When Your Friendly Agent Is More Than Just Friendly: Revealing Side Effects of Using Style Features on LLM

Anonymous ACL submission

Abstract

Many recent studies use style or persona features, such as "empathetic" or "professional", to steer agents' behaviors toward desired styles. However, the unintended stylistic side effects these features introduce into model behavior remain underexplored. This paper identifies and controlls side effects across commonly used style features, revealing significant cross-feature interference. We conducted a comprehensive survey of recent papers to extract widely used style features, and performed empirical analysis using synthetic agent-agent dialogues. Our findings show that many features show strong correlations with others and that their influence can bleed into unrelated traits. We further design and evaluate counter-strategies to neutralize these effects. Our work show the existence of side effects and questions the LLM's faithfulness in following the prompt, offering practical recommendations for safe and targeted style control in LLM-based agents.

1 Introduction

The growing use of style features like *empathetic*, *professional*, and *friendly* has become a standard prompt-based technique for steering conversational agents' tone (Feng et al., 2025; Zhao et al., 2025; Rachidi et al., 2025). A survey of 2023–2025 *CL* papers shows these prompts (e.g., "Please be helpful") are guiding systems from mental-health assistants to task-oriented bots. (Njifenjou et al., 2025; Lee et al., 2025; Lechner et al., 2023; Lee et al., 2023; Wang et al., 2023). Yet style cues also trigger *side effects*: asking a model to be "empathetic" may heighten supportive language while eroding logical precision. In domains where reliability is paramount, such as legal or collaborative decision-making, the assumption that a style prompt alters only its target trait is untenable, underscoring the need for systematic study of downstream impacts.

Most work on prompt side effects emphasizes accuracy loss and emergent bias rather than stylistic drift (Luz de Araujo and Roth, 2025; Gupta et al., 2024). Zheng et al. (2024) tested 162 persona prompts across 2,410 factual questions on four open-source LLMs and found no accuracy gains, while Lutz et al. (2025) showed that such prompts can amplify stereotypes toward marginalized groups.

Studies that do examine style transitions largely focus on demographic personas rather than generic style adjectives (Malik et al., 2024; Lutz et al., 2025). Malik et al. (2024) reported distinct writing shifts across age, profession, and political personas . Multi-turn analyses also can lead rapid persona drift in dialogue (Kovač et al., 2024; Li et al., 2024). Additionally, these publications focus on end-to-end evaluations, rather than statistical analysis with more explainability. Thus, a comprehensive statistical investigation of the stylistic side effects of common style features remains absent.

In this work, we address this gap by identifying and quantifying the cross-feature effects and trade-offs associated with popular style features used in LLM-driven conversational agents. We first curate and analyze all conversational agent-related papers from the ACL anthology from 2023 to 2025 that use prompt-based conversational agent design, extracting and clustering 12 commonly used style features. We then generate synthetic conversations guided by controlled conversation agent prompts and use LLM-as-a-Judge to measure the degree to which each style feature affects not only its intended trait but also others, both positively and negatively.

Our findings indicate that *style features do not act independently*, but instead exhibit structured interference patterns: applying one feature often impacts several others and has unexpected impacts. To address this, we use simplest prompting strategy to evaluate the controllability of the side effect. Our

contributions are threefold:

- **We conduct a comprehensive survey** of style feature usage across 588 ACL Anthology papers and identify frequently used features.
- **We demonstrate empirically** that style features produce measurable side effects across unrelated traits, confirming the presence of cross-feature behavioral entanglement.
- **Our mitigation results show that** simple prompt-based methods are not powerful enough to reduce side effects while maintaining main effects.

2 Feature Extraction - A Survey

To ground our study in contemporary practice, we performed a systematic survey of conversational-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 shown in Figure 1.

2.1 Method

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 two authors annotated all style features used and mentioned in the papers. Next, we transformed these features to adjectives, to get a list of unique features. The distribution of extracted features is shown in Figure E. Next, we used frequency >=5 to get the most frequent list of features as candidates, then we grouped hierarchical clustering based on the cosine similarity >0.5 with embeddings obtained with text-embedding-3-small (OpenAI, 2024). As the result, we extracted 12 distinct style features that typify how recent papers employ prompt-based conversational agent control.

