Colloquial Singaporean English Style Transfer with Fine-Grained Explainable Control

Anonymous ARR submission

Abstract

Colloquial Singaporean English, better known as Singlish, is a distinctive blend of languages that reflects Singapore's multicultural identity. Style transfer between Singlish and Standard English is important for various applications, yet existing approaches often lack fine-grained control and explainability. To address this gap, we contribute in two key ways. First, we construct a large, high-quality dataset of formal and informal sentences, annotated across six linguistic aspects—Syntax, Lexical Borrowing, Pragmatics, Prosody/Phonology, Emoticons/Punctuation, and Code-Switching—with detailed explanations. Starting with manually annotated cases, we scaled the dataset to 140K with ensured quality. Second, inspired by the "Society of Mind" theory, we propose a novel multi-agent framework where large language models (LLMs) act as expert agents for each linguistic aspect. These agents collaborate by iteratively generating, critiquing, and refining responses to achieve controlled, explainable style transfer. Both automatic metrics and human evaluations confirm that our approach enables precise, interpretable transformations, advancing explainability in NLP for Singlish[1].

1 Introduction

Colloquial Singaporean English, or Singlish, is a unique linguistic blend shaped by Singapore's multicultural heritage, incorporating elements from English, Malay, Tamil, and Chinese dialects (Wang et al., 2017; Foo et al., 2024). While Singlish is common in informal contexts, Standard English dominates formal communication (Yunick, 1995; Bajpai et al., 2016). Effective style transfer between these two forms is crucial for applications such as education, cross-cultural communication, and content localization (Liu et al., 2022). However, the complex structure of Singlish, with its blend of syntax, vocabulary, and code-switching *etc.*, presents challenges (Chow and Bond, 2022; Pham et al., 2024). Fine-grained control over this transfer is essential to ensure accurate and context-sensitive transformations.

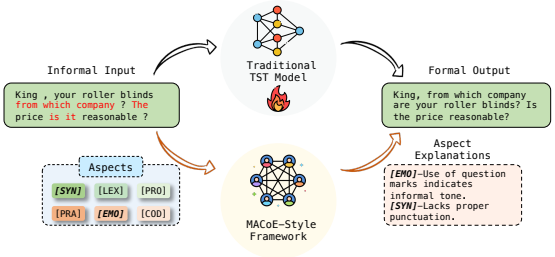

Figure 1: Comparison between traditional TST and our fine-grained controllable and explainable TST.

Existing style transfer methods fall into three categories: parallel supervised, non-parallel supervised, and unsupervised (Mukherjee et al., 2024). Parallel supervised approaches use paired texts in different styles for direct transformations but are limited by the lack of high-quality parallel datasets (Xu et al., 2012; Rao and Tetreault, 2018a). Non-parallel supervised methods employ signals like style labels and adversarial learning to guide transfer without paired data but often struggle with fine-grained control (John et al., 2019). Unsupervised methods, such as those using cycle consistency and disentangled representations (Gatys et al., 2016; Chen et al., 2016), separate content from style without labeled data but typically lack explainability. Across all three approaches, the main challenges remain: (1) limited explainability and (2) insufficient fine-grained control over specific linguistic aspects during the transformation.

Our goal is to address the limitations of existing methods by developing an approach that ensures both explainability and fine-grained control over style transfer for Colloquial Singlish. This is particularly challenging due to the absence of large-scale datasets annotated with linguistic explanations, as well as the lack of methods capable of handling such fine-grained stylistic transforma-

[1] https://anonymous.4open.science/r/colloquial_tst-D1BB

tions. Singlish, with its complex interplay of syntax, lexical borrowing, and code-switching *etc.*, poses additional difficulties for models typically designed for simpler, more homogenous language pairs. Therefore, a comprehensive framework is needed to manage the intricacies of this colloquial variant and provide interpretable transformations.

To tackle these challenges, we propose two key contributions. First, we construct a large dataset of formal and informal sentences with annotations across six linguistic aspects: Syntax, Lexical Borrowing, Pragmatics, Prosody/Phonology, Emoticons/Punctuation, and Code-Switching. This dataset is designed to provide explanations for style differences, enabling models to learn not only how to perform style transfer but also why certain stylistic choices are made. Second, we introduce a multi-agent framework inspired by the Natural Language-based "Society of Mind" (NLSOM) theory (Zhuge et al., 2023; Hong et al., 2024), where multiple large language models (LLMs) act as specialized agents, each responsible for one aspect of the style transfer. These agents collaborate by iteratively generating and refining their transformations, producing a final output that is both controlled and explainable. Through comprehensive experiments, we validate the efficacy of our approach in achieving nuanced, interpretable style transfers.

To sum up, our contributions are threefold:

- We construct and annotate a dataset of 140K sentences with fine-grained explanations across six linguistic aspects, providing a foundation for controlled, explainable style transfer in Singlish.
- We introduce a novel multi-agent framework, where specialized LLMs collaborate to achieve fine-grained, explainable transfer.
- We conduct both automatic and human evaluations to demonstrate that our approach achieves precise, interpretable style transfer, advancing the field of explainable NLP for Colloquial Singlish.

2 Related Work

2.1 Style Transfer

Text Style Transfer (TST) involves modifying the stylistic features of text while preserving its meaning. **Parallel supervised TST** rely on paired datasets of sentences in different styles to guide transformations. (Jhamtani et al., 2017; Rao and Tetreault, 2018b; Lai et al., 2021) Innovations have led to a series of effective TST methods, including data augmentation (Zhang et al., 2020), multi-task learning (Niu et al., 2018), and reinforcement learning (RL)-based approaches (Lai et al., 2021). Despite the progress, a significant challenge remains for such methods due to the scarcity of parallel data (Hu et al., 2022).

Non-parallel supervised TST methods (Liao et al., 2018; Shang et al., 2019) alleviate this by using style-specific corpora without parallel data. It eliminates the need for paired data by learning from style-specific corpora. Three main strategies are used: (1) Explicit style-content disentanglement (Li et al., 2018; Mukherjee et al., 2023), which identifies and substitutes style-specific phrases for style conversion; (2) Implicit style-content disentanglement (Shen et al., 2017; Prabhumoye et al., 2018), which separates latent representations of style and content, then injects target style features during generation; and (3) No style-content disentanglement (Lample et al., 2019; He et al., 2020), where models incorporate style controls without separating content. While more flexible, these methods can produce inconsistent outputs due to data variations and often require large, labeled corpora, which are unavailable for many styles.

