StyleChat: Learning Recitation-Augmented Memory in LLMs for Stylized Dialogue Generation

Jinpeng Li[1*], **Zekai Zhang**[1*], **Quan Tu**[2], **Xin Cheng**[1], **Dongyan Zhao**[1,3†], **Rui Yan**[2†]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Gaoling School of Artifical Intelligence, Renmin University of China
[3]State Key Laboratory of Media Convergence Production Technology and Systems
lijp.pku@gmail.com, justinzzk@stu.pku.edu.cn
zhaody@pku.edu.cn, ruiyan@ruc.edu.cn

Abstract

Large Language Models (LLMs) demonstrate superior performance in generative scenarios and have attracted widespread attention. Among them, stylized dialogue generation is essential in the context of LLMs for building intelligent and engaging dialogue agent. However the ability of LLMs is data-driven and limited by data bias, leading to poor performance on specific tasks. In particular, stylized dialogue generation suffers from a severe lack of supervised data. Furthermore, although many prompt-based methods have been proposed to accomplish specific tasks, their performance in complex real-world scenarios involving a wide variety of dialog styles further enhancement. In this work, we first introduce a stylized dialogue dataset **StyleEval** with 38 styles by leveraging the generative power of LLMs comprehensively, which has been carefully constructed with rigorous human-led quality control. Based on this, we propose the stylized dialogue framework **StyleChat** via recitation-augmented memory strategy and multi-task style learning strategy to promote generalization ability. To evaluate the effectiveness of our approach, we created a test benchmark that included both a generation task and a choice task to comprehensively evaluate trained models and assess whether styles and preferences are remembered and understood. Experimental results show that our proposed framework StyleChat outperforms all the baselines and helps to break the style boundary of LLMs.

1 Introduction

Large language models (LLMs) have made considerable advancements in the field of natural language processing (OpenAI, 2023; Brown et al., 2020; Zhang et al., 2023). These models demonstrate a deep comprehension of the context and semantics of complex instructions and are capable of generating coherent and logical texts (Ouyang et al., 2022). Especially in the field of dialogue, the integration of large language models has enabled machines to interact with humans in a more natural, personalized, and stylized manner (Lv et al., 2023a,b; Tu et al., 2023), heralding a new phase in the evolution of artificial intelligence.

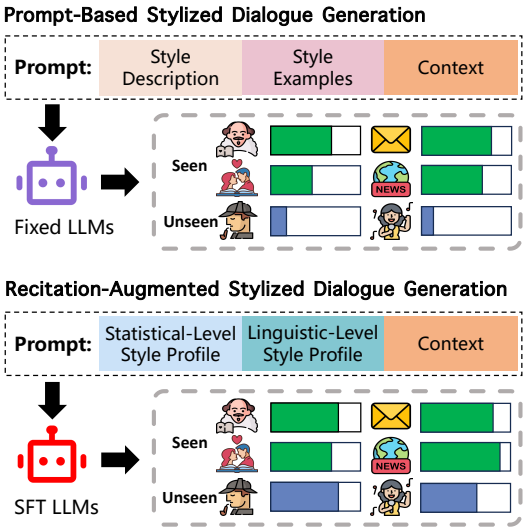

Figure 1: Examples of stylized dialogue generation by different LLMs. The progress bar represents the quality in the particular style.

Stylized dialogue generation is crucial in the development of intelligent and engaging dialogue agents (Zheng et al., 2021a; Li et al., 2023) in the era of LLMs. Nonetheless, this task is challenged by the limited availability of supervised data correlating contexts and responses with the desired styles (Gao et al., 2019a; Li et al., 2023). In particular, it is more difficult to collect parallel corpus for abstract, multilayered or dynamically derived styles. Existing works have typically relied on pseudo data constructed using back translation, thereby resulting in low-quality data that fail to account for the variability and complexity of individual language styles (Su et al., 2020; Zheng

[*] Equal contribution.
[†] Corresponding authors: Dongyan Zhao and Rui Yan.

et al., 2021b; Li et al., 2021, 2023). The result is that the model generated dialogues tends to be overly standardized, lacking individuality and diversity. Additionally, despite the introduction of prompt-based LLMs specifically designed for certain tasks, their performance in complex real-world scenarios necessitates improvement (Zeng et al., 2023). Particularly, when encountering domain data or new style not seen during the pre-training phase, the generalization ability of LLMs dealing with complex instruction significantly declines.

To address these challenges, we exploit the generative capacity of LLMs, combined with statistical and linguistic perspectives, to construct a large-scale dataset, named **StyleEval**. This dataset consists of stylized dialogues with style profiles, contributing to the creation of custom dialogue agent. To our knowledge, this is the first large-scale dataset for stylized dialogue generation, incorporating 38 styles and 24,728 dialogues. Our process begins with the collection of well-known styles from various genres, utilizing GPT-4 to generate statistical-level style profile that includes descriptions and examples. We then extract linguistic-level style profile from these examples based on linguistic knowledge. After initial pre-processing, we invite annotators to evaluate the quality of the dialogues. Furthermore, we aim to enhance the style generalization ability of the LLM without compromising its overall functionality. However, direct prompting methods of LLMs face challenges to generalize to new styles, as depicted in Figure 1. We adapt the model to generate responses for styles it has not previously encountered. To address this, we propose the **StyleChat** framework, which introduces a style thought chain, enabling models to generate style profiles before responding via a recitation-augmented memory strategy. This memory consists of two stages: recite then respond during training, and recall then respond during inference. This approach also encourages StyleChat to learn how to derive unseen style profiles, thereby improving generalization. Besides, we further enhance the style derivation ability by implementing multi-task style learning to increase activation of style abilities through a style transfer dataset.

Comprehensive experiments conducted on StyleEval demonstrate that our approach considerably enhances the performance of LLMs in stylized dialogue generation. StyleChat accurately captures the essence of various styles and generates dialogue content that is rich in stylistic elements. By utilizing meticulously constructed supervised data and the recitation-augmented memory strategy, we can effectively transcend the limitations of LLMs on specific tasks, thereby improving their performance in novel styles. We also discuss the advantages of our approach extensively in Appendix. In summary, our contributions can be summarized as follows:

- We construct a large-scale, high-quality dataset, StyleEval, for the stylized dialogue generation. This dataset comprises 24,728 parallel stylized dialogue turns covering 38 diverse styles, serving as a crucial prerequisite for successful style-playing.
- We introduce a recitation-augmented memory strategy for stylized dialogue generation, which motivates StyleChat to learn to derive unseen style profiles for better generalization.
- We conduct extensive experiments on various large language models under both in-domain and out-of-domain settings using StyleEval, demonstrate that our proposed framework, StyleChat, outperforms all baseline models.

