How Stylistic Similarity Shapes Preferences in Dialogue Dataset with User and Third Party Evaluations

Ikumi Numaya[1], Shoji Moriya[1], Shiki Sato[1,2], Reina Akama[1,3,4], Jun Suzuki[1,4]

[1]Tohoku University, [2]CyberAgent, [3]NINJAL, [4]RIKEN

{numaya.ikumi.t4, shoji.moriya.q7}@dc.tohoku.ac.jp,

sato_shiki@cyberagent.co.jp, {akama, jun.suzuki}@tohoku.ac.jp

Abstract

Recent advancements in dialogue generation have broadened the scope of human–bot interactions, enabling not only contextually appropriate responses but also the analysis of human affect and sensitivity. While prior work has suggested that stylistic similarity between user and system may enhance user impressions, the distinction between subjective and objective similarity is often overlooked. To investigate this issue, we introduce a novel dataset that includes users' preferences, subjective stylistic similarity based on users' own perceptions, and objective stylistic similarity annotated by third party evaluators in open-domain dialogue settings. Analysis using the constructed dataset reveals a strong positive correlation between subjective stylistic similarity and user preference. Furthermore, our analysis suggests an important finding: users' subjective stylistic similarity differs from third party objective similarity. This underscores the importance of distinguishing between subjective and objective evaluations and understanding the distinct aspects each captures when analyzing the relationship between stylistic similarity and user preferences. The dataset presented in this paper is available online.[1]

1 Introduction

Recent advances in dialogue generation have markedly improved the ability of systems to produce appropriate and natural responses (OpenAI, 2024; Shuster et al., 2022; Zhang et al., 2020). Building on this progress, personalization has become a central focus in dialogue system research, wherein system outputs are tailored to individual user preferences (Tsuta et al., 2023; Du et al., 2024; Chen et al., 2024; Firdaus et al., 2021). Unlike standardized evaluations such as appropriateness, individual user preferences are inherently difficult to generalize, highlighting the need for system design that accounts for user-specific characteristics.

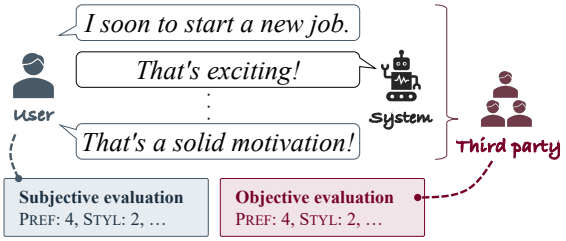


Figure 1: Overview of constructing our dialogue dataset with users' subjective evaluations and objective evaluations averaged over three third party annotators.

Prior studies suggest that, beyond semantic content, stylistic similarity–the extent to which a system's utterances resemble a user's speaking style–may contribute to more positive impressions toward dialogue systems (Kanezaki et al., 2024; Nenkova et al., 2008). However, a critical question is often overlooked : whose perception of stylistic similarity matters, and how it affects users' impressions. While prior studies often rely on third party annotators to judge stylistic similarity and user engagement, it remains unclear whether users' own perceptions of stylistic similarity or third party evaluations of it accurately capture users' preferences.

To investigate this question, we constructed a novel dataset incorporating users' preferences, subjective stylistic similarity based on users' perception, and objective ratings provided by third party annotators. An overview of the dataset construction is shown in Figure 1. The dataset was based on two widely studied open-domain settings: EmpatheticDialogues (Rashkin et al., 2019), focusing on empathetic communication, and Wizard of Wikipedia (Dinan et al., 2018), involving knowledge-grounded conversation. In open-domain dialogues, the primary goal is not explicit task completion but the cultivation of positive

[1]https://github.com/ikuminumaya/duo-dataset

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 192–205
Aug 25–27, 2025.

user impressions through conversational engagement (Yamashita et al., 2023; Qian et al., 2023; Li et al., 2024; Zhang et al., 2018). Focusing on this setting, we use the newly constructed dataset to analyze the factors that influence users' subjective preferences.

Correlations analysis of the subjective evaluations revealed a strong positive relationship between subjective stylistic similarity and user dialogue preference, indicating that stylistic alignment plays a key role in shaping positive user impressions. In contrast, the correlation between objective stylistic similarity and user dialogue preference was weak, and no clear relationship emerged between subjective and objective stylistic similarity. These findings underscore the importance of distinguishing between subjective and objective stylistic similarity in analyzing dialogue quality and user preferences.

The key contributions of this study are:

- We develop a multi-turn, open-domain dialogue dataset that includes both subjective evaluations by users and objective evaluations by third party annotators, covering dialogue preference and stylistic similarity.
- We empirically demonstrate a strong association between user-perceived stylistic similarity and users' subjective preferences.
- We identify a discrepancy between subjective and objective stylistic similarity, emphasizing the need to consider these perspectives separately in dialogue evaluation.

2 Related Work

Open-domain dialogue. Advancements in response generation have greatly enhanced the capacity of dialogue systems to engage in coherent, contextually appropriate multi-turn interactions (OpenAI, 2024; Shuster et al., 2022; Zhang et al., 2020). In open-domain dialogue research, the primary objective is not task completion but the facilitation of user-centered, engaging communication. The Blended Skill Talk (BST) framework (Smith et al., 2020) was introduced to evaluate dialogue systems that can blend multiple conversational skills in open-domain settings. The framework includes personalized dialogue based on profiles (Zhang et al., 2018; Yamashita et al., 2023; Sugiyama et al., 2023), emotion-aware responses that reflect users' affective states (Rashkin et al., 2019; Qian et al., 2023; Wang et al., 2025), and knowledge-grounded dialogue designed to support in-depth topic exploration (Dinan et al., 2018; Li et al., 2024; Zhang and Yongmei, 2024). These settings are widely used in multi-turn response generation studies. In such interactions, systems are implicitly expected to maintain appropriate communication and foster user preferences.