In light of the above issues, **unsupervised TST** (Xu et al., 2020; Shen et al., 2020) seeks to overcome data constraints by using techniques like back-translation and cycle-consistency (Chen et al., 2016), as well as recent advances in prompting large language models (LLMs) for zero- and few-shot style transfer (Luo et al., 2023; Saakyan and Muresan, 2024). While these methods improve flexibility, they often lack precise control over linguistic features and are not inherently explainable—two crucial aspects for transferring between styles like Singlish and Standard English.

Our work addresses these limitations by constructing a large-scale dataset annotated with detailed explanations across six linguistic aspects. We further introduce a multi-agent framework that facilitates fine-grained, explainable style transfer, providing a novel approach to addressing these key challenges in TST.

2.2 Multi-agent Collaboration

Multi-Agent Collaboration (MAC) is rooted in Distributed Artificial Intelligence (DAI) (Chaib-draa et al., 1992) and involves the coordination of autonomous agents toward shared goals (Hong et al., 2024; Wang et al., 2024). In recent years, LLMs have shown promise in collaborative problem-solving, where agents, each specializing in distinct

tasks, engage in iterative dialogues (Zhang et al., 2024) or debates (Du et al., 2024) to solve complex issues more efficiently. These collaborative frameworks have been employed in areas such as strategic decision-making (do Nascimento et al., 2023), planning (Singh et al., 2024), and language interaction, leveraging the unique expertise of each agent to contribute to the overall task.

Our work applies the multi-agent paradigm, where specialized LLM agents focus on distinct linguistic aspects like syntax, lexical borrowing, and pragmatics. These agents collaborate through generation and critique, enabling fine-grained control and explainability in style transfer. This novel approach offers precise, interpretable style transfer, advancing beyond current methods.

3 Construction of the *ExpCSEST* Dataset

In this section, we outline the construction of our Explainable Colloquial Singaporean English Style Transfer (ExpCSEST) dataset. Initially, we establish a taxonomy of fine-grained stylistic aspect explanations (§3.1). We then identify sources for collecting examples and specify pre-processing details (§3.2). Finally, we elaborate on ExpCSEST's explanation annotation (§3.3) and data analysis (§3.4).

3.1 Explanation Taxonomy

Creating a structured labeling taxonomy is essential for enabling explainable, fine-grained TST between Singlish and English. To capture Singlish's unique features while enable precise and interpretable style transfer, we define six key aspects as follows:

- **Syntax:** *This aspect assesses differences in word order, grammatical relations, agreement, and hierarchical sentence structure between formal and informal texts.*
- **Lexical Borrowing:** *This aspect checks for the existence of loanwords from other languages commonly found in Singlish, making the sentence informal.*
- **Pragmatics**: *This aspect examines the presence of pragmatic particles frequently used in Singlish, serving as indicators of sentence informality.*
- **Prosody/Phonology**: *This aspect identifies textual representations of prosodic and phonological features—such as elongated vowels, non-standard spellings, and stress indicators—that indicate informal English or Singlish usage in online conversations.*
- **Emoticons/Punctuation**: *This aspect checks for the use of emoticons, emojis, or non-standard punctuation suggesting informal language usage.*
- **Code-Switching**: *This aspect looks for instances where the speaker switches between different languages or dialects within the same sentence or conversation, a typical feature of Singlish where English is mixed with words or phrases from Malay, Chinese dialects, or Tamil.*

Notably, each of the above stylistic aspect explanations is expressed in free-form natural language rather than being confined to predefined static categories, offering greater flexibility in capturing stylistic nuances. More details of each aspect are provided in Appendix A.1.

3.2 Data Collection and Pre-processing

To approach the finely controllable and explainable TST between Singlish and standard English, we opt to assemble a substantial corpus of formal and informal sentences and then annotate fine-grained stylistic aspect explanations. We prioritize this data collection by scraping user utterances and content from popular Singaporean websites, including (1) the Eat-Drink-Man-Woman forum (EDMW)[2], (2) the Straits Times[3], and (3) official communications from the Prime Minister's Office (PMO)[4]. These sources are meticulously selected to capture a broad spectrum of formal and informal expressions representative of Singlish, facilitating a comprehensive analysis of the distinct linguistic characteristics.

After scraping the raw corpus containing both formal and informal utterances, we perform further processing to smoothly adopt it for effective transformation across colloquial Singaporean and standard English. The data preprocessing details are provided in Appendix A.2.

3.3 Aspect Explanation Annotation

To make the processed utterances appropriate for finely controllable and explainable TST, we explore equipping them with explanations of fine-grained linguistic aspects referring to the taxonomy outlined in Section 3.1. Recently, the rise of LLMs has ushered in a new frontier in automatic annotation, positioning LLMs as cost-effective, labor-efficient tools for annotation tasks. To minimize the high costs and specialized expertise requirements associated with manual explanation annotations, we thereby investigate an in-context prompting-based approach, leveraging LLMs to generate detailed explanations that identify linguistic features, primarily focusing on the presence of Singlish or informal aspects within given utterances. We start by establishing a candidate seed pool, which functions as the basis for in-context demonstrations to steer LLMs toward generating the desired explana-

[2] https://forums.hardwarezone.com.sg/forums/eat-drink-man-woman.16/

[3] https://www.straitstimes.com/

[4] https://www.pmo.gov.sg/

Dataset	ExpCSEST	
	Informal	Formal
#Utterances	104,601	37,676
BiGram/UniGram	12.16	9.26
Avg Len	23.41	31.10
#Syntax	87,404	-
#Lexical Borrowings	40,622	-
#Pragmatics	15,712	-
#Prosody/Phonology	11,082	-
#Emoticons/Punctuation	29,175	-
#Code Switching	4,529	-

Table 1: Statistics of the ExpCSEST dataset.

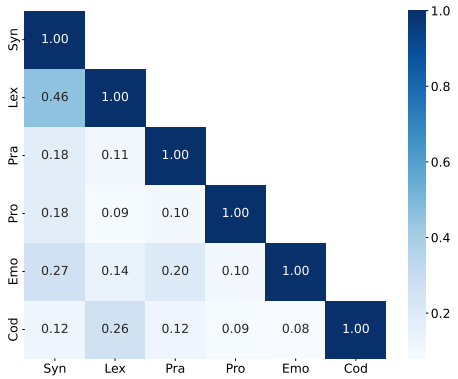

Figure 2: Correlation matrix between the different aspects of informal English in Singapore.

tions. Specifically, we select 100 utterances from the collected corpus, denoted as $\mathcal{S} = \{u_i\}_{i=1}^{100}$, ensuring coverage of all six defined stylistic aspects. For each $u_i \in \mathcal{S}$, we employ domain experts to meticulously annotate the corresponding aspect explanations $\boldsymbol{e}_i = \langle e_i^{\text{Syn}}, e_i^{\text{Lex}}, e_i^{\text{Pra}}, e_i^{\text{Pro}}, e_i^{\text{Emo}}, e_i^{\text{Cod}} \rangle$, forming the final candidate pool $\mathcal{S} = \{u_i, \boldsymbol{e}_i\}_{i=1}^{100}$.