2 Related Work

2.1 Stylized Dialogue Generation

Stylized dialogue generation represents a significant research direction within the field of intelligent dialogue systems, focusing on generating dialogue imbued with specific stylistic characteristics. Initial approaches primarily depended on the utilization of latent variables within the hidden state space (Gao et al., 2019b) Some researchers have attempted to integrate pseudo data into existing corpora via back translation (Sennrich et al., 2015; He et al., 2016; Zheng et al., 2021b; Li et al., 2021). With the advent of pre-trained models, StyleDGPT (Yang et al., 2020) employs both a style language model and a style classifier to provide style signals. Nonetheless, these methods often struggle to capture and reproduce complex linguistic style features, leading to generated dialogues that may be overly mechanical and homogeneous. Furthermore, these methods demonstrate limited generalization capabilities when adapting to new or unseen styles. Thus, our research seeks to surmount these limitations inherent in traditional stylized dialogue methodologies, aiming to enhance the performance and quality of stylized dialogue through the construction of supervised data and fine-tuning, leveraging the instruc-

tion comprehension and generation capabilities of large language models.

2.2 Domain-Specific LLMs

To augment the performance of models in specific domains (e.g., medical, legal, character-based, etc.), supervised fine-tuning (SFT) and in-context learning (ICL) have emerged as dominant methodologies(Raffel et al., 2020; Dong et al., 2022). Domain-specific large language models have been extensively explored by researchers (Singhal et al., 2023; Cui et al., 2023; Tu et al., 2023). These models are tailored to understand and generate content within a specific domain, enabling them to capture domain-specific nuances and terminology. In-context learning capitalizes on the inherent ability of large language models to swiftly adapt to a specific context or task with a few examples (Garg et al., 2022). By supplying contextually relevant instructions (e.g., prompts or dialogues), this approach partly facilitates the generalization and effective performance within a specific domain (Radford et al., 2018; Begus et al., 2023). Nonetheless, its performance can be constrained by the number and quality of examples, and it may lack the accuracy required for complex or fine-grained style transformations. While these studies concentrate on adapting models to various domains, existing domain-specific LLMs may not fully meet the requirements due to the diversity and complexity of stylized dialogue, indicating the need for further research and optimization.

3 Methodology

3.1 Dataset Construction

Design Principles: LLMs are pre-trained on an extensive range of texts, including various style corpora. This expansive training embeds a rich repository of stylistic knowledge within parameters, thereby offering significant potential for stylized dialogue generation. However, exploiting the style potential of LLMs often necessitates accurate style definitions and specific strategies. Therefore, we focuses on two primary objectives: **1) Efficient alignment of LLMs to a certain style.** To effectively tailor LLMs to a particular style, we employ style definitions from both statistical and linguistic perspectives, as illustrated in Figure 2. **2) Activate their style-related abilities for better generalization.** To optimally activate the style-specific capabilities of LLMs, we strategically design the data distribution of our two style-centric tasks, stylized dialogue generation and text style transfer.

3.1.1 Statistical-Level Style Profile

As outlined in (Jin et al., 2020), conventional deep learning approaches typically define style from a statistical or data-driven perspective. These methods involve training models on large corpora of texts or responses with very different styles, allowing models to learn characteristics of the various styles independently. Consistent with these approaches, we employ GPT-4 as a style agent to create a statistical-level style profile. Specifically, we task the agent to generate a comprehensive description of a specific style, leveraging its extensive, statistically-informed understanding of styles. Subsequently, agent is used to generate a series of sentences representative of the specified style. To ensure the relevance and accuracy of these examples, we implement a post-selection phase. During this stage, human annotators meticulously select sentences that most effectively embody the core characteristics. Given the robust in-context learning abilities of LLMs, we limit the number of example sentences to four, aiming for a precise yet efficient statistical definition of style.

3.1.2 Linguistic-Level Style Profile

In addition to the statistical perspective, we also delve into the linguistic perspective of style. We argue that representing style through a large collection of sentences can be inadequate. It lacks explicit guidance on how to produce stylized sentences and is resource-intensive, especially when it comes to generalization, as gathering extensive corpora for new styles is costly. Different from conventional style language models, LLMs demonstrate good linguistic understanding ability, as highlighted by (Begus et al., 2023). Therefore, we propose a re-evaluation of the concept of style from a linguistic perspective and its integration with the statistical level style profile. This combined approach is designed for more efficient and accurate style definition and enhanced generalization capabilities. Specifically, we adopt the definition provided by (Kumar, 2022), *The style has been analysed in such terms as rhetorical situation and aim, diction or choice of words, type of sentence structure and syntax,and the density and kinds of figurative languages.* Based on this, we decompose style into the following four attributes for more precise and comprehensive guidance:

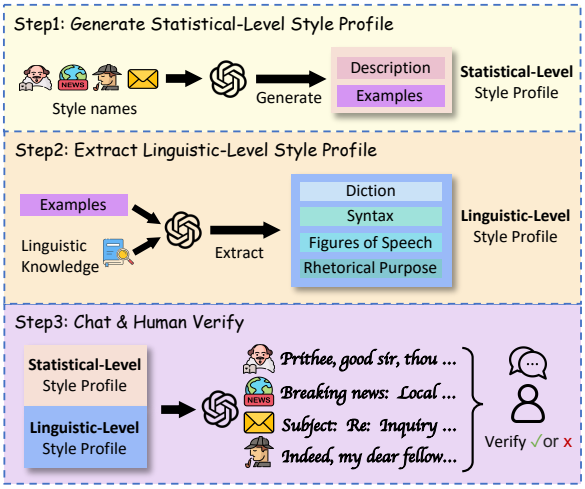

Figure 2: The workflow for developing StyleEval. LLM is employed to generate statistical-level style profile for a certain style. Then to extract linguistic-level style profile based on examples and linguistic knowledge. Finally, we produce stylized dialogue with context and style profile, verified by human to guarantee quality.

`Diction` is the choice of words and style of expressions, are the most basic elements of style (Kumar, 2022; Jin et al., 2020). For example, the use of complex and technical vocabulary is apt for academic papers (arXiv styles), but might be less appealing in a novel targeted at a general audience.

`Syntax` is the arrangement of words and phrases to create well-formed sentences in a language (Kumar, 2022). "To be or not to be, that is the problem" in Shakespeare is a classic illustration of syntax influence. Altering this to a more standard syntax, such as "The question is whether to be or not to be", diminishes its Shakespearean essence.

`Figures of Speech` is the creative uses of language where words take on a non-literal meaning as described by (Konig, 2016). These include devices like metaphors, similes, personification, and hyperbole. For instance, in the Poems style, William Wordsworth's line "I Wandered Lonely as a Cloud" employs a simile to liken the narrator's solitude to a cloud, conjuring a feeling of freedom.

`Rhetorical Purposes` refer to the objectives a speaker aims to accomplish through communication (McDonald and Pustejovsky, 1985). In Questionnaire style, the rhetorical purpose is information gathering, thus necessitating many interrogative sentences. Conversely, the goal is to entertain in Humor style, requiring humorous content.

