TSST: A Benchmark for Evaluation-driven Text Speech-style Transfer

Anonymous ACL submission

Abstract

Text style is highly abstract, as it encompasses various aspects of a speaker's characteristics, habits, logical thinking, and the content they express. However, previous text style transfer tasks primarily focused on data-driven approaches, lacking in-depth analysis and research from the perspectives of psycholinguistic and cognitive science, which results in rapid saturation and the inability to distinguish models. In this paper, we introduce a novel task called Text Speech-Style Transfer (TSST). The main objective is to further explore topics related to the alignment of language processing between human and large language models (LLMs). Considering the objective of our task and the distinctive characteristics of oral speech in real-life scenarios, we train fine-grained ensemble multi-dimension (i.e. *emotionality, vividness, interactivity, filler words*) evaluation models for the TSST task and validate their correlation with human assessments. Experiments demonstrate its superior effectiveness compared to LLM-based methods. We thoroughly analyze the performance of several LLMs and identify areas where further improvement is needed. In summary, we present the TSST task, a new benchmark for style transfer and emphasizing human-oriented evaluation, exploring and advancing the performance of current LLMs.

1 Introduction

Natural Language Processing (NLP) aims to understand and process human language, emulate human cognition, and enable automatic interaction between humans and computers (Manaris, 1998). Recently, the emergence of Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023) has significantly propelled the advancement of automated text-related technologies, such as text summarization and human-computer dialogues. These data-driven LLMs have demonstrated outstanding performance in traditional NLP tasks, thanks to their ability to leverage massive amounts of data.

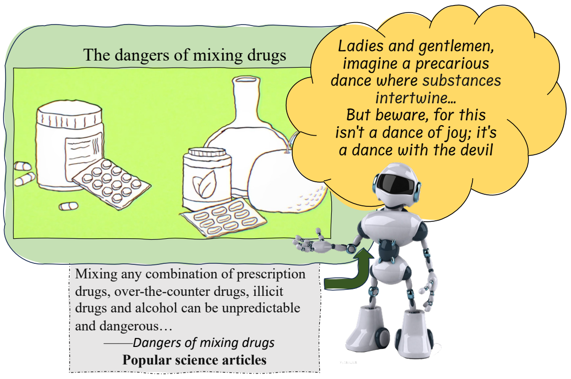

The dangers of mixing drugs
Ladies and gentlemen, imagine a precarious dance where substances intertwine...
But beware, for this isn't a dance of joy; it's a dance with the devil
Mixing any combination of prescription drugs, over-the-counter drugs, illicit drugs and alcohol can be unpredictable and dangerous…
——Dangers of mixing drugs
Popular science articles

Figure 1: Illustration of text speech-style transfer task (TSST). A robot or language model should be able to present an ordinary article, such as a popular science paper, to audiences in a captivating and engaging manner.

Researchers are extensively investigating the alignment between LLMs and human cognition, as well as the extent to which LLMs emulate human language processing (Akata et al., 2023; Aher et al., 2023; Park et al., 2023). Text Style Transfer (TST) is a task that can reflect this ability of LLMs to some extent, which aims to transform text-based language into another style of language. The style of language is intricately related to human cognition, reflecting various aspects of the speaker's characteristics, habits, logical thinking, and specific communication contexts, which exhibits a higher level of abstractness (Jin et al., 2022). As a result, "style" transcends mere text and approaches the realm of computational psycholinguistic modelling (Ratner and Gleason, 2004; Crocker and Brouwer, 2023).

However, challenges have arisen in previous TST tasks (Rao and Tetreault, 2018; Shen et al., 2017a; Xu et al., 2012; Dathathri et al., 2020), preventing the attainment of the aforementioned objectives.

Firstly, Previous definitions of “style” primarily rely on datasets (Rao and Tetreault, 2018; Gan et al., 2017; Mou and Vechtomova, 2020), making them susceptible to other text attributes, such as semantic content, which results in ambiguity and uncertainty in the evaluation stage. The second issue is the scarcity of parallel corpora for many “styles”. Even when parallel corpora are available, constructing the dataset requires substantial manual resources and incurs high costs. Lastly, in tasks like formal-informal (Rao and Tetreault, 2018), the style features are closely tied to the text and have a relatively low level of abstractness (Shown in Table 2). This may result in rapid task saturation, rendering it difficult to discern distinctions between models (Kiela et al., 2021).