2.2 Result

Our pipeline extracts 12 high-frequency style features: *concise*, *expert*, *helpful*, *empathetic*, *friendly*, *detailed*, *engaging*, *curious*, *polite*, *impartial*, *outgoing*, *efficient*. Out of these features, *helpful*, *empathetic*, and *friendly* are the most frequently used. Left skewed distribution shows style feature usages are concentrated into a few terms

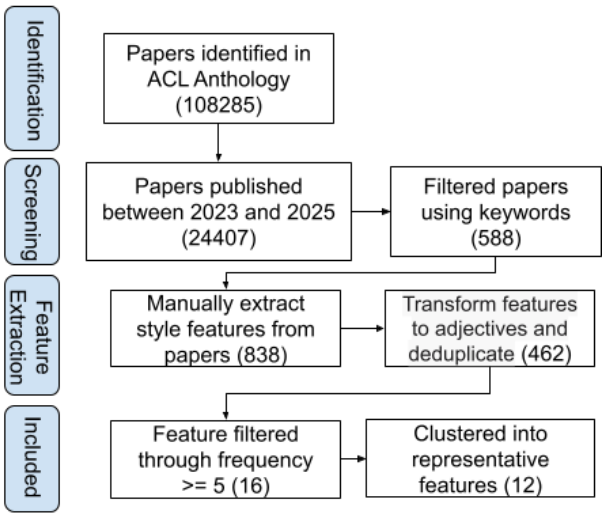

Figure 1: Data collection for papers and style features

3 Identifying Side Effects

In this section, we systematically explore the side effects of prompting with each style feature using a behavioral difference matrix that's constructed using ratings on synthetic conversations that simulate common prompt usage in daily conversations and task-oriented conversations.

3.1 Method+Experiment

To construct synthetic conversations for statistical analysis, we sampled dialogue topics from two datasets: LMSYS-Chat-1M (Zheng et al., 2023) and DailyDialog (Li et al., 2017). From each dataset, we extracted 10 distinct topics—task-oriented topics from LMSYS-Chat-1M and daily-life topics from DailyDialog. For each topic, we selected 10 representative opening messages in English based on the first user turn.

We then generated conversations using pairs of LLM-powered agents: a *user* agent and an *assistant* agent. Both agents were initialized with a shared system prompt specifying the topic. In addition, the assistant agent received an augmented system prompt containing a style feature from one of the 12 extracted through the pipeline in Figure 1. See Appendix C and D for prompt templates.

Each conversation consisted of three turns, initiated by the user agent with a preselected first message. We retained only the assistant agent's responses for downstream analysis. Each response was rated on a scale of 1 to 5 across 12 stylistic features using an LLM-as-a-judge framework (see Appendix A for a template). This setup enables controlled and scalable measurement of how specific

style prompts influence assistant behavior across diverse conversational contexts.

3.2 Result

With rated responses, we construct a Figure 2 to show correlation between features. Figure 2 is a "prompt→rating" heatmap. Each row corresponds to the feature used in the prompt. Each column shows how the resulting responses were rated on that same set of features. The colour scale encodes the mean change in the rated score relative to an baseline model response without style features in its prompts. Asterisks mark statistically significant differences ($p < 0.05$).

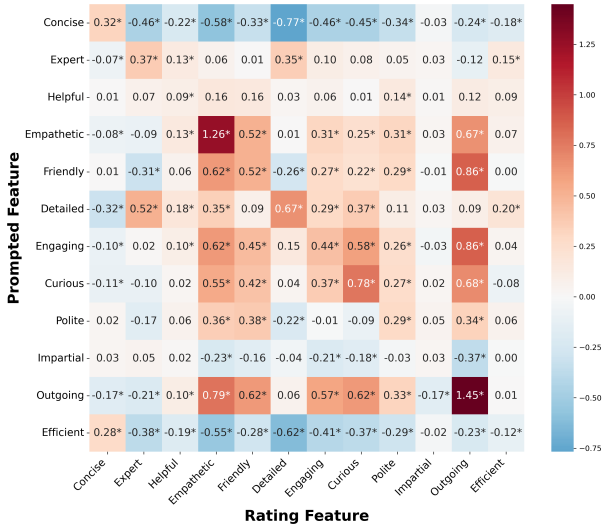

Figure 2: **Style Feature Correlation Matrix.** The y-axis shows prompted style features, and the x-axis shows features rated by LLM-as-a-Judge. Each cell is the average rating for the column feature when prompted with the row feature. Red indicates positive alignment; blue indicates negative correlation. Asterisks (*) denote statistical significance.