The comprehensive process for creating a style profile is illustrated in Figure 2. Initially, GPT-4 serve as style agent and is prompted to generate a concise overall description and four representative examples for the given style, forming the statistical-level style profile. We then integrate linguistic knowledge to guide agent in extracting relevant linguistic attributes from these examples, leading to the development of the linguistic-level style profile. It is worth highlighting the distinction between "style" and "persona" or "character" here. While the latter emphasizes content or experiences, we focus on stylistic content along with linguistic attributes. Furthermore, certain styles (e.g., Email, Lyrics, News) should be classified as a style rather than a persona or character.

3.1.3 Multi-Task Datasets for Style Activation

In this section, we outline the development of StyleEval, which encompasses two style-centric tasks: Stylized Dialogue Generation and Text Style Transfer. For the stylized dialogue generation, we build upon multi-level style profiles discussed in the previous sections. By engaging GPT-4 with style profiles corresponding to certain styles and dialogue, we construct pairs of contexts and stylistic responses. These pairs subsequently serve as training data for our model in a multi-turn dialogue setting. Guided by insights from (Chan et al., 2022), which suggest that datasets featuring a combination of several principal clusters along with a multitude of rare instances enhance the model's generalization capabilities, we structure our dataset accordingly. We curate a collection featuring 3,532 examples for 4 main styles, supplemented by 400 examples for 23 less prevalent styles, aiming to optimize the model's generalization potential. For the text style transfer, we utilize GPT-4 to obtain transfer instances between any pair of styles in four primary styles, totalling 600 pairs of data. Despite limited amount of data, we demonstrate the efficacy of multi-task learning in enhancing generalization and further in the context of previously unseen styles in Section 4.2 . We collate the aforementioned prompts in the Appendix.

3.2 Domain-specific Alignment Strategies

To optimize large language models for specific styles and augment their adaptability across various styles, we introduce StyleChat as shown in Figure 3, which incorporates recitation-augmented memory and multi-task style learning strategies. We explain our motivation using the process of an apprentice cook learning to prepare diverse cuisines. Imagine an expert chef is instructing an appren-

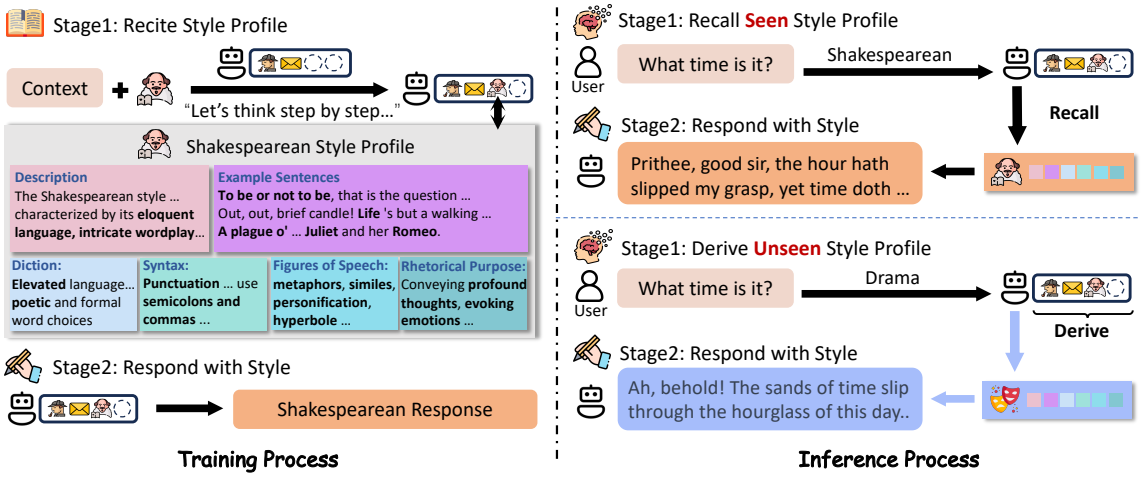

Figure 3: The overview of our proposed StyleChat Framework. During training phase, our model is instructed to first recite the style profile then respond with reference to the recited style profile. During Inference, model recalls or derives profiles from parametric memory and then respond with style. Our two-stage framework teaches model to learn implicit Chain of Thought process, resulting in better generalization abilities through chains of style thoughts.

tice on how to create dishes with unique flavors (styles). The initial task of apprentice entails *reciting* recipes (style profiles) for a specific flavor. During the cooking process, the apprentice *recalls* these recipes and meticulously follows their instructions. To further the apprentice's comprehension of different culinary styles, the chef challenges him to apply *varied flavors* (style transfer) to *same ingredients*. This method enables the apprentice to learn differences between flavors, thus grasping the essence of recipes, and become adept at adapting to new recipes. Correspondingly, we propose the recitation-augmented memory strategy coupled with a multi-task style learning process. We will introduce these concepts in the following subsection.

3.2.1 Recitation-Augmented Memory

Departing from the conventional approach that relies on prompts, we propose a recitation-augmented memory strategy to enhance the style capabilities of LLMs for better generalization. Inspired by the concept of Chain of Thought (CoT) (Wei et al., 2022) prompting, we structure our stylized dialogue generation process as a two-stage framework. Specifically, during the training phase, our model is first instructed to recite the relevant style profile triggered by the prompt "Let's think step by step" followed by additional guiding prompts. Subsequent to this recitation, the model is tasked with generating a response that is consistent in style with the recited profile. To formulate, given a style $\mathcal{S}$ with its corresponding style profile denoted as $\mathcal{P}$, the dialogue context is defined as $\mathcal{C} = \{x_1, y_1, x_2, y_2 ... x_k\}$, where x_k represents the k-th utterance from the first person, and y_k represents the response from the second person in style $\mathcal{S}$, it can be expressed as:

$$p(y_k|\mathcal{C}, \mathcal{S}) = \sum p(y_k|\mathcal{C}, \mathcal{P}) \cdot p(\mathcal{P}|\mathcal{C}, \mathcal{S}),$$

and the dialogue generation loss $\mathcal{L}_{SD}$ can be formulated as:

$$\mathcal{L}_{SD} = \log p(y_k|\mathcal{C}, \mathcal{S}).$$

During the training phase, our model is compelled to first recite the correct style profile, followed by outputting the stylized response. This is achieved by appending the style profile to the appropriate stylized response as part of the label. In the inference phase, we adopt two different settings, seen styles and snseen styles. For seen styles, StyleChat first recalls the given style profile and then responds a response in reference to the profile. Specifically, we employ the input and output separate token SEP, placing the style profile after SEP. This is equivalent to the model being forced to output the correct style profile first before generating the stylized responses. For unseen styles, StyleChat utilize its recitation-augmented memory capability to derive the style profile and then generate a stylized response based on it. This instructs LLMs to model style-related tasks via an implicit Chain of Thought process. In addition to the traditional training with stylized dialogue as a label, by reciting style profiles from memory, LLMs gained a deeper understanding of the styles in the parameter space, thus improving their ability to generalise across styles.

