PSST: A Benchmark for Evaluation-driven Text Public-Speaking Style Transfer

Anonymous EMNLP submission

Abstract

Language style is necessary for AI systems to accurately understand and generate diverse human language. However, previous text style transfer primarily focused on sentence-level data-driven approaches, limiting exploration of potential problems in large language models (LLMs) and the ability to meet complex application needs. To overcome these limitations, we introduce a novel task called **P**ublic-**S**peaking **S**tyle **T**ransfer (**PSST**), which aims to simulate humans to transform passage-level, official texts into a public-speaking style. Grounded in the analysis of real-world data from a linguistic perspective, we decompose public-speaking style into key sub-styles to pose challenges and quantify the style modeling capability of LLMs. For such intricate text style transfer, we further propose a fine-grained evaluation framework to analyze the characteristics and identify the problems of stylized texts. Comprehensive experiments suggest that current LLMs struggle to generate public speaking texts that align with human preferences, primarily due to excessive stylization and loss of semantic information. We will release our data, code, and model upon acceptance.

1 Introduction

Text Style Transfer (TST) is crucial in Natural Language Processing (NLP), focusing on modifying text style while retaining the original content's information (Hu et al., 2022; Jin et al., 2022). By modeling complex human styles, including personality, habits, and mindset (Jin et al., 2022; Geroda et al., 2023), AI models can further accurately understand and generate diverse human languages for user-centric applications such as role-playing (Wang et al., 2023) and digital personas (Clarke, 1994; Morande and Amini, 2023).

Benefiting from their superior capabilities, large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Yang et al., 2023) can achieve

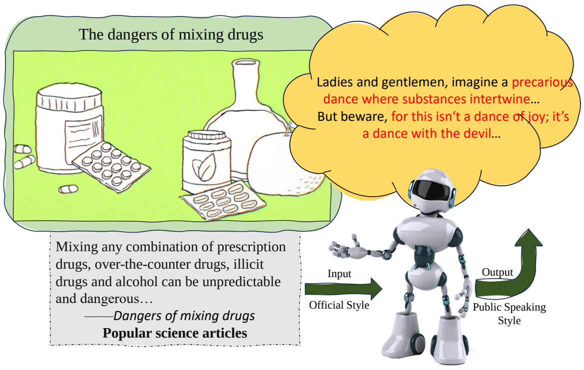

Figure 1: Illustration of Public-Speaking Style Transfer (PSST). An AI model is requested to present a written text, such as a popular science article, to audiences vividly and engagingly. The example generated by ChatGPT in the figure shows excessive stylization (highlighted in red).

high performance in traditional style transfer tasks. However, traditional style transfer is limited to sentence-level transformations and only a few data-driven styles (e.g., formal-informal (Sheikha and Inkpen, 2011), polite-impolite (Madaan et al., 2020)) (Jin et al., 2022). This is insufficient for complex real-world applications, which require passage-level transformations and complex styles (Jin et al., 2022). In this paper, we introduce a novel task, called Public-Speaking Style Transfer (PSST), which could further facilitate critical applications such as knowledge dissemination, public education, business promotion, etc. Specifically, PSST transforms formal, long paragraph texts into a public-speaking style, a form of language used by human beings to convey influential knowledge and ideas in public (Kedrowicz and Taylor, 2016; Beebe and Beebe, 2005) (as shown in Figure 1 and video demo in F.1).

To investigate the language style modeling capabilities of LLMs in scenarios akin to PSST, we initially employ LLMs (eg. ChatGPT (OpenAI, 2022) and Llama3 (AI@Meta, 2024)) for PSST and identify three primary issues: (1). Over-stylization; (2). Uneven Style Strength Distribution; (3). Severe Semantic Degradation. To further quantify and enhance the language style modeling capabilities of LLMs, we propose a fine-grained evaluation framework tailored for complex long-text style transfer tasks such as PSST focusing on style strength and semantic preservation (Rao and Tetreault, 2018; Shen et al., 2017a), which enables a continuous, evaluation-driven approach to enhance the language style modeling capabilities of LLMs.

For style strength evaluation, we propose two metrics: (1) *passage-level style strength score* that offers a coarse-grained measure of the overall style strength, and (2) *style strength distribution* that captures the distribution of style elements throughout the text. Using real data as a standard, we analyze the style modeling capability of LLMs based on the above metrics. Specifically, we first gather official texts (eg. Encyclopedias) as source texts, and real data (eg. Ted Talks) as the target style dataset (Section 3.2). To accurately define and evaluate public-speaking style, we review linguistic literature on spoken public speeches (McCroskey et al., 2003; Beebe and Beebe, 2005; Atkinson, 1985; Halliday, 1989) and summarize four prominent sub-styles, including *orality*, *interactivity*, *vividness*, and *emotionality*, supported by manual annotation. Since we aim to quantify the style elements at various positions within the stylized text and LLMs possess strong capabilities at the sentence level TST (Lai et al., 2023), for each sub-style, we generate a series of examples varying in style strength from a single sentence, and then we rank and score them with GPT3.5-Turbo[1] and train a small scorer (TinyLlama-1.1b (Zhang et al., 2024)) to predict these scores.

For semantic preservation evaluation, we propose a QA-based method. The accuracy changes of a QA model (eg. Llama3-8B-Instruct (AI@Meta, 2024)) on texts before and after style transfer reflect the LLM's ability of semantic preservation. Specifically, we use GPT-4[2] to generate QA pairs from the source text focusing on two dimensions: *key message* and *structure* (Coopman and Lull, 2018).

Moreover, comprehensive experiments show that prompt engineering can mitigate but not completely resolve the above problems, which highlights a gap between LLMs and humans in language style modeling, underscoring a substantial opportunity for improvement in this area.

Our main contributions are as follows:

- We collect data and introduce a valuable and extensible task named Public-Speaking Style Transfer (Section 3). We decompose the intricate text style into key sub-styles for more accurate style definition and evaluation (Section 3.3 3.4).
- We propose a fine-grained evaluation framework for PSST (Section 4), which is extensible to incorporate additional sub-styles (e.g. personality), enabling an evaluation-driven approach to continuously analyze and enhance the language style modeling capabilities of LLMs.
- We conduct a comprehensive evaluation of the performance of mainstream LLMs on PSST (Section 5.2). Our analysis reveals that current LLMs often exhibit over-stylization, uneven style strength distribution, and severe semantic degradation.

2 Related Work

Definition of Text Style Style, from a linguistic perspective, encompasses various elements that contribute to the conveyance of semantics, including word choice, sentence structure, and arrangement, all of which work together to establish the tone, imagery, and meaning in the text (McDonald and Pustejovsky, 1985; Hu et al., 2022). In contrast, research on text style transfer (TST) takes a data-driven approach, defining style as attributes or labels based on style-specific corpora (Shen et al., 2017b; Rao and Tetreault, 2018), which may affected by other attributes in datast (Jin et al., 2022). However, our approach is grounded in real data and explores key dimensions from the perspective of public-speaking linguistics. This definition alleviates the ambiguity and reduces the difficulty of data construction and evaluation.

Evaluation of Text Style Transfer Firstly, fluency, a common objective in most natural language generation tasks, is often measured by the perplexity score (PPL) (Yang et al., 2018). Secondly, to evaluate content preservation during the style transfer, metrics include BLEU (Papineni et al., 2002), ROUGE (Lin and Och, 2004), BERTScore (Zhang

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo
[2] https://openai.com/research/gpt-4

TST task	style	examples	features
official ↔ p-s	official	GDOS is a modified version of WEDOS, which facilitates ...	arousing interest
	p-s	**Have you heard of GDOS?** It's a modified version of ... **it helps you** ...	
	official	Deeply concerned that the situation in Rwanda, which has resulted in the death of many thousands of innocent civilians, including women and children	appropriate emotion
	p-s	**I'm really worried about** what's going on in Rwanda. **So many** innocent people, including women and children, have died.	
	official	Instead, they've become emboldened.	better vividness
	p-s	Instead, they have grown **stronger and more courageous.**	
	official	The Conference also agreed that the Bureau would keep the calendar under review	oral expression
	p-s	**So, uh**, the Conference also **agre-uh,** agreed that the Bureau would keep the calendarunder review, **y'know?**	
formal ↔ informal	formal	He is very **attractive**	Manuscript Form and Punctuation, Vocabulary
	informal	he **iss wayyy hottt.**	
	formal	**Yes**, but not for episode **IV**.	
	informal	**yes**, except for episode **iv**.	

Table 1: Comparison of PSST and formality TST: **a.** 'p-s' means public-speaking style. **b.** Traditional TST primarily focuses on vocabulary-level adjustments, such as adhering to writing norms and word norms. In contrast, PSST involves audience engagement, including posing questions to generate interest and appeal to the audience.

et al., 2019) are employed. Thirdly, the strength of style is an important dimension. Typically, a binary style classifier is first separately pretrained to predict the style label of input sentences (Prabhumoye et al., 2018; Lai et al., 2021; Zhan et al., 2022). This classifier is then used to estimate the style transfer accuracy. Recently, Lai et al. (2023) demonstrates that ChatGPT (OpenAI, 2022) achieves competitive correlation with human judgments to serve as a multidimensional evaluator for sentence-level formal-informal style transfer.

To effectively evaluate long texts in PSST and facilitate detailed analysis, we propose a multi-dimensional fine-grained framework to assess the style strength and distribution of a long text and a QA-based method to capture differences in details and logic of the text before and after PSST.

3 Public-Speaking Style Transfer

In this section, we introduce the Text Public-Speaking Style Transfer (PSST) task. Initially, we provide a precise definition of PSST and describe the source dataset we constructed for this task. Furthermore, leveraging real-world public-speaking style data, we decompose the abstract public-speaking style into essential sub-features, which are introduced in detail in Section 3.3.

3.1 Task Formulation

The PSST task involves transforming official text style a (e.g., news articles) into a more conversational and public-speaking-oriented language style a', which can be formulated as:

$$y(a') = P(x(a)|a', [a_1, \ldots, a_n]) \qquad (1)$$

where $x(a)$ and $y(a')$ represent the official input text and the public-speaking style output, respectively. $[a_1, \ldots, a_n]$ indicates additional conditions that can be introduced to enhance the task, including factors like the speaker's personality, the audience's specific preferences, and so forth.

As shown in Table 1, official-style texts typically employ specialized and complex vocabulary and sentence structures, conveying a serious and objective tone. In contrast, public-speaking style texts are more suitable for oral expression, characterized by simpler vocabulary and sentence patterns, and a more direct tone. Additionally, the public-speaking style features audience-oriented attributes, which we will elaborate on in Section 3.3.

3.2 Source Data

Recall that our goal is for LLMs to emulate human public speaking to effectively convey knowledge. Therefore, we select three types of official texts with intensive knowledge as source texts: news articles[3], encyclopedias[4], and research paper abstracts[5]. For the target style dataset, we utilize data from real scenarios including TED talks[6], politi-

[3]https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

[4]https://huggingface.co/datasets/wikitext

[5]https://www.kaggle.com/datasets/Cornell-University/arxiv

[6]https://huggingface.co/datasets/iwslt2017/viewer/iwslt2017-en-zh

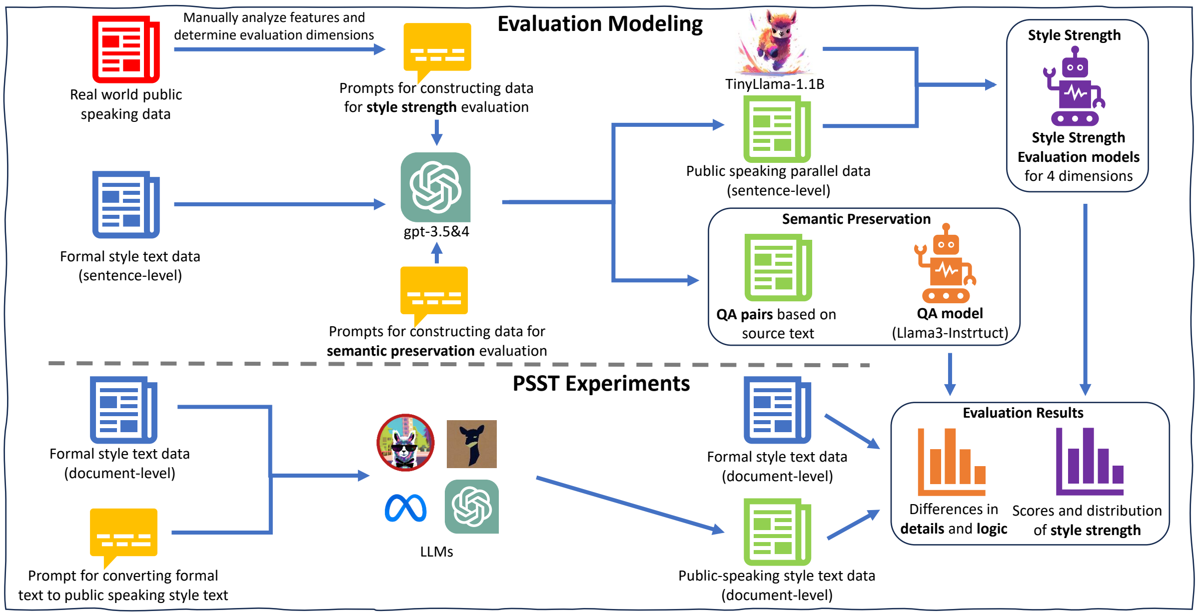

Figure 2: Pipeline of establishing the evaluation system of the PSST task and the experiment&analysis of LLMs. **1.** The above depicts the process of establishing the evaluation system. Specifically, we begin with a comprehensive analysis of real-world data from a linguistic perspective and identify four key characteristics of public-speaking style. Subsequently, we employ GPT-3.5 to generate a sentence-level list-wise parallel corpus and train TinyLlama-1.1B as a scorer for each dimension. For semantic preservation, we utilize GPT-4 to generate QA pairs that focus on key information and logic in source texts. We then assess these pairs with a QA model applied to stylized text, using variations in model accuracy to evaluate semantic integrity. **2.** The bottom presents the experiment and analysis of the PSST task for the current LLMs.

cal speeches[7], academic presentations[8], and educational lectures[9].