Effects of entrainment on speakers. Entrainment, which is also referred to as accommodation or convergence, describes the gradual alignment of linguistic and prosodic features (e.g., lexical choices, response timing, and vocal pitch) between interlocutors during dialogue. Within the framework of Communication Accommodation Theory (CAT), Giles et al. (2007) presented concrete examples of entrainment in human–human interaction, emphasizing its positive social effects. Prior research has shown that entrainment can increase perceived attractiveness (Street et al., 1983), foster interpersonal involvement (LaFrance, 1979), and strengthen relational bonds (Imamura et al., 2011). Beyond these social outcomes, Nenkova et al. (2008) introduced a method for quantifying entrainment in dialogue by analyzing word frequency patterns, demonstrating its role in enhancing perceived naturalness and task success in human dialogue. Building on these findings, recent research has investigated the incorporation of entrainment into dialogue systems to enhance user experience (Kanezaki et al., 2024). Aligned with this line of research, our study explores how entrainment between users and systems influences users' subjective evaluations.

Style definitions. Text style refers to the manner in which information is conveyed beyond its literal content (Ostheimer et al., 2024; Kawano et al., 2023). Prior research has examined a wide range of stylistic dimensions (Lai et al., 2024), including formality (formal vs. informal) (Rao and Tetreault, 2018; Liu et al., 2022), toxicity (toxic vs. nontoxic) (Logacheva et al., 2022; Pour et al., 2023), political ideology (liberal vs. conservative) (Voigt et al., 2018), politeness (impolite vs. polite) (Madaan et al., 2020; Mukherjee et al., 2023), authorship style (e.g., Early Modern English vs. contemporary English) (Xu et al., 2012; Dinu and Uban, 2023), and sentiment (positive vs. negative) (Shen et al., 2017).

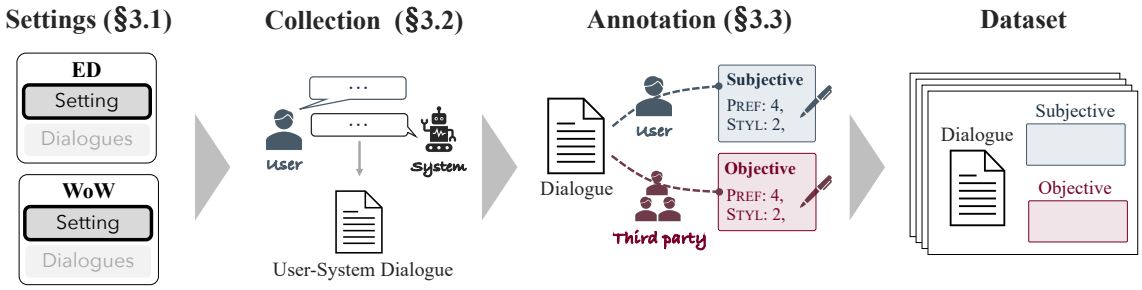


Figure 2: Overview of the entire dataset construction process.

Methods for quantifying stylistic entrainment. As mentioned above, entrainment in dialogue may influence impressions of the partner. In the case of stylistic entrainment, the diversity of relevant features makes consistent and automatic evaluation important. Among the various aspects of stylistic entrainment, Nasir et al. (2019) proposed Local Interpersonal Distance (LID), a metric for quantifying lexical entrainment–specifically, the similarity in word choice and conveyed meaning between interlocutors. LID employs Word Mover's Distance (WMD) (Kusner et al., 2015; Mir et al., 2019; Song et al., 2021) to measure the distance between word embeddings of participant utterances. Building on work, Kanezaki et al. (2024) extended LID by incorporating BERTScore, allowing for more semantically sensitive quantification of lexical alignment. In a related effort, style-sensitive word vectors have been proposed to capture lexical stylistic features (Akama et al., 2018), and sentence-level embeddings have also been explored by fine-tuning BERT (Zenimoto et al., 2023). Moreover, LLM-based evaluation methods, such as LLM-Eval (Lin and Chen, 2023), have recently gained traction as a mainstream approach for assessing dialogue aspects like stylistic similarity.

3 Dataset Construction

This study introduces an open-domain dialogue dataset designed to support subjective and objective evaluations. To facilitate the analysis of the relationship between dialogue preference and stylistic similarity, data were collected under two open-domain settings, followed by annotation from both users and third party evaluators. We named this dataset DUO (Dialogue dataset with User subjective and Objective evaluations). Figure 2 summarizes the overall dataset construction process described in the remainder of this section.

Setting	EmpatheticDialogues
Emotion **Episode**	Excited I'm moving out west, I'm excited about it.
Dialogue	**User**: *I'm going to be making a big move in 2 weeks.* **System**: *That's exciting! How are you feeling about it?* ...
Ratings	PREF.$_{sb}$: 1, CONS.$_{sb}$: 5, STYL.$_{sb}$: 1, EMP.$_{sb}$: 4
	PREF.$_{ob}$: 3.67, CONS.$_{ob}$: 4.33, STYL.$_{ob}$: 3.67, EMP.$_{ob}$: 4.00

Table 1: An example dialogue segment from the EmpatheticDialogues setting. The user selects an emotion label and writes a short personal episode which serves as the context for starting the dialogue. In the dataset, the ratings include both subjective and objective evaluations.

3.1 Dialogue Settings

We adopted two widely used open-domain dialogue settings from the Blended Skill Talk (BST) framework (Smith et al., 2020): EmpatheticDialogues (Rashkin et al., 2019) and Wizard of Wikipedia (Dinan et al., 2018). These complementary settings enable the examination of stylistic entrainment across both emotional and informational dialogue types. Representative examples of the collected dialogues for each setting are shown in Tables 1 and 2.