With this seed pool $\mathcal{S}$, we then construct the in-context prompt to query LLMs as follows:

{**System Prompt**} You are an analyst of language styles. Using these linguistic aspects of style analysis as a guideline:
{**Taxonomy**} Definitions of linguistic aspects.
{**Demonstrations**} $(u_1, \boldsymbol{e}_1), (u_2, \boldsymbol{e}_2), \ldots, (u_p, \boldsymbol{e}_p)$.
{**Instruction**} Now, given the following new utterances, generate stylistic aspect explanations for each one below:
{**Input**} List of utterances to be annotated.

where p denotes the number of in-context demonstrations incorporated in the prompt. We empirically include as numerous demonstrations as the LLM's context length permits, maximizing the richness of the provided information. With this design, we can effectively guide LLMs in identifying Singlish and informal elements within the input utterances, thereby generating precise stylistic aspect explanations.

3.4 Data Analysis

Statistics. As shown in Table 1, the ExpCSEST dataset is a comprehensive collection of 142,277 utterances from three distinct sources, annotated for various linguistic phenomena with detailed explanations. Specifically, it encompasses 104,601 informal samples, reflecting the linguistic diversity of Singlish, and 37,676 formal samples, complementing the standard English features. In addition, formal utterances contain an average of 31.10 words, while informal utterances have significantly fewer, averaging 23.41 words. This suggests that informal communication tends to be less verbose compared to standard English, highlighting the simpler and more straightforward nature of colloquial Singaporean English. Further details on dataset composition are discussed in Appendix A.3.

Linguistic Aspect Correlations. To gain insights into how specific linguistic aspects interact to shape the informal and hybrid nature of Singlish, we further examine the co-occurrence matrix for these aspects flagged in utterances, as depicted in Figure 2. The matrix reveals weak overall co-occurrence, suggesting that each aspect contributes distinctly to Singlish. Syntax and lexical borrowings show a moderate co-occurrence of 0.46, indicating their prevalence in informal language, especially considering the dominant role of syntax in the dataset. In contrast, pragmatics exhibit weaker co-occurrence with syntax (0.18) and lexical borrowings (0.11), implying that pragmatic markers often function independently in informal utterances. This highlights the flexibility and adaptability of Singlish across different contexts.

4 The MACoE-Style Framework

This work studies the task of finely controllable, explainable style transfer between Singlish and standard English formulated as follows: Considering a corpus $\mathcal{D} = \{(u_i, s_i)\}_{i=1}^{N}$, where N denotes the total number of utterances, u_i represents an input utterance, and s_i is its style label. Let $\boldsymbol{c} = \langle c^{\text{Syn}}, c^{\text{Lex}}, c^{\text{Pra}}, c^{\text{Pro}}, c^{\text{Emo}}, c^{\text{Cod}} \rangle$ be the fine-grained control signal that aligns with the linguistic aspect taxonomy (as defined in Section 3.1), where each component is a binary value in $[0, 1]$. Given an arbitrary $(u, s) \in \mathcal{D}$ along with the control signal $\boldsymbol{c}$ as input, the primary goal of the task is to learn a model $\mathcal{M}$ to generate the precise explanation $\boldsymbol{e}$ and the new utterance $\hat{u}$ that adheres to the target style $\hat{s}$ while preserving the semantic integrity of

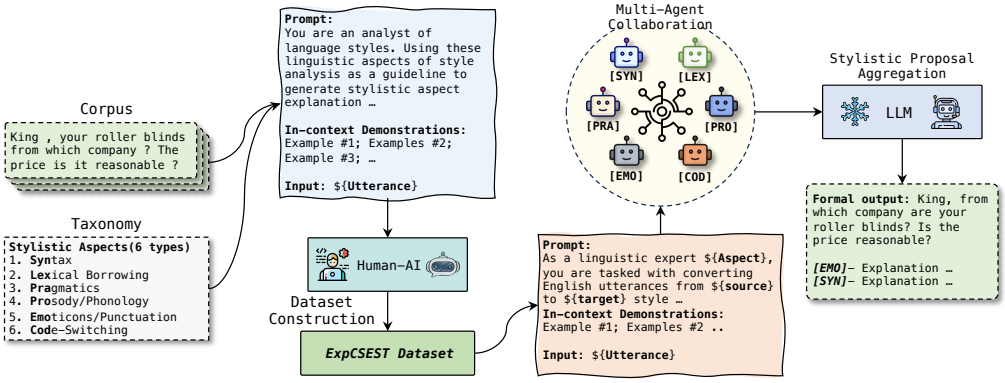

Figure 3: Detailed overview of the ExpCSEST dataset construction and the MACoE-Style approach.

the original utterance.

4.1 Framework Overview

The ExpCSEST dataset, with its well-defined linguistic aspect taxonomy, detailed aspect explanations, and diverse utterances, laid the groundwork for style transfer across Singlish and Standard English. However, controllable and explainable TST in this scenario continues to pose significant challenges, given the complex interplay of informal aspects characteristic of Singlish. Existing methods, which rely on a single model, typically falter with such complex and heterogeneous linguistic elements, struggling to navigate the nuances required for effective transformation.

To this end, we draw inspiration from NLSOM theory to introduce a novel ***M***ulti-***A***gent collaboration framework for effective ***Co***ntrollable and ***E***xplainable ***Style*** transfer (**MACoE-Style**). As shown in Figure x, MACoE-Style comprises three key designs: (1) **Specialized Agent Construction** for tailoring LLMs as specialized agents to handle distinct informal linguistic aspects of the utterances; (2) **Stylistic Proposal Generation** for collaborating these specialized agents over multiple rounds to generate aspect-transformed stylistic proposals; and (3) **Stylistic Proposal Aggregation** for aggregating individual proposals into a cohesive output, harmonizing the various stylistic transformations into a unified utterance. In what follows, we will detail these designs separately.

4.2 Specialized Agent Construction

This stage aims for unambiguous role specialization of multiple LLMs to break down complex style transfer tasks into smaller, fine-grained components, with each agent responsible for a distinct linguistic aspect. To achieve this, we start by developing n LLMs, $\mathcal{A}_1(\theta_1), \ldots, \mathcal{A}_n(\theta_n)$, within the MACoE-Style framework, which can either have parameters ($\theta_1 = \theta_2 = \cdots = \theta_n$) or distinct parameters ($\theta_1 \neq \theta_2 \neq \ldots \neq \theta_n$). To specialize these agents, we further design customized prompting functions $f^i_{prompt}(\cdot)$ to define their roles and operational aspects, contributing outputs tailored to specific stylistic features. In doing so, we can lay the foundation for achieving nuanced control and interpretable adaptations across various linguistic elements throughout the style transfer process.

