Colloquial Singaporean English Style Transfer with Fine-Grained Explainable Control

Jinggui Liang[1], **Dung Vo**[1], **Yap Hong Xian**[1],
Hai Leong Chieu[2], **Kian Ming A. Chai**[2], **Jing Jiang**[1,3], **Lizi Liao**[1]
[1]Singapore Management University [2]DSO National Laboratories [3]Australian National University
jg.liang.2023@phdcs.smu.edu.sg {tiendungvo,jingjiang,lzliao}@smu.edu.sg
hxyap.2021@scis.smu.edu.sg {chaileon,ckianmin}@dso.org.sg

Abstract

Colloquial Singaporean English (Singlish) is an informal English marked by a unique blend of languages reflecting Singapore's multicultural identity. Style transfer between Singlish and Standard (formal) English is vital for various applications, yet existing methods often lack explainability and fine-grained control. To fill 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 method enables precise, interpretable transformations, advancing explainability in NLP for Singlish.[1]

1 Introduction

Colloquial Singaporean English (Singlish) is a distinctive linguistic blend shaped by Singapore's multicultural heritage, incorporating non-standard English features and elements from 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 text 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, 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.

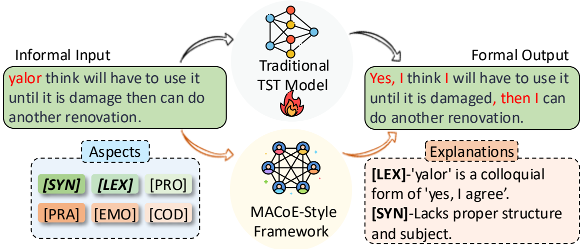
Informal Input
yalor think will have to use it until it is damage then can do another renovation.
Traditional TST Model
Formal Output
Yes, I think I will have to use it until it is damaged, then I can do another renovation.
Aspects
[SYN]
[LEX]
[PRO]
[PRA]
[EMO]
[COD]
MACoE-Style Framework
Explanations
[LEX]-'yalor' is a colloquial form of 'yes, I agree'.
[SYN]-Lacks proper structure and subject.

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

Existing Text Style Transfer (TST) methods fall into three categories: parallel supervised, non-parallel supervised, and unsupervised (Mukherjee et al., 2024a). 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 ex-

[1] Dataset and code are available at https://github.com/dungxibo123/colloquial_tst

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26962–26983
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

planations, as well as the lack of methods capable of handling such fine-grained stylistic transformations. Singlish, with its complex interplay of syntax, lexical borrowing, code-switching, etc., poses additional difficulties for models typically designed for simpler, more homogeneous 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 essential contributions. First, we construct a large-scale non-parallel dataset of formal and informal sentences annotated with six linguistic aspects: Syntax (SYN), Lexical Borrowing (LEX), Pragmatics (PRA), Prosody/Phonology (PRO), Emoticons/Punctuation (EMO), and Code-Switching (COD). 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 MACoE-Style, a novel multi-agent collaboration framework for controllable and explainable style transfer, as shown in Figure 1. Inspired by the Natural Language-based "Society of Mind" (NLSOM) theory (Zhuge et al., 2023; Hong et al., 2024), MACoE-Style specializes multiple large language models (LLMs) as distinct stylistic agents, each responsible for a linguistic 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 non-parallel dataset of 140K sentences with fine-grained explanations across six linguistic aspects, providing a foundation for controlled, explainable TST 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 involves altering the text style while preserving its meaning. **Parallel supervised TST** relies 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. To achieve this, three main strategies are used: (1) Explicit style-content disentanglement (Li et al., 2018; Mukherjee et al., 2023), which aims to identify and substitute style-specific phrases; (2) Implicit style-content disentanglement (Shen et al., 2017; Zhao et al., 2018; 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; Li et al., 2019; He et al., 2020), where models incorporate style controls without separating content. While more flexible, these methods can yield inconsistent outputs due to data variations and require large, labeled corpora that are unavailable for many styles.

In light of the above issues, **unsupervised TST** (Xu et al., 2020; Shen et al., 2020; Carlson et al., 2021) seeks to overcome data constraints by using techniques like back-translation (Bandyopadhyay and Ekbal, 2023) and cycle-consistency (Huang et al., 2020). With the rise of LLMs, prompting-based methods (Reif et al., 2022; Luo et al., 2023; Zhang et al., 2024a) have emerged as a novel paradigm, steering LLMs to generate style-altered texts. Inspired by this trend, the recent ICLEF method (Saakyan and Muresan, 2024) has been developed to enhance the explainability of TST, utilizing LLMs to generate informal attribute terms and then prompting a single LLM to execute all required style adjustments. However, these methods fall short of simultaneously delivering fine-grained control and explainability—two crucial aspects for effective Singlish-English TST.

Our work addresses this 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.

2.2 Multi-agent Collaboration

Multi-Agent Collaboration (MAC) is rooted in distributed artificial intelligence (Chaib-draa et al., 1992) and coordinates autonomous agents toward shared goals (Hong et al., 2024; Wang et al., 2024a). Recently, LLMs have shown promise in collaborative problem-solving, where agents, each specializing in distinct tasks, engage in iterative dialogues (Zhang et al., 2024b) or debates (Du et al., 2024) to solve complex issues more efficiently. These collaborative frameworks have been applied 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.

This work applies the multi-agent paradigm, where specialized LLM agents focus on distinct linguistic aspects like syntax and lexical borrowing. These agents collaborate by generation and critique to enable fine-grained control and explainability in style transfer, advancing beyond existing methods.

3 Construction of the *ExpCSEST* Dataset

This section outlines 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 pivotal for building datasets. To capture Singlish's unique features while enabling precise and interpretable Singlish-English TST, we introduce six key stylistic aspects (Strunk, 2017; Pham et al., 2024):

- **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.*

While this taxonomy defines the linguistics of Singlish via a finite set of stylistic aspects, each of the above aspects is explained in free-form natural language rather than being confined to predefined informal classes, which offers greater flexibility in capturing potential stylistic nuances.