EmpatheticDialogues (ED). In the ED setting, the worker assumes the role of the speaker, sharing a personal emotional experience, while the dialogue system acts as the role of the empathetic listener. Workers are presented with five randomly selected emotion labels from the set provided in the original ED framework (Rashkin et al., 2019), and asked to choose one (e.g., Excited in Table 1). They are then instructed to briefly describe an emotional experience (1–3 sentences) that corresponds

Setting	Wizard of Wikipedia
Topic	Piano
Dialogue	**System**: *The piano, invented by Bartolomeo Cristofori around 1700, is a fascinating instrument where the strings are struck by hammers controlled by a keyboard.* **User**: *Was there anything similar to the piano before this time?* ...
Ratings	PREF.$_{sb}$: 4, CONS.$_{sb}$: 4, STYL.$_{sb}$: 4, ENGAG.$_{sb}$: 3
	PREF.$_{ob}$: 4.00, CONS.$_{ob}$: 5.00, STYL.$_{ob}$: 2.67, ENGAG.$_{ob}$: 4.00

Table 2: An example dialogue segment from the Wizard of Wikipedia setting. Before the dialogue begins, a topic is presented to both participants. The system initiates the dialogue based on this given topic. In the dataset, the ratings include both subjective and objective evaluations.

to the selected emotion (shown under Episode). The dialogue begins with the worker's utterance describing the emotional experience, followed by system responses designed to demonstrate understanding and emotional resonance. We use this setup to assess whether the system effectively captures the user's emotional state and responds with appropriate empathy.

Wizard of Wikipedia (WoW). In the WoW setting, the worker takes on the role of an apprentice, posting questions about a randomly assigned topic (e.g., Piano in Table 2), while the dialogue system acts as the knowledgeable wizard. The system always initiates the dialogue, drawing on background knowledge from the Multi-Source Wizard of Wikipedia dataset (Li et al., 2024), which includes information from Wikipedia, Wikidata (Vrandečić and Krötzsch, 2014), Semantic Frames, and OPIEC (Gashteovski et al., 2019). Workers are instructed to explore the assigned topic in depth by asking questions or sharing their thoughts through the dialogue. We use this to evaluate the system's ability to provide informative, grounded responses and to assess its impact on user satisfaction.

3.2 Collecting Dialogues

Recruitment of participants. We recruited a total of 39 unique crowd workers through Amazon Mechanical Turk.[2] Participants were limited to individuals from the United States, Canada, or the United Kingdom, where English is the primary language. To ensure high-quality dialogue, only workers who achieved high scores in a preliminary annotation task were invited to participate in the main study. All workers participated anonymously, provided informed consent, and were notified that their responses might be publicly released, with all personal information fully protected.

Dialogue systems. We employed two dialogue models: GPT-4o[3] and Llama-3.1-70B-Instruct (Grattafiori et al., 2024). GPT-4o was chosen as a representative high-performing proprietary model, while Llama-3.1-70B-Instruct was selected for its strong performance as a state-of-the-art open-weight model(Suresh et al., 2025). Prompts were designed in accordance with prior studies (Qian et al., 2023; Li et al., 2024) and tailored to each dialogue setting. To explore stylistic alignment, three style control conditions were introduced: (1) instructing the system to match the user's speaking style, (2) instructing it to adopt a style different from the user's, and (3) providing no stylistic instruction. The influence of these conditions on stylistic similarity evaluations is demonstrated in Appendix A. The combination of the two dialogue settings, two models, and three style conditions resulted in 12 experimental configurations, enabling the collection of dialogues exhibiting a diverse range of entrainment phenomena. For the ED setting, we employed a few-shot prompting, a method shown to yield strong performance in dialogue systems. For the WoW setting, we employed a zero-shot prompting, which is optimized to generate knowledge-grounded responses. [4]

Dialogue collection. Based on the dialogue settings and systems described above, we collected human-bot dialogues through interactions with crowd workers. In principle, to minimize potential bias in evaluations, we restricted each worker to participating only once in each of the 12 experimental conditions.

3.3 Annotation

Subjective and objective stylistic similarity. While previous researches employed third party annotators to assess stylistic similarity and engagement, it remains unclear whether users' own perceptions of stylistic similarity or third party evaluations of it accurately reflect users' preferences. To

[2] https://www.mturk.com/

[3] https://platform.openai.com/docs/overview

[4] Other details on the prompts are provided in Appendix B.

	ED	WoW
No. of dialogues	157	157
No. of objective evaluations	50	46
No. of participants	34	34
No. of Unique Topics	–	133
No. of Unique Emotions	30	–
Dialogue length (min–max)	20–23	20–23
Avg. of dialogue length	20.05	21.03
Language	English	English

Table 3: Statistics of the constructed dataset.

	Subjective		Objective	
Label	**ED**	**WoW**	**ED**	**WoW**
PREF.	3.47 ± 1.31	3.96 ± 1.11	3.50 ± 0.79	3.56 ± 0.72
CONS.	4.40 ± 0.80	4.57 ± 0.68	4.55 ± 0.42	4.49 ± 0.57
STYL.	3.32 ± 1.31	3.86 ± 1.06	3.51 ± 0.54	3.18 ± 0.73
EMP.	3.87 ± 1.19	–	4.10 ± 0.58	–
ENGAG.	–	3.87 ± 1.18	–	3.62 ± 0.75

Table 4: Mean and standard deviation of subjective and objective evaluation scores in the ED and WoW settings.

clearly distinguish between these approaches, this study defines stylistic similarity based on who performs the evaluation. We refer to *subjective* stylistic similarity as the degree of stylistic closeness perceived by the user who actually participated in the dialogue. *Objective* stylistic similarity refers to similarity as evaluated by a third party human annotator who did not take part in the dialogue.

Annotating by user. Following each dialogue, workers completed a *subjective evaluation* questionnaire assessing three aspects: Preference (PREF.), Consistency (CONS.), and Stylistic similarity (STYL.). For condition-specific assessments, additional labels were included: Empathy (EMP.) for ED, and Engagingness (ENGAG.) for WoW. All items were rated on a five-point Likert scale, with higher scores indicating more positive evaluations.[5] To ensure data quality, attention check questions with verifiable answers were incorporated into the questionnaire. Responses from workers who failed these checks were excluded from the final dataset.