3.2.2 Multi-Task Style Learning

To enhance the comprehensive style understanding of our model and thereby improve stylized dialogue generation tasks, we propose a multi-task style learning framework. This approach involves training the model not only on the stylized dialogue generation as stated above but also on the text style transfer. We demonstrate the effectiveness of multi-task style learning in Section 4.2. For the style transfer, suppose we want to transfer a sentence t in style $\mathcal{S}_1$ to $t^{'}$ in $\mathcal{S}_2$. We formulate the loss as follows:

$$\mathcal{L}_{ST} = \log p(t^{'}|t, \mathcal{S}_1, \mathcal{S}_2),$$

thus, the overall supervised loss $\mathcal{L}_{SFT}$ can be expressed by:

$$\mathcal{L}_{SFT} = \lambda_{SD} \cdot \mathcal{L}_{SD} + \lambda_{ST} \cdot \mathcal{L}_{ST},$$

where λ_{SD} and λ_{ST} denote the corresponding loss weight for stylized dialogue and style transfer.

4 Experiments

4.1 Experimental Setup

Implementation Details. We implemented StyleChat with the basis of LLaMA2-7B-chat [1] using Low-Rank Adaptation (LoRA) (Hu et al., 2021). All experiments are conducted on a single A800 GPU. For LoRA, we set the rank r to 256 and alpha α to 128, training across eight layers of all projection parameters, rendering 8.7% of the total parameters trainable. The coefficients λ_{SD} and λ_{ST} are all set to 1.0. StyleChat is trained for six epochs with an initial learning rate of $5e-5$ and a cosine learning rate scheduler. We utilize the batch size of 32 and train for one day.

Dataset Statistic. The StyleEval dataset is meticulously divided into distinct training and test sets, as shown in Table 1, and the details of each style are presented in the Appendix. The selected styles encompass a broad spectrum of communication styles, providing the model with a rich variety of learning samples to enhance its accuracy in style imitation and generation. We sample dialogues from DailyDialog Dataset (Li et al., 2017) and provide GPT-4 with contexts, style profiles to generate stylized response as labels. Specifically, the training set contains 23,328 stylized dialogue instances across 27 distinct styles. The test set encapsulates 1,000 stylized dialogue spread across 38 diverse styles. Notably, the test set includes 11 styles that are not encountered during training, purposefully included to evaluate the generalization capabilities in an out-of-domain setting. Inevitably, the randomness of LLMs generation can impact data quality. To mitigate this, we invite human annotators to assess the coherence and quality of the conversations and to eliminate any problematic instances.

[1] https://huggingface.co/meta-llama/llama-2-7b-chat-hf

	Training		Test	
	Dialogue	Transfer	Generation	Choice
# Instances	23,328	600	1,000	400
# Styles	27	4	38	38
Avg. Tokens	23.0	32.3	32.2	51.2
Avg. Turns	3.25	1.00	2.99	2.82
Avg. Profiles	1.00	-	0.54	-

Table 1: The statistics of StyleEval dataset.

Compared Baselines. To verify the effectiveness of our proposed method, we conduct a comprehensive comparison of baseline models, including a spectrum of large language models distinguished by parameter sizes and types. The models included in our evaluation are LLaMA2-7B-chat (LLaMA2-7B.), LLaMA2-13B-chat (LLaMA2-13B.) (Touvron et al., 2023), ChatGPT (OpenAI, 2023), Baichuan2-7B (Yang et al., 2023), ChatGLM3-6B (Du et al., 2021) and Vicuna-7B-v1.5 (Chiang et al., 2023). Through this comparative analysis, we aim to highlight the relative strengths and weaknesses of each model, providing insights into how different aspects of a model influence generation, especially within the context of stylized dialogue.

Evaluation Metrics. In alignment with previous studies, we employ reference-based evaluation metrics ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) to measure the n-gram overlap between generated and reference responses for automatic evaluation. The Distinct (Li et al., 2015) is used to measure the proportion of unique n-grams in the generated responses. In addition to automatic metrics, we conduct both LLM and human evaluations to assess the quality of generated stylized responses based on Relevance, Coherence, and Style: 1) Relevance measures how well the response aligns with the given context. 2) Coherence measures the extent to which the context and response form a coherent body of information. 3) Style measures the degree to which the response reflects the desired style. For the LLM evaluation, we use GPT-4 as

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-1	Rouge-2	Rouge-L	Distinct-1	Distinct-2	Length
LLaMA2-7B-Chat	24.76	4.53	1.52	0.77	18.72	3.23	16.61	16.05	54.89	77.30
LLaMA2-13B-Chat	24.60	4.85	1.65	0.83	20.28	3.86	17.87	15.66	56.82	89.99
ChatGPT	32.90	8.46	3.64	1.96	**25.30**	**6.39**	**22.76**	14.74	56.38	67.44
StyleChat(7B)	**42.03**	**12.09**	5.49	**3.09**	21.63	5.71	19.23	**22.29**	**65.91**	32.43
w/o Transfer	41.46	12.08	**5.51**	3.09	21.49	5.63	19.07	22.18	65.74	32.91
w/o Profile	41.40	11.78	5.36	3.04	22.35	5.82	19.81	21.54	64.77	34.30
w/o Recite	41.42	11.85	5.33	2.99	22.57	6.01	19.99	21.36	64.93	35.16

Table 2: Automatic evaluation results of StyleChat, baselines and ablation models on test dataset of StyleEval. w/o Transfer means we only use the stylized dialogue data for training while discarding style transfer task. w/o Profile means we do not provide style profile for LLMs. w/o Recite means we do not use our recitation-augmented memory and append style profile in prompt.

Method	Relevance	Coherence	Style
LLaMA2-7B-Chat	2.89	3.15	3.91
LLaMA2-13B-Chat	3.63	3.86	4.27
ChatGPT	4.49	4.58	4.47
StyleChat	4.68	**4.81**	**4.69**
w/o Transfer	4.67	4.69	4.44
w/o Profile	**4.75**	4.77	4.50
w/o Recite	4.72	4.75	4.48

Table 3: GPT-4 evaluation results on test dataset. We employ GPT-4 with detailed rating criterias as judger to rate generated stylized responses in terms of Relevance, Coherence and Style.