4.3 Stylistic Proposal Generation

Upon constructing the specialized agents, we then devise a stylistic proposal generation mechanism, which involves synergizing these agents through multi-round interactions to generate and iteratively refine aspect-transformed proposals adhering to the target style, enabling a finely controllable and explainable style transfer process. Specifically, this can be formulated into the following two strategies:

Agent Activation. Given an utterance-style pair (u, s) and its fine-grained control signal $\boldsymbol{c}$, this step first activates the agents corresponding to the aspects where $\boldsymbol{c}$ is marked as 1, allowing them to independently perform aspect-specific style transfers as a preliminary step before further collaboration. By initiating the collaboration with only the specialized agents necessitated by $\boldsymbol{c}$, we aim to reduce potential chaos as the number of participating agents increases, while maintaining precise control to prevent over-transformation.

Upon activating the aspect agents, we then incorporate the input (u, s) into different prompts designed to query the selected agents for eliciting their primary stylistic proposals $\mathcal{P}_0 = \{\boldsymbol{p}_i^0\}_{i=1}^k$, where k is the number of activated agents. Each proposal includes two parts: an aspect-transformed sentence tailored to the target style and an explanation of the corresponding stylistic aspect.

Multi-agent Debate. Given the initial stylistic proposals $\mathcal{P}_0$, we initiate a collaborative round of

Methods	HR@1↑	MRR↑	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERTScore ↑	BARTScore ↓
Direct Prompt	0.1150	0.3328	0.5961	0.5109	0.3150	0.6629	-1.9071
- *w/* Explanation	0.1225	0.3538	0.6011	0.5195	0.3207	0.6742	-1.8807
Duplicate Prompt	0.1550	0.3864	0.5799	0.4904	0.2925	0.6453	-1.9366
- *w/* Explanation	0.1725	0.4154	0.5836	0.4998	0.3023	0.6497	-1.9244
ICLEF	0.1938	0.3490	0.6285	0.5124	0.3482	0.6401	-1.9681
MACoE-Style	0.1650	0.4270	0.6148	0.5272	0.3340	0.6670	-1.9299
- w/ Explanation	**0.4388**	**0.6133**	**0.6394**	**0.5492**	**0.3611**	**0.6899**	**-1.8204**

Table 2: Automatic evaluation of style transfer performance. Direct/Duplicate Prompt and MACoE-Style represent methods without aspect explanations, while *-w/* refers to settings that include aspect explanations.

debate where agents exchange ideas. Each agent $\mathcal{A}_i(\theta_i)$ is provided with a combination of proposals from the other agents as context and is tasked with generating a revised proposal that adheres to its assigned stylistic aspect. During this process, each agent is responsible for critically regulating its peers to facilitate a non-deviated transfer to the target style, while refining its own proposal based on collective input for a more precise transformation. Notably, this procedure can be iteratively repeated over multiple rounds to enhance performance.

4.4 Stylistic Proposal Aggregation

Through multi-agent debate, we can obtain multiple stylistic proposals, with each reflecting a controlled and explainable transformation of a specific linguistic aspect in the style transfer process. Yet, we aim to provide a definitive utterance in the target style rather than multiple options transformed by individual aspects. To achieve this, we facilitate up to r rounds of debate among the specialized agents and then prompt an additional LLM to combine the individual proposals as follows:

$$\hat{u} = \text{LLM}(\mathcal{P}_r, u, s, \texttt{inst}) \tag{1}$$

where $\mathcal{P}_r$ is the stylistic proposals after r rounds of debate, and `inst` is the instruction guiding the LLM to aggregate these proposals into a cohesive output, seamlessly integrating the various stylistic refinements into a unified transformation.

5 Experiments

5.1 Evaluation Metrics

We employ both automatic and human evaluation metrics to assess TST performance. The automatic metrics include: (1) rank-based metrics such as **Mean Reciprocal Rank** (**MRR**) and **HitRate@1** (**HR@1**), (2) content-based metrics like **F1** and **BLEU-1/2**, and (3) similarity-based metrics such as **BERTScore** and **BARTScore**. For human evaluation, we assess **Style Control** (**SC**), **Content Preservation** (**CP**), and **Fluency** (**FL**). The specific definitions and additional details regarding evaluation metrics are provided in Appendix B.1.

5.2 Baselines

We compare the MACoE-Style method with the following baselines: (1) **Direct Prompt *w/o*** and ***w/*** Explanation. (2) **Duplicate Prompt *w/o*** and ***w/*** Explanation. (3) **ICLEF** (Saakyan and Muresan, 2024). More details of the baselines and experiments are provided in Appendix B.2 and B.3.

5.3 Main Results

5.3.1 Automatic Evaluation Results

Table 2 presents the main style transfer outcomes of the MACoE-Style framework compared to existing baselines, highlighting peak performance in **bold**. Generally speaking, MACoE-Style markedly surpasses all baseline methods across multiple evaluation metrics. We analyze the results as follows:

Integrating nuanced stylistic aspect explanations significantly boosts the performance of TST between Singlish and English. As reported in Table 2, it is saliently observable that each method featuring explanations consistently outperforms its counterpart without explanations. For example, Direct Prompt with explanations significantly exceeds its variant without explanations by 0.5% in F1, 0.86% in BLEU1, and 1.13% in BERTScore. A similar trend is also observed for Duplicate Prompt and MACoE-Style. Moreover, it is worth noting that MACoE-Style with explanations not only showcases the most significant improvements but also achieves state-of-the-art performance in style transfer. This underscores MACoE-Style's ability to leverage detailed, aspect-specific explanations to achieve superior stylistic alignment and coherence, effectively navigating the complexities of linguistic transformations.

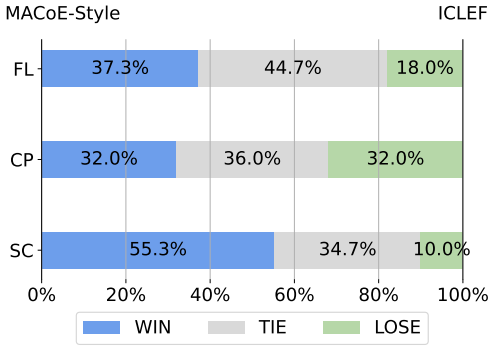

Figure 4: Human evaluation results in style transfer.