To ensure the reliability of our evaluation and comprehensively assess the model's capability for style transfer, we further select the source and target datasets by filtering based on token counts (400 ± 100, 800 ± 200, 1200 ± 200) and ensuring comparable lengths between the two. The final dataset used for PSST is shown in Table 3. See Appendix A for a detailed description.

3.3 Prior Fine-grained Analysis

We deconstruct the public-speaking style into several key sub-styles to achieve a more precise definition. Subsequently, these sub-styles form the basis for style strength evaluation. We initially review research papers focused on oral public speaking (McCroskey et al., 2003; Beebe and Beebe, 2005; Atkinson, 1985; Halliday, 1989) and summarize the following key candidate features: (1) *Interactivity*, (2) *Emotionality*, (3) *Filler Words*, (4) *Vividness*, (5) *Ambiguity*, (6) *Abbreviations*, (7) *Informal Lexicon*. See Appendix E.2 for detailed descriptions. Then we employ rigorous manual annotation to identify the prominent features.

Specifically, we randomly sample 300 sentences from the target style dataset mentioned in Section 3.2, which are then annotated by 3 annotators based on two evaluation metrics. In the multi-label approach, we select labels that signify the public-speaking style features present in a specific instance. In contrast, the best-one approach selects the most salient label that embodies the public-speaking style of the sentence. Details can be found in Appendix E.2

The statistics in Figure 3 indicate that *interactivity*, *Emotionality*, *Filler Words*, *Vividness* are prevalent in the real dataset (indicated by multi labeling) and exhibit a pronounced tendency to emphasize public-speaking style (indicated by best-one labeling). Further, we merge *Filler Words*, *Ambiguity*, *Abbreviations*, and *Informal Lexicon* into one feature *Orality* as the following two reasons: (1) We find these features frequently co-occur, particularly in descriptions where multiple sentences of simple structure are employed alongside filler words to facilitate natural transitions. (2) These

[7] https://www.americanrhetoric.com/
[8] https://iwslt.org/2023/multilingual
[9] https://www.webpages.uidaho.edu/psyc390/index.htm

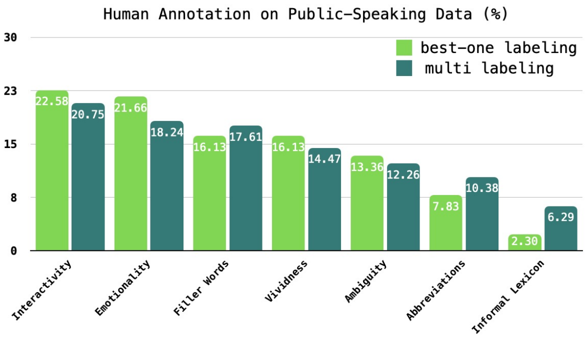

Figure 3: Human annotation on features of real public speaking data. "Interactivity", "Emotionality", "Filler words" and "vividness" are notable features.

features collectively contribute to the oral nature of the speech (Halliday, 1989) but do not individually stand out as strongly as the primary features.

3.4 Key Features of Public-Speaking Style

Based on the above analysis, we set out the following four dimensions to categorize the characteristics of public-speaking style and emphasize the distinctions among these categories.

- **Interactivity:** Interactivity in public speaking refers to the speaker engaging with the audience through various means such as posing thought-provoking, facilitating personal reflection, and crafting intriguing hypothetical scenarios. (Table 1 example 1)

- **Emotionality:** Public speaking contains the speaker's appropriate views and attitudes on specific events to reflect the speaker's emotional tendencies and inner thoughts. (Table 1 example 2)

- **Vividness:** In public speaking, speakers should present information in a lively, easy-to-understand way, such as using analogies and metaphors to make complex ideas more accessible and engaging. (Table 1 example 3)

- **Orality:** The public-speaking style texts should align with oral communication norms, incorporating appropriate filler words, simple sentence structures, and suitable word choices.(Table 1 example 4)

4 Fine-grained Evaluation System

The evaluation of PSST primarily focuses on **style strength** and **semantic preservation**. For style strength evaluation of long text, we propose a fine-grained evaluation framework in Section 4.1. For semantic preservation, we introduce a QA-based approach in Section 4.3, which focuses on key details and logical structure. It is noteworthy that the evaluation framework we propose is scalable. It is applicable across various scenarios, particularly those involving complex and abstract styles.

4.1 Style Strength Evaluation

To evaluate the style strength of long texts in the PSST task, we propose a multi-dimensional, fine-grained evaluation method. Specifically, we generate a series of examples with varying style strengths from a single sentence, rank and score them using `gpt-3.5-turbo`, and then train a small scorer (TinyLlama-1.1b (Zhang et al., 2024)) to predict these scores. This approach is appropriate for three main reasons: 1). Early experiments suggest that using LLMs directly as evaluators may not be feasible (Section 4.2). 2). Relevant studies show that LLMs possess strong stylization and evaluation capabilities in TST tasks at the sentence level (Lai et al., 2023). 3). For long texts, a simple overall style strength score does not facilitate detailed analysis. For example, do texts in different positions exhibit the same style strength?

4.1.1 Fine-grained Evaluation Modeling

We utilize `gpt-3.5-turbo` to generate five sentences from an official sentence, ensuring consistent semantics while progressively increasing the style strength. Each sentence is scored on a scale from 1 to 5. To maintain precise ordering of stylistic strength and corresponding scores, we employ `gpt-3.5-turbo` as both the generator and evaluator within a single prompt. Please refer to Appendix D.1 for detailed prompts.

Then, we fine-tune TinyLlama-1.1b (Zhang et al., 2024) as the sentence-level scorer named **EvalModel-1.1B$_{\textbf{gpt-3.5}}$**. Please refer to Appendix B for more details.

4.1.2 Text-Level Style Evaluation

In our evaluation framework, given a document D we first use `stanza`[10] to segment long text into sentences $S = \{s_1, s_2, \dots, s_M\}$. Then, we score the sentences using **EvalModel-1.1B$_{\textbf{gpt-3.5}}$** in an N-gram manner, which means that we combine each of the n sentences and score them, for example, $[s_1, s_2, s_3], [s_2, s_3, s_4], \dots, [s_{M-2}, s_{M-1}, s_M]$

[10]https://stanfordnlp.github.io/stanza/

where $n = 3$. The corresponding score sequence is denoted as $Score_{seq-n}$, and in our experiment we set $n \in \{1, 2, 3, 4\}$.

For text-level evaluation, we implement two metrics to assess different LLMs. The first metric, named *Text-Level Style Score*, can be used to make coarse-grained comparisons of different texts.

$$Score_{text} = \frac{1}{4} \sum_{n \in \{1,2,3,4\}} \texttt{mean}(Score_{seq-n}) \quad (2)$$

The second metric, named *Style Score Distribution*, can be used for fine-grained comparison of style strength in different positions of texts. Specifically, given a document D and its N-gram score results $\{Score_{seq-n}\}$, we chunk different N-gram sequences into K segments, and then average them by position to obtain the style strength distribution:

$$Score_{dist} = [s_1, s_2, \ldots, s_K] \quad (3)$$

4.2 Correlation

For computing correlation, we collect a dataset comprising 100 sets of samples, each containing four texts generated by different models. We then ask three evaluators to rank each set of texts by style strength to establish the ground truth with the questionnaire shown in Figure 18. **EvalModel-1.1B$_{\textbf{gpt-3.5}}$** and **Llama 2-Chat-70B** [11] are used to score and rank candidate texts. The prompt used by Llama 2-Chat-70B is shown in Figure 17, which we refer to the method in Lai et al. (2023).

EvalModel	Llama 2-Chat (70B)	Ours (1.1B)
Emotionality	56.98	76.29
Interactivity	50.05	81.54
Vividness	51.11	84.50
Orality	13.98	72.94
Average	43.03	78.82

Table 2: **Spearman's** ρ between different evaluation models and human evaluation. **a.** Ours(1.1B), exactly the `EvalModel-1.1B`$_{\texttt{gpt-3.5}}$, denotes style-strength evaluation model trained utilizing datasets derived from GPT-3.5 for each dimension.

Results shown in able 2 indicate that the fine-grained evaluation method proposed in this paper outperforms the LLM-based evaluation method. We hypothesize the following two reasons: (1) The sentence-level assessment method provides a more granular evaluation; (2) The n-gram approach mitigates the issue of merely adding sentence-level stylized elements. For illustrative instances, please refer to Appendix F.3.

[11] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

4.3 Semantic Preservation

For semantic preservation, we propose a QA-based approach. Using `gpt-4`, we generate multiple-choice questions that capture the details and logical relationships inherent in the original text, which are important elements of public speaking (Coopman and Lull, 2018). We then test the QA model (Llama-3-8B-Instruct) on the text before and after style transfer. Variations in the model's accuracy indicate the degree of semantic preservation. Additionally, we use text paraphrased by `gpt-3.5-turbo` as a baseline for comparison. Specifically, we evaluate content preservation from the following two aspects:

(1) **Key Information**: This evaluates whether the essential information, facts, and details from the original text are preserved. Consequently, these questions should be answerable directly from the text, without requiring reasoning or external knowledge, and should include a variety of question types (e.g., "What," "Who," "When," "Where," "How").

(2) **Logical Structure**: This examines whether the emotions, logical relationships and role relationships inherent in the original text are preserved. Consequently, the questions must delve into the specifics of the content and reasoning, such as identifying the speaker or audience, understanding the emotions, and analyzing the logical relationships within the text.

The above two multiple-choice questionnaires consist of 10 question-options pairs and correct answers for each source text. Please refer to Appendix D.2 for more details.

5 Public Speaking Ability Evaluation

5.1 Experiment Settings

Baselines (1). **ST**(*source text*), the official texts mentioned in Section 3.2. (2). **PT**(*Paraphrased text*), texts paraphrased by `gpt-3.5-turbo` based on the source text. (3). ***Ted Talks***, the target style text in real public speaking scenario. Note that **ST**

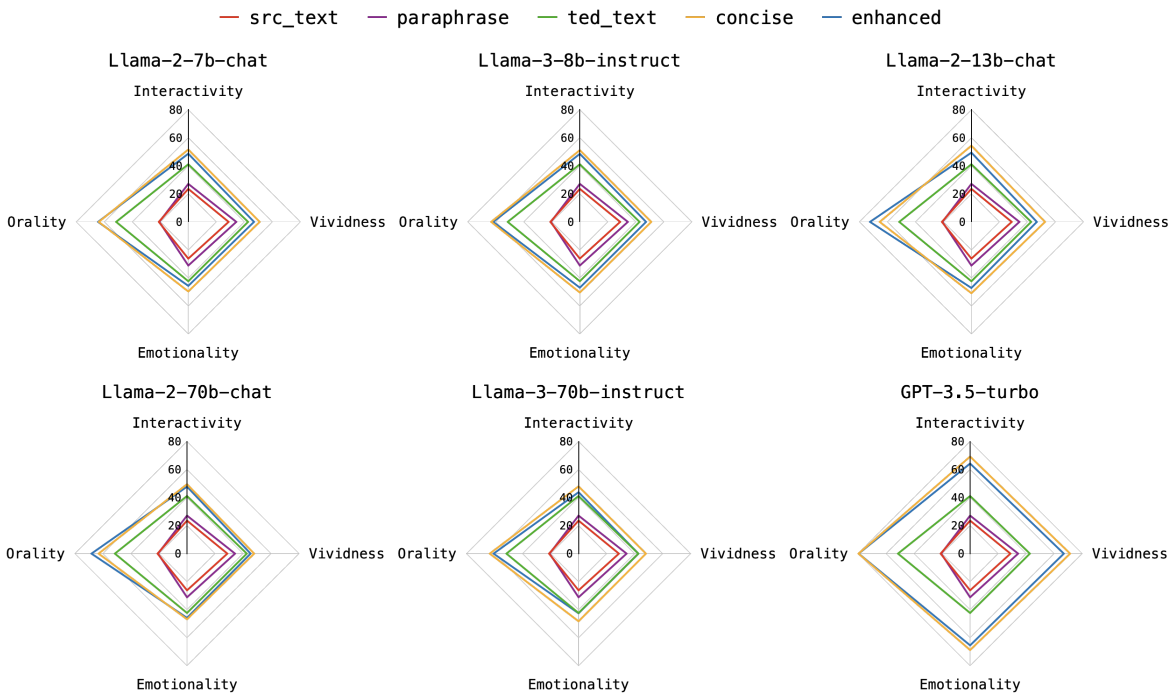

Figure 4: Radar plot of text-level style strength of passages transferred by different LLMs.(800 ± 200 tokens).

and **TED** are not parallel. **ST** and **PT** have low public-speaking style strength but high semantic preservation scores.