In this paper, we introduce a novel and more abstract task called Text Speech-Style Transfer (**TSST**), which we hope can serve as a starting point to drive LLMs towards a more advanced emulation of human cognitive processes. The goal of TSST is to enable models to leverage a given passage to emulate human oral presentations, such as TED talks[1], producing engaging and captivating spoken content (shown in Figure 1). As shonw in Figure 2, to precisely define speech-style, we perform a thorough linguistic and cognitive analysis, leveraging real-world speech data from Ted Talks. We propose a multidimensional definition of speech-style, which encompasses *emotionality*, *vividness*, *interactivity*, and *filler words*. This approach involves decomposing the higher-level abstract style into more specific “substyles,” reducing the ambiguity associated with the notion of “style”. To enhance the effectiveness of our evaluation system, we employ meticulously designed prompts for LLMs, enabling the generation of sentence-level parallel high-quality spoken content with reduced reliance on extensive human resources. Subsequently, we train fine-grained ensemble models (4×GPT-Neo-1.3B (Gao et al., 2020)) to evaluate the speech-style strength for each dimension, presenting a simple but more effective alternative to LLM-based methods (Lai et al., 2023) (Llama 2-Chat-70B (Touvron et al., 2023)). Finally, based on TSST, we conduct a thorough quantitative and qualitative evaluation analysis of mainstream LLMs (including ChatGPT (AI, 2022) and Llama 2-Chat (Touvron et al., 2023)), summarizing the primary challenges. TSST is flexible and dynamic. By adding extra conditions or “substyles,” we can better evaluate how well the model aligns with human judgment, enhancing its ability to differentiate between different models.

one can experience certain speech-style text in audio version here.

Our main contributions are as follows:

- We introduce a novel dynamic task named TSST to enhance the alignment of LLMs with human processing in psycholinguistic modeling, which allows for the introduction of additional constraints to continually assess the capabilities of stronger models.
- We propose fine-grained ensemble evaluation models (4×GPT-Neo-1.3B) for the TSST task. Results show it is more effective compared to LLM-based methods (Llama 2-Chat-70B).
- We assess the performance of LLMs in TSST and summarize the encountered challenges based on qualitative and quantitative analyses using the proposed evaluation models.

2 Related Work

Text style transfer (TST) aims to automatically convert a text from its source style to a target style while preserving the content (Jin et al., 2022; Hu et al., 2022). In this regard, we first provide a clear definition of style, then we delve into the dimensions of TST evaluation, which include fluency, content preservation, and transfer strength (Ostheimer et al., 2023).

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 to linguistic studies, research on text style transfer (TST) takes a data-driven approach, defining style as attributes or labels based on style-specific corpora. These corpora often consist of attributes that can be effectively modeled using neural machine learning techniques, such as sentiment transfer and formality transfer (Shen et al., 2017b; Rao and Tetreault, 2018). However, we propose an aspect that serves as a bridge between the two aforementioned definitions. Our approach is grounded in real data, specifically TED talks, and examines

---

[1] https://www.ted.com/talks

language dimensions from a speech linguistics perspective. The definition of speech-style extends beyond data-driven informality and encompasses abstract, linguistic, and cognitive concepts such as emotionality, vividness, and interactivity.

Automatic Evaluation of TST Several metrics have been proposed to assess the effectiveness of style transfer based on three criteria. 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 score (Papineni et al., 2002), ROUGE score (Lin and Och, 2004), BERTScore (Zhang et al., 2019) are employed. Furthermore, 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. However, there are several shortcomings with the above metrics when evaluating passage-level and speech-style text. For example, the PPL tends to favor shorter texts (Jin et al., 2022), and a single classifier may not accurately differentiate between formal text and speech-style. To address these limitations, we propose a fine-grained ensemble of classifiers as a metric to evaluate the transfer strength of the speech-style text in TSST.