MACoE-Style's expertise role-play effectively unleashes the power of LLM collaboration for precise style transfer. Among the baselines, the Duplicate Prompt approaches, which engage multiple agents without specialization to simulate a multi-agent cooperation scenario, unexpectedly underperform compared to the Direct Prompt baselines that employ a single agent. This underscores the inefficacy of merely increasing agent numbers without clearly defined roles, leading to redundant or conflicting outputs that compromise the coherence and effectiveness of the style transfer. Conversely, MACoE-Style with explanations achieves significant gains across all automatic evaluation metrics over all baselines. This observation suggests that our MACoE-Style framework, by strategically assigning distinct linguistic tasks to each agent, ensures focused attention on specific stylistic aspects, which not only circumvents the pitfalls of redundancy but also fosters effective synergy among the agents, confirming its superior ability to manage style transfers with precision and clarity.

5.3.2 Human Evaluation Results

To comprehensively validate the efficacy of our method to complement the automatic evaluation, we further conduct a rigorous human evaluation. We engage three well-educated students and randomly sample 50 utterance pairs from the ExpCSEST dataset (i.e., outputs generated by our MACoE-Style and the baseline ICLEF). The evaluators are asked to indicate which utterance in each pair performs better by assigning 1 (WIN), 0 (TIE), or -1 (LOSE), considering the SC, CP, and FL perspectives, without exposing the source of the generated utterances. As presented in Figure 4, MACoE-Style outperforms ICLEF in almost all perspectives of the human evaluation, except in content preservation, where both methods deliver comparable results. By examining the qualitative cases, we hypothesize that ICLEF's minor modifications contribute to its performance in content preservation, whereas MACoE-Style excels in precise stylistic control and fluency, effectively transforming utterances while maintaining high coherence.

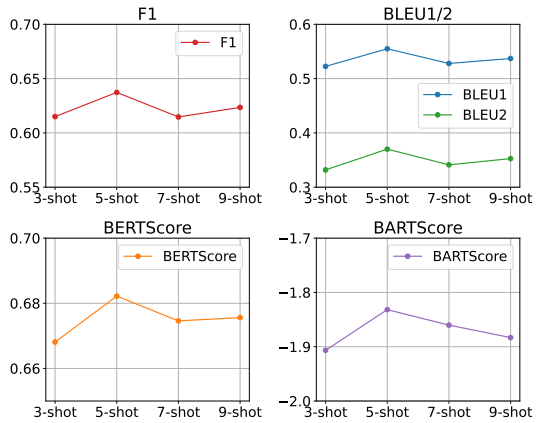

Figure 5: The impact of the shots of the in-context exemplars, ranging from 3 to 9.

r	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERTScore ↑	BARTScore ↓
1	**0.6394**	**0.5492**	**0.3611**	**0.6899**	**-1.8204**
2	0.6059	0.5127	0.3152	0.6643	-1.9215
3	0.6077	0.5199	0.3277	0.6631	-1. 9398

Table 3: Effect of multi-agent collaboration rounds.

5.4 In-depth Analyses

5.4.1 Impact of In-context Exemplar Shots

We analyze the impact of in-context exemplar quantity on MACoE-Style's style transfer performance. In the standard setting, we incorporate 5-shot demonstrations into the context for prompting specialized agents within our framework. Considering the pivotal role this high-quality data plays in providing guidance to agents for generating finely controlled stylistic transformations, we experimentally vary the number of exemplars to further assess its impact. Figure 5 presents the performance trends across different numbers of in-context exemplars. It is noted that optimal performance is achieved with 5 representative examples, with diminishing returns observed when altering the number to 3, 7, or 9 examples. We hypothesize that this can be attributed to fewer in-context examples that limit the comprehensiveness of the information provided to the stylistic agents, while a larger number of examples extends the prompt length, diminishing the agents' learning efficiency.

5.4.2 Effect of Collaboration Rounds

To investigate the efficacy of multi-agent collaboration within the MACoE-Style framework, we examine the effect of varying discussion rounds—from 1

Method	Agents	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERTScore ↑	BARTScore ↓
MACoE-Style	*ChatGPT only*	0.6394	0.5492	0.3611	0.6899	-1.8204
	Mistral only	**0.7099**	**0.6318**	**0.4647**	**0.7354**	**-1.6574**
	Claude only	0.5497	0.5533	0.2685	0.6242	-2.0320
	ChatGPT+Mistral	0.6620	0.5770	0.4010	0.7000	-1.7610
	ChatGPT+Claude	0.5640	0.4700	0.2890	0.6180	-1.9680
	Mistral+Claude	0.5760	0.4810	0.3000	0.6060	-1.9630
	ChatGPT+Mistral+Claude	0.6020	0.5130	0.3320	0.6430	-1.9220

Table 4: Effect of different backbone LLMs for expertise role-play.

Methods	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERTScore ↑	BARTScore ↓
Direct Prompt	0.5667	0.4211	0.2149	0.5194	-3.5795
- *w/* Explanation	0.5598	0.4142	0.2088	0.5005	**-3.5218**
Duplicate Prompt	0.5270	0.3773	0.1764	0.4782	-3.8066
- *w/* Explanation	0.5527	0.4059	0.2076	0.4812	-3.5787
MACoE-Style	0.5400	0.3835	0.1901	0.4718	-3.7088
- w/ Explanation	**0.5906**	**0.4392**	**0.2402**	**0.5233**	-3.5339

Table 5: Automatic evaluation of style transfer performance from formal to informal English.

to 3—on the model performance in executing style transfer. Results in Table 3 reveal that performance peaks during the initial round, with no further improvements in subsequent rounds of discussions. This is understandable since, with the support of specialized agents for transforming fine-grained stylistic aspects, the MACoE-Style framework can achieve the most effective stylistic changes early in the process. Further iterations may focus primarily on refining previously transformed elements, a practice that does not consistently contribute to enhanced performance.

5.5 Effect of LLMs on Expertise Role-Play

We examine the effect of different backbone LLMs' specialization on style transfer performance within our MACoE-Style. Specifically, we compare the performance of various LLMs—ChatGPT, Claude, and Mistral Large—in both isolated and combined settings. In the combined setting, LLMs are randomly distributed across six stylistic aspects, ensuring that each LLM type handles an equal number of aspects. As shown in Table 4, integrating only Mistral outperforms all other setups, demonstrating its robustness in handling nuanced aspects of style transfer. Moreover, combinations such as ChatGPT+Mistral exceed the performance of the standard MACoE-Style that uses ChatGPT only. This suggests that our framework stands to gain from synergizing more advanced LLMs, thereby enhancing overall effectiveness. These results further indicate the importance of carefully selecting and combining LLMs based on their complementary strengths to optimize style transfer within multi-agent collaboration.