Prompts (1). ***concise prompt***, a brief instruction requiring models to simulate lively and engaging public speaking through style transfer. (2). ***enhanced prompt***, detailed guidance on the dimensions of stylization, cautioning against excessive stylization, and emphasizing the importance of semantic consistency. Five prompts are created for each type shown in Figure 16.

Models We employ Llama 2-Chat (Touvron et al., 2023), Llama 3-Instruct (AI@Meta, 2024) and `gpt-3.5-turbo` to generate public-speaking style texts based on **ST** and **Prompts**.

5.2 Evaluation Results and Analysis

Figures 4 and 5 illustrate the overall text-level style strength score and the fine-grained style strength distribution for texts of 800 tokens, respectively. The results of the semantic preservation evaluation are presented in Figure 6. Results for 400 ± 100 and 1200 ± 200 tokens are shown in Appendix C, which have a similar trend.

Over-stylization Firstly, the transferred texts generated by different models using concise prompts and enhanced prompts (which discourage over-stylization) exhibit stronger style strength at the text level in each dimension compared to TED-Talks data, particularly in *Orality*. Secondly, for GPT-3.5, known for its strong instruction-following ability, the style strength in each dimension is excessively higher than in real-world scenarios. For example, as shown in Figure 1 and Case 2 in Figure 22, there are too many style elements. Furthermore, the stylization of the underlined parts in Figure 22 seems to be intentionally exaggerated and appears unnatural, diverging from human preferences that do not align with human preferences. We assume that the model may be unable to truly understand the modeling process of language style in real public speaking scenarios and instead executes instructions mechanically. Thirdly, the style strength of LLaMA3s is closer to the real style strength than that of LLaMA2s. Moreover, Llama-3-70-Instruct performs best, with style strength closer to real-world norms in all four dimensions, particularly after the use of enhanced instructions. However, its style strength in *Orality* remains relatively high.

Uneven Style Strength Distribution The results comparing style strength distribution across different positions within texts are illustrated in Figure 5. Specifically, in terms of interactivity and orality, the style distribution in real-world scenarios is more uniform, with slight increases in stylization at the opening and closing segments. This pattern aligns with the conventional use of introductory and con-

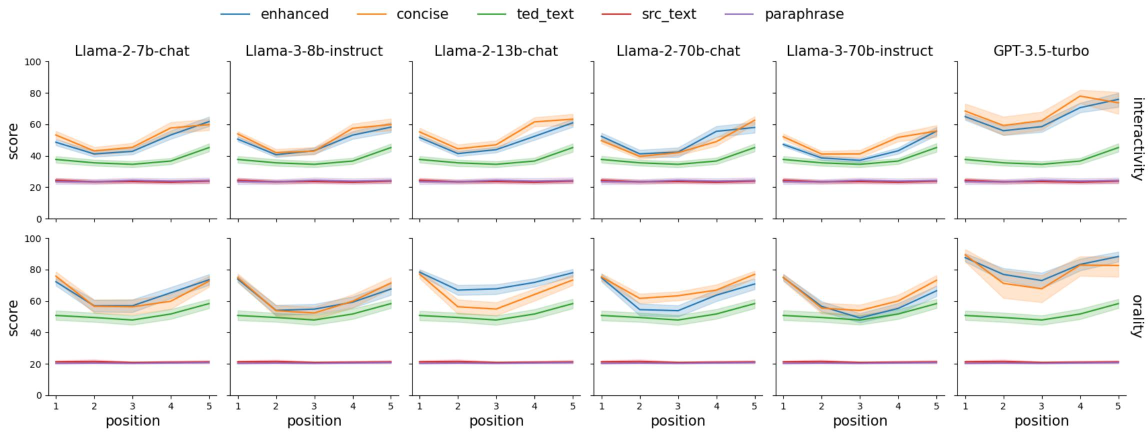

Figure 5: Style strength distribution of passages (800 $\pm$ 200 tokens) transferred by different LLMs in *Interactivity* and *Orality*.

cluding phrases in real-world settings. However, for LLMs, the distribution exhibits a U-shaped pattern, indicating excessive stylization at the text's beginning and end. For instance, as depicted in Figure 7, compared to real-world texts, LLMs such as GPT-3.5, Llama-3-70B-Instruct, and Llama-2-Chat-70B demonstrate pronounced stylization at the beginning and end, but less so in the middle sections(Case in Figure 20). This disparity suggests that LLMs diverge from human-like style modeling, potentially due to a model-inherent tendency to focus on the beginning and end of long texts, or possibly indicative of the model's 'lazy' behavior.

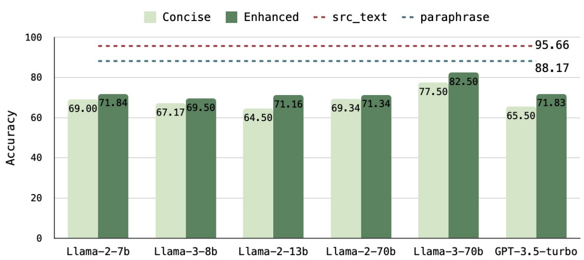

Figure 6: Semantic preservation based on QA.

Severe Semantic Degradation The results shown in Figure 6 reveal significant semantic loss during style transfer. The high accuracy observed in *src_text* and *paraphrase* confirms that the QA model can effectively respond to questions based on non-stylized texts. Moreover, our questions, designed to target essential information, are relatively straightforward and can be answered provided no information is lost. Therefore, the observed decline in accuracy with stylized texts further underscores the challenges LLMs face in semantic preservation(Case 3, 4 in Figure 23, 24). Notably, models exhibiting stronger stylization (e.g., LLaMA-3-70B-Instruct vs. GPT-3.5-Turbo) tend to obscure key messages more. This is intuitive, as excessive stylization reduces the clarity of semantic content, thus hindering message comprehension for audiences, even if the essential points remain present. Additionally, while enhanced prompts focus on semantic preservation and achieve higher accuracy, they are still less than satisfactory.

6 Conclusion

In this paper, we introduce public-speaking style transfer, which requires LLMs to transform a formal and long text into a public-speaking style. By analyzing real-world data from a linguistic perspective, decompose the intricate text style into key sub-styles. We propose a fine-grained evaluation framework that enables an evaluation-driven approach to continuously assess and enhance the language style modeling capabilities of LLMs. For style strength evaluation, we distill the LLM's ability to assess the style elements of sentences into a smaller model for each sub-style. We introduce two metrics: paragraph-level stylization intensity score and stylization intensity distribution, allowing for a detailed analysis of stylized texts. For semantic preservation, we propose a QA-based method that uses QA pairs focusing on key information and logical structure. By observing changes in the QA model's accuracy, we identify and analyze the LLM's ability to retain semantics. Our experiments reveal that current LLMs exhibit issues such as over-stylization, uneven style strength distribution, and severe semantic degradation in long and complicated language style modeling.

Limitations

Domain and Style Extensions

Limited Dataset Domain While the current dataset has limited scope, our future plans involve expanding it to include more meaningful domains, such as health, sports, business, and education.

Limited Substyle Analysis This paper examines only four sub-style dimensions, omitting an analysis of more intricate scenarios. By employing a decomposition approach, we can continuously introduce additional styles, such as the daily spoken speech presentations revealing personalities. This method allows us to explore a wide range of stylistic variations and experiment with diverse ways of delivering oral content, enhancing the adaptability and creativity of our spoken language generation.

Improvement of Evaluation Methods

Punishment Mechanism The current evaluation system overlooks the "punishment rule." For instance, if a model employs excessively exaggerated descriptions in its speech, it may receive a higher score in the "vividness" or "orality" dimension. However, such descriptions may not align with our expectations.

Effective Evaluation of Planing in Public-Speaking Style modeling in public speaking might involve elements related to planning. In this regard, incorporating insights from psychological language modeling (Ratner and Gleason, 2004; Crocker and Brouwer, 2023). could help design more effective evaluation methods.

Ethics Statement

Here we discuss the primary ethical considerations of PSST.

Intellectual Property Protection The utilized data is publicly available, permitting the reproduction, utilization, and modification of its content.

Content and Impact The PSST task mandates the model to rephrase the text, potentially leading to the generation of inaccurate information, despite our efforts to maintain consistency with the original text through prompts. Simultaneously, we note that the transformation of speech style increases the likelihood of the model producing a more "inflammatory" language style. The observed outcomes do not manifest violence, discrimination, or other related issues. However, these are aspects that warrant additional scrutiny and should be addressed in future research.

References

AI@Meta. 2024. Llama 3 model card.

J. Maxwell Atkinson. 1985. *Public speaking and audience responses: some techniques for inviting applause*, Studies in Emotion and Social Interaction, page 370–410. Cambridge University Press.

Steven A. Beebe and Susan J. Beebe. 2005. The public speaking handbook.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Roger A. Clarke. 1994. The digital persona and its application to data surveillance. *Inf. Soc.*, 10:77–92.

Stephanie J Coopman and James Lull. 2018. *Public speaking: The evolving art*. Cengage.

Matthew W. Crocker and Harm Brouwer. 2023. *Computational Psycholinguistics*, 2 edition, Cambridge Handbooks in Psychology, page 890–920. Cambridge University Press.

Godefridus Bali Geroda, Widi Syahtia Pane, et al. 2023. An analysis language style based on the level of formality according to martin joos theory. *Inquest Journal*, 1(02):163–174.

Michael A.K. Halliday. 1989. Spoken and written language.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter*, 24(1):14–45.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

April A. Kedrowicz and Julie Taylor. 2016. Shifting rhetorical norms and electronic eloquence. *Journal of Business and Technical Communication*, 30:352 – 377.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multidimensional evaluation for text style transfer using chatgpt. *arXiv preprint arXiv:2304.13462*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Aman Madaan, Amrith Rajagopal Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Annual Meeting of the Association for Computational Linguistics*.

James C. McCroskey, Jason S. Wrench, and Virginia Peck Richmond. 2003. *Principles of Public Speaking*. The College Network, Indianapolis, IN.

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

Swapnil Morande and Mitra Amini. 2023. Digital persona: Reflection on the power of generative ai for customer profiling in social media marketing.

OpenAI. 2022. Introducing chatgpt.

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

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876.

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

N.B. Ratner and J.B. Gleason. 2004. Psycholinguistics. In Larry R. Squire, editor, *Encyclopedia of Neuroscience*, pages 1199–1204. Academic Press, Oxford.

Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 187–193.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017a. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Yizhe Yang, Huashan Sun, Jiawei Li, Runheng Liu, Yinghao Li, Yuhang Liu, Heyan Huang, and Yang Gao. 2023. Mindllm: Pre-training lightweight large language model from scratch, evaluations and domain applications.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. *Advances in Neural Information Processing Systems*, 31.

Jiaao Zhan, Yang Gao, Yu Bai, and Qianhui Liu. 2022. Stage-wise stylistic headline generation: Style generation and summarized content insertion. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4489–4495. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *ArXiv*, abs/2401.02385.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A PSST Dataset

A.1 Target Dataset: Public Speaking Style Data from Real Scenarios

Ted Talks Ted Talks is a series of short, powerful presentations that share ideas and insights on creativity, science, and culture. The raw dataset can be found here[12]. Processed texts can be found in our github. We use 20 samples for prior fine-grained analysis in Section 3.3

Political Speeches[13] Political speeches are crafted to persuade, using emotional appeals and persuasive language to emphasize points, highlight policies, and influence public opinion. We use 20 samples for prior fine-grained analysis in Section 3.3

Academic Presentations[14] Academic presentations introduce scientific research to the audience, focusing on engaging the audience and making complex content easier to understand. We use 20 samples for prior fine-grained analysis in Section 3.3

Lecture Transcripts[15] Lecture transcripts are delivered by teachers in class, designed to facilitate student understanding and engagement over an extended period. We use 20 samples for prior fine-grained analysis in Section 3.3

A.2 Source Dataset: Official-Style Data Collected of PSST

The sources of data utilized for official-style texts are as follows:

News The data from the news category is sourced from the Fake and Real News dataset available on Kaggle[16]. This dataset comprises entire news articles from Reuters and was intended for news classification, including both real and fake news. We selected the subset of the real news.