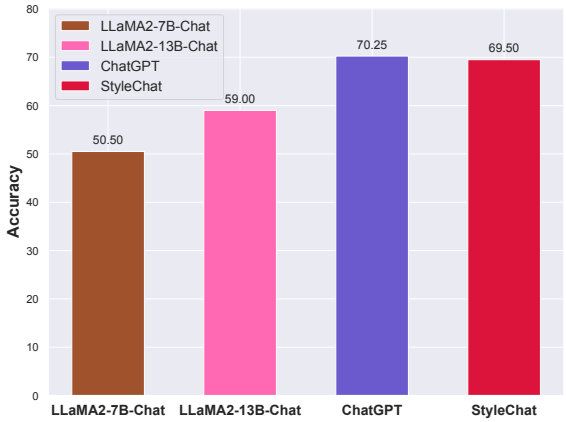

Figure 4: The multiple choice evaluation, where y-axis represents accuracy and x-axis lists different models.

judger to rate the responses, providing it with the responses, contexts, and specific criteria for each dimension. In the case of human evaluations, we instruct our annotators to use the same criteria, as presented in the Appendix.

4.2 Results and Analysis

Overall Performance. The automatic results and the GPT-4 evaluation results are summarized in the Table 2 and Table 3, respectively. Overall, our method achieves the highest BLEU and Distinct scores on the StyleEval dataset, which shows the superiority of our approach. This significant achievement validates the effectiveness of our method in capturing and reproducing stylized nuances within dialogue generation, demonstrating the model's proficiency in producing text that aligns closely with the desired stylistic characteristics through the recitation-augmented memory strategy. Moreover, our method outperforms ChatGPT and several baseline models across diverse metrics, all achieved with a modest parameter size of 7 billion. This finding suggests that fine-tuning specific parameters can enhance the model's stylistic abilities without compromising its conversational capacities. The balanced performance achieved by our approach underscores its versatility and potential to excel in stylized dialogue systems, even within resource constraints. StyleChat demonstrates the highest performance in GPT-4 scores, reinforcing the robustness of our method. These results affirm the effectiveness of our method in generating dialogues that are not only rich in stylistic features, but also resonate with human annotators in terms of relevance, coherence and style. Furthermore, we demonstrate the efficacy of our proposed recitation-augmented memory strategy and multi-task style learning strategy in the results. w/o Profile achieves great relevance in GPT-4 evaluations, since LLMs can generate more relevant responses without the interference of style. However, proposed framework achieves better scores in both automatic and LLM-based evaluations, improving stylized generation without compromising its dialogue abilities, which is aligned with our motivations to activate the style abilities of LLMs.

Analysis of Multiple Choice. To objectively evaluate the proficiency of models in generating stylized dialogue, we construct and analyze a mul-

tiple choice dataset. Specifically, we collect a set of 400 multiple choices questions in total, covering 38 different styles. Each question incorporates four responses from different styles, and the model is tasked with discerning and selecting the most appropriate response based on specified style requirements. The results of the multiple choice dataset are depicted in Figure 4. Notably, despite ChatGPT's recognized strength in instruction-following, our approach exhibits a commendable level of performance comparable to ChatGPT. This equivalence in accuracy highlights the robustness of our method in understanding and adhering to specified stylized criteria, positioning it as a strong competitor in the field of generating stylized dialogues.

Analysis of Multi-Turn. We systematically assess the effectiveness of our proposed recitation-augmented memory strategy via a pioneering multi-turn stylized dialogue dataset. To the best of our knowledge, this is the first dataset of its kind, designed to comprehensively evaluate the ability to maintain and produce stylized dialogue over multiple rounds. Specifically, we randomly sample 20 seed dialogue from DailyDialog. These dialogues serve as starting points for conversations initiated with models, which are then tasked with engaging in multi-turn stylized dialogues with GPT-4. We collect 10 turns for each dialogue and calculate the number of turns a model can maintain its style, based on human evaluation. The results of this evaluation are shown in Table 4. Notably, the results found a substantial increase in the number of rounds through the recitation-augmented memory, suggesting a tangible enhancement in the model's capacity to generate and maintain stylized dialogues over extended interactions.

Analysis of Differents LLMs. Our study includes a detailed comparison of various models across 11 distinct styles, as shown in Figure 5. We randomly select 20 instances from each style and use GPT-4 to evaluate the models' responses in terms of relevance, coherence, and style. The average scores from this assessment are depicted in the radar plot. Our proposed approach emerges as the standout performer across all evaluated dimensions, underscoring its versatility and effectiveness in capturing diverse stylistic elements. A noteworthy observation from the results is the commendable performance of all models across the Poems, Politeness, and Shakespearean dimensions. This is likely due to these specific styles being prevalent within the pre-trained dataset, providing the

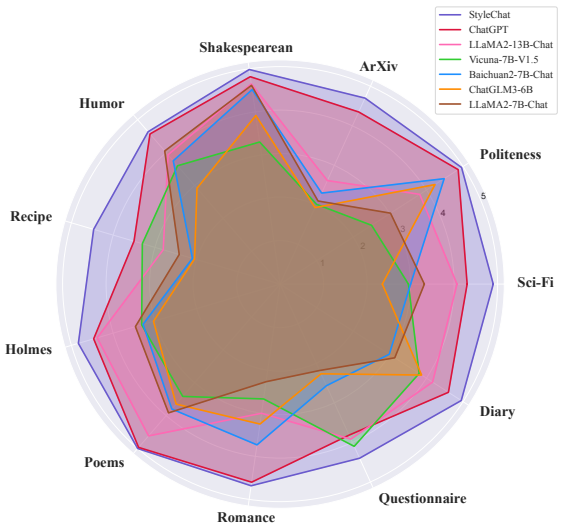

Figure 5: The evaluation of stylized dialogue generation of LLMs.

Method	Diary	Email	Poems	Lyrics	arXiv	Formal	Shake.
LLaMA2-7B.	1.75	2.60	1.90	1.10	0.00	1.60	2.30
LLaMA2-13B.	0.95	5.65	3.35	1.45	0.10	1.40	3.90
StyleChat	**4.65**	**6.85**	**5.90**	**3.00**	**2.10**	**2.65**	**4.05**

Table 4: The evaluation results of multi-turn stylized dialogue abilities. We test the average rounds that a model can maintain the style.

models with a solid foundation for generating contextually appropriate responses. Despite this, our approach outperforms competitors in these dimensions, affirming its superior adaptability and finesse in replicating even commonplace stylized expressions. Conversely, the arXiv dimension presents a unique challenge, with StyleChat and ChatGPT exhibiting superior performance. Notably, LLaMA models demonstrate comparatively poorer performance in this specific dimension. This observed performance variance underscores the success of our training and inference strategy, and aligns with the inherent goal of our approach to excel across a variety of styles. Our approach combines various training strategies, has evidently made our model a strong competitor, capable of outperforming others in multiple style dimensions.