Human Evaluation of TST Human-based evaluation serves as a gold standard for assessing the performance of TST models. Evaluators are asked to assess the style-transferred sentences based on the three criteria discussed in the earlier paragraphs (Xu et al., 2012; Niu et al., 2017). Recently, Lai et al. (2023) demonstrates that ChatGPT (AI, 2022) achieves competitive correlation with human judgments to serve as a multidimensional evaluator for formal-informal style transfer. However, based on our preliminary experiments, we have observed that LLMs (eg. ChatGPT (AI, 2022) or Llama 2-Chat-70B (Touvron et al., 2023)) is not effective in evaluating TSST and lacks credibility.

3 Text Speech-Style Transfer

In this section, we introduce the Text Speech-Style Transfer (TSST) task, which serves as a platform for studying speech-style in real-world scenarios. To begin, we offer a clear definition of the TSST and introduce the source dataset we compile for TSST. We then offer a concise overview of the challenges that differentiate this task from its original counterpart. Furthermore, using real-world speech-style data (TED talks), we break down the abstract speech-style, highlighting its essential subfeatures in Section 3.4. This helps us to identify and summarize the essential evaluation dimensions of speech-style, contributing to the advancement of LLMs capturing and transferring speech-style.

3.1 Task Formulation

The TSST task involves transforming official text style a (e.g., news articles or academic abstracts) into a more conversational and speech-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 speech-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 habits, the audience's specific preferences, and so forth.

Style	Examples
official	The company has **implemented** a series of **cost-cutting** measures to improve its profitability.
speech	The company has **done a lot of things** to **save money** and make more profit.

Table 1: Examples of official style and speech style.

As shown in Table 1, official-style texts, like news reports and business letters, typically employ professional and intricate vocabulary and sentence structures, conveying a serious and objective tone. On the contrary, speech-style texts are more suitable for oral expression, utilizing simpler vocabulary and sentence patterns, and adopting a more direct tone. Additionally, the speech style possesses audience-oriented characteristics, which we will elaborate on in Section3.4.

3.2 Source Data

we select three categories to facilitate the training of evaluation models and establish a test set for TSST: official news, paper abstracts, and Wikipedia articles See Appendix A.1 for a detailed description.