3.2 Data Collection and Pre-processing

To achieve finely controllable and explainable Singlish-English TST, we assemble a substantial corpus of formal and informal sentences and annotate it with fine-grained stylistic aspect explanations. We collect the data exclusively by scraping user utterances and content from the following three prominent Singaporean websites: (1) the Eat-Drink-Man-Woman forum (EDMW), (2) the Straits Times, and (3) official communications from the Prime Minister's Office (PMO). These sources are meticulously selected to capture a broad spectrum of formal and informal expressions representative of Singlish, thus facilitating a comprehensive analysis of its distinct linguistic characteristics (see Appendix A.2). After scraping the above raw corpus, we perform further processing to facilitate its effective use for Singlish-English transfer. More details about the preprocessing procedure can be found in Appendix A.3.

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 (Zhang et al., 2023; Xiao et al., 2023; Wang et al., 2024b). To minimize the high costs and specialized expertise requirements associated with manual explanation annotations, we 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 begin by establishing a candidate

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.

seed pool, which serves as the basis for in-context demonstrations to steer LLMs toward generating the desired explanations for the informal aspects. 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 $\mathcal{S} = \{u_i, \boldsymbol{e}_i\}_{i=1}^{100}$. Subsequently, we construct the in-context prompt (see the full prompt in Appendix A.1) to query LLMs:

{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} $(u_1, \boldsymbol{e}_1)$, $(u_2, \boldsymbol{e}_2)$, ... ,$(u_p, \boldsymbol{e}_p)$.
{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 used in the prompt. Practically, we include 90 randomly selected demonstrations in the prompt when querying GPT-4o-mini, maximizing the richness of the provided information. As such, we can effectively guide LLMs in identifying Singlish and informal elements within the input utterances, generating precise stylistic aspect explanations.

3.4 Data Analysis

Statistics. As shown in Table 1, ExpCSEST 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 are significantly shorter, averaging 23.41 words. This suggests that informal communication tends to be less verbose than standard English, highlighting the more straightforward nature of Singlish. Further details on dataset composition are discussed in Appendix A.4 and A.5.

	Ove.	SYN	LEX	PRA	PRO	EMO	COD
AIA	0.90	0.94	0.93	0.88	0.92	0.90	0.85
$\mathcal{K}$	0.51	0.55	0.49	0.45	0.48	0.57	0.44
AEV	0.82	0.86	0.84	0.80	0.83	0.82	0.79
$\mathcal{K}$	0.50	0.52	0.43	0.48	0.53	0.59	0.42

Table 2: Human evaluation results. Scores (0 to 1) are averaged across all samples rated by evaluators. **Ove.** indicates overall performance across all six aspects, and $\mathcal{K}$ denotes Fleiss' Kappa score (Fleiss and Cohen, 1973).

Explanation Annotation Quality. To enhance the practical applicability of the constructed ExpCSEST dataset, it is important to ensure the reliability and quality of the aspect explanations annotated by LLMs. We address this issue by conducting two human evaluations from the following perspectives. First, we engage two human evaluators to directly assess the LLM-annotated aspect explanations using two criteria: (1) Aspect Identification Accuracy (AIA) and (2) Aspect Explanation Validity (AEV). Detailed instructions for each criterion are provided in Appendix A.6. Evaluators are tasked with reviewing 200 randomly selected samples to assign binary AIA and AEV labels to each of the defined aspects for each sample. We present the experimental results in Table 2.

In addition, we evaluate the annotation quality by measuring the consistency between aspect explanations generated by LLMs and humans. Detailed evaluation procedures and results are presented in Appendix A.7. Notably, the above outcomes reveal a moderate inter-annotator agreement (around 0.5), supporting the reliability of the evaluation process. These results demonstrate that our LLM annotation process not only accurately identifies the informal aspects in the given utterances but also provides appropriate explanations that closely align with human-annotated ones, affirming the high quality and practicality of the ExpCSEST dataset.

4 Methodology

4.1 Problem Formulation

We study the task of finely controllable, explainable Singlish-English style transfer 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

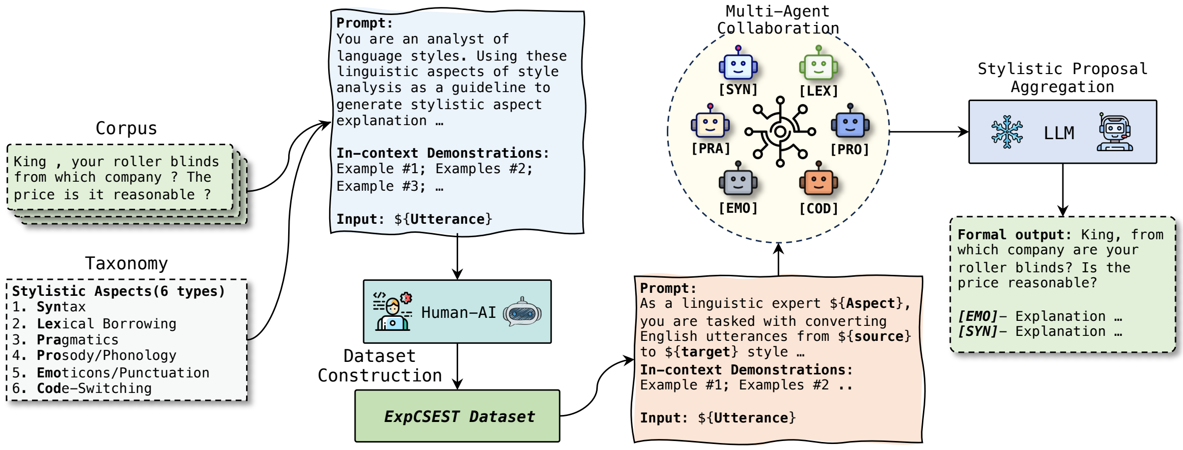
Corpus
King , your roller blinds from which company ? The price is it reasonable ?
Taxonomy
Stylistic Aspects(6 types)
1. Syntax
2. Lexical Borrowing
3. Pragmatics
4. Prosody/Phonology
5. Emoticons/Punctuation
6. Code-Switching
Prompt: You are an analyst of language styles. Using these linguistic aspects of style analysis as a guideline to generate stylistic aspect explanation …
In-context Demonstrations: Example #1; Examples #2; Example #3; …
Input: ${Utterance}
Human-AI
Dataset Construction
ExpCSEST Dataset
Prompt: As a linguistic expert ${Aspect}, you are tasked with converting English utterances from ${source} to ${target} style …
In-context Demonstrations: Example #1; Examples #2 ..
Input: ${Utterance}
Multi-Agent Collaboration
[SYN]
[LEX]
[PRA]
[PRO]
[EMO]
[COD]
Stylistic Proposal Aggregation
LLM
Formal output: King, from which company are your roller blinds? Is the price reasonable?
[EMO]- Explanation …
[SYN]- Explanation …