4 Counter Side Effects

While style prompting succeeds at amplifying its target trait, it simultaneously increases unexpected side effect, questioning models' ability to follow prompts faithfully. Therefore, the goal of our Counter Side Effect (CSE) Experiments is: to test whether simple prompt engineering interventions can mitigate these unexpected spillovers.

4.1 Method and Experiment

Balanced-prompt strategy: For each of the 12 primary style features, we construct a *balanced prompt* that pairs the original instruction with its most positively correlated side effect identified in Section 3.2 (See Appendix B for a template).

Conversations are generated with topic and messages from Section 3.1, with same hyperparameters.

Evaluation protocol: We collect the assistant's replies and rate them with the same LLM-as-judge rubric on all 12 features (1–5 scale). We report (i) **Change in main feature strength**, and (ii) **Change in side effect feature strength**.

4.2 Result

Table 1 shows the results for the Counter Side Effect Experiments. The table shows, for each main style feature (orange header) and its strongest side-effect feature (purple header beneath), how ratings change under three conditions. "Original" is the baseline with no style prompt; "Main" applies the main prompt alone; "Main ¬Side" adds our counter-prompt aimed at reducing the side effect. The upper block ("Main Feature Ratings") reports scores on the main trait, while the lower block ("Side Effect Ratings") reports the side-effect scores. Positive or negative deltas indicate how much each setting raises or lowers the respective rating relative to baseline; asterisks denote significance ($p < 0.05$).

5 Discussion

5.1 Analyzing Side Effects

The correlation matrix reveals two opposing yet interlocked style dimensions: an *informational* axis, anchored by brevity, neutrality, and analytical depth, and a *social* axis, defined by empathy, friendliness, and engagement. Steering the model decisively along one axis predictably pulls it back on the other, with bidirectional penalties of roughly 0.3–0.8 rating points. Task-utility prompts such as *Concise*, *Efficient*, and especially *Impartial* dampen social warmth, whereas persona-oriented cues like *Empathetic*, *Friendly*, and *Engaging* erode informational economy. For instance, "Impartial" boosts its own rating by +0.03 but cuts down "Outgoing" (-0.37*) and "Empathetic" (-0.23*). The reverse holds for persona cues: "Empathetic" raises empathy (+1.26*) yet drags "Concise" down (-0.08*). "Friendly" follows the same pattern (+0.52* on itself, -0.31* on "Expert").

The trade-off, however, is not absolute: *Detailed* simultaneously elevates perceptions of expertise (+0.52*), and *Engaging* bolsters outgoingness

Ratings	Used Features	Polite Friendly	Friendly Polite	Helpful Efficient	Impartial Polite	Efficient Helpful	Curious Engaging	Engaging Helpful	Detailed Expert	Empathetic Friendly	Expert Detailed	Outgoing Friendly
Main	Original	4.60	4.30	4.76	4.92	4.61	3.72	4.22	3.92	2.96	3.97	2.95
Features	Main	+0.29*	+0.52*	+0.10*	+0.03	-0.13	+0.78*	+0.44*	+0.67*	+1.26	+0.38*	+1.45*
Ratings	Main, ¬Side	+0.12*	+0.35*	+0.07	+0.02	-0.04	+0.47*	+0.28*	+0.58*	+0.81*	+0.18*	+0.86*
Side	Original	4.30	4.60	4.61	4.60	4.76	4.22	4.76	3.97	4.30	3.92	4.30
Effects	Main	+0.38*	+0.29	+0.09	-0.03	-0.19*	+0.37*	+0.10*	+0.53*	+0.52*	+0.35*	+0.63
Ratings	Main, ¬Side	+0.21*	+0.16*	+0.06	+0.00	-0.06	+0.19*	+0.04	+0.43*	+0.38*	+0.12	+0.45*

Table 1: The Counter Side Effects Experiemtal Results. The first feature of each column is the main style feature, and the second feature in second line is the side effect feature whose strength this experiment attempts to reduce.

