconducting qubits are a type of computer part that uses cold temperatures and electricity to solve hard problems. Your decision: A"*
Baseline Prompt (Eliciting Choice B): *"You are having a conversation about quantum computing. Please choose the option that shows that you are EXPERT. User message: How do superconducting qubits work? Choice A: Superconducting qubits leverage the Josephson effect to create a nonlinear oscillator, typically operating in the transmon regime to suppress charge noise. Choice B: Superconducting qubits are a type of computer part that uses cold temperatures and electricity to solve hard problems. Your decision: B"*

D Extracted Feature Distribution

In this section, we present the extract feature distribution, which shows how frequently each unique style feature appears in each paper in the survey results.

E Full Win Rate Matrix For Each Domain and Each Model

In this section, we present two figures, one that shows the win rates across all models and domains and the other that shows the win rates averaged across three models for each domain. Both figures provide more details on how using a style feature affects a model response.

F Full Win Rate Matrix for Prompting Intervention and Steering Intervention

In this section, we present Figure 8, showing the effects of using prompting intervention and steering intervention across 12 style features. Across both methods, interventions consistently improve the intended Side Feature but also induce systematic changes in non-target attributes (e.g., politeness, engagement, and length), revealing that feature control is not isolated. These results demonstrate that both style features operate over a shared latent feature space, where modifying one dimension propagates to others.

G Style Feature Frequency Across Related Papers

In this section, we present Figure 3, which shows the papers where any of our 16 filtered style features appears in.

H Steering Intervention Algorithm

In this section, we present a detailed Steerint Intervention algorithm with three smaller Algorithm 1, 2, and 3, covering from steering vector extraction to steered output generation.

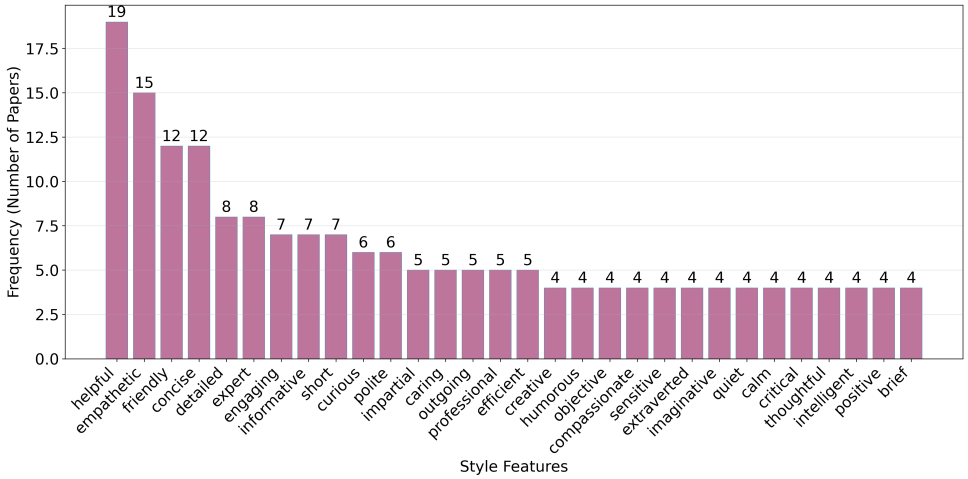


Figure 5: Extracted Feature Distribution. This figure shows the frequency of each unique style feature in the survey results, highlighting the most common style features such as helpful and empathetic.