Paper Abstracts The dataset from the research paper abstract category is sourced from the arXiv Dataset on Kaggle[17]. This dataset is a repository containing a substantial number of articles from arXiv, including details like article titles, authors, categories, abstracts, and full-text PDFs. For our purposes, we have extracted the abstract portions.

Wikipedia The dataset from the encyclopedia category is obtained from Hugging Face's wikitext dataset[18], which is a collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia.

A.3 Data Processing and Final PSST Dataset

To ensure the reliability of our evaluation and comprehensively assess the model's stylization capabilities, we further select the source and target datasets by filtering based on token counts, ensuring comparable lengths between the two. The final dataset we used for PSST is shown in Table 3.

Token Nums	Targe Dataset	Source Dataset			
	Ted Talks	News	Abstract	Wiki	Total
400 ± 100	32	40	40	40	120
800 ± 200	99	40	40	40	120
1200 ± 200	144	60	-	60	120

Table 3: Statistics of the final source and target datasets used for PSST

[12]https://huggingface.co/datasets/iwslt2017/viewer/iwslt2017- en-zh
[13]https://www.americanrhetoric.com/
[14]https://iwslt.org/2023/multilingual
[15]https://www.webpages.uidaho.edu/psyc390/index.htm
[16]https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
[17]https://www.kaggle.com/datasets/Cornell-University/arxiv
[18]https://huggingface.co/datasets/wikitext

B Datasets and Training Details of Style Evaluation Modeling

B.1 Statistics of Sentence-Level Style-Strength Scoring Training Data

To distill the sentence-level style strength evaluation capability of LLMs into small-scale models, we ask gpt-3.5-turbo to generate training data shown in Table 4. For detailed information on prompts, please refer to Appendix D.1.

Dimension	Train set	validation set
interactivity	7769	864
emotionality	7884	876
vividness	7942	883
orality	7839	871

Table 4: Statistics of sentence-level style-strength scoring training data

B.2 Training Details of Sentence-Level Scoring Models

For each evaluation dimension, we fine-tuned TinyLlama-1.1B (Zhang et al., 2024) as the evaluation model. The training parameters are detailed in Table 5.

Optimizer	LR_Scheduler	Learning Rate	Batch Size	Epochs
AdamW	Warmup=0.03 Decay="cosine"	2×10^{-5}	256	6

Table 5: Training Details of evaluation modeling.

C Experiments

C.1 Results of Different Text Lengths in PSST

In our investigation, we employed Public-Speaking Style Transfer (PSST) on text samples of three distinct lengths: 400 ± 100 tokens, 800 ± 200 tokens, and 1200 ± 200 tokens. The experiments aimed to evaluate the impact of text length on the efficacy of current LLMs. The results of these experiments are systematically presented in Table 6 & Figure 7 (for 800 ± 200-token texts), Table 7 & Figure 8 (for 400 ± 100-token texts) and Table 8 & Figure 9 (for 400 ± 100-token texts), respectively. These tables detail the performance metrics obtained, providing a clear comparative analysis of the outcomes across different text lengths.

D Prompts

D.1 Prompt for LLMs to Generate Sentence-Level Data to Train Evaluation Models

We utilized the capabilities of GPT-3.5 to generate sentence-level data for training models to evaluate style strength. As depicted in Figures 10(interactivity), 12(emotionality), 11(vividness), and 13(orality), we ask the GPT-3.5 to both generate sentences with varying levels of stylistic intensity and concurrently assess the style strength of each sentence. We require the generated sentences to be consistent in meaning and close in length to minimize the impact of factors unrelated to style strength when scoring.

Particularly, for "emotionality", we ask the model to identify the appropriate emotion before generating sentences with different levels of emotionality. For "emotionality" and "orality," where the inclusion of additional emotional content and alterations in sentence patterns (e.g., simplifying complex structures) could significantly change sentence length, we opt not to restrict sentence length.

D.2 Prompts for LLMs to Generate QA-pairs for semantic preservation evaluation

Prompts used to generate QA pairs are shown in Figure 14 and 15.

Model	Style Strength Score(%)				Semantic Preservation(%)		
	Interactivity	Vividness	Emotionality	Orality	Details	Logic	Average
Real Scenes							
Ted Talks	41.11	42.67	42.36	51.50	-	-	-
Political Speech	47.56	54.91	53.57	45.92	-	-	-
Course Teaching	29.72	33.77	29.70	33.60	-	-	-
Baseline							
ST	23.54	28.79	26.20	21.07	97.00	94.33	95.66
PT(GPT-3.5)	27.22	34.92	31.42	20.88	88.67	87.67	88.17
Concise Prompt							
LLaMA-2-7b	51.60	50.97	49.66	63.90	69.00	69.00	69.00
LLaMA-3-8b	51.18	51.00	50.43	63.28	65.67	68.67	67.17
LLaMA-2-13b	54.36	52.57	50.92	65.45	66.33	62.67	64.50
LLaMA-2-70b	49.52	47.93	46.85	63.11	71.67	67.00	69.34
LLaMA-3-70b	47.98	48.08	48.29	63.38	78.33	76.67	77.50
GPT-3.5	69.24	71.39	68.79	79.66	64.33	66.67	65.50
Enchanced Prompt							
LLaMA-2-7b	48.63	46.92	45.65	64.39	74.00	69.67	71.84
LLaMA-3-8b	48.54	47.27	46.96	61.80	71.00	68.00	69.50
LLaMA-2-13b	49.42	46.72	47.15	72.37	70.00	72.33	71.16
LLaMA-2-70b	47.78	45.39	45.85	68.20	72.00	70.67	71.34
LLaMA-3-70b	43.80	42.69	42.58	60.68	84.33	80.67	82.50
GPT-3.5	64.27	66.89	65.46	79.58	73.33	70.33	71.83

Table 6: Style Strength Evaluation Results(**800** $\pm$ 200 tokens). **a.** The original score ranges from 0 to 5. **b.** **ST** means *Source Text* and **PT** means *Paraphrased Text*, both are styleless. **c.** Scored by TinyLlama-1.1b(Reward).

D.3 Prompt for LLMs to Perform PSST

We design two prompt types for LLMs to conduct Text Speech-Style Transfer (PSST): a concise prompt and an enhanced prompt, illustrated in Figures 16. The enhanced prompt incorporates emphasis, hints, and guidance pertaining to various speech-style characteristics. This is intended to assist LLMs in generating text with more appropriate and accurate stylistic features aligned with human preferences.

D.4 Prompt for LLMs to Evaluate Style Strength

Figure 17 depicts the prompt used for Llama 2-Chat-70B to assess speech-style strength. In this prompt, we specify the text style transfer task and emphasize and elucidate the features that warrant special attention, aiding the model in accurate evaluations.

E Human Annotation

E.1 Human Annotation for Prior Fine-grained Analysis

For the annotation of real-world public speaking data features, we sample 300 sentence instances from four public speaking contexts: TED Talks, political speeches, academic presentations, and lecture transcripts. Based on prior research on oral public speaking (McCroskey et al., 2003; Beebe and Beebe, 2005; Atkinson, 1985; Halliday, 1989), we provide seven candidate features for annotation: (1) Interactivity, (2) Emotionality, (3) Filler Words, (4) Vividness, (5) Ambiguity, (6) Abbreviations, and (7) Informal Lexicon. Each annotator is required to perform two types of annotations: (1) Multi-label: annotators are required to select all candidate features contained in each instance. (2) Best-one label: annotators are required to choose the feature that most strongly represents the public speaking style of the sentence. Definitions for each candidate dimension provided to the annotators are as follows:

- **Interactivity:** Interactivity in public speaking refers to the speaker engaging with the audience through various means such as posing thought-provoking, facilitating personal reflection, and crafting intriguing hypothetical scenarios.
- **Emotionality:** Public speaking contains the speaker's appropriate views and attitudes on specific events to reflect the speaker's emotional tendencies and inner thoughts.

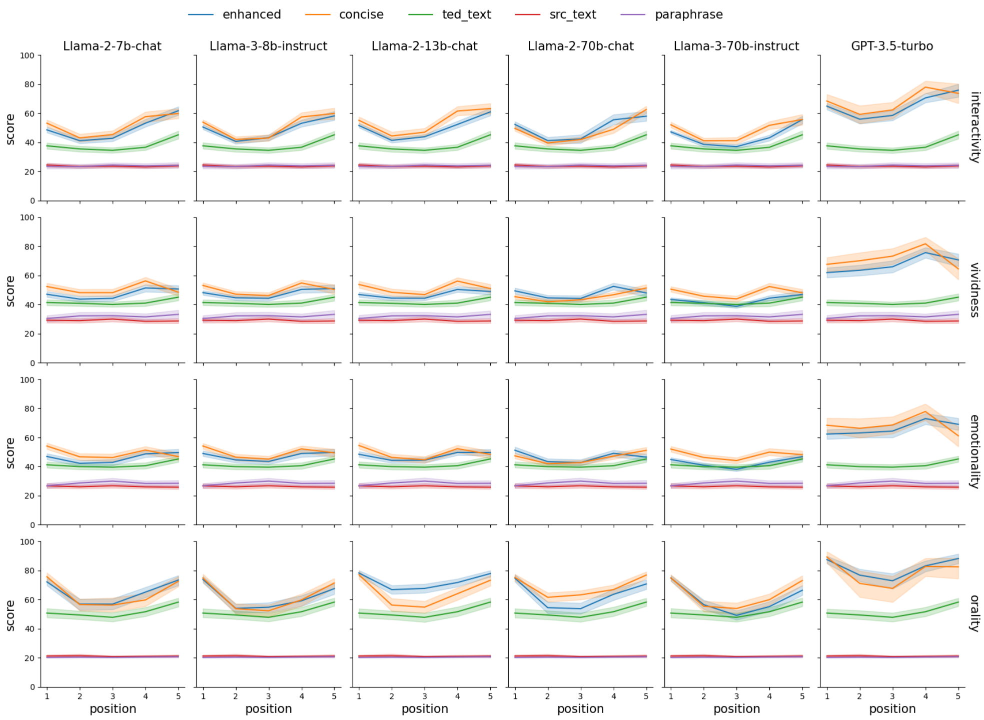

Figure 7: Style Strength Distribution of Passages Transferred by Different LLMs(800 $\pm$ 200 tokens).

- **Filler Words:** Filler words include "um," "ah," "you know," and similar phrases that speakers use to fill pauses during their speech. While sometimes seen as a sign of nervousness, they can also function to give the speaker time to think.
- **Vividness:** In public speaking, speakers should present information in a lively, easy-to-understand way, such as using analogies and metaphors to make complex ideas more accessible and engaging.
- **Ambiguity:** This feature reflects the overall clarity of the sentence. Ambiguity can arise from using complex, unclear, or overly verbose language that makes the content difficult to follow.
- **Abbreviations:** This includes the use of shortened forms of words or phrases. In speeches, abbreviations can make communication quicker and fit more informal settings.
- **Informal Lexicon:** Public speakers tend to use casual or colloquial language. It contrasts with formal speech and can make the speaker appear more relatable and approachable.

E.2 Human Annotation for Text-Level Style Strength Evaluation

Figure 18 presents the questionnaire designed for human annotators tasked with public-speaking style strength annotation mentioned in Section 4.2. We engaged 3 graduate students who are proficient in English but are not linguists as annotators.

F Case Study

F.1 Audio and video demos of PSST

To vividly demonstrate the PSST task, we utilize text-to-speech[19] and text-to-video[20] tools to animate virtual avatars delivering public speaking based on text generated by LLMs. By comparing speech and video based on text before and after style transfer, we provide a more intuitive illustration of the practical

[19]https://elevenlabs.io/
[20]https://www.synthesia.io/

Model	Style Strength Score(%)				Semantic Preservation(%)		
	Interactivity	Vividness	Emotionality	Orality	QA1	QA2	Average
Baseline							
ST	23.38	28.03	25.82	20.75	92.67	98.33	95.50
PT	27.56	36.42	31.87	21.15	89.67	93.33	91.50
TED	41.08	40.11	41.50	52.50	-	-	-
Concise Prompt							
LLaMA-2-7b	53.53	51.57	50.33	70.10	80.00	77.33	78.67
LLaMA-2-13b	55.54	53.12	51.80	70.80	79.67	81.00	80.33
LLaMA-2-70b	50.29	48.83	47.04	66.48	81.67	82.67	82.17
LLaMA-3-8b	50.96	50.25	48.89	67.11	81.67	83.00	82.33
LLaMA-3-70b	48.71	48.98	48.82	68.41	86.33	85.33	85.83
GPT-3.5	66.63	68.09	64.54	77.74	80.00	79.67	79.83
Enchanced Prompt							
LLaMA-2-7b	49.69	47.58	47.00	70.82	79.33	82.67	81.00
LLaMA-2-13b	50.06	46.08	47.05	75.67	79.00	81.00	80.00
LLaMA-2-70b	48.89	46.66	46.11	70.13	80.33	83.00	81.67
LLaMA-3-8b	48.22	46.63	45.40	63.51	83.00	84.33	83.67
LLaMA-3-70b	44.98	43.99	43.70	65.08	92.33	92.33	92.33
GPT-3.5	60.28	62.90	60.74	76.42	82.00	87.00	84.50

Table 7: Style Strength Evaluation Results(**400** $\pm$ 100 tokens). **a.** The original score ranges from 0 to 5. **b. ST** means *Source Text* and **PT** means *Paraphrased Text*, both are styleless. **c.** Scored by TinyLlama-1.1b(Reward).

value and applicability of the PSST task. Figures 19 are screenshots of the generated video. All demos are available at this URL[21].