Annotating by third party. In addition to the subjective evaluations collected from workers after each dialogue, we also gathered *objective evaluations* from independent third party annotators who did not participate in the dialogues. For each stylistic similarity score (1–5) from the subjective evaluation, 20 dialogues (10 from each of the WoW and ED settings) were randomly selected and independently annotated by three different workers. The final dialogue-level score was calculated by averaging the annotations from the three annotators. To enable a direct comparison between subjective and objective evaluations, the annotation procedure (e.g., evaluation criteria and question items) followed the original subjective evaluation protocol. Annotators were given a description of the dialogue setting and were instructed to assess the dialogue as if they were the users.

[5]Evaluation questions are provided in Appendix C.

3.4 Statistics of Dataset

Table 3 shows basic statistical information of our analysis dataset DUO. A total of 314 dialogues were gathered, evenly distributed between the two settings: 157 from ED and 157 from WoW. Furthermore, the number of dialogues annotated by third party evaluators is 50 from ED and 46 from WoW. Each dialogue consists of approximately 10 turns. In the ED setting, this results in 20 utterances. In contrast, in the WoW setting, the system is designed to initiate the dialogue, leading to one additional utterance (21 utterances in total). The ED setting includes 30 distinct emotion labels, whereas the WoW setting covers 133 unique topics, reflecting the diversity of the collected dialogues.

The mean and standard deviation for each subjective and objective evaluation label in both dialogue settings are provided in Table 4. PREF. and STYL. exhibited relatively high standard deviations, indicating considerable variability in user impressions, suggesting that the dataset captures a range of user-perceived preference and stylistic similarity. In contrast, CONS. demonstrated high mean scores and low variance, suggesting that the dialogues were generally perceived as coherent and contextually appropriate. These patterns were consistently observed across both dialogue settings and for both subjective and objective evaluations. The lower standard deviations in the objective evaluations are attributed to the averaging of scores from three independent annotators.

The inter-annotator agreement (IAA) among the three annotators was measured using Krippendorff's α, and the results showed that the α values for all labels were below 0.25.[6] Two factors are considered to be the cause of this low agreement. First, the evaluation metrics, such as preference and style, were highly dependent on individual sensibilities. Second, because the evaluation target

[6]Other statistics and IAA are provided in Appendix D.

Label	ED				WoW			
	PREF.$_{sb}$	CONS.$_{sb}$	STYL.$_{sb}$	EMP.$_{sb}$	PREF.$_{sb}$	CONS.$_{sb}$	STYL.$_{sb}$	ENGAG.$_{sb}$
PREF.$_{sb}$	1.00	0.44	**0.75**	0.74	1.00	0.50	**0.67**	0.80
CONS.$_{sb}$	–	1.00	0.46	0.54	–	1.00	0.38	0.44
STYL.$_{sb}$	–	–	1.00	0.66	–	–	1.00	0.59
EMP.$_{sb}$/ENGAG.$_{sb}$	–	–	–	1.00	–	–	–	1.00

Table 5: Spearman correlation coefficients between subjective labels in the ED and WoW settings. All correlations are statistically significant with $p < 0.001$.

was the entire dialogue, consisting of about 20 utterances, judgments were prone to differ depending on which utterances the annotators focused on.

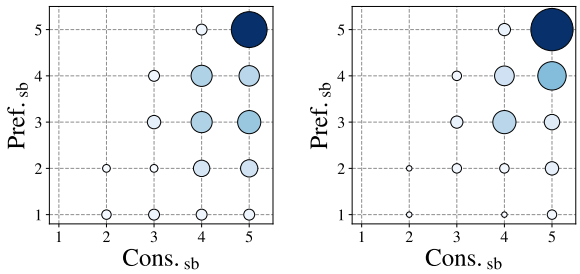
4 Dataset Analysis

This section presents analyses based on evaluations collected during the dataset construction. Hereafter, subjective (sb) and objective (ob) evaluation labels are abbreviated as follows: Preference (PREF.$_{sb}$ / PREF.$_{ob}$), Consistency (CONS.$_{sb}$ / CONS.$_{ob}$)Stylistic similarity (STYL.$_{sb}$ / STYL.$_{ob}$), Empathy for ED (EMP.$_{sb}$ / EMP.$_{ob}$), and Engagingness for WoW (ENGAG.$_{sb}$ / ENGAG.$_{ob}$).

4.1 Factors Influencing User's Preference

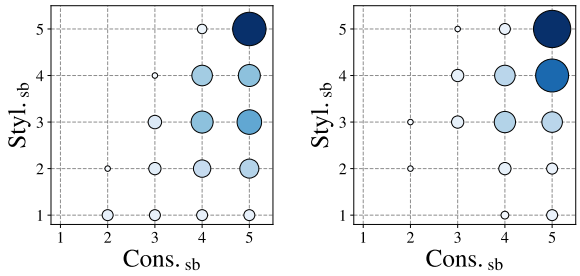
Preference–stylistic similarity association. Table 5 presents the Spearman correlations, r, among the subjective evaluations in the dataset. All correlation coefficients were statistically significant ($p < 0.001$). Notably, STYL.$_{sb}$ exhibited strong positive correlations with PREF.$_{sb}$ in both dialogue settings ($r = 0.75$ for ED; $r = 0.67$ for WoW). These results suggest that the degree of subjective similarity between a user's and the system's speaking styles influences users' favorable impressions of the dialogue. These empirical results strongly support one of our main claims, namely, the existence of a strong association between stylistic similarity (STYL.$_{sb}$) and users' subjective preferences (PREF.$_{sb}$), which is one of the main claims of this paper as mentioned at the end of Section 1.