5.6 Evaluation of Formal to Informal TST

To further assess the MACoE-Style framework, we examine its performance in transferring style from standard English to Singlish, complementing our previous studies. As shown in Table 5, MACoE-Style generally outperforms other baseline methods in style adaptation. However, in terms of BARTScore, it falls slightly behind the Direct Prompt with explanation. This discrepancy may be attributed to MACoE-Style's broader focus on capturing diverse stylistic nuances, which can introduce minor inconsistencies in semantic coherence. In contrast, the Direct Prompt, utilizing a single LLM for transforming input utterances, potentially enhances coherence and context retention, which are critical for BARTScore.

6 Conclusion

In this work, we proposed a novel approach to style transfer between colloquial Singaporean English and standard English, addressing the challenges of fine-grained control and explainability. We contributed ExpCSEST, a large-scale annotated dataset of 140K utterances, offering detailed explanations across six linguistic aspects. Additionally, we introduced the MACoE-Style framework, a multi-agent system inspired by the "Society of Mind" theory, in which specialized LLMs collaborate to generate controlled, explainable style transfers. Experimental results, evaluated through both automatic metrics and human assessments, validate the advantages of the ExpCSEST dataset as well as the superiority of MACoE-Style in producing precise and interpretable transformations.

7 Limitations

While the ExpCSEST dataset and the MACoE-Style framework make significant strides in style transfer between colloquial Singaporean English and standard English, this work still exhibits certain limitations. First, both the construction of the ExpCSEST dataset and the implementation of MACoE-Style rely heavily on LLMs, making them susceptible to inherent limitations such as biases in the training data and the potential for generating hallucinated or inaccurate outputs. Furthermore, while we collaborate with multiple LLMs to enable finely controlled and explainable TST, our approach does not delve into modifications of the underlying LLM architectures, and the inherent capabilities of LLMs may therefore constrain its effectiveness.

Additionally, our ExpCSEST dataset is confined to English varieties (Singlish and standard English) and a pure text modality, which constrains the generalization of the framework to other languages and multimodal contexts. Expanding the dataset to include more languages, dialects, and multimodal inputs could enhance the framework's versatility and adaptability across a broader range of real-world applications.

References

Rajiv Bajpai, Danyuan Ho, and Erik Cambria. 2016. Developing a concept-level knowledge base for sentiment analysis in singlish. In *CICLing*, pages 347–361.

Brahim Chaib-draa, Bernard Moulin, René Mandiau, and P. Millot. 1992. Trends in distributed artificial intelligence. *Artif. Intell. Rev.*, pages 35–66.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29.

Siew Yeng Chow and Francis Bond. 2022. Singlish where got rules one? constructing a computational grammar for singlish. In *LREC*, pages 5243–5250.

Nathalia Moraes do Nascimento, Paulo S. C. Alencar, and Donald D. Cowan. 2023. Gpt-in-the-loop: Adaptive decision-making for multiagent systems. *CoRR*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *ICML*.

Linus Tze En Foo, Lynnette Hui Xian Ng, and Peter Bell. 2024. Disentangling singlish discourse particles with task-driven representation.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *ICLR*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for A multi-agent collaborative framework. In *ICLR*. OpenReview.net.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C. Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *SIGKDD Explor.*, 24:14–45.

Harsh Jhamtani, Varun Gangal, Eduard H. Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *CoRR*, abs/1707.01161.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you bart! rewarding pre-trained models improves formality style transfer. In *ACL/IJCNLP*, pages 484–494.

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *ICLR*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *NAACL-HLT*, pages 1865–1874.

Yi Liao, Lidong Bing, Piji Li, Shuming Shi, Wai Lam, and Tong Zhang. 2018. Quase: Sequence editing under quantifiable guidance. In *EMNLP*, pages 3855–3864.

Zhengyuan Liu, Shikang Ni, Ai Ti Aw, and Nancy F. Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *COLING*, pages 3924–3936.

Guoqing Luo, Yutong Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. In *Findings of EMNLP*, pages 5740–5750.

Sourabrata Mukherjee, Zdenek Kasner, and Ondrej Dusek. 2023. Balancing the style-content trade-off in sentiment transfer using polarity-aware denoising. *CoRR*, abs/2312.14708.

Sourabrata Mukherjee, Mateusz Lango, Zdenek Kasner, and Ondrej Dusek. 2024. A survey of text style transfer: Applications and ethical implications. *CoRR*, abs/2407.16737.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *COLING*, pages 1008–1021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Nhi Pham, Lachlan Pham, and Adam L. Meyers. 2024. Towards better inclusivity: A diverse tweet corpus of english varieties. *CoRR*, abs/2401.11487.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *ACL*, pages 866–876.

Sudha Rao and Joel Tetreault. 2018a. 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.

Sudha Rao and Joel R. Tetreault. 2018b. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT*, pages 129–140.

Arkadiy Saakyan and Smaranda Muresan. 2024. ICLEF: in-context learning with expert feedback for explainable style transfer. In *ACL*, pages 16141–16163.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. In *EMNLP-IJCNLP*, pages 4936–4945.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NeurIPS*, pages 6830–6841.

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi S. Jaakkola. 2020. Educating text autoencoders: Latent representation guidance via denoising. In *ICML*, pages 8719–8729.

Ishika Singh, David Traum, and Jesse Thomason. 2024. Twostep: Multi-agent task planning using classical planners and large language models. *CoRR*, abs/2403.17246.

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal dependencies parsing for colloquial singaporean english. In *ACL*, pages 1732–1744.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, page 186345.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2020. On variational learning of controllable representations for text without supervision. In *ICML*, pages 10534–10543.

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

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *NeurIPS*, pages 27263–27277.

Stanley Yunick. 1995. The step-tongue: Children's english in singapore. *World Englishes*, 14(2):304–307.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. 2024. Building cooperative embodied agents modularly with large language models. In *ICLR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *ACL*, pages 11325–11345, Bangkok, Thailand.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piekos, Aditya A. Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanic, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind. *CoRR*, abs/2305.17066.

A Dataset Related

A.1 Annotation Prompt

The prompt used for data generation is provided in the table 6.

A.2 Data Pre-processing

Given the raw scraped corpus containing both formal and informal utterances, we perform further processing to smoothly adopt it for effective transformation across colloquial Singaporean and standard English. Specifically, recognizing the presence of user identifiers—which are biased towards different cultural communities and could potentially mislead models in learning formal and informal linguistic features—we first apply carefully designed regular expressions to remove usernames from the corpus. After that, to guarantee the usability of the scraped content while avoiding non-linguistic markers (*e.g.*, website URLs, Unicode characters, and newlines) that do not provide semantic insights into either informal or formal perspectives, we tokenize the corpus and remove these irregular tokens. The resulting union of formal and informal texts contains about 140K examples.