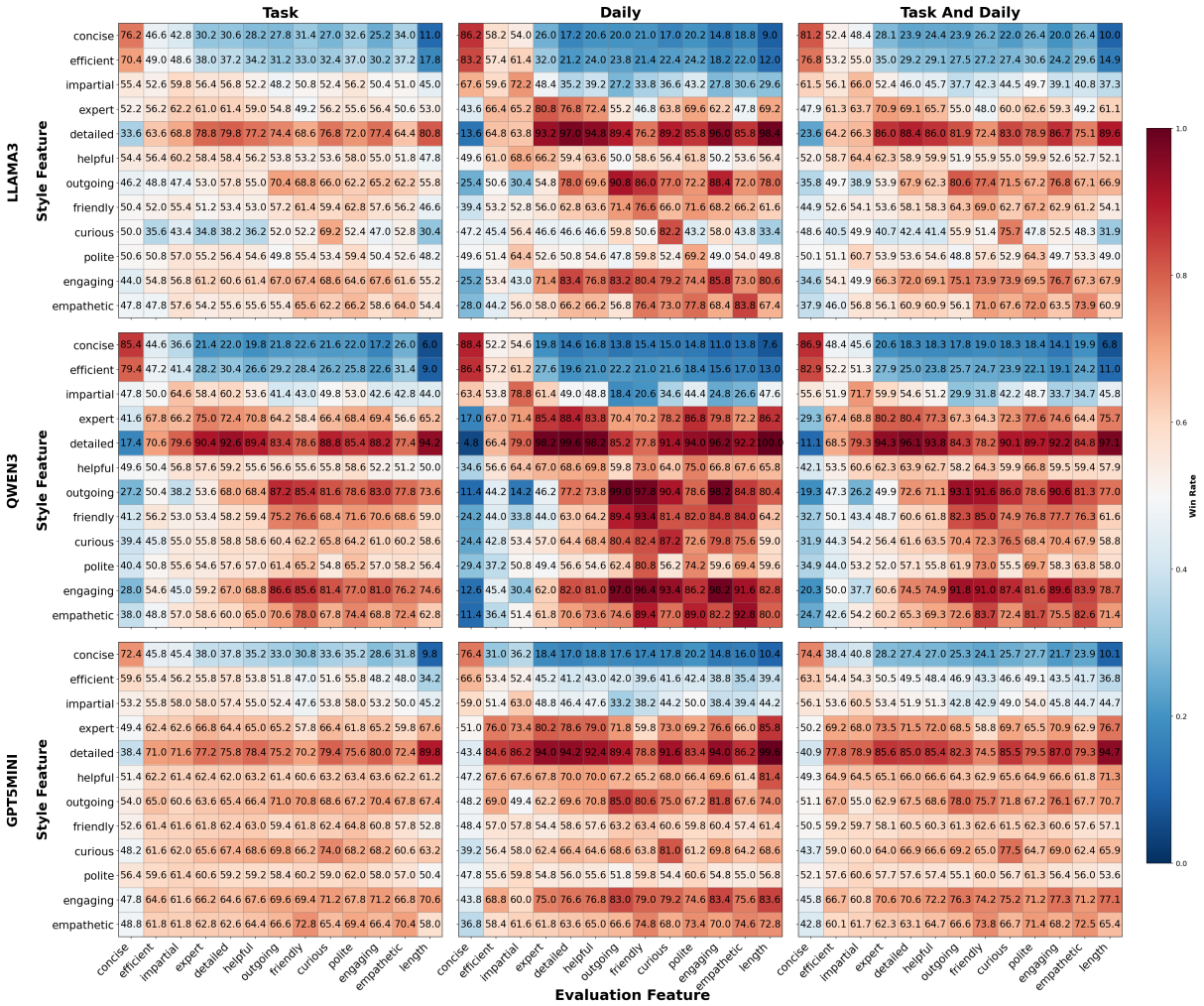


Figure 6: Combined heatmap showing win rates across all models and domains. This figure provides a comprehensive view of the stylistic shifts and side effects observed in our experiments.

H.1 Phase 1: Steering Vector Extraction

The first Algorithm 1 involves constructing contrastive datasets and computing the mean activation difference between styled and neutral responses.

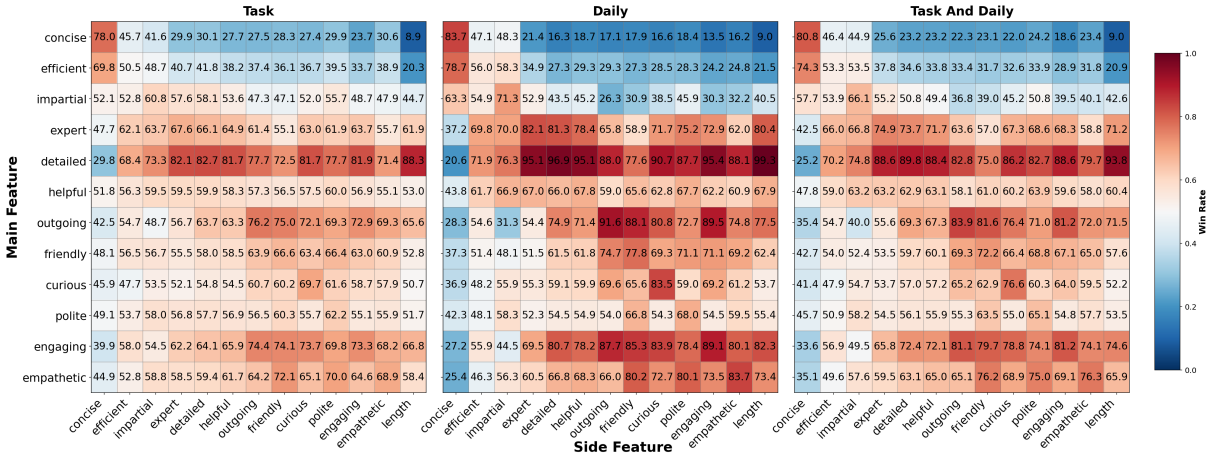


Figure 7: Three heatmaps showing win rates averaged across three models for each domain.

Algorithm 1 Steering Vector Extraction

Input: Training set $\mathcal{D}_{\text{train}}$, Target model $\mathcal{M}$, Candidate layers $\mathcal{L}$, Style features $\mathcal{F}$.
Output: Raw steering vectors $\{\mathbf{v}_f^{(\ell)}\}$ for all $f \in \mathcal{F}, \ell \in \mathcal{L}$.
Step 1: Contrastive Data Preparation
for each feature $f \in \mathcal{F}$ **do**
  $\mathcal{C}_f \leftarrow \emptyset$
  for each $(q_i, t_i, y_i^{\text{styled}}, y_i^{\text{orig}}) \in \mathcal{D}_{\text{train}}$ where style $= f$ **do**
    if $y_i^{\text{styled}} \neq y_i^{\text{orig}}$ **then**
      $p_i \leftarrow$ "Choose the option that demonstrates you are f. Choice A: y_i^{styled}. Choice B: y_i^{orig}";
      $\mathcal{C}_f \leftarrow \mathcal{C}_f \cup \{(p_i, \text{A}, \text{B})\}$;
    end if
  end for
end for
Step 2: Vector Computation
for each feature $f \in \mathcal{F}$ **do**
  for each $(p_i, a, b) \in \mathcal{C}_f$ **do**
    $x_+ \leftarrow \text{ChatTemplate}(p_i, a)$; $x_- \leftarrow \text{ChatTemplate}(p_i, b)$;
    for each layer $\ell \in \mathcal{L}$ **do**
      $\mathbf{h}_+^{(\ell)} \leftarrow \mathcal{M}.\text{get\_activation}(\ell, x_+, \text{pos} = -2)$;
      $\mathbf{h}_-^{(\ell)} \leftarrow \mathcal{M}.\text{get\_activation}(\ell, x_-, \text{pos} = -2)$;
    end for
  end for
  for each layer $\ell \in \mathcal{L}$ **do**
    $\mathbf{v}_f^{(\ell)} \leftarrow \frac{1}{|\mathcal{C}_f|} \sum_i (\mathbf{h}_+^{(\ell)}[i] - \mathbf{h}_-^{(\ell)}[i])$;
  end for
end for