Figure 2: Detailed overview of the ExpCSEST dataset construction and the MACoE-Style approach.

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 the original utterance.

4.2 The MACoE-Style Framework

We present the proposed MACoE-Style framework in Figure 2. It 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. An example illustrating the above process is provided in Appendix B. In what follows, we will detail these designs separately.

4.2.1 Specialized Agent Construction

The specialized agents are crafted to emulate distinct stylistic experts to modify their corresponding linguistic aspects for completing the Singlish-English style transfer. Typically, these specializations are defined by their designated roles, knowledge, and response styles. To construct them, we configure LLMs with customized prompting instructions $\texttt{inst}_{\text{role}}(\cdot)$, each incorporating a stylistic aspect and its definition, along with examples illustrating style adjustments in this specific aspect. We provide more details in Appendix D.3. The goal of these specialized agents is to generate style-altered utterances and aspect explanations that are consistent with their instructions. By specializing agents contributing to distinct linguistic aspects, we can lay the foundation for achieving nuanced control over various linguistic aspects while providing precise explanations throughout the TST process.

4.2.2 Stylistic Proposal Generation

After constructing the specialized agents, we propose a stylistic proposal generation mechanism to synergize their efforts. It involves coordinating the agents via multi-round interactions, enabling them to generate and iteratively refine proposals to produce the final output adhering to the target style with fine-grained control and explainability. Specifically, this can be formulated as 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 associated with 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 activation, we then feed the input utterance u into all these activated agents to elicit their corresponding stylistic proposals as follows:

$$p_{\text{role}} = \mathcal{A}_{\text{role}}(\texttt{inst}_{\text{role}}(u)), \tag{1}$$

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
Agent Duplicates	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
CoTeX	0.1850	0.4258	0.6165	0.5324	0.3376	0.6752	-1.9886
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 3: Automatic evaluation of Singlish to English TST performance. Direct Prompt, Agent Duplicates, and MACoE-Style are methods without aspect explanations, while *-w/* refers to settings that include aspect explanations.

where the proposal $\boldsymbol{p}_{\text{role}}$ consists of two parts: the aspect explanation and the aspect-transformed utterance tailored to the target style. These proposals form the basis for the subsequent multi-agent collaboration that iteratively transforms the input utterance into the target style. For clarity, we denote these proposals as $\mathcal{P}_0 = \{\boldsymbol{p}^0_{\text{role}} | \boldsymbol{c}(\text{role}) = 1\}$.

Multi-agent Debate. Given the proposals $\mathcal{P}_0$, we initiate a collaborative debate round where all activated agents exchange ideas. Specifically, each activated agent $\mathcal{A}_{\text{role}}$ incorporates proposals from other participating agents in the previous round to critique or refine its proposal, ensuring adherence to its assigned stylistic aspect. This brainstorming process enables these agents to critically regulate their peers to facilitate a non-deviated style transfer while refining their proposals based on collective input for a more precise transformation. Notably, this can be iterated over multiple rounds to enhance performance.

4.2.3 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. However, 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:

$$\hat{u} = \text{LLM}(\mathcal{P}_r, u, s, \texttt{inst}), \tag{2}$$

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. It is worth noting that the aspect explanations within the generated proposals can shed light on the modification traits of these agents, offering explainability throughout the style transfer process.

5 Experiments

5.1 Experimental Setup

Evaluation Metrics. We adopt both automatic and human evaluation to assess TST performance. The automatic metrics include: (1) Rank-based metrics (**Mean Reciprocal Rank** (**MRR**) and **HitRate@1** (**HR@1**)), which rank model outputs based on their adherence to target style norms; (2) Content-based metrics (**F1** and **BLEU-1/2**), which assess the overlap between generated outputs and the ground truth; and (3) Similarity-based metrics (**BERTScore** and **BARTScore**), which measure alignment with references at a semantic level. For human evaluation, we assess **Style Control** (**SC**), **Content Preservation** (**CP**), and **Fluency** (**FL**). More details are provided in Appendix C.1.

Baselines. We compare the MACoE-Style framework against the following baselines in our experiments: (1) **Direct Prompt *w/o*** and ***w/*** Explanation. (2) **Agent Duplicates *w/o*** and ***w/*** Explanation, use the same setting as Direct Prompt but allow for more times of computation for fairer comparison with MACoE-Style. (3) **CoTeX** (Zhang et al., 2024a). (4) **ICLEF** (Saakyan and Muresan, 2024), which is a closely related SOTA method. More details about the baselines and experimental settings are provided in Appendix C.2 and C.3.