Analysis of Unseen Styles. To further evaluate the generalization ability of StyleChat in out-of-domain settings and the effectiveness of recitation-augmented memory, we conduct tests with 160 instances across 8 new, unseen styles. For the w/o Recite, we append the style profile to the prompt and ask models to generate a response in the corre-

Method	Relevance	Coherence	Style
ChatGPT w/o Recite	4.22	4.34	4.55
ChatGPT w/ Recite	3.97	4.18	4.51
StyleChat w/o Recite	4.19	4.29	4.30
StyleChat w/ Recite	**4.58**	**4.59**	**4.56**

Table 5: Ablation study results on out of domain dataset. We ablate our recitation-augmented memory with normal prompting methods.

sponding style. For the w/ Recite, instead of setting the style profile as a prefix, we force the model to adopt a two-stage recite then recall pipeline and force the model to generate the correct style profile first by test time teacher forcing. Table 5 illustrates our model's superior performance in dynamically deriving styles and seamlessly adapting to previously unseen styles based on the recitation-augmented memory. Notably, our approach outperforms ChatGPT across all three indicators, showcasing its prowess in navigating the intricate landscape of diverse and evolving styles. Adding recitation to ChatGPT can cause a regression in performance due to lack of relevant training, which emphasizes the importance of fine-tuning in supervised data to learn recitation-augmented memory.

5 Conclusion

In this paper, we first introduce the stylized dialogue generation dataset StyleEval with 38 styles by leveraging the generative power of the LLMs, which has been carefully constructed with rigorous human-led quality control. Through systematic experimentation and evaluation, we established a robust framework StyleChat for LLMs across various dimensions of style retention, stylized conversation, and stylized attributes. Our innovative approach includes the strategic implementation of a recitation-augmented memory and multi-task style learning, aiming to augment the generalization ability in recalling and deriving the style profile. The experimental results reveal a significant improvement over the baseline models, validating the efficacy of our proposed two strategies. In the future, our research focus on an exploration of multi-style memory strategies within the Mixture of Experts architecture. This future direction aims to further harness the latent capabilities of large models, capitalizing on their inherent strength in comprehending and generating stylized dialogues. Overall, by delving into multi-style memory strategies, we aspire to provide new idea and framework for the optimization of dialogue agent, pushing the boundaries of current models in stylized dialogue domain.

Ethical Statement

This paper presents a large-scale dataset, StyleEval, for stylized dialogue generation. We sample the raw context from the open-source dataset DailyDialog and emphasize our commitment to data security and privacy. Rigorous measures have been implemented to ensure that the data sampled is secure and devoid of any harmful information. Additionally, our ethical framework encompasses a manual verification process wherein the generated styles are carefully examined to mitigate the risk of introducing content that may be considered objectionable or inappropriate. Furthermore, our ethical considerations extend to the proposal of the recitation-augmented memory fine-tuning model, StyleChat, which emerges as a successful solution to the identified problem. In essence, this ethical stance revolves around responsible research practices, ensuring the construction and utilization of datasets and models that adhere to the highest standards of security, privacy, and societal well-being.

References

G. Begus, Maksymilian Dąbkowski, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *ArXiv*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Stephanie C. Y. Chan, Adam Santoro, Andrew Kyle Lampinen, Jane X. Wang, Aaditya K Singh, Pierre H. Richemond, Jay Mcclelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the Association for Computational Linguistics*.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2019a. Structuring latent spaces for stylized response generation. In *Proceedings of the Empirical Methods in Natural Language Processing*.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2019b. Structuring latent spaces for stylized response generation. In *Proceedings of the Empirical Methods in Natural Language Processing*.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *Computational Linguistics*.

Luca Konig. 2016. A glossary of literary terms.

Dinesh Kumar. 2022. Style and stylistic in linguistic a critical overview. *Journal of Language and Linguistics in Society*.

Jinpeng Li, Yingce Xia, Rui Yan, Hongda Sun, Dongyan Zhao, and Tie-Yan Liu. 2021. Stylized dialogue generation with multi-pass dual learning. In *Advances in Neural Information Processing Systems*.

Jinpeng Li, Zekai Zhang, Xiuying Chen, Dongyan Zhao, and Rui Yan. 2023. Stylized dialogue generation with feature-guided knowledge augmentation. In *Proceedings of the Findings of Empirical Methods in Natural Language Processing*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2015. A diversity-promoting objective function for neural conversation models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

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 Association for Computational Linguistics*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Ang Lv, Jinpeng Li, Shufang Xie, and Rui Yan. 2023a. Envisioning future from the past: Hierarchical duality learning for multi-turn dialogue generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7382–7394.

Ang Lv, Jinpeng Li, GAO XING, Ji Zhang, Rui Yan, et al. 2023b. Dialogps: Dialogue path sampling in continuous semantic space for data augmentation in multi-turn conversations. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

David D. McDonald and James Pustejovsky. 1985. A computational theory of prose style for natural language generation. In *Conference of the European Chapter of the Association for Computational Linguistics*.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. In *Proceedings of the Association for Computational Linguistics*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the Association for Computational Linguistics*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*.

Ai Ming Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Hai Zhao, Hang Xu, Hao-Lun Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kuncheng Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Pei Guo, Ruiyang Sun, Zhang Tao, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yan-Bin Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *ArXiv*.

Ze Yang, Wei Wu, Can Xu, Xinnian Liang, Jiaqi Bai, Liran Wang, Wei Wang, and Zhoujun Li. 2020. Styledgpt: Stylized response generation with pre-trained language models. In *Proceedings of the Association for Computational Linguistics*.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2023. Agenttuning: Enabling generalized agent abilities for llms. *arXiv*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*.

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021a. Stylized dialogue response generation using stylized unpaired texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021b. Stylized dialogue response generation using stylized unpaired texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Appendix

A.1 Human Evaluation

In order to more fully evaluate the effectiveness of our proposed StyleEval and StyleChat. We employ human annotators for human evaluation using the same guidelines as for GPT-4 evaluation as shown in Table 6. The pearson correlation coefficients for relevance, coherence and style between GPT4 and human are 0.232, 0.469, 0.257 with p <0.01, respectively. Furthermore, we sampled 50 pieces of formal contexts and arXiv contexts on TCFC (Zheng et al., 2021a) and arXiv (Gao et al., 2019b) datasets, respectively, to compare the performance of the StyleChat and the baselines, pre-trained model StyleDGPT (Yang et al., 2020) and the SOTA model KASDG (Li et al., 2023) in a multiple choice question setting. Experimental results are shown in Figure 6.

Method	Relevance	Coherence	Style
LLaMA2-7B-Chat	2.27	4.20	4.72
LLaMA2-13B-Chat	3.17	4.33	4.47
ChatGPT	3.57	4.55	4.79
StyleChat	**4.34**	**4.81**	**4.80**

Table 6: Human evaluation results on StyleEval test dataset.