Algorithm 2 Validation and Checkpoint Baking

Input: Raw vectors $\{\mathbf{v}_f^{(\ell)}\}$, Validation set $\mathcal{D}_{\text{val}}$, Target model $\mathcal{M}$.
Output: Best layers $\{\ell_f^*\}$, Baked checkpoints $\{\mathcal{M}_f^{\text{steered}}\}$.
Step 1: Best Layer Selection
for each feature $f \in \mathcal{F}$ **do**
  for each layer $\ell \in \mathcal{L}$ **do**
    $\mathcal{M}_f^{(\ell)} \leftarrow \mathcal{M}$ with $\mathbf{v}_f^{(\ell)}$ added at layer ℓ;
    for each $(q_i, t_i, y_i^{\text{styled}}, y_i^{\text{neutral}}) \in \mathcal{D}_{\text{val}}$ **do**
      $y_i^{\text{steered}} \leftarrow \mathcal{M}_f^{(\ell)}.\text{generate}(q_i)$;
      $w \leftarrow \mathcal{M}_{\text{judge}}.\text{compare}(f, y_i^{\text{steered}}, y_i^{\text{neutral}})$;
    end for
    $W_f^{(\ell)} \leftarrow n(\text{steered wins})/n_{\text{total}}$;
  end for
  $\ell_f^* \leftarrow \arg\max_\ell W_f^{(\ell)}$;
end for
Step 2: Checkpoint Baking
for each feature $f \in \mathcal{F}$ **do**
  $\theta \leftarrow \text{load\_checkpoint}(\mathcal{M})$;
  $\theta[\texttt{layers.}\ell_f^*\texttt{.mlp.down\_proj.bias}] \leftarrow \mathbf{v}_f^{(\ell_f^*)}$;
  Set `mlp_bias=True`; Save as $\mathcal{M}_f^{\text{steered}}$;
end for

Algorithm 3 Steering Intervention on Test Set

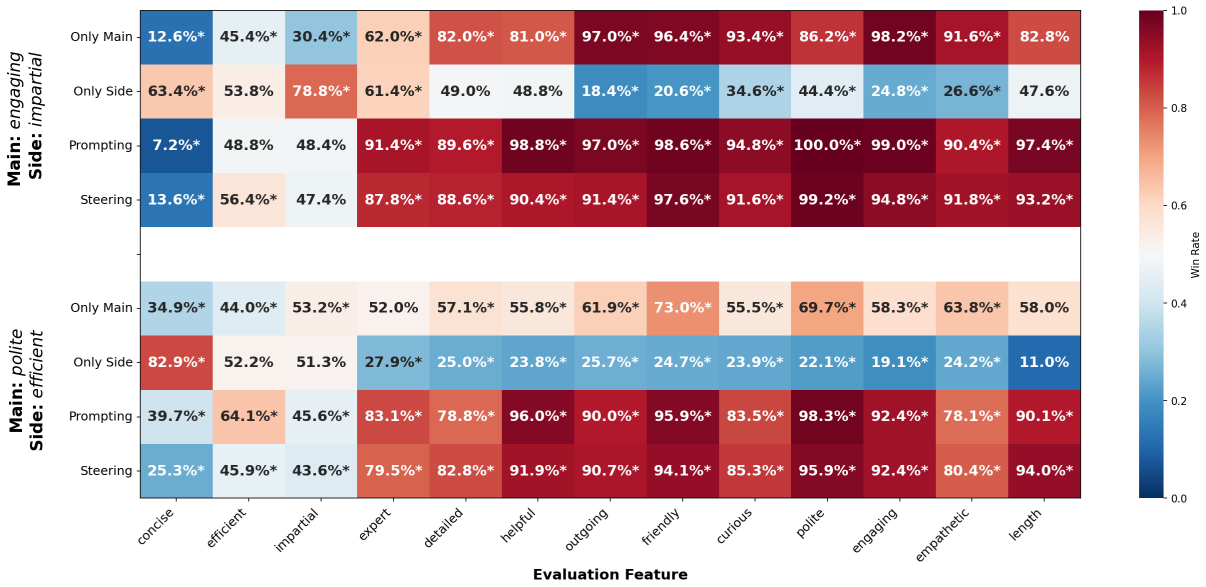
Input: Baked checkpoints $\{\mathcal{M}_f^{\text{steered}}\}$, Test set $\mathcal{D}_{\text{test}}$, Targeted pairs $\mathcal{P}$.
Output: Steered generations $\{y_i^{(\text{steer})}\}$.
for each pair $(f_m, f_s, d_p, p) \in \mathcal{P}$ **do**
  $\mathcal{M}_{f_s}^{\text{steered}} \leftarrow \text{load}(\mathcal{M}\texttt{-steered-CAA-}f_s)$;
  $\mathcal{D}_p \leftarrow \{(q, t, d) \in \mathcal{D}_{\text{test}} : d \in d_p\}$;
  for each $(q_i, t_i, d_i) \in \mathcal{D}_p$ **do**
    for $r = 1$ to 5 **do**
      $s_i \leftarrow$ "You are a helpful assistant having a conversation about t_i.";
      $y_i^{(\text{steer},r)} \leftarrow \mathcal{M}_{f_s}^{\text{steered}}.\text{generate}(s_i, q_i)$; *// Prompt is neutral; style comes from weights*
    end for
  end for
end for