5.2 Main Results

5.2.1 Automatic Evaluation Results

Table 3 presents the main style transfer results of the MACoE-Style framework compared to exist-

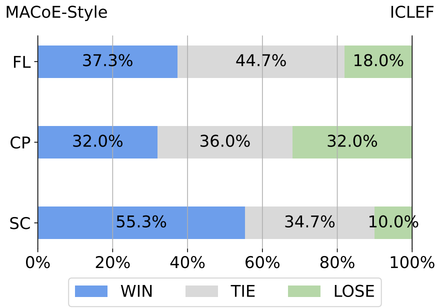

Figure 3: Human evaluation results for Singlish-to-English TST. The Fleiss' Kappa scores are SC=0.79, CP=0.75, and FL=0.82.

ing baselines, with peak performance highlighted in **bold**. Generally speaking, our MACoE-Style consistently surpasses all baseline methods across multiple evaluation metrics by large margins (see Appendix C.4). We analyze the results as follows:

Integrating nuanced stylistic aspect explanations significantly boosts the Singlish to English TST performance. As reported in Table 3, it is saliently observable that each method featuring explanations consistently outperforms its counterpart without explanations. For example, Direct Prompt with explanations exceeds its variant without explanations by 0.5% in F1, 0.86% in BLEU1, and 1.13% in BERTScore. A similar trend is observed for Agent Duplicates 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.

MACoE-Style's expertise role-play effectively unleashes the power of LLM collaboration for precise style transfer. Among the baselines, the Agent Duplicates approaches 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.

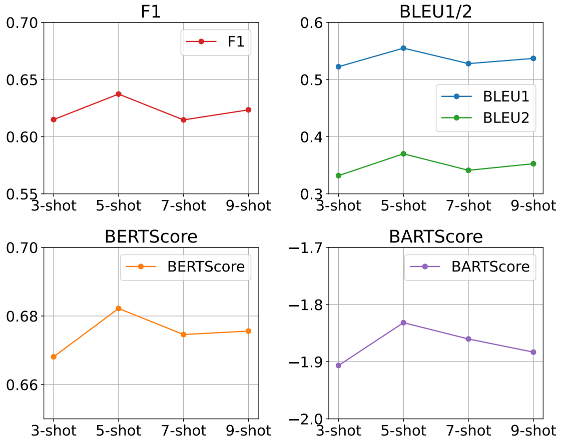

Figure 4: Impact of the shots of the in-context exemplars on Singlish-to-English TST, ranging from 3 to 9.

5.2.2 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 3, 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.

5.3 In-depth Analyses

5.3.1 Impact of In-context Exemplar Shots

We analyze the impact of in-context exemplar quantity on MACoE-Style's TST performance. In the standard setting, we include 5-shot demonstrations

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 on expertise role-play in MACoE-Style for Singlish-to-English TST.

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**
Agent Duplicates	0.5270	0.3773	0.1764	0.4782	-3.8066
- *w/* Explanation	0.5527	0.4059	0.2076	0.4812	-3.5787
ICLEF	0.5463	0.3798	0.1972	0.4758	-3.7250
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 English-to-Singlish TST performance.

in the context for prompting specialized agents within our framework. Given the pivotal role this high-quality data plays in providing guidance to agents for generating finely controlled stylistic transformations, we vary the number of exemplars to further assess its impact. Figure 4 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 limiting the comprehensiveness of the information provided to the stylistic agents, whereas a larger number of examples extends the prompt length, diminishing the agents' learning efficiency. We leave further details on the exploration of the impact of in-context exemplars in Appendix C.5.

5.3.2 Effect of LLMs on Expertise Role-Play

We examine the effect of various specialized LLMs on MACoE-Style's TST performance. We first integrate general-purpose LLMs—ChatGPT, Claude, and Mistral Large—into MACoE-Style as stylistic experts in both isolated and combined settings. In the combined setting, LLMs are randomly distributed across six stylistic aspects, with each type handling an equal number of aspects. Table 4 shows that Mistral integrated in isolation outperforms all other setups, demonstrating its robustness in handling nuanced aspects of style transfer. Combinations like ChatGPT+Mistral exceed the performance of the standard MACoE-Style with ChatGPT alone, suggesting that the proposed framework stands to gain from synergizing advanced LLMs to enhance overall efficacy. Additionally, we further explore the generality of MACoE-Style by assessing its performance when incorporating various multilingual-specific LLMs, with outcomes presented in Appendix C.6. These experimental results demonstrate the importance of strategically selecting and combining LLMs based on their complementary strengths to optimize text style transfer within multi-agent collaboration.

5.3.3 Evaluation of Formal to Informal TST

Note that the above experiments evaluate the proposed MACoE-Style framework from the perspective of Singlish-to-English TST. Therefore, we further assess its performance in English-to-Singlish TST using 200 randomly selected samples. Table 5 shows that MACoE-Style outperforms other baselines 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.

Round	F1 ↑	BLEU1/2 ↑	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 6: Effect of multi-agent collaboration rounds on Singlish-to-English TST.

5.3.4 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 to 3—on model performance in executing style transfer. Results in Table 6 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.

6 Conclusion

This work proposes a new approach to Singlish-English style transfer, addressing the challenges of fine-grained control and explainability. We contribute the ExpCSEST, a large-scale annotated dataset of 140K utterances, offering detailed explanations across six linguistic aspects. Additionally, we introduce 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, demonstrate the advantages of the ExpCSEST dataset as well as the superiority of MACoE-Style in producing precise and interpretable transformations.

Limitations

While the ExpCSEST dataset and the MACoE-Style framework make significant strides in style transfer between Singlish 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 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 method does not involve modifications to the underlying LLM architectures, and its effectiveness may therefore be constrained by the inherent capabilities of those models.

Additionally, our ExpCSEST dataset is confined to English varieties (Singlish and Standard English) and a purely textual modality, which limits the generalizability 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.

Acknowledgments

This research was supported by DSO National Laboratories under agreement DSOCL22233 and the Lee Kong Chian Fellowship. It was also supported by the Ministry of Education, Singapore, under its AcRF Tier 2 Funding (Proposal ID: T2EP20123-0052). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

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.

Dibyanayan Bandyopadhyay and Asif Ekbal. 2023. Unsupervised text style transfer through differentiable back translation and rewards. In *PAKDD*, pages 210–221.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2021. Unsupervised text style transfer with content embeddings. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 226–233.

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*.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613 – 619.

Jessica Foo and Shaun Khoo. 2024. Lionguard: Building a contextualized moderation classifier to tackle localized unsafe content. *CoRR*, abs/2407.10995.

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.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. In *COLING*, pages 2213–2223.

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*.