Consistency–subjective evaluations relationship. Table 5 shows relatively low correlations between CONS.$_{sb}$ and PREF.$_{sb}$ compared to other subjective evaluation pairings in both the ED and the WoW settings. To explore this relationship, we visualized the distributions of subjective CONS.$_{sb}$ and PREF.$_{sb}$ in each setting, as illustrated in Figure 3. Both plots reveal a consistent pattern: when CONS.$_{sb}$ scores are low, PREF.$_{sb}$ scores are similarly low. In contrast, when CONS.$_{sb}$ scores are high, PREF.$_{sb}$ scores exhibit a broader distribution.[7] This pattern suggests that the consistency is a prerequisite for positive user impressions. Inconsistent dialogues tend to result in negative perceptions, regardless of other factors. Similarly, correlations between CONS.$_{sb}$ and STYL.$_{sb}$ were low in Table 5. Figure 4 shows the distributions of subjective CONS.$_{sb}$ and STYL.$_{sb}$, both exhibiting a similar pattern, namely, lower CONS.$_{sb}$ scores are associated with lower STYL.$_{sb}$ scores, whereas higher CONS.$_{sb}$ scores correspond to a wider spread of STYL.$_{sb}$ scores. This observation suggests that users may only perceive stylistic similarity when the system's responses appropriately capture the



(a) ED (b) WoW

Figure 3: Consistency (CONS.$_{sb}$) vs. Preference (PREF.$_{sb}$), where both are subjective evaluations. Larger and darker points indicate higher frequency.



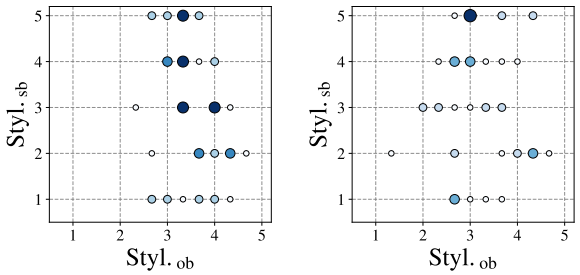
(a) ED (b) WoW

Figure 4: Consistency (CONS.$_{sb}$) vs. Stylistic Similarity (STYL.$_{sb}$), where both are subjective evaluations. Larger and darker points indicate higher frequency.

[7] Other plots are provided in Appendix E.

Obj. \ Subj.	ED				WoW			
	PREF.$_{sb}$	CONS.$_{sb}$	STYL.$_{sb}$	EMP.$_{sb}$	PREF.$_{sb}$	CONS.$_{sb}$	STYL.$_{sb}$	ENGAG.$_{sb}$
PREF.$_{ob}$	0.14	0.16	−0.01	0.25	0.35*	0.25	0.28	0.29*
CONS.$_{ob}$	0.17	0.16	0.10	0.18	0.47***	0.21	0.50***	0.40**
STYL.$_{ob}$	−0.19	−0.05	−0.28	−0.11	0.19	0.22	0.01	0.18
EMP.$_{ob}$/ENGAG.$_{ob}$	0.20	0.15	0.14	0.27	0.14	0.18	0.04	0.14

Table 6: Spearman correlation coefficients between subjective (columns) and objective labels (rows) in the ED and WoW settings. Asterisks denote significance (* $p < .05$, ** $p < .01$, *** $p < .001$).



(a) ED (b) WoW

Figure 5: Objective stylistic similarity (STYL.$_{ob}$) vs. Subjective stylistic similarity (STYL.$_{sb}$). Larger and darker points indicate higher frequency.

Dialogue	**User**: *I was prepared to do my taxes this past weekend. I had all my documents ready.* **System**: ***Ah**, taxes! Something unexpected came up, **huh?*** **User**: ***No**, it was a couple days before tax day.* **System**: ***Ah**, cutting it close, **huh?*** **User**: ***Oh** I don't know, I gave myself a few days buffer.* **...**
Ratings	STYL.$_{sb}$: 2 STYL.$_{ob}$: 4.33 (4,4,5)

Table 7: Example dialogue in the ED setting where both subjective stylistic similarity is low, while objective stylistic similarity is high.

Dialogue	**User**: *I was surprised when my husband gave me a surprise birthday party to celebrate my 50th birthday.* **System**: ***Wow**, that sounds amazing! It must have been a wonderful surprise to celebrate such a special milestone.* **User**: ***Yes**, I had no idea he had been planning this surprise for me and I was truly surprised!* **System**: ***That's awesome!** He must really know how to make you feel special.* **User**: ***Yes**, I feel blessed to have my husband and he is truly my soulmate. I felt so surprised that he planned this special party for me.* **...**
Ratings	STYL.$_{sb}$: 5 STYL.$_{ob}$: 2.67 (2,2,4)

Table 8: Example dialogue in the ED setting where both subjective stylistic similarity is high, while objective stylistic similarity is low.

dialogue context and appear natural.

Empathy/Engagingness–subjective evaluations. The setting-specific evaluation labels, EMP.$_{sb}$ for ED and ENGAG.$_{sb}$ for WoW, also showed strong correlations with PREF.$_{sb}$ ($r = 0.74$ for EMP.$_{sb}$; $r = 0.80$ for ENGAG.$_{sb}$). Notably, the correlation coefficients are nearly equivalent to or even higher than those of the STYL.$_{sb}$ (0.75 vs. 0.74 for ED and 0.67 vs. 0.80 for WoW). These results suggest that achieving the intended communicative goals of each setting is closely linked to users' subjective preferences.

4.2 Gap Between Subjective and Objective

Table 6 presents the Spearman correlation coefficients between subjective and objective evaluations. This section focuses on the correlation between stylistic similarity in third party objective evaluations and subjective evaluations. While this is a slight digression, CONS.$_{ob}$ shows a positive correlation with subjective evaluations in the WoW setting, and this result suggests an interesting direction for future work.

Subjective vs. objective stylistic similarity. In Table 6, a weak but positive correlation was observed between PREF.$_{sb}$ and STYL.$_{ob}$, particularly in the WoW setting ($r = 0.28$). In contrast, the correlation between STYL.$_{sb}$ and STYL.$_{ob}$ remained low. This indicates a gap between the users' perceived degree of stylistic similarity during actual participation in the dialogue and the similarity assessed by third party annotators. Notably, although both subjective and objective stylistic similarity are based on human evaluations, the observed discrepancy suggests they capture different perspectives. These findings underscore the importance of distinguishing between subjective and objective evalua-

	ED				WoW			
	PREF.	CONS.	STYL.	EMP.	PREF.	CONS.	STYL.	ENGAG.
Subjective	0.22^{**}	0.17^{*}	0.17^{*}	0.26^{**}	0.27^{***}	0.20^{*}	0.17^{*}	0.29^{***}
Objective	0.23	0.25	0.07	0.36^{*}	0.47^{**}	0.45^{**}	0.22	0.38^{**}

Table 9: Spearman correlation coefficients between stylistic similarity measures, Subjective and Objective, (rows) and users' subjective evaluations (columns) in the ED and WoW settings. Asterisks indicate significance (* $p < .05$, ** $p < .01$, *** $p < .001$).

tions when analyzing stylistic similarity, which is also one of the main claims of this paper as mentioned at the end of Section 1.