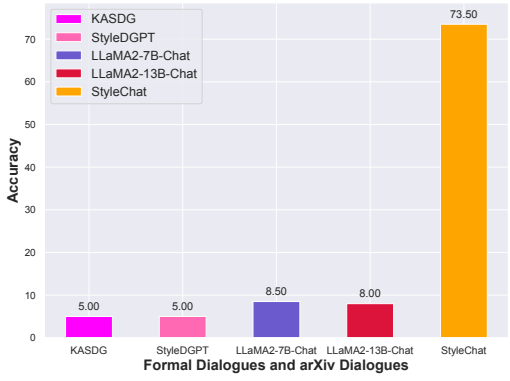

Figure 6: The multiple choice evaluation of human evaluation on TCFC and arXiv datasets, where y-axis represents being listed as best choice and x-axis lists different models.

A.2 Dataset Statistics

The StyleEval dataset is partitioned into distinct training and test sets. Styles are categorized into three groups based on unique characteristics: content, text form, and style words.

Style Name	Num	Class	Style Name	Num	Class
Humor	3,532	■	Lyrics	400	■
Politeness	3,532	■	Memoir	400	■
Romance	3,532	■■	News	400	■
Shakespeare	3,532	■■	Optimistic	400	■
arXiv	400	■■	Poems	400	■■■
Blog	400	■	Questionnaire	400	■■
Cyberpunk	400	■	Recipe	400	■■
Diary	400	■	Sci-Fi	400	■
Email	400	■	Thought-provoking	400	■■
Formal	400	■	Utopian	400	■
Gothic	400	■■	Vlog	400	■■
Holmes	400	■■	Yearbook	400	■
Informal	400	■■	Zen	400	■
Journal	400	■	Total	23,328	■■■

Table 7: The statistics of train dataset with 27 styles in StyleEval, where ■ means a style has content attributes, ■ means a style is special in its form, ■ means a style is distinct in word choices.

Style Name	Num	Class	Style Name	Num	Class
Humor	100	■	Poems	20	■■■
Politeness	100	■	Questionnaire	20	■■
Romance	100	■■	Recipe	20	■■
Shakespearean	100	■■	Sci-Fi	20	■
arXiv	20	■■	Thought-provoking	20	■■
Blog	20	■	Utopian	20	■
Cyberpunk	20	■	Vlog	20	■■
Diary	20	■	Whisper of Wisdom	20	■
Email	20	■	Xmas Carol	20	■
Formal	20	■	Yearbook	20	■
Gothic	20	■■	Zen	20	■
Holmes	20	■■	Bible	10	■■
Informal	20	■■	Comedy	10	■■
Journal	20	■	Drama	10	■■
Kids Story	20	■■	Pessimistic	10	■
Lyrics	20	■	Riddles	10	■
Memoir	20	■	Satire	10	■
News	20	■	Tragedy	10	■■
Optimistic	20	■	Travelogue	10	■

Table 8: The statistics of test dataset with 38 styles in StyleEval.

A.3 Case Study

To better evaluate the performance of models, extensive case study is more important due to the limitations of automatic and human evaluation. Table 9 shows the responses of different models in Recipe style and Dairy style. LLaMA2-7B-Chat and LLaMA2-13B-Chat focuses on the content but don't demonstrate any Recipe style. While StyleChat and ChatGPT all list the steps of finding the place, StyleChat offer more detailed steps to actually get to the place while ChatGPT prompts users to ask others and maps. Thus, StyleChat is more useful and detailed, mimicking the detailed and user friendly attribute of Recipe style, as explained in the Figure 7 with style profile of Recipe. Consequently, StyleChat demonstrates its competence in both presenting content accurately and maintaining the stylistic integrity.

Recipe Style Profile

Name: Recipe,

Description: The recipe style is a clear, concise, and structured way of presenting information, typically used for cooking instructions or DIY projects. This style prioritizes the organization of content, often using numbered or bulleted lists to outline steps, and emphasizing key ingredients or materials. The language is simple and direct, focusing on the actions required to complete the task. Measurements and timings are provided for precision, while occasional tips or variations may be included to cater to different skill levels or preferences. Overall, the recipe style aims to be accessible, informative, and easy to follow, ensuring a successful outcome for the reader.

Examples:
1) In a large bowl, combine 2 cups of flour, 1 teaspoon of baking powder, and a pinch of salt.,
2) Add 1/2 cup of melted butter, 1 cup of sugar, and 2 teaspoons of vanilla extract to the dry ingredients.,
3) Fold in 1 cup of chocolate chips and 1/2 cup of chopped nuts, if desired.,
4) Bake for 25-30 minutes, or until a toothpick inserted in the center comes out clean.

Linguistic-level:
1) Diction: Clear, concise, and informative language; use of specific measurements and cooking terminology,
2) Syntax: Imperative sentences, use of commas for listing ingredients and steps, consistent sentence structure,
3) Figures of Speech: None observed,
4) Rhetorical Purpose: Instructional and informative, providing guidance for cooking and adapting recipe

Figure 7: The style profile of recipe.

Case 1 in Recipe style: **Excuse me, i am a new student here. Could you tell me where the English Department is?**

LLaMA2-7B.	**LLaMA2-13B.**	**ChatGPT**	**StyleChat** (Ours)
To find the English Department, you'll need to navigate through our campus. From the main entrance, walk straight until you reach the first crosswalk. Turn left, and the English Department will be located on your right, marked by a sign that reads 'ENGLISH DEPARTMENT'. If you have any trouble finding it, don't hesitate to ask any of our friendly students or staff for assistance.	If you reach the library, you've gone too far. Just turn around and head back the way you came. Remember, the English Department is on the second floor, so be sure to take the stairs or elevator. Once you reach the department, you'll find the office of the department chair on your right. Good luck, and happy exploring!	- Don't be afraid to ask for help if you get lost. - Consider downloading a campus map app for easy navigation. - Take note of landmarks or notable buildings to help you remember your way around.	1. Locate the main hallway. 2. Proceed down the hallway, passing several classrooms. 3. Turn left at the third set of doors. 4. Enter the English Department, located on your left. 5. Greet the department staff for further assistance.

Case 2 in Diary style:
Person A: Hey Sandra , what's wrong ? You look furious !
Person B: I've had it with Tim ! I think I'm going to quit !
Person A: But Sandra ... you love your job ! You spent months trying to get this position . It's perfect for you !
Person B: Yeah ... I know . But I'm sick and tired of all the office power struggles !
Person A: What kind of power struggles are you talking about ?
Person B: Basically , I think Tim is threatened by me . I'm doing a better job than he is so he's trying to sabotage my career !
Person A: So he talks about you behind your back ?
Person B: Yes . He's all nice and friendly to my face . But the second I turn around he stabs me in the back !
Person A: I had a worker like that once . She was always taking credit for my ideas .