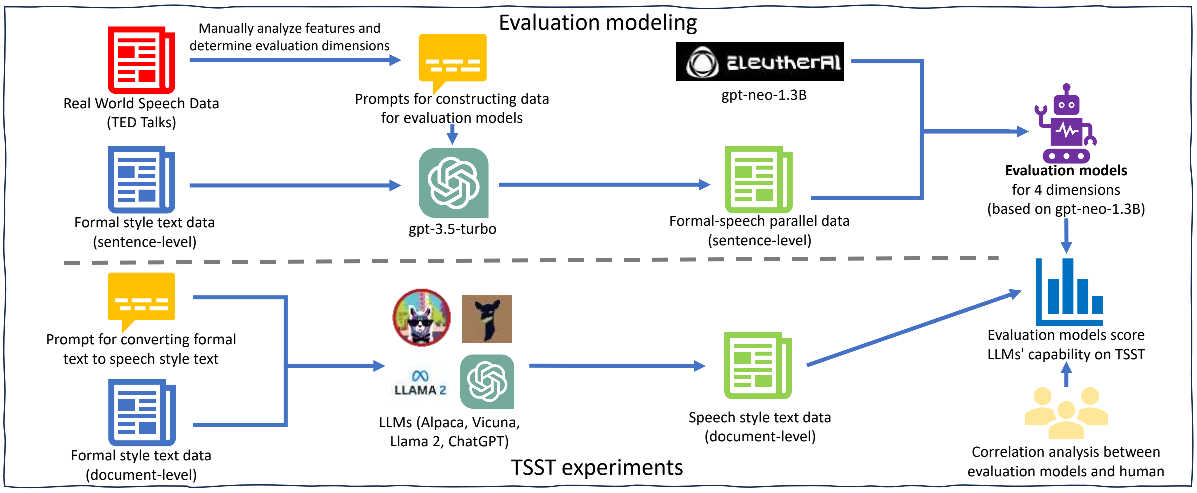

Evaluation modeling
Manually analyze features and determine evaluation dimensions
Real World Speech Data (TED Talks)
Prompts for constructing data for evaluation models
EleutherAI
gpt-neo-1.3B
Formal style text data (sentence-level)
gpt-3.5-turbo
Formal-speech parallel data (sentence-level)
Evaluation models for 4 dimensions (based on gpt-neo-1.3B)
Prompt for converting formal text to speech style text
LLAMA 2
LLMs (Alpaca, Vicuna, Llama 2, ChatGPT)
Speech style text data (document-level)
Evaluation models score LLMs' capability on TSST
Formal style text data (document-level)
TSST experiments
Correlation analysis between evaluation models and human

Figure 2: Pipeline of establishing the evaluation system of the TSST task and the experiment&analysis of LLMs. **1.** The above depicts the process of establishing the evaluation system. Specifically, we begin by conducting a comprehensive analysis of real-world TED data and identify four key characteristics of speech-style. Subsequently, we employ GPT-3.5 to generate a sentence-level list-wise parallel corpus and train GPT-Neos as an evaluator for each dimension. **2.** The bottom presents the experiment and analysis of the TSST task for the current LLMs.

We filter out data instances with token counts outside the range of 400 to 900 tokens. This ensures that our dataset remains manageable and aligned with the context length limitations, enabling us to develop more effective models for the task at hand. As shown in Table 5, after the filter-out operation, 9247, 189298, and 550 instances remained in the three categories, respectively.

3.3 Challenges of TSST

There are three primary challenges in the task, and examples are shown in Table 2:

Non-binary Styles, such as Impolite-Polite (Rao and Tetreault, 2018) and Positive-Negative (Shen et al., 2017a), exhibit clear duality that is easily discernible. However, TSST introduces a level of ambiguity as the boundaries between official and speech-style language can become blurred. As long as the text aligns with spoken expressions or can be readily employed in oral descriptions, it qualifies as speech-style text, even within formal contexts.

Psycholinguistic Modelling Traditional style transfer tasks lack consideration for cognitive science and psychology, hindering their ability to discern novel approaches in text processing by models. However, TSST distinguishes itself by delving into capturing dialogical interactions with the audience, reflecting emotions, adapting language and tone to elicit specific reactions, and effectively conveying perspectives. These aspects encompass the intricate and abstract nature of oral communication.

Long Text Previous text style transfer tasks primarily focused on the sentence level (Rao and Tetreault, 2018; Gan et al., 2017; Shen et al., 2017a), whereas this task aims to conduct style transfer at the document level. We find that LLMs tend to lose portions of original information when performing TSST, indicating persistent challenges in processing long contexts (Liu et al., 2023).

3.4 Prior Fine-grained Analysis

To identify the prominent features of the speech-style text, we employ a rigorous human evaluation process that involves multiple candidates evaluating dimensions. Then we conclude four crucial dimensions that serve as the foundation for training evaluation models in the next section.

3.4.1 Human Evaluation on Speech-Style Data

To establish a comprehensive framework for evaluating speech-style text, we begin by heuristically selecting six features to assess the text at both the sentence and word levels: (1) *Ambiguity*, (2) *interactivity*, (3) *Emotionality*, (4) *Filler Words*, (5) *Abbreviations*, (6) *Informal Lexicon*.