Explaining the gap in stylistic similarity. As discussed earlier, the correlation between STYL._{sb} and STYL._{ob} was not positive, and the two measures themselves showed no significant relationship. To explore this discrepancy, we conducted a qualitative analysis. Figure 5 presents scatter plots comparing STYL._{sb} and STYL._{ob}. Notably, several dialogues rated low on STYL._{sb} (scores of 1 or 2) still received relatively high STYL._{ob} scores (above 2). This suggests that users' perceptions of stylistic mismatch may diverge from those of third party annotators. Additionally, this suggests that users tend to make more polarized, intuitive judgments based on whether the system's style feels personally similar. In contrast, third party annotators, who rely solely on dialogue text, appear to assess stylistic similarity in a more graded and continuous manner. While this analysis highlights a gap between subjective and objective stylistic similarity, similar discrepancies may arise in other tasks due to differing evaluative perspectives, such as those between dialogue participants and external annotators.

Examples of the gap in stylistic similarity. Tables 7 and 8 show examples of dialogues that exhibit gaps between subjective and objective stylistic similarity scores. In Table 7, the objective stylistic similarity score assigned by third party annotators is higher than the subjective rating provided by the user. At the beginning of the dialogue, the user adopts a formal style (e.g., *'No'*, avoiding contractions), while the system responds in a more casual manner (e.g., *'Ah'* or *'huh'*). As the conversation progresses, the user shifts to a more casual tone (e.g., *'Oh'*), seemingly adapting to the system's style. From the user's perspective, the initial stylistic mismatch, and the perceived need to adjust, may have contributed to a lower subjective similarity rating. In contrast, third party annotators may have focused on the later-stage stylistic convergence, resulting in a higher objective rating. In Table 8, the opposite pattern is observed: the user assigns a higher stylistic similarity score than the third party annotators. Here, the system produces positive expressions (e.g., *'Wow'*, *'That's awesome'*), and the user replies with similarly positive responses (e.g., *'Yes'*). This emotional alignment may have led the user to perceive high stylistic similarity. However, third party annotators may have noticed subtle mismatches in formality, such as the difference in tone between *'Wow'* and *'Yes'*, and rated the similarity lower. These examples highlight how discrepancies in evaluation can emerge depending on the rater's perspective. They underscore the challenge of achieving consistent assessments and the importance of distinguishing between subjective evaluations from dialogue users and objective evaluations by third party annotators.

4.3 Automatic Evaluation

The primary focus of this paper is on the human evaluation setting. Additionally, this section briefly discusses the automatic evaluation setting.

Evaluation method. We followed LLM-Eval (Lin and Chen, 2023) as an automatic evaluation method. LLM-Eval has been shown to produce dialogue quality scores that better align with objective human evaluations compared to traditional automatic metrics such as BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004). In this experiment, GPT-4o was used as the underlying LLM to generate LLM-Eval scores, specifically producing scores for the label "stylistic similarity."

Results and discussions. Table 9 shows the correlation coefficients between the quantified similarity score automatically provided by LLM-Eval and both users' subjective and objective evaluations. We find that LLM-Eval's scores correlate more strongly with objective evaluations. This is unsurprising, given that the method itself is objective

rather than subjective. These results again highlight the importance of considering human evaluations, whether obtained in subjective or objective settings. Different settings may lead to different conclusions. They also support our claim that there is a discrepancy between subjective and objective stylistic similarity, emphasizing the need to consider these perspectives separately in dialogue evaluation. We also note that we have never claimed that our results conflict with previous findings of stylistic similarity correlates with human preference, as there is a certain degree of correlation in the objective settings, and prior studies primarily focus on objective stylistic similarity.

5 Conclusion

This study focused on stylistic similarity as a key factor influencing individual users' dialogue preferences. While prior work has pointed to the potential relevance of stylistic similarity, it remained unclear whether users' own perception of stylistic similarity or third party assessments play a more critical role. To explore this question, we collected human–bot dialogues in open-domain settings and obtained both users' subjective preferences and stylistic similarity ratings, as well as third party annotations. Our findings revealed a strong correlation between subjective stylistic similarity and user preference, whereas objective stylistic similarity showed only a weak correlation. Moreover, there was no significant association between subjective and objective stylistic similarity. They suggest that subjective and objective evaluations reflect different aspects of style, underscoring the need to analyze them separately.

The observed discrepancies suggest that whether the evaluation is made by a participant or a third party may matter not only for stylistic similarity but also for other aspects of dialogue assessment. Future work could consider the impact of such perspective differences on system evaluation.

Limitations

In this study, the scope of the dialogues analyzed is restricted to open-domain dialogues, excluding task-oriented dialogues, such as those focused on information provision or problem-solving, which were not sufficiently addressed. As a result, the proposed methodology and evaluation findings may not be directly applicable to task-oriented dialogue systems or other specific conversational contexts.

Ethical Considerations

Participants were recruited via the crowdsourcing platform Amazon Mechanical Turk to collect dialogue data and evaluations. Two tasks were conducted: the subjective evaluation task, where participants interacted with the system and rated their experience, and the objective annotation task, where third party annotators evaluated 10 dialogues per session. A detailed task description, outlining structure and key instructions, was provided to all participants. Informed consent was obtained, including assurances that personal information would be protected and that the collected data might be made publicly available in the future. For third party annotators, additional consent was obtained to inform them that the dialogues may contain potentially sensitive content. The subjective evaluation task and the objective annotation task were designed to take approximately 15 and 20 minutes respectively to complete, with participants compensated at an estimated rate of $15 per hour (i.e., $3.75 and $5.00 per task, respectively).