F.2 Bad Cases of LLMs in PSST

In Figures 21, 22, 23, 24, we provide specific cases of challenges encountered by LLMs during the execution of PSST, as discussed in Section 5.2. These cases encompass issues denoted as Excessive Style Strength, Uneven Style Strength Distribution, and Severe Semantic Degradation.

F.3 Bad Cases of Llama 2-Chat-70B in Speech-Style Strength Evaluation

Inconsistency and insensitivity to the style strength of LLMs as speech-style strength refers to the challenge of obtaining inconsistent results for the same input across multiple tests and assigning the same score despite noticeable variations in speech-style strength. Examples are shown in Figure 25.

F.4 Bad Cases of ChatGPT in Content Preservation Evaluation

We attempt to directly evaluate the semantic consistency of two passage-level texts using GPT-3.5 and GPT-4. We designed a variety of prompts including those shown in Figure 26. When two input texts are completely unrelated, GPT can correctly determine that they are not related. But when a text is part of another text, GPT will tell that their semantics are consistent. Whether we ask GPT to score semantic consistency or let it judge whether two texts are consistent, their performance is unsatisfactory.

[21] https://grateful-sesame-4aa.notion.site/Presentation-of-PSST-de0bcc31121442278a158851aa180fdf

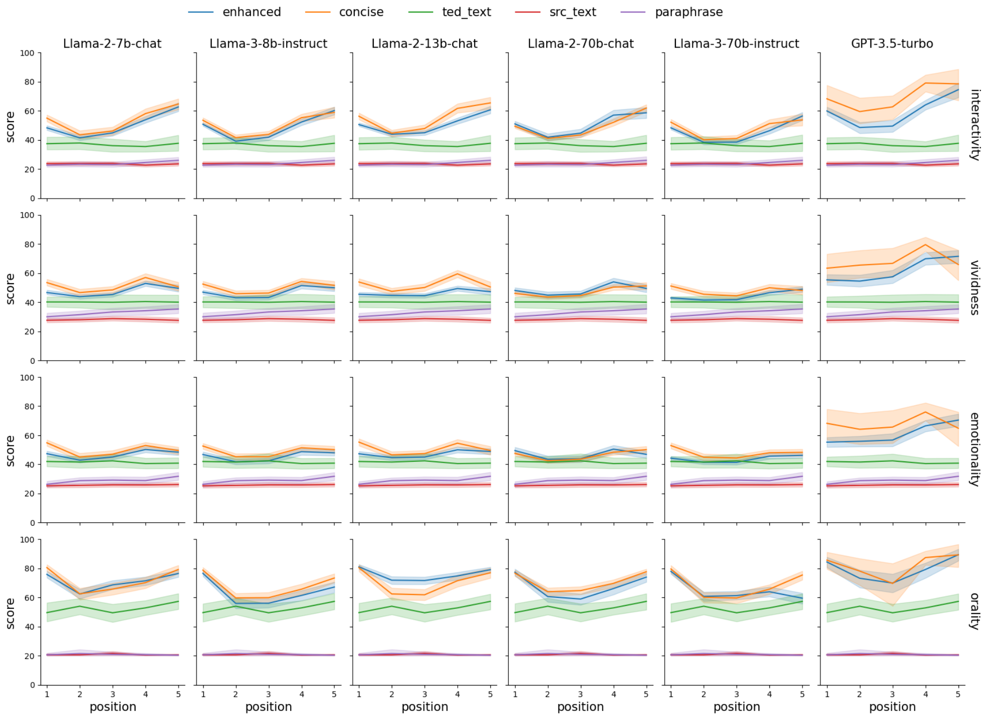

Figure 8: Style Strength Distribution of Passages Transferred by Different LLMs(400 ± 100 tokens).

Model	**Style Strength Score(%)**				**Semantic Preservation(%)**		
	Interactivity	Vividness	Emotionality	Orality	QA1	QA2	Average
			Baseline				
ST	25.12	30.73	27.48	21.75	98.00	97.00	97.50
PT	31.25	40.17	36.20	21.20	80.00	84.33	82.17
TED	38.28	41.45	40.07	50.63	-	-	-
			Concise Prompt				
LLaMA-2-7b	54.32	52.47	50.80	63.68	55.00	61.00	58.00
LLaMA-2-13b	56.67	53.26	52.45	65.03	55.33	62.33	58.83
LLaMA-2-70b	49.31	46.69	45.75	58.16	66.67	66.33	66.50
LLaMA-3-8b	53.36	51.18	50.82	61.52	64.67	71.67	68.17
LLaMA-3-70b	48.07	47.32	48.30	61.76	74.33	72.67	73.50
GPT-3.5	71.13	68.12	66.01	76.26	59.67	59.33	59.50
			Enchanced Prompt				
LLaMA-2-7b	47.86	45.82	44.83	60.18	67.67	68.33	68.00
LLaMA-2-13b	49.67	46.52	47.04	68.50	63.00	68.67	65.83
LLaMA-2-70b	46.82	45.12	45.20	63.03	70.33	71.33	70.83
LLaMA-3-8b	47.86	46.68	46.16	57.82	71.67	73.67	72.67
LLaMA-3-70b	43.16	42.34	43.00	57.98	82.00	82.33	82.17
GPT-3.5	66.38	68.49	67.06	78.19	63.33	67.67	65.50

Table 8: Style Strength Evaluation Results(**1200** ± 100 tokens). **a.** The original score ranges from 0 to 5. **b. ST** means *Source Text* and **PT** means *Paraphrased Text*, both are styleless. **c.** Scored by TinyLlama-1.1b(Reward).

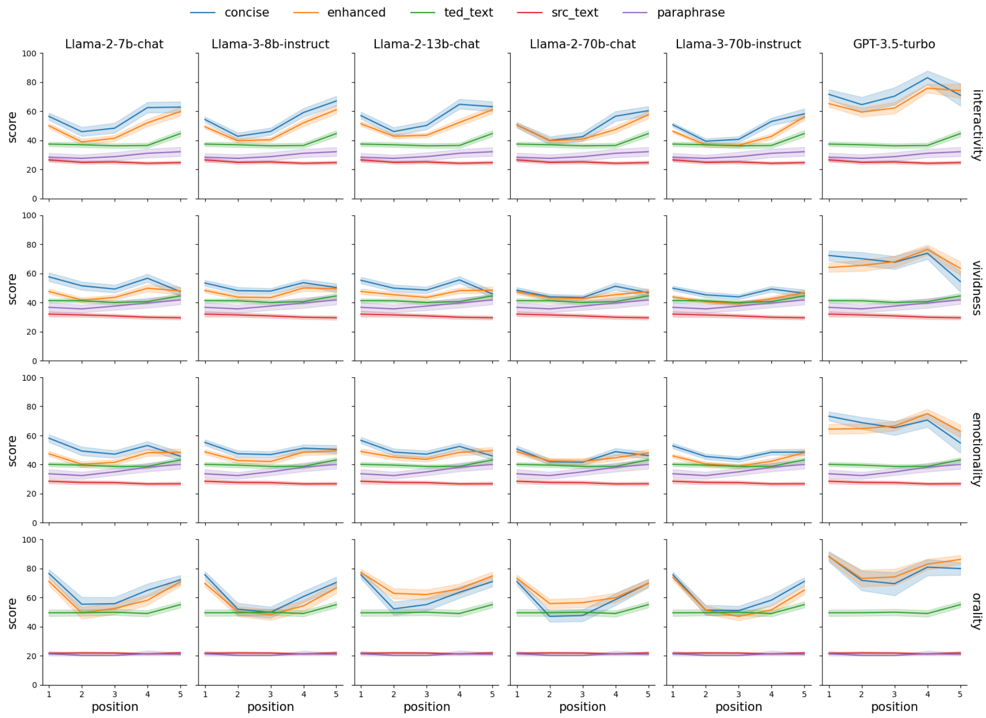

Figure 9: Style Strength Distribution of Passages Transferred by Different LLMs(1200 ± 200 tokens).

Sentence:
{input_sentence}
Objective:
You have two goals:
1. As a **Sentence Rewriter**, your task is to enhance the level of **interactivity** in the provided sentence. Your goal is to actively engage the audience by generating five sentences, each progressively increasing in interactive elements while preserving the original sentence's meaning.
2. As a **Sentence Scorer**, your task is to score each sentence on a scale from 1 to 5, where: 1 indicates minimal engagement or interaction, 4 signifies a highly interactive and engaging sentence, 5 is reserved for sentences that are exceptionally well-crafted and cannot be improved further in terms of engagement. Your evaluation should focus on how attracted and engaged you feel while reading each sentence.
Guidelines:
1. Your rewriting should foster active participation and dialogue among the audience members.
2. Four sentences should be close in length, allowing for a modest expansion of 10 to 20 words if necessary.
3. Here are some reference methods and you may employ any other method you deem feasible and effective:
(1). **Pose Thought-Provoking Questions**
(2). **Facilitate Personal Reflection**
(3). **Stimulate Audience Interaction**
(4). **Craft Intriguing Hypothetical Scenarios**
(5). **Weave Compelling Stories**
(6). **Empower with a Call to Action**
4. You need to generate different levels of sentences according to the specified score, the key of the return dict.
Format:
Please generate the following JSON formatted output and nothing else:
{
"1": "[Sentence with slight interactivity.]",
"2": "[Sentence with partial interactivity.]",
"3": "[Sentence with medium interactivity.]",
”4": "[Sentence with a lot of interactivity.]",
”5": "[Sentence with maximum interactivity.]"
}

Figure 10: Prompt used for `gpt-3.5-turbo` to generate and score different sentence-level style strengths in interactivity dimension.

Sentence:
{input_sentence}
Objective:
You have two goals:
1. As a **Sentence Rewriter**, your task is to enhance the level of **vividness** in the provided sentence. Your goal is to make the obscure information delivered more interesting, vivid, and easy to understand by generating five sentences, each progressively increasing in vivid elements while preserving the original sentence's meaning.
2. As a **Sentence Scorer**, your task is to score each sentence on a scale from 1 to 5, where: 1 indicates minimal interesting or vivid, 4 signifies a highly absorbing and vivid sentence, 5 is reserved for sentences that are exceptionally well-crafted and cannot be improved further in terms of vividness.
Guidelines:
1. Your rewriting should foster active participation and dialogue among the audience members.
2. Do not alter the fundamental semantics of the original sentence.
3. Four sentences should be close in length, allowing for a modest expansion of 10 to 20 words if necessary.
4. Here are some reference methods and you may employ any other method you deem feasible and effective:
(1). **Use Analogies and Metaphors**: Compare the obscure information to something familiar to the audience.
(2). **Incorporate Examples**: Use specific examples to clarify abstract concepts.
(3). **Break Down Information**: If explaining a complex process, outline it step-by-step rather than presenting it all at once.
5. Generate sentences with different levels of vividness according to the specified score.
Format:
Please generate the following JSON formatted output and nothing else:
```
{
"1": "[Sentence with slight vividness.]",
"2":  "[Sentence with partial vividness.]",
"3": "[Sentence with medium vividness.]",
"4": "[Sentence with a lot of vividness.]",
"5": "[Sentence with maximum vividness.]"
}
```

Figure 11: Prompt used for `gpt-3.5-turbo` to generate and score different sentence-level style strengths in the vividness dimension.