A.3 Dataset Composition

To delve deeper into the dataset composition, Table 1 presents the distribution of colloquial Singaporean English in ExpCSEST, highlighting various fine-grained aspects of informal linguistic characteristics. Syntactic annotations are the most prevalent, occurring in 87,404 utterances. Lexical borrowings are also well-represented, with 40,622 instances, evidencing ample language contact of English and other regional languages in Singlish. Additionally, pragmatic features are captured in 15,712 utterances, offering insights into contextual language use. Prosodic and phonological characteristics are annotated in 11,082 cases, capturing textual representations of intonation. Notably, 29,175 occurrences of emoticons and unconventional punctuation reflect the dataset's coverage of informal, computer-mediated communication. Code-switching, while less frequent (4,529 instances), is still substantially represented, allowing for meaningful analysis of multilingual practices.

B Experiments Related

B.1 Evaluation Matrics

To investigate how well the generated utterances align with the target styles while preserving style-independent content, we first employ human annotators to label the target utterances in the test set, establishing the ground truth for performance evaluation. Based on this, we adopt both automatic and human evaluations to assess our experiments. The automatic metrics we utilized comprised: (1) rank-based **Mean Reciprocal Rank** (**MRR**) and **HitRate@1** (**HR@1**), (2) content-based **F1** and **BLEU-1/2**, and (3) similarity-based **BERTScore** and **BARTScore**.

Specifically, MRR evaluates how well a ranking list prioritizes the correct answer by measuring how early the target utterance appears in the list. HR@1 measures the proportion of utterances for which the target utterance is ranked in the top position (rank 1). To assess these metrics, we leverage GPT-4's capabilities to critique the outputs of different methods. We adopt the ranking prompt shown in Table 7 to obtain the final ranking of the outputs. Based on these rankings, we calculate MRR and HR@1 as follows:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}, \tag{2}$$

$$\text{HR@1} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\mathbb{1}(\text{rank}_i = 1), \tag{3}$$

where $|Q|$ denotes the number of queries.

Additionally, the transformation from informal to formal language can fundamentally be viewed as a generative process. Based on this intuition, given that the actual labels are established above, we employ existing automated generation metrics to evaluate the quality of the resulting formal sentences. These metrics include word-level F1 and BLEU-N (N=1, 2) (Papineni et al., 2002), which measure the lexical overlap between the ground-truth labels and the generated sentences. To complement these content-based methods and mitigate potential overestimation or underestimation, we further utilize BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021) to assess semantic similarity, offering additional insights into how well the generated data aligns contextually with the ground truth.

B.2 Baselines

In the experiments, we compare our methods with the following baselines:

Direct Prompt: Directly providing LLMs with necessary instructions as prompts to perform TST.

Prompting without explanation
You are an analyser of language styles. Using these linguistic aspects of style analysis as a guideline: **Syntax**: Check if the word order, grammatical relations, agreement and hierarchical sentence structure match that of Standard English. For example, if it's informal an example would be: "Tomorrow don't need bring camera. " against the standard English "You don't need to bring a camera tomorrow.", due to there being topic prominence like in Singlish instead of subject prominence like in standard English. **Lexical borrowings**: Check if there are loan words used from other languages common in Singlish, making the sentence informal. For example, saying ""Why John always haolian ah?" vs "Why does John always show off?" where Singlish borrows the word "haolian" from Hokkien to mean show-off. **Pragmatics**: Check for pragmatic particles common in Singlish as a signal for whether said sentence is informal. For example, a Singlish informal sentence would be: "So, I applied for Health Visitor lah" as opposed to: "Therefore, I applied for Health Visitor", where the "lah" is a Singlish pragmatic particle which was used to express obviousness. **Prosody/Phonology**: Check for textual representations of prosodic and phonological features (e.g., elongated vowels, non-standard spellings, stress indicators) that suggest informal English or Singlish usage in online conversations. For example, "Owadioooo" has extra o's to express excitement/relief. **Emoticons / Punctation**: Check for the use of emoticons, emojis, or non-standard punctuation that might indicate informal language or Singlish. For example, excessive use of exclamation marks or question marks, or the inclusion of emoticons like ":)" or ";P" could suggest a more informal tone. **Code-switching**: Look for instances where the speaker switches between different languages or dialects within the same sentence or conversation. This is common in Singlish, where speakers might mix English with words or phrases from Malay, Chinese dialects, or Tamil. For example, "I want to makan already, very hungry"combines English with the Malay word "makan" (eat). Here are some human-annotated examples: ${**List of annotated examples**}$ Now, given these new utterances, generate a new explanation for each of the below utterances. Follow the guidelines and the structure given in the human annotated examples. If there are no hints that it would be singlish for that guideline, leave as "no signal". If there are hints, mention the signs of Singlish/informal English writing. Return your response as a python list of annotations, where the response content is only the list. No yapping. ${**List of utterances to be annotated**}$

Table 6: Our prompt for generating the explanation for the our ExpCSEST dataset

This includes settings where LLMs are prompted using in-context exemplars **without** and **with** stylistic aspect explanations. Specifically, the exemplars are randomly selected from a set of manually constructed TST examples.

Duplicate Prompt: Building on the Direct Prompt, prompt LLMs in a duplicated manner to simulate the multi-agent collaboration. Notably, this setting differs from MACoE-Style by simply copying agents without stylistic aspect specialization, also including **without** and **with** explanations in-context prompting.

ICLEF (Saakyan and Muresan, 2024): A novel human-AI collaboration approach for model distillation, integrating scarce expert human feedback with in-context learning and model self-critique to achieve explainable style transfer.

B.3 Implementation Details

Model's endpoint Throughout this study, we harness the capabilities of renowned Large Language Models such as GPT, Mistral, and Claude-sonnet to generate the formal or informal version of a given input sentence. We specifically list the endpoints that we use for this works.

- **OpenAI's ChatGPT:** We use the `gpt-3.5-turbo` endpoint for all the generating and inference experiments including the Direct prompting, Duplicate prompting and Multi-agent scenarios. Additionally, we also use `gpt-4o-mini` for generating the explanation of our **ExpCSEST** dataset. Finally, we utilize the strength of the `gpt-4o` to do the evaluation phrase for ranking between the methods' outputs.
- **Mistral:** We leverage the power of `mistral-small-lastest` to perform the ablation studies that related to using different backbone agents.
- **Claude:** We employ the strongest endpoint of Claude-sonnet 3.5 which is `claude-3-5-sonnet-20240620` for incorporating with others LLM's endpoints in the different backbone agents setting.

Hyperparameters In this study, we also explore different sets of hyperparameters that impact the generation phase of large language models, such as `top_p` and `temperature`. We tested various values for `top_p` and `temperature` during both inference and evaluation. Ultimately, we decided to set `top_p` to 0.9 and use `temperature` values of 0.95 for inference and 0.1 for evaluation.