The collected data may include potentially sensitive content, such as personal views on political orientation or social values. These data are solely used for analyzing evaluations of dialogue and are not employed for model training or response generation. Any utterances containing discriminatory or overtly offensive language were manually filtered. Furthermore, all personally identifiable information, including user names and named entities related to the individual, has been masked to ensure anonymity.

We plan to release the DUO dataset in a format that allows reuse for non-commercial and nonprofit research purposes only. To promote responsible use, we will explicitly note that the dataset may contain culturally or ideologically sensitive content, and encourage users to review it carefully and exercise appropriate judgment before use.

Acknowledgments

This research was supported by JST Moonshot R&D Grant Number JPMJMS2011-35 (fundamental research), JSPS KAKENHI Grants JP25K21263, JST BOOST JPMJBY24A1.

References

Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. 2018. Unsupervised learning of style-sensitive word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 572–578.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Liviu P. Dinu and Ana Sabina Uban. 2023. A computational analysis of the voices of shakespeare's characters. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 295–300.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering.

Mauajama Firdaus, Umang Jain, Asif Ekbal, and Pushpak Bhattacharyya. 2021. SEPRG: Sentiment aware emotion controlled personalized response generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 353–363.

Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, and Rainer Gemulla. 2019. Opiec: An open information extraction corpus.

Howard Giles, Tania Ogay, and 1 others. 2007. Communication accommodation theory. *Explaining communication: Contemporary theories and exemplars*, pages 293–310.

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.

Makiko Imamura, Yan Bing Zhang, and Jake Harwood. 2011. Japanese sojourners' attitudes toward americans: Exploring the influences of communication accommodation, linguistic competence, and relational solidarity in intergroup contact. *Journal of Asian Pacific Communication*, 21(1):115–132.

Shota Kanezaki, Seiya Kawano, Akishige Yuguchi, Marie Katsurai, and Koichiro Yoshino. 2024. Investigation of the influence of different entrainment metrics/strategies on subjective evaluation in reranking dialogue systems. In *Proceedings of International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Seiya Kawano, Shota Kanezaki, Angel Fernando Garcia Contreras, Akishige Yuguchi, Marie Katsurai, and Koichiro Yoshino. 2023. Analysis of style-shifting on social media: Using neural language model conditioned by social meanings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7911–7921.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966.

Marianne LaFrance. 1979. Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. *Social Psychology Quarterly*, pages 66–70.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443.

Xiangci Li, Linfeng Song, Lifeng Jin, Haitao Mi, Jessica Ouyang, and Dong Yu. 2024. A knowledge plug-and-play test bed for open-domain dialogue generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 666–676.

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

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models.

Ao Liu, An Wang, and Naoaki Okazaki. 2022. Semi-supervised formality style transfer with consistency training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4701.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan

Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504.

Sourabrata Mukherjee, Vojtěch Hudeček, and Ondřej Dušek. 2023. Polite chatbot: A text style transfer application. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 87–93.

Md Nasir, Sandeep Nallan Chakravarthula, Brian Baucom, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2019. Modeling interpersonal linguistic coordination in conversations using word mover's distance. In *Interspeech*, volume 2019, page 1423.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL-08: HLT, Short Papers*, pages 169–172.

OpenAI. 2024. Gpt-4 technical report.

Phil Sidney Ostheimer, Mayank Kumar Nagda, Marius Kloft, and Sophie Fellenz. 2024. Text style transfer evaluation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15802–15822.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Manasa Bharadwaj, Nikhil Verma, Ali Pesaranghader, and Scott Sanner. 2023. COUNT: COntrastive UNlikelihood text style transfer for text detoxification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8658–8666.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. SentSim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156.

Richard L. Street, Robert M. Brady, and William Benjamin Putman. 1983. The influence of speech rate stereotypes and rate similarity or listeners' evaluations of speakers. *Journal of Language and Social Psychology*, 2:37 – 56.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 685–691.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and EngSiong Chng. 2025. DiaSynth: Synthetic dialogue generation framework for low resource dialogue applications. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 673–690.

Yuma Tsuta, Naoki Yoshinaga, Shoetsu Sato, and Masashi Toyoda. 2023. Rethinking response evaluation from interlocutor's eye for open-domain dialogue systems. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 55–63.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85.

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.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861.

Yuki Zenimoto, Shinzan Komata, and Takehito Utsuro. 2023. Style-sensitive sentence embeddings for evaluating similarity in speech style of Japanese sentences by contrastive learning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 32–39.

Haoran Zhang and Tan Yongmei. 2024. Enhancing knowledge selection via multi-level document semantic graph. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5996–6006.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

A Influence of Style Control Conditions

Table 10 shows the mean subjective stylistic similarity for each system. We controlled the style of the systems used for dialogue collection with the following three types of prompts: (1) Aligned: Instructing the system to match the user's speaking style. (2) Not Aligned: Instructing it to adopt a style different from the user's. (3) Neutral: Providing no stylistic instruction. The mean values indicate that these prompts successfully generated a range of subjective stylistic similarities. In the WoW setting, in particular, we observed a trend where the mean stylistic similarity increased in the intended order: Not aligned, Neutral, and then Aligned.

	ED		WoW	
Prompt	**GPT-4o**	**Llama-3.1**	**GPT-4o**	**Llama-3.1**
Not Aligned	3.60	3.57	3.71	3.87
Neutral	2.61	3.32	3.74	3.89
Aligned	3.59	3.38	3.86	4.13

Table 10: Mean subjective stylistic similarity for each prompt condition (Aligned, Neutral and Not Aligned).

B Prompts

B.1 Stylistic Similarity Prompt

The dialogue systems used in this study were prompted to respond in either a similar or dissimilar speaking style relative to the user's utterances. Examples of the prompts are shown in Figure 6. The illustrative style categories follow the classification proposed by Lai et al. (2024).

B.2 EmpatheticDialogue

Figure 7 shows the prompt used for the system in the ED setting. This prompt is based on the template proposed by Qian et al. (2023), which demonstrated strong performance for existing dialogue systems using a few-shot prompting approach. The dialogue examples were randomly sampled from the training split of the ED dataset.