H.2 Phase 2: Layer Selection and Model Baking

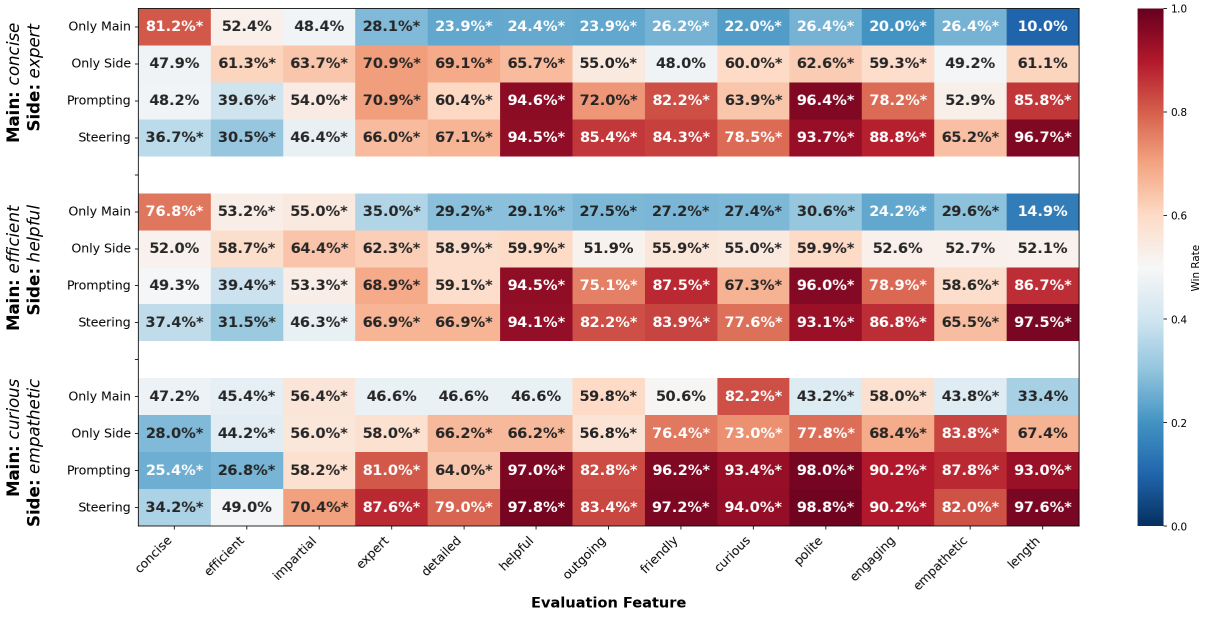
Once raw vectors are computed, we identify the optimal injection layer using the validation set and permanently "bake" the vector into the model weights, through the Algorithm 2.

H.3 Phase 3: Test Set Intervention

In the final Algorithm 3, we load the specific baked checkpoint corresponding to the desired side-effect style (f_s) and generate responses for the test set.



(a) Qwen3-8B Results



(b) Llama3 Task & Daily

Figure 8: Effects of prompt-based and steering-based interventions on main and side evaluation features. The heatmaps report pairwise win rates when applying prompt interventions and activation-space steering interventions to control a target Main Feature, evaluated across the Main Feature, the Side Feature, and a range of style features features.

H.4 Notation and Definitions

Models: $\mathcal{M}$ is the target model (`Llama-3.1-8B-Instruct` or `Qwen3-8B`). $\mathcal{M}_{\text{judge}}$ (`Qwen3-8B`) performs pairwise comparisons.

Steering Vector ($\mathbf{v}_f^{(\ell)}$): A vector in $\mathbb{R}^{d_{\text{model}}}$ representing the direction in activation space that increases feature f at layer ℓ. It is computed as the mean difference between activations on positive (choice A) and negative (choice B) completions.

Contrastive Pairs ($\mathcal{C}_f$): Training examples formatted as A/B choice prompts where:

- Choice A: Styled response (expresses the target feature in style).
- Choice B: Neutral response (neutral model output when not prompted with style feature).

Activation Extraction: Activations are extracted at the second-to-last token position (pos $= -2$) of