(+0.86*), suggesting that certain stylistic clusters co-activate rather than conflict. In practice, these findings underscore that single-feature prompting is intrinsically entangled; developers must therefore balance predictable side effects when optimising LLM personas for both competence and sociability.

5.2 Counter Side Effect Experiments

Attempts to neutralize the trade-offs with lightweight "counter prompts" underscore the limits of prompt engineering. While composite instructions—such as pairing a *Friendly* cue with *Polite* (0.29 down to 0.16*) or adding *Detailed* to *Expert*—do soften the most severe penalties (+0.35* down to +0.12*), they rarely eliminate them and often introduce new, unanticipated shifts elsewhere in the style grid. In several cases the secondary prompt even dilute the desired primary effect, trimming up to 25% of the original gain, while failing to restore more than half of the lost ground on the targeted counter-dimension. These outcomes suggest that simple prompt concatenation cannot disentangle the strongly coupled style axes uncovered earlier; a more principled approach—such as iterative reinforcement learning, targeted fine-tuning, or multi-objective optimization—will be required to balance competence and warmth without collateral drift.

6 Conclusion

This study provides the first large-scale, quantitative look at how single-style prompts reshape LLM behaviour across thousands of synthetic dialogues. By rating assistant responses on a twelve-dimensional rubric, we uncover a robust, bidirectional trade-off between *informational* and *persona* traits: instructions that enhance brevity, precision, or impartiality predictably erode social warmth, while empathy- or friendliness-oriented prompts suppress task utility. Our counter-prompt experiment further reveals the limits of lightweight mitigation: our results show that simple prompt-based methods are not powerful enough to reduce side effects while simultaneously preserving the desired primary effect, and they may themselves introduce new, unintended behaviours. Together, these findings chart the latent structure of stylistic control in LLMs and underscore the need for richer mitigation strategies—such as adaptive prompting or targeted fine-tuning—to balance competence and likeability in real-world conversational agents. Future work will extend evaluation to human judges in open-domain deployments and explore model-level approaches for disentangling stylistic axes.

Limitation

A key limitation of our study is ecological validity: all findings stem from short, synthetic agent-agent exchanges rated by another LLM, 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 evaluate only twelve high-frequency English-language style features on a single base model family, due to limitations in compute resources; less common cues, other languages, and model architectures could yield different interference patterns. Finally, our mitigation test uses simple prompt concatenation, 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

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.

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.

Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2024. Stick to your role! stability of personal values expressed in large language models. *PLOS ONE*, 19(8):e0309114.

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.

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. 2025. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *Preprint*, arXiv:2406.14703.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling instruction (in)stability in language model dialogs. *Preprint*, arXiv:2402.10962.

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.

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.

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.

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.

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

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.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. 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.

Weiliang Zhao, Daniel Ben-Levi, Wei Hao, Junfeng Yang, and Chengzhi Mao. 2025. 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.

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.

A LLM-as-a-Judge Prompt Template

"Given a message, rate {feature} from 1 (not {feature}) to 5 (extreme {feature})"
where the "{feature}" refers to one of the 12 features.

B Counter Side Effect Experiments Prompt Template

"This is a conversation about {topic}. Please be {main feature} and {most correlated side effect feature}."

C System Template for User Agent

: "This is a conversation about {topic}."

D System Template for Assistant Agent

: "This is a conversation about {topic}. Please be {style_features}"

E Extracted Feature Distribution

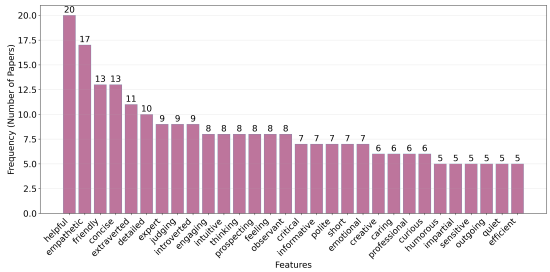

Figure 3: Extracted Features' Frequency Distributions. Only list top 30 most frequent features here for the sake of space