Dianqi Li, Yizhe Zhang, Zhe Gan, Yu Cheng, Chris Brockett, Bill Dolan, and Ming-Ting Sun. 2019. Domain adaptive text style transfer. In *EMNLP-IJCNLP*, pages 3302–3311.

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. 2024a. A survey of text style transfer: Applications and ethical implications. *CoRR*, abs/2407.16737.

Sourabrata Mukherjee, Mateusz Lango, Zdenek Kasner, and Ondrej Dušek. 2024b. A survey of text style transfer: Applications and ethical implications. *Preprint*, arXiv: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.

David Ong and Peerat Limkonchotiwat. 2023. SEA-LION (Southeast Asian languages in one network): A family of Southeast Asian language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 245–245.

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.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *ACL*, pages 837–848.

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.

William Strunk. 2017. The elements of style : fourth edition.

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. 2024a. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, page 186345.

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. 2024b. Effective demonstration annotation for in-context learning via language model-based determinantal point process. In *EMNLP*, pages 1266–1280.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. Freeal: Towards human-free active learning in the era of large language models. In *EMNLP*, pages 14520–14535.

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.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024a. Distilling text style transfer with self-explanation from llms. In *NAACL: Human Language Technologies: Student Research Workshop*, pages 200–211.

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

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. In *Findings of EMNLP*, pages 13088–13103.

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

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024c. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *CoRR*, abs/2407.19672.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *ICML*, pages 5897–5906.

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 Aspect Explanation Annotation Prompt

Table 7 presents the full prompt used for annotating stylistic aspect explanations. This prompt is specifically designed to process multiple utterances in a single query, aiming to improve the cost and time efficiency of the annotation process. In practice, the list of utterances to be annotated includes up to 50 entries and both input and output are formatted as JSON for querying LLMs. By leveraging the LLM batch API with batched utterances, we significantly reduce costs, achieving approximately 50% savings compared to querying each utterance individually. Additionally, we have sampled a subset of utterances to compare the time efficiency of the batch method with individual queries. The results reveal that the batch method is five times faster than processing utterances individually.

A.2 Data Resources

To collect data capturing a broad spectrum of formal and informal expressions representative of Singlish, we meticulously scrape user utterances and content from the following three popular Singaporean websites to construct the ExpCSEST dataset:

The EDMW Forum[2]: A popular Singaporean forum featuring colloquial discussions on major societal issues, trending topics, and various aspects of daily life in Singapore. Notably, the EDMW forum serves as the primary source of informal data for the ExpCSEST dataset due to its wide recognition as a reliable repository of authentic colloquial Singlish. Existing studies (Foo and Khoo, 2024) have also utilized this forum to collect informal utterances, underscoring its effectiveness in capturing informal linguistic features.

The Straits Times[3]: A leading English-language newspaper in Singapore, known for its formal language and comprehensive coverage of local and international news, politics, business, and culture. While this source predominantly provides formal utterances, occasional informal expressions would appear in specific contexts, reflecting the flexibility of journalistic expression.

PMO[4]: The official website of the Prime Minister's Office in Singapore, offering formal content such as speeches, statements, news updates, and government policies. Similar to The Straits Times, this source contributes significantly to the formal utterances in the ExpCSEST dataset due to its highly structured and formal language, with a minimal number of informal utterances included.

After collecting data from the above resources, we can observe a size difference between informal and formal utterances. This disparity is primarily due to the challenges associated with collecting formal sentences, as the two main sources of formal data—The Straits Times and the PMO website—typically implement anti-scraping mechanisms. These restrictions make extracting formal utterances significantly more difficult compared to the informal data from the EDMW forum, which is more accessible and lacks such constraints.

A.3 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.4 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,

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

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

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

In-context Prompt

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 the 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 that 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 / Punctuation: 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 it 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 7: The prompt used to generate the stylistic aspect explanations for the ExpCSEST dataset.

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.

A.5 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 5. 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.

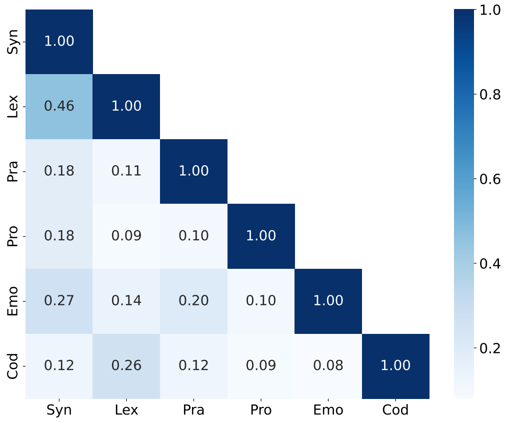

Figure 5: Correlation matrix between the different aspects of informal English in Singapore.

A.6 Human Evaluation Instruction

Regarding the constructed ExpCSEST dataset, We evaluate the LLM-generated stylistic aspect explanations from two primary perspectives, as outlined below:

- **Aspect Identification Accuracy**: This evaluates whether the LLM correctly identifies the presence of each of the six informal aspects in an utterance. If an aspect is identified correctly, it is marked as 1; otherwise, it is marked as 0. The final output is a list of six scores, each indicating whether a specific informal aspect is correctly identified.

- **Aspect Explanation Validity**: This assesses whether the explanation generated by the LLM for each aspect is relevant and adequately explains the informal aspects of the utterance. A score of 1 is assigned if the explanation is appropriate; otherwise, 0. The final output for each utterance is a list of six scores, representing the validity of the explanations for the informal aspects.

A.7 More Human Evaluation

In addition to directly evaluating the quality of the aspect explanation annotations, we further assess annotation quality from a comparative perspective, measuring the consistency between aspect explanations generated by LLMs and humans. To achieve this, we first involve three local students to identify the informal aspects in 50 randomly selected utterances and provide corresponding explanations based on the explanation taxonomy. Then, two human evaluators compare the annotations generated by LLMs and humans to determine whether the LLM annotations outperform those of humans, with possible outcomes being WIN, LOSE, or TIE. The evaluation process yielded WIN = 0.11, TIE = 0.71, and LOSE = 0.18, with $\mathcal{K}$ = 0.52. This indicates that LLM-generated aspect explanations closely align with human annotations, highlighting the effectiveness and reliability of our automated annotation process.