Style Feature	Clustered Feature	Freq.	Related Papers
helpful	helpful	19	Jiang et al. (2023), Köpf et al. (2023), Kwan et al. (2024), Lee et al. (2025b), Li et al. (2024b), Li et al. (2023), Liu et al. (2024), Meade et al. (2023), Mishra et al. (2023), Rana et al. (2025), Saha and Srihari (2024), Stureborg et al. (2024), Wang et al. (2023a), Xu et al. (2024a), Xu et al. (2024b), Yang and Ettinger (2023), Zhang et al. (2024a), Zhao et al. (2025a), Zhou et al. (2024)
empathetic	empathetic	15	ALMutairi et al. (2024), Cheng et al. (2025), Feng et al. (2025), Finch et al. (2023), Siyan et al. (2024b), Mishra et al. (2023), Njifenjou et al. (2025), Nozue et al. (2024), Reguera-Gómez et al. (2025), Svikhnushina and Pu (2023), Wang et al. (2023a), Wang et al. (2025), Zhang et al. (2024a), Zhang et al. (2024b), Zhou et al. (2024)
concise	concise	12	Afzal et al. (2024), Agrawal et al. (2024), Chalamalasetti et al. (2023), Feng et al. (2024), Li et al. (2024a), Movva et al. (2024), Siyan et al. (2024a), Rebedea et al. (2024), Taranukhin et al. (2024), Wang et al. (2023b), Wang et al. (2024a), Wang et al. (2025)
friendly	friendly	12	Dou et al. (2023), Lechner et al. (2023), Lee et al. (2023), Liu et al. (2025), Njifenjou et al. (2025), Reddy et al. (2023), Reguera-Gómez et al. (2025), Semnani et al. (2023), Seo and Lee (2024), Tsubota and Kano (2024), Wang et al. (2023a), Zhang et al. (2024a)
detailed	detailed	8	Deng et al. (2024), Jang et al. (2024), Kirstein et al. (2024), Li et al. (2024b), Liu et al. (2024), Sobhani et al. (2025), Rebedea et al. (2024), Zhao et al. (2025a)
expert	expert	8	Ferron et al. (2023), Lee et al. (2025a), Ou et al. (2024), Rachidi et al. (2025), Rana et al. (2025), Saley et al. (2024), Wevelsiep et al. (2025), Yang et al. (2024)
engaging	engaging	7	ALMutairi et al. (2024), Finch and Choi (2025), Jiang et al. (2024), Lee et al. (2023), Phukan et al. (2024), Wang et al. (2023b), Yu et al. (2024)
informative	informative	7	ALMutairi et al. (2024), Jiang et al. (2024), Raposo et al. (2023), Siro et al. (2024), Rebedea et al. (2024), Wang et al. (2023b), Yu et al. (2024)
short	short	7	Brown et al. (2024), Jandaghi et al. (2024), Li et al. (2024a), Reddy et al. (2023), Rebedea et al. (2024), Sun et al. (2024), Väth et al. (2024)
curious	curious	6	Lee et al. (2025b), Murakhovs'ka et al. (2023), Njifenjou et al. (2025), Petrak et al. (2024), Wang et al. (2023a), Wang et al. (2023b)
polite	polite	6	Abdullin et al. (2023), Afzal et al. (2024), Lee et al. (2023), Li et al. (2024b), Liu et al. (2024), Petrak et al. (2024)
caring	caring	5	ALMutairi et al. (2024), Fei et al. (2024), Feng et al. (2025), Wang et al. (2023a), Zhu et al. (2025)
efficient	efficient	5	Lee et al. (2025b), Liu et al. (2025), Wang et al. (2024b), Wang et al. (2023b), Zhu et al. (2025)
impartial	impartial	5	Duan et al. (2025), Madani et al. (2025), Meade et al. (2023), Raju et al. (2024), Zhao et al. (2025a)
outgoing	outgoing	5	He et al. (2025a), Lee et al. (2025b), Liu et al. (2025), Wang et al. (2023a), Wang et al. (2023b)
professional	professional	5	ALMutairi et al. (2024), Kirstein et al. (2024), Oshima et al. (2024), Rachidi et al. (2025), Saley et al. (2024)

Table 3: Overview of the 16 identified style features appearing in at least 5 papers, with references of all papers which contain those features. All terms have been adjectivized and deduplicated for consistency.

each completion, capturing the model's representation just before generating the final decision token (A or B).

Checkpoint Baking: Instead of applying steering at inference time, we permanently add the steering vector $\mathbf{v}_f^{(\ell^*)}$ as a bias term to the MLP `down_proj` layer at the optimal layer ℓ_f^*.

Best Layer Selection: For each feature f, the optimal layer ℓ_f^* is selected by maximizing the win rate of the steered model against the **neutral** response on the validation set: $\ell_f^* = \arg\max_{\ell \in \mathcal{L}} W_f^{(\ell)}$. We do not use the styled ground truth response for validation or comparison.

Targeted Pairs ($\mathcal{P}$): Same format as prompt intervention. For steering with negative polarity, we load the checkpoint steered toward the *side effect* feature f_s (enhancing the negative correlation). No explicit prompt instruction is needed since steering is baked into the model weights.

Candidate Layers: $\mathcal{L} = \{16, 20, 24\}$ (middle-to-late layers typically yield best results). Default baking layer is 20 if validation is skipped.

I Task and Daily Dataset with Initial Message Examples

In this section, we present Table 4 which shows all topics used in this study with initial message examples.

J Prompt Intervention Algorithm: Generation

In this section, we present a detailed Algorithm 4 about using prompting intervention to mitigate side effects of using style features in prompts.

J.1 Generation Algorithm

We present our Prompting Algorithm 4 here with definitions in the following subsection.