We randomly sample 300 speech-style sentences from TED dataset[2] (Cettolo et al., 2017) and annotate them using two different evaluating metrics by 3 people. In the multi-label approach, we select labels that signify the speech-style features present

---

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

TST task	style	examples	features
official ↔ speech	official	GDOS is a modified version of WEDOS, which facilitates ...	arousing interest
	speech	**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
	speech	**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
	speech	Instead, they have grown **stronger and more courageous.**	
	official	The Conference also agreed that the Bureau would keep the calendar under review	filler words
	speech	**So, uh**, the Conference also **agre-uh,** agreed that the Bureau would keep the calendar under 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 2: Comparison of TSST and formality TST: **a.** TSST introduces certain ambiguous areas. **b.** Formality TST primarily focuses on vocabulary-level adjustments, such as adhering to writing norms and word norms. In contrast, TSST involves audience engagement, including posing questions to generate interest and appeal to the audience.

in a specific instance. In contrast, the best-one approach selects the most salient label that embodies the speech-style of the sentence. The result is demonstrated in Figure 3.

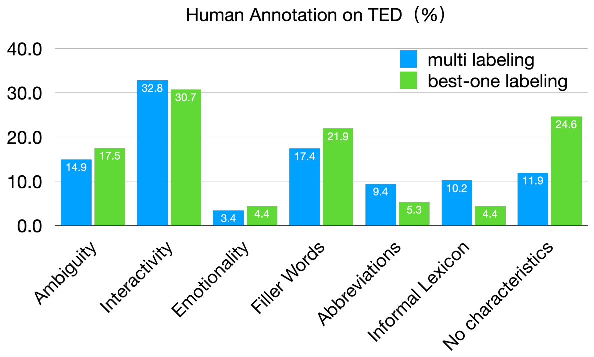


Figure 3: Manually annotating oral attributes on the TED dataset. The proportion of "interactivity," "filler words," and "ambiguity" is notable.

The statistics in Figure 3 indicate that interactivity, filler words, and ambiguity are the primary characteristics. For abbreviations and informal lexicons, the multi-labeling proportion is approximately twice that of the best one labeling, suggesting their widespread usage while their ability to highlight oral characteristics is weak. However, ambiguity predominantly encompasses the simplification and expansion of complex sentences, and its transformation is also subject to certain conditions. Thus, we focus on the two significant characteristics of *Interactivity* and *Filler Words*.

Our objective is to create a more engaging and impactful experience for the listeners. Therefore, we emphasize the inclusion of vividness and emotionality. The observed low emotionality in Figure 3 may since the annotation is conducted at the sentence level, whereas emotions are typically associated with specific events, encompassing multiple sentences or paragraphs, and thus occur less frequently.

3.4.2 Evaluation Dimensions

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

- **Interactivity:** Interactivity in speech refers to the speaker engaging with the audience through various means such as asking questions, making appeals, and emphasizing key points. (Table 2 example 1)
- **Emotionality:** It is desirable to generate appropriate perspectives and attitudes towards specific events in the original text to reflect the speaker's thought. (Table 2 example 2)
- **Vividness:** Official text typically adopts a neutral and objective tone. However, in oral speech, it is essential for the speaker to present information in a lively and easily comprehensible manner. (Table 2 example 3)
- **Filler Words:** Filler Words feature is strongly related to semantic content and speaker habits and represents a prominent characteristic of oral speech. (Table 2 example 4)

4 Evaluation System

In this section, we present the evaluation system for the TSST task. Given the high fluency of text generated by large language models, our primarily focus on style strength and content preservation.

We introduce a more style strength evaluation in Section 4.1. Specifically, we present methods for LLM-based high-quality data generation and model training in Section 4.1.1 and Section 4.1.2, respectively. In Section 4.2, we examine the correlation between the evaluation models and human preference, highlighting its advantage compared to LLM-based approaches. In Section 4.3, we introduce the evaluation method of content preservation.

4.1 Style Evaluation Models

Our early experiments suggest that using LLMs directly as evaluators may not be feasible. (analyzed in Section 4.2). However, it is noteworthy that LLMs exhibit commendable sentence-level style transfer capabilities. Therefore, we achieve transfer strength evaluation by training small-scale fine-grained ensemble evaluation models using stylistic sentence-level parallel data generated from LLMs.