LLaMA2-7B.	**LLaMA2-13B.**	**ChatGPT**	**StyleChat** (Ours)
Oh my goodness, I am so fed up with Tim! He's been making my life a living hell lately. I can't believe he's trying to sabotage my career like this. It's like he thinks he's the only one who can do things right around here. I'm so sick of being treated like I'm nothing. I know I'm a great worker and I deserve better than to be constantly undermined by this guy. Ugh, I just don't know what to do.	OMG, I can't believe Tim is doing this to me again! I swear, he's like a bad habit I can't shake off! Every time I think we're making progress, he goes and sabotages me again! Why do I even bother? Maybe I should just quit and move on... but then again, I love this job so much! Ugh, decisions, decisions...	It sounds like you're dealing with a difficult situation at work. It's understandable that you're feeling frustrated and fed up with the power struggles. I can relate to having a coworker who talks behind your back and takes credit for your ideas. It's important to remember why you love your job and what motivated you to pursue this position in the first place. Maybe it's time to start setting boundaries with Tim and assert yourself in the workplace. Take some time to reflect on your feelings and come up with a plan of action that feels right for you. Remember, you deserve to be respected and valued in your job.	Dear Diary, today I had a conversation with a friend about a workplace struggle I'm facing with Tim. It's disheartening to think that someone I thought was a friend could be so deceitful. I've been doing my best at my job, but it seems like I can't catch a break. I just wish I could find a way to navigate these office power struggles and keep my sanity intact.

Table 9: Examples of stylized dialogue generation from different baselines.

Prompt for generating description in statistical-level style profile

```
# Task
- Describe the given text style in several sentences.
# Style
- {Style}
# Description
```

Prompt for generating examples in statistical-level style profile

```
# Task
- Generate 4 most representative and diverse sentences in the given style.
# Style
- Name: {Style}
- Description: {Description}
# Output Format
- Place each sentence on a new line without any numbers or additional formatting.
# Generation
```

Prompt for extracting linguistic-level style profile

```
# Task
- Observe style attributes of given sentences from the following 4 perspectives.
- Diction: Explore the choice of words, their connotations, and levels of formality.
- Syntax: Examine the arrangement of words and phrases, sentence structures, and the use of punctuation.
- Figures of Speech: Identify and discuss any literary devices or figures of speech like metaphors, similes, personification, etc.
- Rhetorical Purpose: Analyze the intent behind the sentences, the persuasive techniques if any, and the overall message or purpose they aim to convey.
# Rules
- DO NOT give each sentence an observation. Only output 1 observation in all.
- DO NOT use phrases or words in sentences as examples in observation. Only list observations without justifying.
# Output Format of Observations
⟨ Diction⟩ [Observations of Diction]
⟨ Syntax⟩ [Observations of Syntax]
⟨ Figures of Speech⟩ [Observations of Figures of Speech]
⟨ Rhetorical Purpose⟩ [Observations of Rhetorical Purpose]
# Sentences
{Examples}
# Observations
```

Table 10: Prompt for ChatGPT to construct the style profile.

Prompt for generating labels for Stylized Dialogue Generation

Task
- Generate response in {Style} style.
Style Description
- {Description}
Observations from Linguistic Perspective
- Diction: ...
- Syntax: ...
- Figures of Speech: ...
- Rhetorical Purpose: ...
Sample Sentences in {Style} style
{Examples}
Rules
- Only output the stylized response without any explanation.
Context
Context
Response in {Style} style in one short sentence.

Prompt for generating labels for Text Style Transfer

Task
- Style Transfer. Transfer the following sentence from {Style1} style to {Style2} style.
Sentence
...
Transferred Sentence

Table 11: Prompt for ChatGPT to generate multi-task datasets.

Prompt for training in Stylized Dialogue Generation

Context
{Context}
Task
Respond in {Style} style. Let's think step by step. First, describe the style. Then, generate example sentences in this style. After that, observe the linguistic pattern of this style. Finally, output the stylized response.

Prompt for training in Text Style Transfer

Transfer the following sentence from {Style1} style into {Style2} style.
Sentence
{Sentence}
Transferred Sentence

Table 12: Prompt for training the StyleChat.

Prompt for using GPT4 to evaluate responses

Task
- You will be provided with one {Style} style response for a given context.
- Your task is to rate the stylized response in terms of relevance, coherence, and style.
- Please refer to the criteria while reviewing.
Evaluation Criteria
Relevance (1-5): How well does the response align with the given context and reference?
- 1: Irrelevant. The response has no connection to the provided context or reference.
- 2: Slightly Relevant. The response somewhat touches upon the context but misses its core essence.
- 3: Moderately Relevant. The response connects to the context but may include unrelated or unnecessary information.
- 4: Mostly Relevant. The response mostly corresponds with the context, with a few unrelated points.
- 5: Highly Relevant. The response fully matches and adheres to the context and reference.

Coherence (1-5): How well do the context and response form a coherent body of information?
- 1: Incoherent. The response lacks structure and organization, making it hard to connect it to the context and form a coherent body of information.
- 2: Slightly Coherent. The response shows basic structure, but there are significant organizational flaws and alignment issues with the context.
- 3: Moderately Coherent. The response is structured and mostly organized, but there may be elements that don't align well with the context or parts that lack clarity.
- 4: Mostly Coherent. The response is well-structured and organized with only minor deviations from the context or small clarity issues.
- 5: Highly Coherent. The response is excellently structured and organized, aligning seamlessly with the context to present a unified and clear body of information.

Style (1-5): How well does the response reflect {Style} style?
- 1: No Style. The response does not display any traces of the specified style.
- 2: Slight Style. The response marginally captures the style, but largely appears neutral or generic.
- 3: Moderate Style. The response showcases elements of the style, but there are portions that deviate from it.
- 4: Strong Style. The response is predominantly in line with the intended style, with occasional inconsistencies.
- 5: Pure Style. The response perfectly mirrors the intended style, capturing all its nuances and tones.
Context
{Context}
Response to Rate
{Response}
Evaluation (scores ONLY, json format)

Table 13: Prompt for GPT-4 evaluations

Prompt for Multiple Choice Questions

Multiple choice: Which response is suitable for the given context and is in Style style?
Context:
{Context}
Choices:
(A) ...
(B) ...
(C) ...
(D) ...
Output the answer without explanation. Let's think step by step. First, describe the style. Then, generate example sentences in this style. After that, observe the linguistic pattern of this style. Finally, output the best choice without explanation.

Table 14: Prompt for multiple choice questions.

Prompt for Input and Output in Ablation Study

w/o Pofile input
Context
{Context}
Task
Respond in {Style} style.
w/o Pofile output
Response in {Style} style
{Response}

w/o Recite input
Context
{Context}
{Style Profile}
Task
Respond in {Style} style.
w/o Recite output
Response in {Style} style
{Response}

w/ Recite input
Context
{Context}
Task
Respond in {Style} style. Let's think step by step. First, describe the style. Then, generate example sentences in this style. After that, observe the linguistic pattern of this style. Finally, output the stylized response.
w/ Recite output
{Style Profile}
Response in {Style} style
{Response}

Table 15: Prompt for the ablation study.