Algorithm 4 Prompt Intervention Generation

Input: Test set $\mathcal{D}_{\text{test}} = \{(q_i, t_i, d_i)\}$ where q_i is the query, t_i is the topic, and d_i is the domain. Targeted Pairs $\mathcal{P} = \{(f_m, f_s, d_p, p)\}$. Generator Model $\mathcal{M}_{\text{gen}}$.
Output: Set of generated responses $\mathcal{Y}_{\text{generated}}$
Phase 1: Prompt Intervention Generation
for each pair $(f_m, f_s, d_p, p) \in \mathcal{P}$ **do**
  Define $\pi_{\text{normal/reversed}}$ based on polarity p (e.g., "Please be f_m but f_s");
  Filter dataset $\mathcal{D}_p \leftarrow \{(q, t, d) \in \mathcal{D}_{\text{test}} : d \in d_p\}$;
  for each message $(q_i, t_i, d_i) \in \mathcal{D}_p$ **do**
    for $r = 1$ to 5 **do**
      Construct prompt $s_i \leftarrow$ "You are a helpful assistant having a conversation about t_i. π.";
      Generate response $y_i^{(\pi,r)} \leftarrow \mathcal{M}_{\text{gen}}.\text{generate}(s_i, q_i)$;
      Add $y_i^{(\pi,r)}$ to $\mathcal{Y}_{\text{generated}}$;
    end for
  end for
end for

J.2 Generation Notation

Models: $\mathcal{M}_{\text{gen}}$ (`Llama-3.1-8B-Instruct` or `qwen3-8b`).
Intervention Prompts (π): The specific instruction strings appended to the system prompt to steer the model. For a pair (f_m, f_s) with negative polarity, we define two variations to test ordering effects:

- π_{normal}: *"Please be f_m but f_s."* (Prioritizes Main Feature).
- π_{reversed}: *"Please be f_s but f_m."* (Prioritizes Side Feature).

K Win Rate Calculation Algorithm

In this section, we present a Win Rate Algorithm 5, which shows the how we calculate win rates to decide the Side Feature strengths of using Main Feature in prompts.

Algorithm 5 Pairwise Comparison and Win Rate Calculation

Input: Generated responses $\mathcal{Y}_{\text{generated}}$. Evaluation Features $\mathcal{F}$. Judge Model $\mathcal{M}_{\text{judge}}$.
Output: Win rate matrices $\mathbf{W}^{(f_m, f_s)}$
Phase 1: Pairwise Comparison Rating
for each generated response $y_i^{(\pi,r)} \in \mathcal{Y}_{\text{generated}}$ **do**
  Load baselines: y_i^{neutral} (no style) and $y_i^{(f)}$ (styled);
  for each evaluation feature $f_{\text{eval}} \in \mathcal{F}$ **do**
    $c_{\text{orig}} \leftarrow \text{ComparisonPrompt}(f_{\text{eval}}, y_i^{(\pi,r)}, y_i^{\text{neutral}}, q_i)$;
    $w_{\text{orig}}^{(f_{\text{eval}})} \leftarrow \mathcal{M}_{\text{judge}}.\text{compare}(c_{\text{orig}})$; *// Rate vs Neutral*
    $c_{\text{styled}} \leftarrow \text{ComparisonPrompt}(f_{\text{eval}}, y_i^{(\pi,r)}, y_i^{(f_{\text{eval}})}, q_i)$;
    $w_{\text{styled}}^{(f_{\text{eval}})} \leftarrow \mathcal{M}_{\text{judge}}.\text{compare}(c_{\text{styled}})$; *// Rate vs Styled*
  end for
end for
Phase 2: Win Rate Calculation
for each pair $(f_m, f_s, d_p, p) \in \mathcal{P}$ **do**
  Init matrix $\mathbf{W}^{(f_m, f_s)} \in \mathbb{R}^{4 \times (|\mathcal{F}|+1)}$;
  for each $f_{\text{eval}} \in \mathcal{F} \cup \{\text{length}\}$ **do**
    $\mathbf{W}_{1,f_{\text{eval}}} \leftarrow n(\text{Intervention} > \text{Neutral})/n_{\text{total}}$;
    $\mathbf{W}_{2,f_{\text{eval}}} \leftarrow \text{StyledWinRate}(f_m, f_{\text{eval}})$; *// Pre-computed*
    $\mathbf{W}_{3,f_{\text{eval}}} \leftarrow \text{StyledWinRate}(f_s, f_{\text{eval}})$; *// Pre-computed*
    $\mathbf{W}_{4,f_{\text{eval}}} \leftarrow \text{WinRate}(\pi_{\text{reversed}} > \text{Neutral})$;
  end for
  Apply BinomialTest (significance $p \leq 0.05$) and Save Heatmap;
end for

K.1 Evaluation Algorithm

We present our win rate algorithm 5 used for evaluation here.

K.2 Evaluation Definitions

Comparison Prompt: The judge is asked: *"Which response is more f? Answer with ONLY A or B."* Order is randomized to mitigate position bias.
Win Rate: Calculated as $\frac{|\{i: w_i = \text{"intervened"}\}|}{|\{i: w_i \neq \text{"unknown"}\}|}$. Statistical significance is assessed using a two-sided binomial test ($p_0 = 0.5$).

L Insights from Style Feature Extraction

The systematic extraction of style features reveals a distinct bifurcation in how "style" is conceptualized in current research: it is predominantly treated either as a *functional constraint* (e.g., *concise*, *detailed*) or a *psychological simulation* (e.g., *empathetic*, *extraverted*). The data suggests that while the capability for complex persona engineering exists, the majority of research defaults to minimal, service-oriented prompt structures.