B Interaction Example for MACoE-Style

Table 8 presents an interaction example generated by MACoE-Style during the collaborative process of transferring an informal utterance into its formal version, with key adjustments made by specialized agents highlighted in red.

C Experiments Related

C.1 Evaluation Metrics

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. Following Saakyan and Muresan (2024) and Mukherjee et al. (2024b), the automatic metrics we utilized included: (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**. In what follows, we detail these evaluation metrics separately.

Rank-based Metrics: To evaluate the effectiveness of various methods, one of the most straightforward approaches is to rank their generated outputs based on adherence to target style norms, with higher rankings indicating better style transfer performance. Motivated by this, we thus adopt MRR and HitRate@1 as metrics to evaluate how effectively these methods perform.

Specifically, MRR evaluates how effectively a method's outputs are prioritized within the ranking list. A higher MRR value indicates that a method's results consistently rank closer to the top of the list compared to other methods. HitRate@1 measures the proportion of a method's outputs that are ranked in the top-1 position compared to all other methods. A higher HitRate@1 score indicates that the outputs from this method more frequently rank as the best results. To assess these metrics, we leverage GPT-4 to rank the outputs of different methods according to their adherence to targeted style norms across specific linguistic aspects (see the full ranking prompt in Table 17). From this ranking list, we compute the MRR and HitRate@1 for each method as follows:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i} \tag{3}$$

$$\text{HR@1} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\mathbb{1}(\text{rank}_i = 1) \tag{4}$$

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

Content-based Metrics: 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 the F1 score and BLEU-N (N=1, 2) (Papineni et al., 2002), which measure the lexical overlap between the ground-truth labels and the generated sentences. Specifically, the F1 score in this context reflects a balance of token-level precision and recall, quantifying the token overlap between model

Model Input Phase	
Raw Sample	**Utterance**: Sibei sian .... Barton trying to be hero again U r the 5th option when we healthy KNN just play with the team can a not??? **Informal Aspects**: [Syntax], [Lexical_borrowings], [Emoticons/Punctuation]
Specialized Agent Construction	
Specialized Agents	$\{\mathcal{A}_{\text{role}} = \text{LLM}(\texttt{inst}_{\text{role}}) \mid \text{role} \in \{\text{Syn}, \text{Lex}, \text{Pra}, \text{Pro}, \text{Emo}, \text{Cod}\}\}$
Stylistic Proposal Generation	
Input	**Utterance** u: Sibei sian .... Barton trying to be hero again U r the 5th option when we healthy KNN just play with the team can a not??? **Control Signal** c: $[1_{\text{Syn}}, 1_{\text{Lex}}, 0_{\text{Pra}}, 0_{\text{Pro}}, 1_{\text{Emo}}, 0_{\text{Cod}}]$
$\mathcal{A}_{\textbf{Syn}}$	**Explanation** $e_{\textbf{Syn}}$: Fragmented structure, lack of subject-verb agreement, standard grammatical forms, proper punctuation and capitalization. Use of informal "U r" instead of "You are". **Utterance** $u_{\textbf{Syn}}$: Sibei sian .... Barton is trying to be a hero again. You are the 5th option when we are healthy KNN. Can you just play with the team???
$\mathcal{A}_{\textbf{Lex}}$	**Explanation** $e_{\textbf{Lex}}$: "Sibei sian" is a phrase borrowed from the Hokkien dialect, meaning "very boring" or "extremely tedious.". "KNN" is an abbreviation for a Hokkien profanity. **Utterance** $u_{\textbf{Lex}}$: Very boring .... Barton trying to be hero again U r the 5th option when we healthy just play with the team can a not???
$\mathcal{A}_{\textbf{Emo}}$	**Explanation** $e_{\textbf{Emo}}$: Use of ellipsis (...) suggests a pause and multiple question marks (???) indicate increased inquiry intensity, common in informal writing. **Utterance** $u_{\textbf{Emo}}$: Sibei sian. Barton trying to be hero again U r the 5th option when we healthy KNN just play with the team can a not?
Multi-agent Collaboration	
Input	**Utterance** u: Sibei sian .... Barton trying to be hero again U r the 5th option when we healthy KNN just play with the team can a not??? **Initial Proposals**: $[e_{\text{Syn}}, u_{\text{Syn}}, e_{\text{Lex}}, u_{\text{Lex}}, e_{\text{Emo}}, u_{\text{Emo}}]$
$\mathcal{A}_{\textbf{Syn}}$	**Utterance** $u_{\textbf{Syn}}$: It is very boring. Barton is trying to be a hero again. You are the 5th option when we are healthy. Can you just play with the team?
$\mathcal{A}_{\textbf{Lex}}$	**Utterance** $u_{\textbf{Lex}}$: Very boring. Barton is trying to be a hero again. You are the 5th option when we are healthy. Can you just play with the team?
$\mathcal{A}_{\textbf{Emo}}$	**Utterance** $u_{\textbf{Emo}}$: Very boring. Barton is trying to be a hero again. You are the 5th option when we are healthy. Can you just play with the team?
Stylistic Proposal Aggregation	
Input	**Utterance**: u and **Updated Proposals**: $[e_{\text{Syn}}, u_{\text{Syn}}, e_{\text{Lex}}, u_{\text{Lex}}, e_{\text{Emo}}, u_{\text{Emo}}]$
Formal Output	It is very boring. Barton is trying to be a hero again. You are the fifth option when we are healthy. can you just play with the team?

Table 8: An example generated by the proposed MACoE-Style framework during the interaction process.

predictions and references. Letting s denote the model-predicted sentence and r the ground-truth sentence, the F1 score is computed as follows:

$$\text{Precision} = \frac{|s \cap r|}{|s|} \tag{5}$$

$$\text{Recall} = \frac{|s \cap r|}{|r|} \tag{6}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Similarity-based Metrics: To complement the above 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.