4.1.1 Data Generation for Evaluation Models

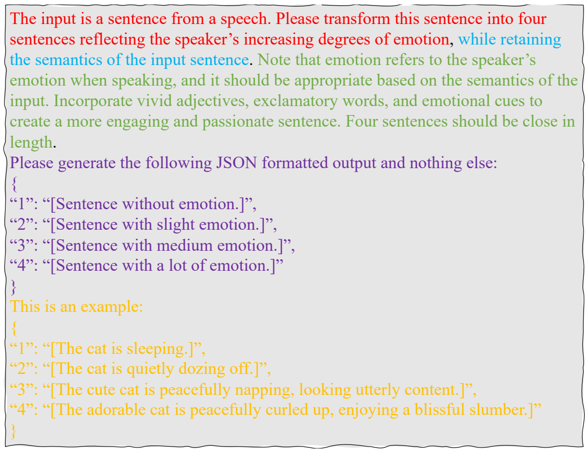
LLMs are utilized to construct list-wise data data for training evaluation models for each dimension. For each official text input, we generate four sentences with consistent semantics, orderly escalating the degree of speech-style strength.

Specifically, we design a prompt for each dimension. Figure 4 shows the prompt for emotionality. Prompts for other dimensions are shown in Appendix C.2. When designing the prompt, three issues are mainly considered: First, the data style should meet our requirements and have the correct degree order. Second, in order to eliminate the influence of semantics and facilitate the model to learn stylistic features, the semantic before and after the transformation should be consistent. Third, the format of output should be standardized.

We sample 4000 sentences from the formal-style texts. Combining the three designed prompts, we use LLM (gpt-3.5-turbo[3]) to generate 4000 sets of sentence lists for each dimension.

4.1.2 Training Evaluation Models

We fine-tune the decoder-only model GPT-Neo-1.3B (Gao et al., 2020) as the evaluation model for each dimension. LoRA (Hu et al., 2021) technique was employed for streamlined training and deployment. We used a list-wise ranking loss function shown below:



The input is a sentence from a speech. Please transform this sentence into four sentences reflecting the speaker’s increasing degrees of emotion, while retaining the semantics of the input sentence. Note that emotion refers to the speaker’s emotion when speaking, and it should be appropriate based on the semantics of the input. Incorporate vivid adjectives, exclamatory words, and emotional cues to create a more engaging and passionate sentence. Four sentences should be close in length.
Please generate the following JSON formatted output and nothing else:
{
“1”: “[Sentence without emotion.]”,
“2”: “[Sentence with slight emotion.]”,
“3”: “[Sentence with medium emotion.]”,
“4”: “[Sentence with a lot of emotion.]”
}
This is an example:
{
“1”: “[The cat is sleeping.]”,
“2”: “[The cat is quietly dozing off.]”,
“3”: “[The cute cat is peacefully napping, looking utterly content.]”,
“4”: “[The adorable cat is peacefully curled up, enjoying a blissful slumber.]”
}

Figure 4: Prompt for constructing data for emotionality evaluation model. The red part is the task description. The blue part asks semantic consistency. The green part indicates further requirements and provides feasible suggestions. The purple part specifies the format of the output. The orange part provides an example.

$$\mathbb{L}_{ranking} = -\sum_{C_j \in D_i} \sum_{x_j, x_k \in C_j} \log(\sigma(r_\theta(x_j) - r_\theta(x_k))) \quad (2)$$

where D_i means the training dataste of dimension i, C_j is a candidates set which contains four sentence with different style strength but same meaning, and x_j is better than x_k that exhibiting a specific target style.

Please refer to Table 8 for more training details.

We train two series of models using data generated by GPT-3.5 and LLaMA 2-Chat-13B, respectively, **EvalModel-1.3B$_{\textbf{gpt-3.5}}$** and **EvalModel-1.3B$_{\textbf{llama2-13B}}$**.

4.2 Correlation

Dataset For each evaluation dimension, we collected a dataset containing 50 samples, each consist of four texts: an original formal text, a speech-style text generated by GPT-3.5, and two speech-style texts generated by Llama 2-Chat-13B (using both concise and enhanced prompts mentioned in Section 5.1, which can help model to generate text with different speech-style strength). We annotate the testset manually to establish ground truth.