B.4 Direct/Duplicate Prompt

Prompting without explanation
As a linguistic expert, you are tasked with converting English utterances from an informal to a formal style. Below are examples that illustrate the transformation from informal to formal usage. Use these examples as a guide to help you understand common patterns in style adjustment. Here is some examples: Informal sentence: ${**Informal input**}$ Formal sentence: ${**Formal output**}$ Based on the provided examples, transform the following input informal utterance into a formal style. Input Informal sentence: ${**Input Sentence**}$ Final Output:

Prompting with explanation
You are a linguistic expert tasked with converting English utterances from formal to informal styles. Below are examples that not only show transformations but also provide explanations for stylistic changes. Use these examples to understand common patterns in style adjustments. Here is some examples: Informal sentence: ${**Informal input**}$ Explanation: {$**Informal input's explanation**$} Formal sentence: ${**Formal output**}$ Task: Using the provided examples as a guide, transform the following input formal utterance into an informal style. Focus specifically on the indicated style aspects. Input formal sentence: ${**Input sentence**}$ Style Aspects to Focus On: ${**Informal aspects**}$ Final Output:

B.5 MACoE-Style Prompt

In the MACoE-Style paradigm, we employ six distinct agents delineated by input prompts. Given the similarity of these prompts, we solely provide the prompts for using LLMs as a **SYNTAX** expert.

Below is the prompt that we use for syntax-expertise agent in the multi-agent scenario. The *italic* styled text denotes sections that are subject to modification based on the agent's expertise.

SYNTAX agent's prompt without explanation
As a linguistic expert specializing in *syntax*, you are tasked with converting English utterances from an informal to a formal style, focusing specifically on the aspect of *syntax* as defined below. *Syntax* Definition: *Syntax assesses differences in word order, grammatical relations, agreement, and hierarchical sentence structure between formal and informal texts.* Below are examples that illustrate the transformation from informal to formal usage. Use these examples as a guide to help you understand common patterns in style adjustment. Examples: Informal sentence: ${**Informal input**}$ Formal sentence: ${**Formal output**}$ Task: Based on the provided examples, transform the following informal utterance into a formal style. Focus on *syntactic* aspects, and provide a clear explanation of the changes made. Output: - Explanation: (Provide a detailed explanation of the *syntactic* changes made.) - Formal sentence: (Provide the formally corrected sentence.) Input informal sentence: ${**Input sentence**}$

SYNTAX agent's prompt with explanation
As a linguistic expert specializing in *syntax*, you are tasked with converting English utterances from an informal to a formal style, focusing specifically on the aspect of *syntax* as defined below. *Syntax* Definition: *Syntax assesses differences in word order, grammatical relations, agreement, and hierarchical sentence structure between formal and informal texts.* Below are examples that not only show transformations but also provide explanations for stylistic changes Use these examples as a guide to help you understand common patterns in style adjustment. Examples: Informal sentence: ${**Informal input**}$ Explanation: ${**Input's explanation on SYNTAX**}$ Formal sentence: ${**Formal output**}$ Task: Based on the provided examples, transform the following informal utterance into a formal style. Focus on syntactic aspects, and provide a clear explanation of the changes made. Output: - Explanation: (Provide a detailed explanation of the syntactic changes made.) - Formal sentence: (Provide the formally corrected sentence.) Input informal sentence: ${**Input sentence**}$

B.5.1 Aggregate prompt

Arregate Prompt
You are a linguistic expert. You will be given several utterances that are refined outputs from multiple agents. Your task is to aggregate these refined outputs and derive the final, correct formal version of the sentence. Input: (Multi responses from many agents) Output: - Formal sentence: (Your aggregated version of the input sentences) The agents' outputs are: ${**Agent's final round output**}$ Based on the outputs from the multiple agents, return only the final formal version of the sentence. - Formal sentence: (Your refine sentence go here)

B.5.2 Prompting-based ranking prompt

Prompting-based ranking prompt
You are a linguistic expert. You will be given several utterances that are refined outputs from multiple agents. Your task is to aggregate these refined outputs and derive the final, correct formal version of the sentence. Input: (Multi responses from many agents) Output: - Formal sentence: (Your aggregated version of the input sentences) The agents' outputs are: ${**Agent's final round output**}$ Based on the outputs from the multiple agents, return only the final formal version of the sentence. - Formal sentence: (Your refine sentence go here)

Table 7: Our GPT4o's prompt for ranking between the outputs of methods

C Human evaluation instruction

Guideline of Human Evaluation	
You need to evaluate the results of two different models in the text style transfer task, focusing on converting informal language to formal language. You will receive two formal outputs generated from an informal input sentence. Your task is to determine which method performs better across three specific aspects. Refer to the definitions below to understand the aspects we need to concentrate on. Provided examples will demonstrate how to assess each metric.	
Results	1 (First method), 0 (Equal), -1 (Second method)
(1) Style Control	
Definition	The Style Control (SC) assess the degree to which generated text reflects formal linguistic characteristics—such as elevated vocabulary, structured syntax, and adherence to grammatical conventions
Example	1. Low style control: Realized that many people prefer dark theme gunmetal color walls, but find it problematic when paired with red cabinets in the kitchen or red walls in the living room. (The verb "realized" is at the start of the sentence without a noun, verb "find" have no any subject to which it refers) 2. High style control: It has been observed that many people prefer a dark theme with gunmetal-colored walls. However, what is more extreme are those with red cabinets in the kitchen or a red wall in the living room.
(2) Content Preservation	
Definition	Content Preservation (CP) checks the ability to retain the original meaning, information, and intent of the input text while altering its style.
Example	Original informal sentence: can, the door can be painted if u need some homeowner also throw it away 1. Content preservation: The door can be painted if the homeowner wants it done, and they can also dispose of it if needed. 2. Content changed: Can the door be painted? If needed, the homeowner can also dispose of it. (Originally, "can" is an affirmative. It has been transformed into a question format.)
(3) Fluency	
Definition	Fluency: Check natural flow, coherence, vocabulary appropriateness, and proper punctuation.
Example	1. High fluency: Yes, Arenas was an exceptional player and arguably the most prolific scoring point guard of that era. Surprisingly, I never anticipated him to emerge as the standout player. (smoother flow, better coherence, and more varied vocabulary) 2. Low fluency: Yes, Arenas was an awesome player. He was literally the best scoring point guard during that period. Wow.. Never expected this guy to be the one. (contains abrupt shifts in expression that can disrupt the flow)

Table 8: Guideline for human evaluation of the informal to formal style transfer task inspired from Zheng et al., 2024