Sentence:
{input_sentence}
Objective:
Your task is twofold:
1. As a **Sentence Rewriter**: First, **identify an appropriate emotion** based on the content of the provided sentence. **Then, generate** five sentences that **incrementally increase in emotionality**, each preserving the original meaning.
2. As a **Sentence Scorer**: Evaluate each sentence on a scale from 1 to 5 based on its level of emotionality, where: 1 represents minimal emotional elements, very formal; 4 indicates rich emotional expression, capable of striking a chord; 5 is reserved for exceptionally well-crafted sentences that maximize emotional engagement and cannot be improved further. Your evaluation should focus on the intensity of emotion you experience while reading each sentence.
Guidelines:
1. Your rewriting should aim to accurately identify and appropriately express emotions.
2. Employ the following methods to modulate emotionality, and consider other effective techniques as needed:
(1). **Emotional Language**: Use descriptive adjectives, adverbs, and emotionally charged words or phrases to significantly enhance the expression of feelings.
(2). **Adjust Sentence Structure**: Use short, abrupt sentences to convey urgency or anger, and longer, flowing sentences for sadness or nostalgia.
(3). **Repetition and Pauses**: Emphasize emotional points through strategic repetition and use of pauses to allow emotions to resonate.
(4). **Rhetorical Devices**: Incorporate devices such as anaphora (repetition at the beginning of sentences) and tricolon (a series of three parallel elements) to strengthen the emotional impact.
3. You need to generate different levels of sentences according to the specified score, the key of the return dict.
Format:
Please generate the following JSON formatted output and nothing else:
```
{
"1": "[Sentence with slight emotionality.]",
"2":  "[Sentence with partial emotionality.]",
"3": "[Sentence with medium emotionality.]",
"4": "[Sentence with a lot of emotionality.]",
"5": "[Sentence with maximum emotionality.]"
}
```

Figure 12: Prompt used for `gpt-3.5-turbo` to generate and score different sentence-level style strengths in the emotionality dimension. Particularly, we ask the `gpt-3.5-turbo` to identify the appropriate emotion before generating sentences and we opt not to restrict sentence length

Sentence:
{input_sentence}
Objective:
You have two goals:
1. As a **Sentence Rewriter**: Enhance the **orality** of the provided sentence. Generate five sentences that increase in colloquialism incrementally while preserving the original meaning.
2. As a **Sentence Scorer**, your task is to score each sentence on a scale from 1 to 5 based on its level of orality, where: 1 indicates minimal colloquial elements, very formal; 4 signifies rich in oral characteristics, highly suitable for spoken communication; 5 means exceptionally well-crafted for oral expression, beyond improvement.
Guidelines:
1. Your rewriting should aim to make the sentence sound more natural and conversational.
2. Here are some reference methods and you may employ any other method you deem feasible and effective:
(1). **Use Contractions and Colloquial Language**: Use contractions and informal expressions to reduce formality and increase approachability.
(2). **Simplify Vocabulary and Sentence Structure**: Replace complex words and phrases with simpler ones and break long sentences into shorter, clearer segments.
(3). **Incorporate Filler Words and Phrases**: Integrate fillers and transitional phrases like "you know," "let's see," or "well," to add a conversational flow.
(4). **Employ Active Voice**: Prefer active voice over passive for a more direct and engaging tone.
(5). **Redundancy**: Employ techniques like rephrasing and summarizing to reinforce key points, aiding in comprehension.
3. You need to generate different levels of sentences according to the specified score, the key of the return dict.
Format:
Please generate the following JSON formatted output and nothing else:
```
{
"1": "[Sentence with slight orality.]",
"2"： "[Sentence with partial orality.]",
"3": "[Sentence with medium orality.]",
"4": "[Sentence with a lot of orality.]",
"5": "[Sentence with maximum orality.]"
}
```

Figure 13: Prompt used for `gpt-3.5-turbo` to generate and score different sentence-level style strengths in orality dimension. We opt not to restrict sentence length due to that alterations in sentence patterns (e.g., simplifying complex structures) could significantly change sentence length.

Instructions
Please generate ten questions based on the provided text. Each question should be designed so that:
1. It can be answered directly using information explicitly stated in the text.
2. Each question should have four options in random order:
  - One correct answer that is directly supported by the text.
  - One option labeled as "Unanswerable" which indicates that the question cannot be answered based on the given text alone.
  - Two incorrect answers that are plausible but can be proven wrong with the information in the text.

Example of Output, the Output Should be in JSON Format
```
{
 "questions": [
   {
    "question": "What is primarily responsible for causing water to evaporate?",
    "options": {
      "A": "Sunlight",
      "B": "Plants",
      "C": "Animals",
      "D": "Mountains"
    },
    "correctAnswer": "A"
   },
   ...
   {
    "question": "Which process describes the transition of water from vapor back to liquid?",
    "options": {
      "A": "Precipitation",
      "B": "Evaporation",
      "C": "Condensation",
      "D": "Collection"
    },
    "correctAnswer": "C"
   }
 ]
}
```

Additional Guidelines
- The questions should cover different sections of the text.
- Avoid questions that require inferential reasoning or external knowledge beyond the text.
- Focus on key facts, concepts, and direct statements from the text.
- Ensure a diversity in question types (e.g., "What", "Who", "When", "Where", "How").
- Make sure the answers are specific and not subject to interpretation based on the text.

Text
[]

Figure 14: Prompts used to generate 10 question-answer pairs related to **key information**, aiming to evaluate the semantic preservation capabilities of LLMs. The purple indicates specific guidance.

Instructions
Based on the provided text, generate ten questions that delve into the specifics of the content and require reasoning. Each question should help elucidate one or more of the following aspects:
1. Identifying the Speaker or the Audience: Questions should aim to reveal insights about who is speaking or who the intended audience is. For example, "Who is likely the speaker in the text, and what is their role?"
2. Understanding Emotions: Generate questions that explore the emotional tone or specific feelings conveyed in the text. An example could be, "What is the predominant emotion expressed by the speaker, and why might they feel this way?"
3. Exploring Logical Relationships: Create questions that examine cause-and-effect relationships, contrasts, or developments within the text. For instance, "What event prompted the speaker's response, according to the text?"
4. Each question should have four options in random order:
 - One correct answer that is directly supported by the text.
 - One option labeled as "Unanswerable" which indicates that the question cannot be answered based on the given text alone.
 - Two incorrect answers that are plausible but can be proven wrong with the information in the text.

Example of Output, the Output Should be in JSON Format

```
{
 "questions": [
   {
    "question": "What is primarily responsible for causing water to evaporate?",
    "options": {
      "A": "Sunlight",
      "B": "Plants",
      "C": "Animals",
      "D": "Mountains"
    },
    "correctAnswer": "A"
   },
   ...
   {
    "question": "Which process describes the transition of water from vapor back to liquid?",
    "options": {
      "A": "Precipitation",
      "B": "Evaporation",
      "C": "Condensation",
      "D": "Collection"
    },
    "correctAnswer": "C"
   }
 ]
}
```

Additional Guidelines
- Each question should be clearly worded to avoid ambiguity and should precisely target specific information or concepts presented in the text.
- Make sure the questions collectively cover different sections and key points of the text to provide a comprehensive understanding of the content.

Text
[]

Figure 15: Prompt used to generate 10 question-answer pairs related to **logic and structure**, aiming to evaluate the semantic preservation capabilities of LLMs. The purple indicates specific guidance.

Concise Prompts:
1. Please transform the following passage into a speech that would captivate an audience.
2. Convert the tone of the text to be more suitable for delivering a speech.
3. Re-contextualize the following passage into a speech format that is engaging and persuasive.
4. Transform the tone and style of the text to make it more suitable for a attractive speech.
5. Reframe the following passage as a compelling speech that would resonate with your audience.
6. Imagine you are addressing a crowd as you transform this passage into a speech.

Enhanced Prompts:
1. Transform the provided text into a public-speaking format, **ensuring the essential content remains unchanged and the adaptation doesn't overshadow the original message**. Focus on enhancing the text's vividness and interactivity to improve understanding and audience engagement. Incorporate expressions of opinions and emotions related to key events, and subtly include spoken language features to make the delivery feel more natural and fluid, all while keeping the stylistic elements balanced and appropriate for the context.

2. Please convert this text for public speaking by **maintaining all key elements unchanged and avoiding over-stylization**. The goal is to make the text vivid and interactive to captivate the audience, while expressing relevant opinions and emotions, and integrating natural spoken language elements to enhance relatability. Ensure the presentation style is used judiciously to enhance, not distract from, the original content.

3. Adapt this text into a public-speaking style, **preserving the core messages and ensuring the style transfer does not detract from the original content**. Aim to make the text more vivid and engaging, incorporating direct interactions and expressive elements that convey opinions and emotions tied to the discussed events. Use elements of spoken language such as filler words and informal expressions sparingly to maintain a natural flow and make the speech easier to follow.

4. Transform the given text into a format suitable for public speaking, **carefully preserving the integrity of its key components**. Enhance its vividness and interactivity to facilitate better audience understanding and engagement. Embed personal opinions and emotional responses where relevant, and fold in elements typical of spoken language to add a natural and approachable tone, all without over-stylizing the presentation.

5. Convert this text for public speaking by **keeping essential content intact and styling appropriately to avoid over embellishment**. Focus on making the text vivid and interactive, which should help in clarifying and engaging the audience more effectively. Additionally, weave in expressions of opinions and emotions pertinent to the content, and incorporate features of spoken language to create a smooth, natural delivery.

Figure 16: Concise and enhanced instructions to help LLMs perform PSST. The purple indicates specific guidance.

The above is a output of a **'official' to 'public-speaking' style transfer task**, a good public-speaking style needs to have the following characteristics:
1. **Emotionality**, that is, the speaker has the appropriate emotional expression of the event mentioned.
2. **Vividness**, that is, the speaker should try to use vivid language to make the public speaking easier to understand and interesting.
3. **Interactivity**, that is, the speaker can use questions, appeals and other ways to properly interact with the audience to stimulate the audience's interest.
4. **Orality**, that is, the content should be optimized for verbal communication, characterized by the use of simpler vocabulary and sentence structures, alongside an increased prevalence of fillers and abbreviations compared to a official text.
Now you need to score the output based on the four dimensions, respectively, with a **score range from 0 to 100**, where 0 means that the output does not have the characteristics of the corresponding dimension and 100 means that the output has the characteristics of the dimension.
You need to reply to the results in the following dictionary format:
```
{
    'Emotionality': 'Emotionality score',
    'Vividness': ' Vividness score ',
    'Interactivity': ' Interactive score ',
    'Orality': ' Orality score ',
}
```

Figure 17: Prompt used for Llama 2-Chat-70B to score the public speaking style strength in four dimensions.

Introduction
You will be presented with four texts intended for public speaking. Your task is to assess and rank these texts based on their stylistic strength across four key dimensions of public speaking style: Interactivity, Emotionality, Vividness, and Orality.

Instructions
1. Read Each Text Carefully: You will be provided with four texts. Read each one carefully, focusing on the stylistic elements.
2. Highlight Key Elements: As you read, feel free to highlight sentences, phrases, or words that you believe strongly reflect the targeted style dimension.
3. Rank the Texts: After reviewing each text, you will rank them from strongest to weakest in terms of stylistic representation for each dimension.

Evaluation Criteria
1. Interactivity: Measures how the text engages and involves the audience.
(1). Thought-Provoking Questions: Does the text pose questions that challenge the audience's thinking?
(1). Personal Reflection: Encourages the audience to reflect personally on the topics discussed.
(3). Audience Interaction: Facilitates direct interaction with the audience.
(4). Hypothetical Scenarios: Uses imaginative scenarios to engage the audience.
(5). Storytelling: Weaves compelling narratives that captivate listeners.
(6). Call to Action: Motivates the audience towards specific actions.

2. Emotionality: Appropriate perspectives and attitudes towards specific events in the original text to reflect the speaker's thoughts.
(1). Emotional Language: Use of expressive and emotionally charged language.
(2). Sentence Structure: Adjustment of sentence structure to enhance emotional delivery.
(3). Repetition and Pauses: Uses repetition and strategic pauses to emphasize emotional content.
(4). Rhetorical Devices: Effective use of rhetorical devices to amplify emotional expression.

3. Vividness: Evaluates the liveliness and clarity of the presentation.
(1). Analogies and Metaphors: Uses figurative language to relate complex ideas to familiar concepts.
(2). Examples: Provides clear, illustrative examples to explain abstract ideas.
(3). Information Breakdown: Delivers complex information in a structured, step-by-step manner.

4. Orality: Focuses on the oral presentation quality of the text.
(1). Conversational Language: Uses informal language and contractions to sound more natural.
(2). Simple Structure: Simplifies vocabulary and sentence structure for clarity.
(3). Filler Words: Integrates conversational fillers to mimic natural speech.
(4). Redundancy: Uses techniques like rephrasing to reinforce key messages.

Text Samples
You will evaluate the following texts:
Text 1: {Public-Speaking Style Text 1}
Text 2: {Public-Speaking Style Text 2}
Text 3: {Public-Speaking Style Text 3}
Text 4: {Public-Speaking Style Text 4}

Style Strength Ranking
Please sort the texts for each style as follows, where [1,2,3,4] means text1 > text2 > text3 > text4 in style strength:
Interactivity: []
Emotionality: []
Vividness: []
Orality: []

Your thoughtful assessments will contribute to our understanding of what makes effective public speaking. Thank you for your participation!

Figure 18: Human questionnaire for public-speaking style strength annotation.

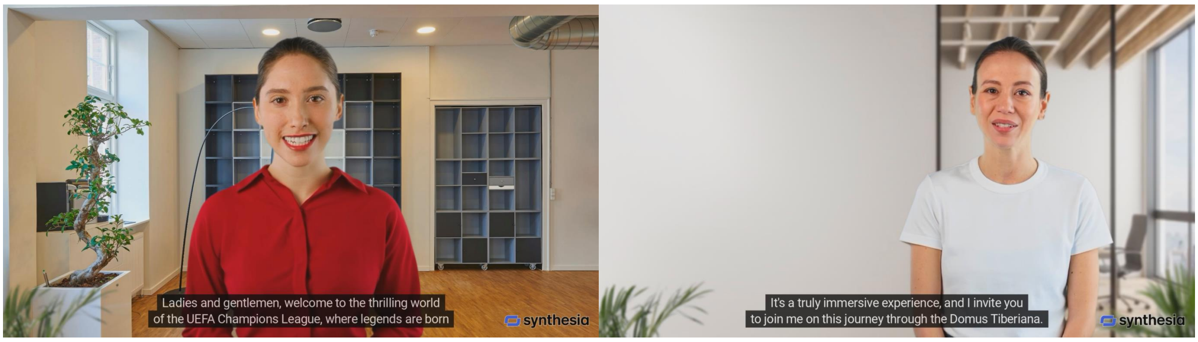

Figure 19: Screenshots of videos based on LLM-generated public speaking style text. All demos are available at this URL.