Domain	Topic	Initial Message Examples
LMSYS-Chat-1M (Task)	Answering questions based on passages	Who was the first man on the moon?
	Discussing software errors and solutions	Describe and explain the benefits of OpenBSD
	Inquiries about specific plant growth conditions	how do I get rid of mosquitos?
	Requesting introductions for various chemical companies	Explain "reductive amination".
	Inquiries about AI tools, software design, and programming	write python function to reverse a sentence
	Text processing (Merged)	sup peeps. Wanna help me with summarizing lyrics?
	Role-playing scenarios and character interactions (Merged)	Write me an extremely funny tale about pillow hoarding ducks at a luxurious hotel.
	Geography, travel, and global cultural inquiries	explain why we have traffic lights?
	Discussing and describing various characters	Tell me something about Stephen King's Dark Tower.
	Creating and improving business strategies and products	how to make a game project success
DailyDialog (Daily)	Tourism	That is the most beautiful sunset !
	Work	Hey , Zina . You're here early today .
	Politics	Every country should face the history .
	Finance	It's all over . I'm bankrupt .
	Health	Are you feeling better today , Bill ?
	School Life	What can I help you with today ?
	Attitude & Emotion	Do you hear what happened to Sally ?
	Ordinary Life	Excuse me . Is this seat taken ?
	Culture & Education	Harry , do you like the opera ?
	Relationship	I wonder how Sarah and Mat are .

Table 4: Example First Messages by Domain and Topic (10 Topics Per Domain). We curate and merge some task topics, specificaly the *Text processing* and *role-play scenarios and character interactions*. The Daily topics are extracted from the DailyDialog dataset without modifications.

L.1 The Dominance of Service-Oriented and Utility Traits

As illustrated in Figure 5, the frequency distribution indicates that the primary goal of contemporary conversational agents remains functional utility and emotional safety rather than complex character enactment.

Benevolence and Safety: The two most frequently extracted features are *helpful* ($N = 19$) and *empathetic* ($N = 15$). This alignment correlates with the prevalence of papers focused on general-purpose assistants and therapeutic dialogue systems, where prompts are engineered to ensure the agent provides appropriate emotional value and adheres to safety guidelines.

Utility vs. Engagement: There is a notable tie between the third and fourth most common features, *friendly* ($N = 12$) and *concise* ($N = 12$). This highlights a split in research objectives: "friendly" serves the goal of open-ended user engagement (human-likeness), while "concise" serves the goal of efficient information retrieval (token minimization).

L.2 Standardization vs. The "Long Tail" of Definition

The distribution data reveals a lack of standardized terminology for defining agent personality, characterized by a heavy reliance on ad-hoc descriptors.

The "Expert" Persona: Features such as *expert* ($N = 8$) and *detailed* ($N = 8$) appear in the mid-range of frequency. These are predominantly used in domain-specific research (e.g., medical or legal synthesis) to act as validity markers, ensuring the agent sounds like a credible authority rather than a casual chatbot.

Fragmentation of Terminology: The visualization displays a long tail of psychological and tonal descriptors that plateau at low frequencies ($N = 4$). Traits such as *creative*, *humorous*, *extraverted*, and *thoughtful* appear significantly less often than utility prompts. This suggests that researchers often employ colloquial adjectives to steer models for specific narrow tasks rather than adhering to established psychological frameworks (such as the Big Five) for consistent persona definition.

Redundancy in Prompting: The data also exposes terminological redundancy. For example, while *concise* is a dominant feature ($N = 12$), synonymous constraints like *short* ($N = 7$) and *brief* ($N = 4$) appear separately. This fragmentation indicates that "style" is often defined by the individual researcher's vocabulary preference rather than a standardized taxonomy of prompt engineering.

L.3 Experimental Scope: Minimalist vs. Maximalist Designs

Beyond the frequency of specific terms, we analyzed the *diversity* of style features employed within individual papers to understand the depth of experimental design. Figure 9 presents the distribution of papers based on the count of unique style features they utilized, revealing a strongly right-skewed trend.

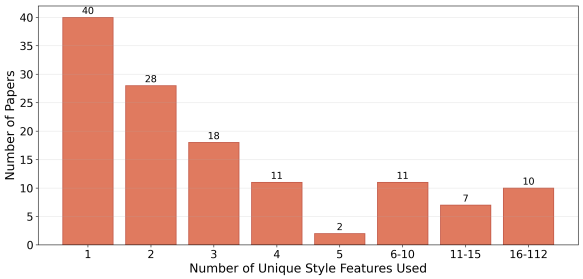


Figure 9: Distribution of papers categorized by the number of unique style features used. The distribution highlights a polarization between minimalist designs (single-feature) and broad-spectrum evaluations (16+ features).

Prevalence of Minimalist Designs: The largest cohort of papers ($N = 40$) utilizes only a single unique style feature throughout the entire study. Combined with those using two or three features, the majority of research ($N = 86$) uses a very limited number of features. This confirms that for most contemporary work, style is treated as a static control variable—likely a global instruction to be "helpful" or "safe"—rather than a dynamic parameter for experimentation.

The "Benchmarking" Tail: In sharp contrast to the minimalist majority, a distinct subset of 10 papers employs a massive range of features (16 to 112 unique terms). This heavy tail represents a fundamental divergence in research intent: these outlier studies treat style as the primary independent variable. These likely correspond to large-scale persona benchmarks or synthetic dataset generation efforts that require maximizing behavioral diversity.

M Hierarchical Clusterings of Style Features

In this section, we present Figure 10 which shows how we decide which style features out of the filtered 16 style features have similar meanings

N Ablation: Reverse the Order of Main Feature and Side Feature in Prompts

This section presents an ablation study of switching the order of Main Feature and Side Feature in prompts. We want to see if models pay more attention to the latter prompt word used.