C.2 Baselines

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

Direct Prompt: Directly providing the necessary instructions as prompts to a single LLM to query it to transfer utterances into the target style, requiring only prompting once to execute all necessary style adjustments. 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.

Agent Duplicates: Building upon the Direct Prompt, this baseline prompts LLMs in a duplicated manner. Notably, the MACoE-Style framework prompts multiple distinct LLMs, inherently consuming more computational resources than the Direct Prompt. To ensure a fair comparison, the Agent Duplicates baseline uses the same prompt to query the same LLM multiple times without sharing their outputs as new input. This approach can simulate the computational cost of the multi-agent setup, enabling a performance comparison under equivalent resource usage. This baseline includes both **without** and **with** explanations in in-context prompting settings. For the Agent Duplicates with Explanation, we provide all the required style aspects to these duplicated agents during the utterance transformation.

CoTeX (Zhang et al., 2024a): A novel framework that utilizes LLMs alongside chain-of-thought prompting to distill complex rewriting and reasoning capabilities for effective style transfer.

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. Specifically, the ICLEF workflow consists of the three key steps:

- **Informal attributes generation**: An LLM is used to identify the informal attributes present in the input sentence. For example, given the informal sentence, *I would throw them out asap!*, the LLM is prompted to output a list of informal attributes like textese ("*asap*"), colloquialism (“*throw out*”), exclamation mark, abbreviated language ("*I would*").
- **In-Context Learning from Expert Feedback**: Given that the LLM-generated informal attributes may contain errors, this step involves combining human expert feedback with the in-context learning and self-critique capabilities of LLMs to refine the initial informal attributes. As a result, the incorrect attribute abbreviated language ("*I would*") would be removed.
- **Paraphrasing**: This step prompts the LLM to paraphrase the informal sentence based on the refined informal attributes, thereby obtaining the formal version of the input sentence.

As a strong baseline in the experiments, a key difference between ICLEF and the proposed MACoE-Style framework is that MACoE-Style leverages linguistic knowledge to define a structured framework with six informal aspects for describing informality, whereas ICLEF relies on model-generated informal attribute terms without a predefined structure. Additionally, MACoE-Style employs a multi-agent framework to achieve greater controllability in style transfer.

In our experiments, we implement the ICLEF baseline strictly according to its released code[5]. Notably, all prompts used to query the LLMs are identical to those specified in the paper.

C.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, Agent Duplicates, and Multi-agent scenarios. Additionally, we use `gpt-4o-mini` to generate explanations for our **ExpCSEST** dataset, offering a cost-effective solution that strikes a balance between annotation quality and affordability. 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.

[5] https://github.com/asaakyan/explain-st

Method	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERTScore ↑	BARTScore ↑
MACoE-Style	0.6394	0.5492	0.3611	0.6899	-1.8204
	0.6385	0.5497	0.362	0.6829	-1.8368
	0.6469	0.5547	0.3668	0.6888	-1.8161
	0.6443	0.5597	0.3715	0.6854	-1.8243

Table 9: Effect of different paraphrased prompting instructions in MACoE-Style for Singlish-to-English TST.

Methods	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERTScore ↑	BARTScore ↑
MACoE-Style *w/ Full Exemplars*	**0.6394**	**0.5492**	**0.3611**	**0.6899**	**-1.8204**
- *w/o Explanation*	0.6148	0.5272	0.3340	0.6670	-1.9299
- *w/o Style Sentences*	0.5868	0.5031	0.3162	0.6454	-1.9602

Table 10: Impact of style sentences in in-context exemplars on Singlish-to-English TST.

Method	Agents	F1 ↑	BLEU1 ↑	BLEU2 ↑	BERTScore ↑	BARTScore ↑
MACoE-Style	*ChatGPT*	0.6394	0.5492	0.3611	0.6899	-1.8204
	Mistral	**0.7099**	**0.6318**	**0.4647**	**0.7354**	**-1.6574**
	Claude	0.5497	0.5533	0.2685	0.6242	-2.0320
	SeaLLM	0.6112	0.5162	0.3366	0.635	-2.0264
	SeaLion	0.4249	0.3467	0.1626	0.5191	-2.4515
	Qwen2.5-plus	0.4851	0.3977	0.2100	0.5709	-2.1767

Table 11: Effect of different multilingual backbone LLMs on expertise role-play in MACoE-Style for Singlish-to-English TST.

C.4 Effect of Prompting Instructions

In the MACoE-Style framework, we leverage customized prompting instructions to configure LLMs as specialized agents. To thoroughly evaluate the stability of the framework, we conduct extensive experiments using various paraphrased prompting instructions. As shown in Table 9, paraphrasing the prompting instructions for configuring LLMs does not lead to significant changes in performance, indicating the robustness of our framework to prompt variation. In light of these results, we ultimately chose one of the prompts at random without deliberation to better ensure generalization.

C.5 Impact of In-context Exemplars

In-context exemplars play a crucial role in guiding the behavior of specialized agents in the MACoE-Style framework. Each exemplar consists of a pair of informal and formal style sentences, along with the corresponding aspect explanation. To investigate the relative impact of style sentences versus aspect explanations on the style transfer process, we conduct additional experiments in which the LLMs are prompted using only aspect explanations, omitting the corresponding style sentence pairs. Table 10 reveals a significant performance decline when style sentences are removed from the prompting setup, dropping even more sharply than with the removal of the corresponding aspect explanations. This suggests that the style sentences serve as essential reference points for the LLMs, enabling them to effectively learn the mapping from informal to formal styles. The aspect explanations function as enhancers, allowing the LLMs to modify specific aspects of the informal style more precisely, thereby improving both fine-grained control and explainability in the style transfer process.