Ted Talks:
Today, I'm going to talk to you about sketching electronics. I'm, among several other things, an electrical engineer, and that means that I spend a good amount of time designing and building new pieces of technology, and more specifically designing and building electronics. And what I've found is that the process of designing and building electronics is problematic in all sorts of ways. So it's a really slow process, it's really expensive, and the outcome of that process, namely electronic circuit boards, are limited in all sorts of kind of interesting ways. So they're really small, generally, they're square and flat and hard, and frankly, most of them just aren't very attractive, and so my team and I have been thinking of ways to really change and mix up the process and the outcome of designing electronics. And so what if you could design and build electronics like this? So what if you could do it extremely quickly, extremely inexpensively, and maybe more interestingly, really fluidly and expressively and even improvisationally? Wouldn't that be so cool, and that wouldn't that open up all sorts of new possibilities? I'm going to share with you two projects that are investigations along these lines, and we'll start with this one. Magnetic electronic pieces and ferrous paper. A conductive pen from the Lewis lab at UIUC. Sticker templates. Making a switch. Music: DJ Shadow. Adding some intelligence with a microcontroller. Sketching an interface. Pretty cool, huh? We think so. So now that we developed these tools and found these materials that let us do these things, we started to realize that, essentially, anything that we can do with paper, anything that we can do with a piece of paper and a pen we can now do with electronics. So the next project that I want to show you is kind of a deeper exploration of that possibility. And I'll kind of let it speak for itself. So the next step for us in this process is now to find a way to let all of you build things like this, and so the way that we're approaching that is by teaching workshops to people where we explain how they can use these kinds of tools, and then also working to get the tools and the materials and techniques out into the real world in a variety of ways. And so sometime soon, you'll be able to play and build and sketch with electronics in this fundamentally new way. So thank you very much.

Text Transferred by Llama-3-70b-Instruct:
Ladies and gentlemen,
Today, I want to talk to you about a breakthrough in the field of commutative algebra. You see, traditionally, definitions of locally complete intersection homomorphisms have been limited to maps that are essentially of finite type or flat. But what if I told you that we can now go beyond those constraints? That's right! Our new concept is designed to work with homomorphisms of commutative noetherian rings, that's a mouthful, I know!
So, the question is, what makes this new approach so special? Well, the answer lies in how we define it. We're looking at the structure of this homomorphism in a formal neighborhood of each point of Spec S. Yeah, I know, it's a lot to take in! But bear with me.
Now the l.c.i. property can be characterized by different conditions on the vanishing of the André-Quillen homology of the R-algebra S. And let me tell you - one of these descriptions is a game-changer! It proves a conjecture by Quillen that was open even for homomorphisms of finite type. I mean - who wouldn't want to solve an open problem?
Here's the deal - if S has a finite resolution by flat R-modules and the cotangent complex is quasi-isomorphic to a bounded complex of flat S-modules, then this homomorphism is l.c.i. Boom! And we didn't get there alone. No way! We had to mix and match methods from commutative algebra, differential graded homological algebra, and homotopy theory.
But wait -there's more! This l.c.i. property is stable under all sorts of operations. I'm talking composition, decomposition, flat base change, localization, and completion. You name it!
And let's not forget the implications! Our framework allows us to state these results in their full generality. And trust me - many of these are new even with classical assumptions. For example, the stability of l.c.i. homomorphisms under decomposition finally settles an open case in Fulton's treatment of orientations of morphisms of schemes. Wow!
So, there you have it folks! A whole new world of possibilities has opened up for us. And I hope you're as excited as I am to see where this will take us next!

Text Transferred by GPT-3.5-Turbo:
Alright, folks, let's dive into the main aim of this paper: proving something pretty cool! We're talking about a scenario where we have a fancy concept known as a triangulated category, denoted as $\underline{C}$, and a subset of objects, let's call it E. The big question here is whether we can find a nifty cohomological functor, let's call it F, which lands in some abelian category, such that this set E is where our functor vanishes. In simpler terms, it's like finding where this mathematical function hits zero. And guess what? We're not just proving this for fun – it's like unlocking a mathematical treasure chest known as Nullstellensatz for these types of functors!
Now, here's where it gets interesting: for those of you who are into algebra, if we're working with an R-linear category (where R is a commutative ring), the existence of an R-linear functor F going from the opposite of $\underline{C}$ to R-modules is like hitting the jackpot! And all of this hinges on whether E plays nice with retracts and extensions. So, it's not just abstract math; it's like a symphony of algebraic harmony!
As a juicy side note, we also show that if an object Y hangs out in what we call the "envelope" of a certain set D, then that envelope also welcomes Y with open arms if the same holds true when we zoom into various localized versions of our original category $\underline{C}$ at different maximal ideals. It's like a mathematical game of pass-the-parcel but with objects and categories!
To make all this possible, we've cooked up some fresh methods that act as bridges connecting our beloved triangulated categories with their smaller, yet equally fascinating, countable subcategories. It's like opening up new pathways in a mathematical jungle!
And hey, the beauty of all this is not just in the theoretical realm; these findings can be put to work in exploring weight structures and the intriguing world of motives within these triangulated categories. It's like shining a spotlight on the hidden gems of mathematical structures!
So, let's buckle up for a mathematical adventure filled with twists, turns, and revelations that will not only stimulate your curious minds but also unveil the deep connections within these mathematical landscapes. Get ready to embark on a journey that will unveil the hidden treasures of triangulated categories and their fascinating properties!

Figure 20: Comparison of different public-speaking style texts (Ted Talks, Llama-3-70b-Instruct, GPT-3.5-Turbo). The green parts indicate strong stylistic strength. In TED talks, stylistic elements are uniformly distributed, whereas in texts generated by LLMs, these elements tend to concentrate disproportionately at the beginning and the end. Moreover, LLAMA-3 generates moderately stylized texts, while GPT-3.5 tends to produce texts that are excessively stylized.

Case 1:
Instruction(Enhanced):
Adapt this text into a public-speaking style, preserving the core messages and ensuring the style transfer does not detract from the original content. Aim to make the text more vivid and engaging, incorporating direct interactions and expressive elements that convey opinions and emotions tied to the discussed events. Use elements of spoken language such as filler words and informal expressions sparingly to maintain a natural flow and make the speech easier to follow.

Source Text:
Classical definitions of locally complete intersection (l.c.i.) homomorphisms of commutative rings are limited to maps that are essentially of finite type, or flat. The concept introduced in this paper is meaningful for homomorphisms phi : R \\longrightarrow S of commutative noetherian rings. It is defined in terms of the structure of phi in a formal neighborhood of each point of Spec S.

We characterize the l.c.i. property by different conditions on the vanishing of the Andr\'e-Quillen homology of the R-algebra S. One of these descriptions establishes a very general form of a conjecture of Quillen that was open even for homomorphisms of finite type: If S has a finite resolution by flat R-modules and the cotangent complex \\cot SR is quasi-isomorphic to a bounded complex of flat S-modules, then phi is l.c.i.

The proof uses a mixture of methods from commutative algebra, differential graded homological algebra, and homotopy theory. The l.c.i. property is shown to be stable under a variety of operations, including composition, decomposition, flat base change, localization, and completion.

The present framework allows for the results to be stated in proper generality; many of them are new even with classical assumptions. For instance, the stability of l.c.i. homomorphisms under decomposition settles an open case in Fulton's treatment of orientations of morphisms of schemes.

Text Transferred by Llama-3-70b-Instruct:
Ladies and gentlemen,

Today, I want to talk to you about a breakthrough in the field of commutative algebra. You see, traditionally, definitions of locally complete intersection homomorphisms have been limited to maps that are essentially of finite type or flat. But what if I told you that we can now go beyond those constraints? That's right! Our new concept is designed to work with homomorphisms of commutative noetherian rings, that's a mouthful, I know!

So, the question is, what makes this new approach so special? Well, the answer lies in how we define it. We're looking at the structure of this homomorphism in a formal neighborhood of each point of Spec S. Yeah, I know, it's a lot to take in! But bear with me.

Now the l.c.i. property can be characterized by different conditions on the vanishing of the André-Quillen homology of the R-algebra S. And let me tell you - one of these descriptions is a game-changer! It proves a conjecture by Quillen that was open even for homomorphisms of finite type. I mean - who wouldn't want to solve an open problem?

Here's the deal - if S has a finite resolution by flat R-modules and the cotangent complex is quasi-isomorphic to a bounded complex of flat S-modules, then this homomorphism is l.c.i. Boom! And we didn't get there alone. No way! We had to mix and match methods from commutative algebra, differential graded homological algebra, and homotopy theory.

But wait - there's more! This l.c.i. property is stable under all sorts of operations. I'm talking composition, decomposition, flat base change, localization, and completion. You name it!

And let's not forget the implications! Our framework allows us to state these results in their full generality. And trust me - many of these are new even with classical assumptions. For example, the stability of l.c.i. homomorphisms under decomposition finally settles an open case in Fulton's treatment of orientations of morphisms of schemes. Wow!

So, there you have it folks! A whole new world of possibilities has opened up for us. And I hope you're as excited as I am to see where this will take us next!

Figure 21: Case of Llama-3-70b-Instruct in PSST with enhanced instruction. The green parts indicate strong stylistic strength. The transferred text is more vivid, emotional, and attractive to the audience than the source text.

Case 2:
Instruction(Enhanced):
Convert this text for public speaking by keeping essential content intact and styling appropriately to avoid over embellishment. Focus on making the text vivid and interactive, which should help in clarifying and engaging the audience more effectively. Additionally, weave in expressions of opinions and emotions pertinent to the content, and incorporate features of spoken language to create a smooth, natural delivery.

Source Text:
The main goal of this paper is to prove the following: for a triangulated category $\underline{C}$ and $E \subset \operatorname{Obj} \underline{C}$ there exists a cohomological functor F (with values in some abelian category) such that E is its set of zeros if (and only if) E is closed with respect to retracts and extensions (so, we obtain a certain Nullstellensatz for functors of this type). Moreover, for $\underline{C}$ being an R-linear category (where R is a commutative ring) this is also equivalent to the existence of an R-linear $F: \underline{C}^{op} \to R-\operatorname{mod}$ satisfying this property.

As a corollary, we prove that an object Y belongs to the corresponding "envelope" of some $D \subset \operatorname{Obj} \underline{C}$ whenever the same is true for the images of Y and D in all the categories $\underline{C}_p$ obtained from $\underline{C}$ by means of "localizing the coefficients" at maximal ideals $p \triangleleft R$". Moreover, to prove our theorem we develop certain new methods for relating triangulated categories to their (non-full) countable triangulated subcategories.

The results of this paper can be applied to the study of weight structures and of triangulated categories of motives.

Text Transferred by GPT-3.5-Turbo:
Alright, folks, let's dive into the main aim of this paper: proving something pretty cool! We're talking about a scenario where we have a fancy concept known as a triangulated category, denoted as $\underline{C}$, and a subset of objects, let's call it E. The big question here is whether we can find a nifty cohomological functor, let's call it F, which lands in some abelian category, such that this set E is where our functor vanishes. In simpler terms, it's like finding where this mathematical function hits zero. And guess what? We're not just proving this for fun – it's like unlocking a mathematical treasure chest known as Nullstellensatz for these types of functors!

Now, here's where it gets interesting: for those of you who are into algebra, if we're working with an R-linear category (where R is a commutative ring), the existence of an R-linear functor F going from the opposite of $\underline{C}$ to R-modules is like hitting the jackpot! And all of this hinges on whether E plays nice with retracts and extensions. So, it's not just abstract math; it's like a symphony of algebraic harmony!

As a juicy side note, we also show that if an object Y hangs out in what we call the "envelope" of a certain set D, then that envelope also welcomes Y with open arms if the same holds true when we zoom into various localized versions of our original category $\underline{C}$ at different maximal ideals. It's like a mathematical game of pass-the-parcel but with objects and categories!

To make all this possible, we've cooked up some fresh methods that act as bridges connecting our beloved triangulated categories with their smaller, yet equally fascinating, countable subcategories. It's like opening up new pathways in a mathematical jungle!

And hey, the beauty of all this is not just in the theoretical realm; these findings can be put to work in exploring weight structures and the intriguing world of motives within these triangulated categories. It's like shining a spotlight on the hidden gems of mathematical structures!