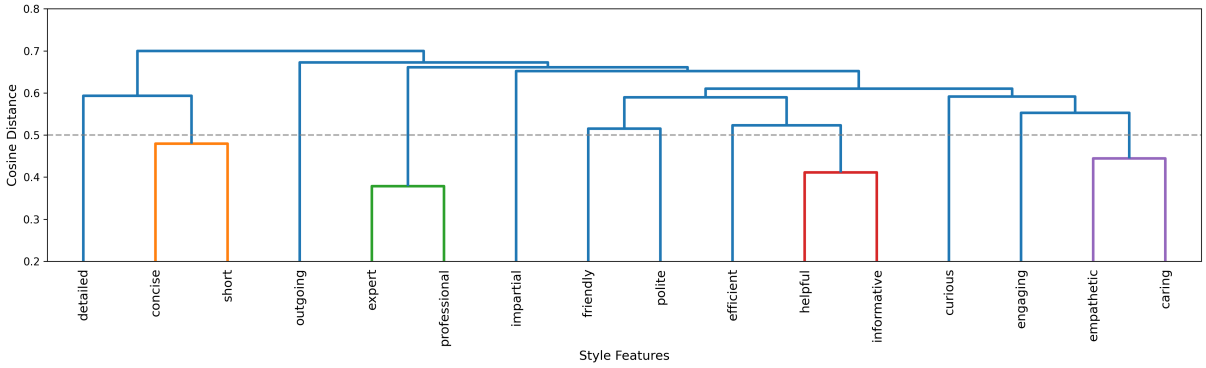
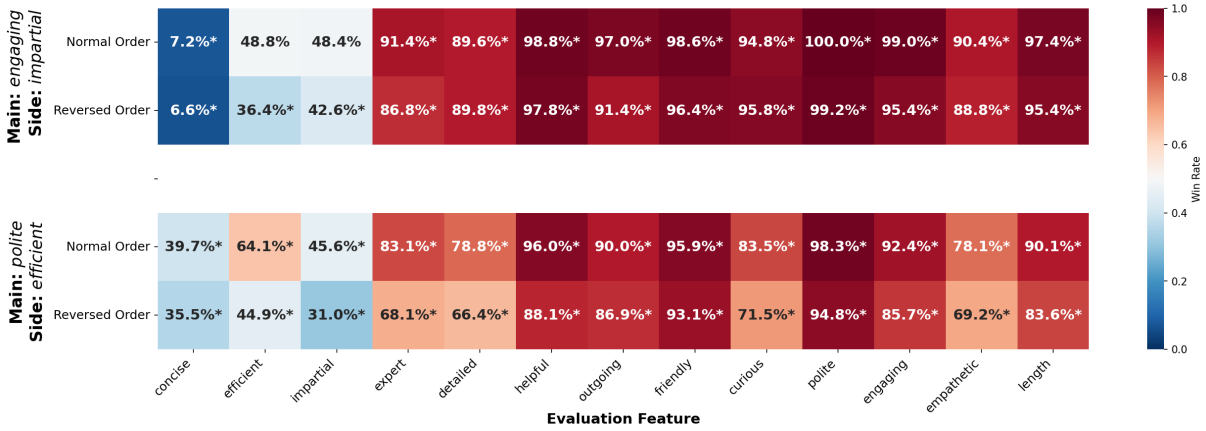
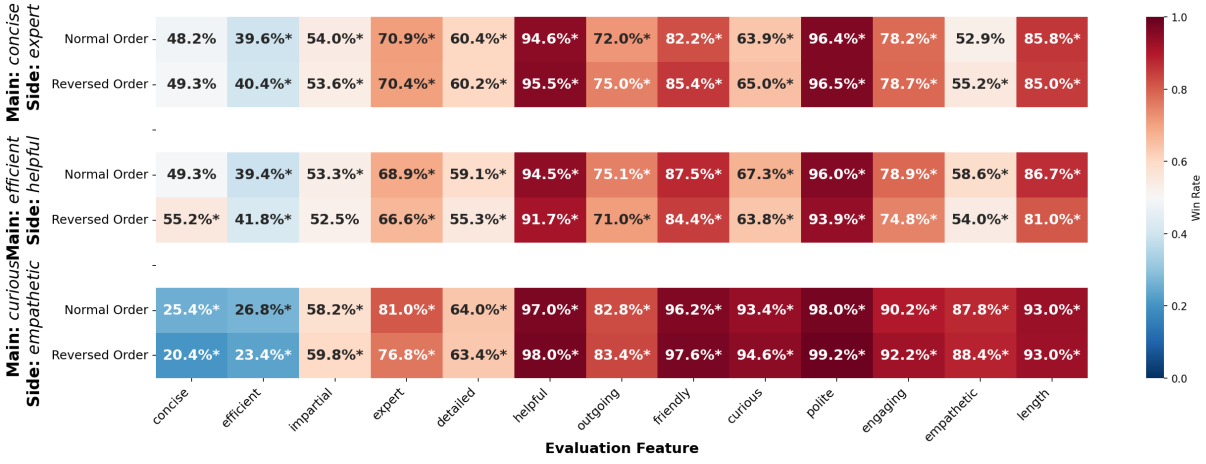


Figure 10: Hierarchical clustering of style features based on semantic similarity. Features are embedded using OpenAI's `text-embedding-3-small` model and clustered via average linkage on cosine distances. Semantically related features cluster together, like (concise, short), (expert, professional), (helpful, informative), and (empathetic, caring).



(a) `Qwen3-8B` Results



(b) `Llama3` Task & Daily

Figure 11: Effects of reversing order of Main Feature and Side Feature in Prompting Intervention. The heatmaps report pairwise win rates when applying prompt interventions in different orders of Main Feature and Side Feature.