C.6 Effect of Multilingual LLMs on Expertise Role-Play

Table 11 reports the comparison of incorporating various multilingual-specific LLMs into MACoE-Style for performing style transfer from Singlish to Standard English. Following existing works, we explore the inclusion of the general multilingual LLM Qwen (Qwen Team, 2024), as well as the Southeast Asia-specific LLMs Sea-Lion (Ong and Limkonchotiwat, 2023) and SeaLLM (Zhang et al., 2024c). The experimental results in Table 11 reveal that MACoE-Style based on general LLMs achieves superior performance. For instance, incorporating ChatGPT and Mistral into MACoE-Style consistently outperforms those methods relying on multilingual-specific LLMs. We suggest that this can be attributed to two key factors: (1) Existing general large-scale language models typically exhibit strong generalization capabilities and broad language coverage, making them particularly effective for the Singlish-English style transfer task, especially given the inherent similarities between Singlish and English. (2) Multilingual-specific LLMs generally have fewer model parameters, which limits their ability to perform Singlish-English TST as effectively as general LLMs.

Additionally, fine-tuning on Southeast Asia-specific languages significantly impacts performance. Notably, MACoE-Style based on SeaLLM surpasses the performance of MACoE-Style with Claude, underscoring the importance of fine-tuning LLMs on Southeast Asia-specific languages for effective style transfer between Singlish and English.

D Prompting Details

In this section, we present the prompting details in our experiments.

D.1 Direct Prompt

The prompts used for implementing the Direct Prompt baseline are presented in Table 12 and Table 13, including Direct Prompt without Explanation and Direct Prompt with Explanation.

D.2 Agent Duplicates

As discussed in Appendix C.2, the Agent Duplicates baseline can be considered a variant of the Direct Prompt method used in a duplicated manner. Thus, the prompts for implementing the Agent Duplicates baseline are the same as those reported in Table 12 and Table 13. Notably, we use these prompts to query the same LLM multiple times without sharing their respective outputs.

D.3 MACoE-Style Prompt

In the MACoE-Style framework, we employ six distinct specialized agents delineated by input prompts. Given the similarity of these prompts, we provide the prompt for using the LLM as a **SYNTAX** expert for illustration. Table 14 and Table 15 present the prompts used for implementing the syntax-expertise agent in the multi-agent scenario. The *italic*-styled text denotes sections that can be modified based on the agent's expertise.

D.4 Aggregation prompt

After obtaining multiple stylistic proposals, each reflecting a controlled and explainable transformation of a specific linguistic aspect in the style transfer process, we utilize an aggregator to produce a definitive sentence in the target style. Table 16 shows the prompt for performing stylistic proposal aggregation.

D.5 Ranking prompt

As discussed in Appendix C.1, we leverage GPT-4's capabilities to critique and rank the outputs of various methods to assess their style transfer performance. Table 17 presents the prompt for implementing the ranking agent.

E Human evaluation instruction

The detailed guideline for human evaluation of the informal to formal style transfer task is illustrated in Table 18.

Direct Prompt 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 are 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**: [Provide the final sentence here]

Table 12: The prompt for implementing Direct Prompt baseline without providing stylistic aspect explanations.

Direct Prompt with Explanation
As a linguistic expert, you are tasked with converting English utterances from an informal to a formal style. Below are examples that not only show transformations but also provide explanations for stylistic changes. Use these examples as a guide to understand common patterns in style adjustments. Here are some examples: **Informal sentence**: ${**Informal input**}$ **Aspect Explanations**: {$**Input's all informal aspect explanations**$} **Formal sentence**: ${**Formal output**}$ Based on the provided examples, transform the following input informal utterance into a formal style. Focus specifically on the indicated style aspects. **Input formal sentence**: ${**Input sentence**}$ **Style Aspects to Focus On**: ${**Input's all informal aspects**}$ **Final Output**: [Provide the final sentence here]

Table 13: The prompt for implementing Direct Prompt baseline with stylistic aspect explanations.

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* aspect, and provide a clear explanation of the changes made. **Output Format**: - **Explanation**: (Provide a detailed explanation of the *syntactic* changes made.) - **Formal sentence**: (Provide the formally corrected sentence.) **Input informal sentence**: ${**Input sentence**}$

Table 14: The prompt for implementing the specialized syntax agent without explanations within the MACoE-Style framework.

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: ${**SYNTAX aspect explanation**}$

Formal sentence: ${**Formal output**}$

Task: Based on the provided examples, transform the following informal utterance into a formal style. Focus on *syntactic* aspect, and provide a clear explanation of the changes made.

Output Format:

- **Explanation**: (Provide a detailed explanation of the syntactic changes made)
- **Formal sentence**: (Provide the formally corrected sentence)

Input informal sentence: ${**Input sentence**}$

Table 15: The prompt for implementing the specialized syntax agent with explanations within the MACoE-Style framework.

Aggregation 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 distinct agents)

Output: Formal sentence: (Your aggregated version of the input sentences)

Agents' outputs: ${**List of agents' final round outputs**}$

Based on the outputs from the multiple agents, return only the final formal version of the sentence.

Formal sentence: (Provide your refined sentence here)

Table 16: The prompt for constructing the stylistic proposal aggregator.

Ranking Prompt

Role: You are a linguistic expert specializing in text style transfer. Your task is to evaluate the effectiveness of different methods in transforming Singlish into Standard English.

Task Overview:

Your task is to review and assess the quality of outputs from various methods. Focus on how well each method adheres to Standard English norms across specific linguistic aspects.

Original Singlish Sentence: The provided Singlish sentence serves as the baseline for each transformation.

Critical Aspects for Evaluation: ${**Definitions of aspects**}$

Critically assess each method's output by focusing on the six key aspects mentioned above. After completing the evaluation, rank the models from the most to the least effective in achieving a high-quality transformation from Singlish to Standard English.

Final Task: Provide a final ranking for each method. Assign 1 for the highest quality and increase for the lower quality.

Input Sentence: ${**Informal input**}$

Explanation of Key Aspects: ${**Ground-truth explanation for informal input**}$

Outputs from the Methods: ${**Final outputs from distinct methods**}$

Final Ranking: (All mapping "Method x: y" have to be placed on a single line, separated by a comma)

Table 17: The prompt used to rank the outputs of various methods for assessing their style transfer performance.

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) assesses 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 18: Guideline for human evaluation of the informal to formal style transfer task.