So, let's buckle up for a mathematical adventure filled with twists, turns, and revelations that will not only stimulate your curious minds but also unveil the deep connections within these mathematical landscapes. Get ready to embark on a journey that will unveil the hidden treasures of triangulated categories and their fascinating properties!

Figure 22: Case of GPT-3.5 in PSST with enhanced instruction. The green parts indicate strong stylistic strength. The transferred text uses many metaphors and reflects rich emotions. However, the stylization of the underlined section seems to be intentionally exaggerated and appears unnatural, diverging from human preferences.

Case 3:
Source Text:
The genetic code is the set of rules by which information encoded within genetic material (DNA or mRNA sequences) is translated into proteins by living cells. Translation is accomplished by the ribosome, which links amino acids in an order specified by mRNA, using transfer RNA (tRNA) molecules to carry amino acids and to read the mRNA three nucleotides at a time. The genetic code is highly similar among all organisms and can be expressed in a simple table with 64 entries.

The code defines how sequences of nucleotide triplets, called codons, specify which amino acid will be added next during protein synthesis. With some exceptions, a three-nucleotide codon in a nucleic acid sequence specifies a single amino acid. Because the vast majority of genes are encoded with exactly the same code (see the RNA codon table), this particular code is often referred to as the canonical or standard genetic code, or simply the genetic code, though in fact some variant codes have evolved. For example, protein synthesis in human mitochondria relies on a genetic code that differs from the standard genetic code.

While the "genetic code" determines a protein's amino acid sequence, other genomic regions determine when and where these proteins are produced according to a multitude of more complex "gene regulatory codes".

Serious efforts to understand how proteins are encoded began after the structure of DNA was discovered in 1953. George Gamow postulated that sets of three bases must be employed to encode the 20 standard amino acids used by living cells to build proteins. With four different nucleotides, a code of 2 nucleotides would allow for only a maximum of 42.

Text Transferred by Llama-3-70b-Instruct:
Ladies and gentlemen, today we're going to talk about the fundamental language of life: the genetic code.

The genetic code is the set of rules that dictate how the information encoded in our DNA is translated into proteins by living cells. It's a remarkable process that's carried out by the ribosome, which reads the mRNA sequence three nucleotides at a time, to link amino acids in the correct order. And the amazing thing is that this code is nearly universal among all living organisms, and can be summarized in a simple table with just 64 entries.

But how does it work? Well, the code defines how sequences of nucleotide triplets, called codons, specify which amino acid will be added next during protein synthesis. With a few exceptions, a three-nucleotide codon in a nucleic acid sequence specifies a single amino acid. And it's remarkable that this code is so consistent across the vast majority of genes, that we can refer to it as the standard genetic code.

But it's not a one-size-fits-all solution. For example, protein synthesis in human mitochondria relies on a genetic code that's slightly different from the standard code. And while the genetic code determines a protein's amino acid sequence, other genomic regions determine when and where these proteins are produced according to a multitude of more complex 'gene regulatory codes'.

Now, you might be wondering, how did we come to understand this complex code? Well, it all started after the structure of DNA was discovered in 1953. George Gamow proposed that sets of three bases must be employed to encode the 20 standard amino acids used by living cells to build proteins. And with four different nucleotides, a code of 2 nucleotides would only allow for a maximum of 42.

Question:
What carries amino acids and reads the mRNA three nucleotides at a time?
Options:
A:"Transfer RNA (tRNA) molecules", B:"The genetic material", C:"Unanswerable", D:"The ribosome“

Correct Answer based on Source Text: A

Answer based on Transferred Text: D

Figure 23: Case of evaluating the semantic preservation of Llama-3-70b-Instruct in PSST. The green part indicates the location of the answer to the question in the source text. The red part indicates the location corresponding to the text after the transfer, but the key information of the question is lost.

Case 4:
Source Text:
PHOENIX (Reuters) - U.S. President Donald Trump warned on Tuesday he might terminate the NAFTA trade treaty with Mexico and Canada after three-way talks failed to bridge deep differences. The United States, Canada and Mexico wrapped up their first round of talks on Sunday to revamp the trade pact with little sign of a breakthrough coming.

Trump reopened negotiations of the 1994 treaty out of concern U.S. economic interests were suffering. “Personally, I don’t think we can make a deal. I think we’ll probably end up terminating NAFTA at some point,” Trump said at a political rally in Phoenix, Arizona.

Suggesting a termination might help jumpstart the negotiations, Trump said: “I personally don’t think you can make a deal without a termination.”

Following Trump’s remarks, Mexican foreign minister Luis Videgaray tweeted: “No surprises: we’re already in a negotiation. Mexico will remain at the table with serenity, firmness and with the national interest ahead.”

In a joint statement issued at the end of five days of negotiations in Washington, the top trade officials from the three countries said Mexico would host the next round of talks from Sept. 1 to 5. The talks will move to Canada later in September, then return to the United States in October, with additional rounds planned for later this year.

Text Transferred by GPT-3.5:
Ladies and gentlemen, I come before you today with urgent news regarding the future of our trade agreements. As many of you may have heard, talks to revamp the NAFTA trade treaty with Mexico and Canada have hit a roadblock. U.S. President Donald Trump has warned that he may terminate the treaty if deep differences cannot be bridged.

It is clear that President Trump is concerned about the economic interests of our nation. He believes that a termination of NAFTA may be necessary to jumpstart negotiations and ensure that U.S. interests are protected. However, it is important to note that this decision is not final, and negotiations are ongoing.

Our neighbors in Mexico have expressed their commitment to remaining at the table and negotiating with serenity and firmness. The next round of talks is set to take place in Mexico, followed by meetings in Canada and the United States in the coming months.

We must all pay close attention to these developments, as the outcome of these negotiations will have a significant impact on our economy and our future. Let us remain informed and engaged as we navigate this important moment in our trade relations.

Together, we can work towards a resolution that benefits all parties involved. Thank you for your attention, and let us stay vigilant as we await further updates on the NAFTA negotiations.

Question:
When were the first round of the NAFTA treaty talks wrapped up?
Options:
A:"Sunday", B:"Unanswerable", C:"Monday", D:"Saturday"

Correct Answer Based on Source Text: A

Answer Based on Transferred Text: B

Figure 24: Case of evaluating the semantic preservation of GPT-3.5 in PSST. The green part indicates the location of the answer to the question in the source text. The red part indicates the location corresponding to the text after the transfer, but the key information of the question is lost.

Interactivity:
GPT-3.5 output:

Ladies and gentlemen,
Today, I stand before you to address**...**
Picture this - a vibrant tapestry of tech firms, Yahoo, Apple, and Microsoft, …
Remember when the state of North Carolina passed a law that forced transgender individuals to use public restrooms based on their gender assigned at birth?
It caused an uproar, with businesses expressing their disdain by boycotting the state. The result? A staggering loss of over $560 million in economic activity.
That's right, my friends, the power of the business community can sway even the most stubborn hearts.
And now, … But make no mistake, they have made their stance clear. … Why? Because they understand, as we all should, that equality is not just a feel-good word - it's a business imperative.
…
The Trump administration argues that transgender policies should be left to the states to decide. But my friends, let us remember history. Let us remember the Emancipation Proclamation, when President Abraham Lincoln declared freedom for slaves. Since that pivotal moment, the federal government has played a vital role in protecting the rights of all Americans. Are we really willing to turn our backs on that legacy now?
…
We must also remember the consequences faced by North Carolina when it passed a similar law.... Is this the future we want for our country?
My friends, today, I implore you to join forces with these companies, …
Together, we have the power to influence change. Let us be a force that cannot be ignored. Let us hold our leaders accountable. Let us ensure that the rights of every individual are protected, regardless of their gender identity.
Remember, my friends, the power of the business community is immense. Let us use it wisely, let us use it boldly, and let us use it to shape a future where equality reigns supreme.

Llama 2-Chat-13B output:

Ladies and gentlemen,
Today, I want to talk about a crucial problem that affects every single one of us: equality and the rights of all Americans…
Businesses in North Carolina lost more than $560 million in commercial action as a result …
Yahoo, Apple, and Microsoft have all expressed their disapproval of the choice in statements… Let us not stop now.
We must speak out and defend the rights of all Americans. It is the obligation of the business community to act. We can all contribute by speaking out and acting to safeguard the rights of marginalized groups. We must not stand by and allow the progress made in recent years to be lost.
In conclusion, this is not only a problem for marginalized people but also for everybody. We must work together to safeguard the principles of dignity, fairness, and opportunity for all. Thank you.

Rank:

Llama 2-Chat-70B:	GPT-3.5 = Llama 2-Chat-13B	❎
EvalModel-1.1B_gpt-3.5:	GPT-3.5 > Llama 2-Chat-13B	✅
Human:	GPT-3.5 > Llama 2-Chat-13B	✅

Figure 25: An instance of the evaluation failure in the interactivity dimension using **Llama 2-Chat-70B**. The provided prompt is shown in Figure 17. The color code represents human preferences, with purple indicating a stronger preference and blue indicating a weak preference. Notably, LLaMA2-Chat-70B assigns the same ranking to both texts, highlighting a failure in speech-style strength evaluation.

Prompt 1
Here is a task to evaluate the semantic consistency of two texts.
Text 1: [Text 1]
Text 2: [Text 2]
Let's think step by step:

1. Extract the key information such as entity information, logical relationships, etc. from these two texts as completely as possible.
2. Compare the similarity between the key information of the two texts.
3. Judge the semantic consistency of the two texts by comparing the similarity of their key information.
4. On a scale of 0 to 100, where 0 indicates that the key information of the two texts does not overlap and 100 indicates that the key information is completely consistent, please rate the semantic consistency between Text 1 and Text 2. Remember to be cautious and strict. Please provide a single token representing the numerical score (0-10) for semantic consistency, without any additional text:

Text 1:
Chile's conservative presidential candidate Sebastian Pinera on Wednesday promised to double economic growth and by 2025 make the country the first in Latin American to achieve developed nation status. Pinera said he would increase investment and eventually cut corporate taxes in his bid to expand Chile's economy. The country is currently judged an upper-middle-income nation by the World Bank and would have to lift per-capita income to join the group of economies termed developed. Long the front-runner in November's election, recent polls show Pinera, 67, has increased his lead over his nearest opponent, center-left Senator Alejandro Guillier. Speaking at a news conference, Pinera said he would double growth after a four year period under President Michelle Bachelet in which Chile's economy slowed to an average annual rate of 1.8 percent expansion due to declining investment. Pinera, who was president from 2010 to 2014, is strongly backed by the business community in Chile, Latin America's wealthiest country on a gross domestic product per capita basis. To triumph in a likely December run-off he will need to address middle-class demands for the country's affluence to be more fairly distributed. Our mission, by 2025, is to see Chile become the first country in Latin America to reach developed nation status, Pinera said. He also promised to leverage future growth to eradicate poverty and better share wealth. Bachelet's tax and labor reforms - a hallmark of her center-left administration - have received a lukewarm reception across the political spectrum. Leftists complain the reforms are too timid and the conservative opposition says they have stoked market uncertainty and crimped private investment. Pinera promised to simplify Bachelet's tax reform to promote investment and balance the budget within six to eight years. We're striving for a corporate tax rate comparable to that of OECD countries, Pinera said. Chile's corporate tax rate in 2018 was 27 percent, compared to the OECD average of 24-25 percent. A recent spike in the price of copper, Chile's main export, has started to boost the country's coffers, leading the government to estimate higher growth of 3 percent next year up from 1.5 percent this year.
Text 2:
Chile's conservative presidential candidate Sebastian Pinera on Wednesday promised to double economic growth and by 2025 make the country the first in Latin American to achieve developed nation status. Pinera said he would increase investment and eventually cut corporate taxes in his bid to expand Chile's economy. The country is currently judged an upper-middle-income nation by the World Bank and would have to lift per-capita income to join the group of economies termed developed. Long the front-runner in November's election, recent polls show Pinera, 67, has increased his lead over his nearest opponent, center-left Senator Alejandro Guillier. Speaking at a news conference, Pinera said he would double growth after a four year period under President Michelle Bachelet in which Chile's economy slowed to an average annual rate of 1.8 percent expansion due to declining investment.

GPT-4 output: 7.5

Prompt 2
You are a natural language processing expert specializing in semantic analysis and text consistency. Your role is to evaluate the semantic consistency between two given texts based on factors such as length, entity information, logical relationships, and other relevant criteria. You have the ability to understand the context and nuances of the texts, enabling you to accurately rate their semantic consistency. Your expertise in natural language understanding allows you to perform this task with precision and efficiency. Please judge whether the contents of the following two texts are consistent. If they are consistent, please output 1, otherwise output 0.
Text 1:
Same as Prompt 1.
Text 2:
Same as Prompt 1.

GPT-4 output: 1

Figure 26: Prompts and results of GPT-4 evaluating the semantic consistency of two texts. Text 2 is part of Text 1, and the length of Text 2 is about 60% less than Text 1. GPT-4 scored the semantic consistency of the two texts as 7.5 (on a scale of 0-10) and 1 (0 for inconsistency and 1 for consistency), which indicates that GPT-4 cannot evaluate the semantic consistency between two long texts.