B.3 Wizard of Wikipedia

Figure 8 shows the prompt used for the system in the Wizard of Wikipedia setting. This prompt is based on the template proposed by Li et al. (2024), which demonstrated strong performance in generating responses grounded in factual knowledge for existing dialogue systems. For each topic, knowledge entries were sourced from the Multi-Source Wizard of Wikipedia dataset (Li et al., 2024), which aggregates information from multiple sources, and incorporated into the prompt.

```
Please make the style match from (or be
different) user's Style includes things like
the following:
    - Formality: informal vs formal
    - Politeness: impolite vs polite
    - Gender: masculine vs feminine
    - Biasedness: biased vs neutral
    - Toxicity: offensive vs non-offensive
```

Figure 6: Prompt snippet for controlling stylistic alignment in system responses.

```
This is an empathetic dialogue task: The first
worker (Speaker) is given an emotion label
and writes his own description of a situation
when he has felt that way. Then, Speaker tells
his story in a conversation with a second
worker (Listener). The emotion label and
situation of Speaker are invisible to Listener.
Listener should recognize and acknowledge
others' feelings in a conversation as much as
possible.

Now you play the role of Listener, please give
the corresponding response according to the
existing context. You only need to provide the
next round of response of Listener.

The following is the existing dialogue context:

Instance 1:
...
Instance 2:
...
Instance 3:
...
Instance 4:
...
Instance 5:
...
"""
{Dialog history}
"""
"""
{Stylistic similarity prompts}
"""
You MUST keep response as short as possible.
```

Figure 7: Prompt for the system in the ED setting.

C Post-Dialogue Questionnaire

Table 11 presents the specific question items used to collect both the subjective evaluations from dialogue participants and the objective evaluations from third party annotators for each label in the dataset. For the objective evaluations, annotators were instructed to answer the questions from the perspective of a user interacting with the system. The question for Consistency was adapted from the

```
The following is the conversation between the "Wizard", a knowledgeable speaker who can access Wikipedia knowledge sentences to chat to with the "Apprentice", who does not have access to Wikipedia. The conversation topic is {Topic}.

This is their conversation history:
"""
{Dialog history}
"""

Here is some retrieved Wikipedia knowledge for the Wizard. The Wizard can choose any subset of the following knowledge. It's also allowed to not choose any of them.
"""
{Knowledge}
"""
"""
{Stylistic similarity prompts}
"""
Given the knowledge above, make a very brief, such as one sentence, natural response for the Wizard. Not all information in the chosen knowledge has to be used in the response.

The Wizard's response is:
```

Figure 8: Prompt for the system in the WoW setting

Label	Question
Preference	Was the dialogue preferable for you?
Stylistic similarity	Was the system's speaking style similar to yours?
Consistency	Was the system consistent in the information it provided throughout the dialogue?
Empathy	Did the responses show understanding of the feelings of you talking about your experience?
Engagingness	How engaging did you find the conversation?

Table 11: Questions presented to users for evaluating different aspects of the dialogue.

FED metric (Mehri and Eskenazi, 2020), while the Empathy item was based on the original prompt used in ED (Rashkin et al., 2019). All responses were collected on a 5-point Likert scale (1: not at all, 2, 3: somewhat, 4, 5: very much).

D Other Statistics of Dataset

Table 12 presents statistics of our dataset that were not discussed in the main text. Focusing on utterance length, utterances in the WoW setting tend to be longer. This is likely due to the nature of the task, where the Wizard provides informative responses related to the given topic.

	ED	WoW
No. of utterances	3148	3302
No. of tokens	44640	63157
Utterance length	1–76	1–146
Avg. of Utterance length	14.18	19.13
Vocabulary	3819	6439
Type-Token ratio	0.086	0.10
Herdan's C	0.77	0.79

Table 12: Additional statistics of the dataset.

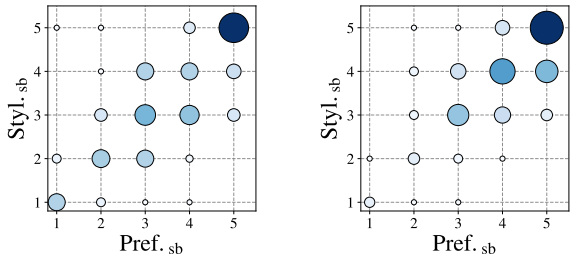
	PREF.$_{ob}$	CONS.$_{ob}$	STYL.$_{ob}$	EMP.$_{ob}$	ENGAG.$_{ob}$
ED	0.20	0.15	−0.07	−0.09	-
WoW	0.11	0.24	0.09	-	0.13

Table 13: Inter-annotator agreement for the objective evaluations, measured by Krippendorff's α.

Table 13 shows the results for the inter-annotator agreement among the three annotators for the objective evaluation, as measured by Krippendorff's α. The evaluation criteria include Pref_{ob}, Cons_{ob}, Styl_{ob}, Emp_{ob}, Engag_{ob}. The results showed that the α values for all labels were below 0.25.

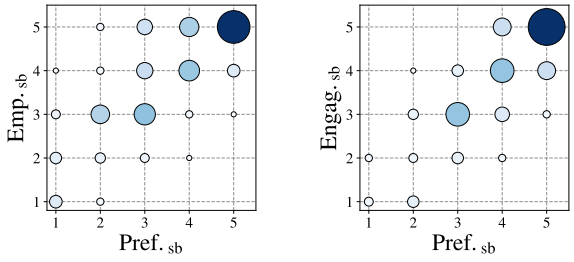
E Scatter Plots of Subjective Evaluations

In Section 4.1, we analyzed the correlations between subjective evaluation labels, as shown in Table 5. Scatter plots for label pairs not discussed in the main text are presented in Figures 9 and 10.



(a) ED (b) WoW

Figure 9: Subjective preference vs. stylistic similarity.



(a) ED (b) WoW

Figure 10: Subjective preference vs. empathy on ED and engagingness on WoW.