---

[3] https://platform.openai.com/docs/api-reference/chat/object

Evaluation Model	Emotionality	Vividness	Interactivity	Average
EvalModel-1.3B$_{\text{gpt-3.5}}$	90.00	84.14	77.78	83.97
EvalModel-1.3B$_{\text{llama2-13B}}$	85.83	82.76	76.67	81.75
Llama 2-Chat-70B	62.37	44.22	64.33	56.97

Table 3: **Pearson's** ρ between different evaluation models and human evaluation in each dimension.**EvalModel-1.3B$_{\text{gpt-3.5}}$** and **EvalModel-1.3B$_{\text{llama2-13B}}$** denote style-strength evaluation models trained utilizing datasets derived from GPT-3.5 and Llama 2-Chat-13B, respectively. For each dimension, we trained an evaluation model, resulting in three distinct evaluation models represented as EM.

Evaluation Models EvalModel-1.3B$_{\text{gpt-3.5}}$, EvalModel-1.3B$_{\text{llama2-13B}}$ and Llama 2-Chat-70B [4] are used to score and rank candidate texts. The prompt used by Llama 2-Chat-70B is shown in Figure 6, which we refer to the method in (Lai et al., 2023).

Results and Analysis We calculate the Spearman correlation coefficient between different evaluation models and human evaluation in each dimension. The correlation results are shown in Table 3.

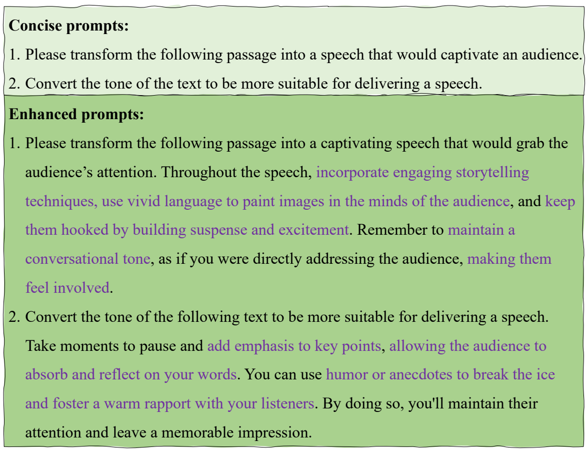
We observe a high average correlation exceeding 80% between the evaluation model, trained using the proposed method in this paper, and human evaluation. In contrast, the correlation score between Llama 2-Chat-70B and humans is merely 56.97. This contrast demonstrates the effectiveness and validity of the sentence-level fine-grained speech-style evaluation method introduced in this paper.

In contrast to the findings of (Lai et al., 2023), who identified the substantial potential of ChatGPT in evaluating TST in sentence level, our experimental results indicate that LLMs, like Llama 2-Chat-70B, encounter challenges when assessing long text with abstract style. For illustrative instances, please refer to Appendix D.2.

4.3 Content Preservation

After empirical exploration, we find it is not feasible to use LLMs to directly evaluate semantic consistency, including methods such as prompt and COT. Details are provided in Appendix D.3.

Considering the effectiveness of current LLMs in summarization, we adapt the following procedure: First, we employ Llama 2-Chat-70B (Touvron et al., 2023) to generate summaries for input and output. Then, we measure the semantic consistency between the summarizations by BLEURT and BertScore.



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.

Enhanced prompts:

1. Please transform the following passage into a captivating speech that would grab the audience's attention. Throughout the speech, incorporate engaging storytelling techniques, use vivid language to paint images in the minds of the audience, and keep them hooked by building suspense and excitement. Remember to maintain a conversational tone, as if you were directly addressing the audience, making them feel involved.
2. Convert the tone of the following text to be more suitable for delivering a speech. Take moments to pause and add emphasis to key points, allowing the audience to absorb and reflect on your words. You can use humor or anecdotes to break the ice and foster a warm rapport with your listeners. By doing so, you'll maintain their attention and leave a memorable impression.

Figure 5: Examples of concise and enhanced instructions, respectively. All instructions are presented in Appendix C.3

5 Speech-style Data Generation and Analysis

We performed a comprehensive quantitative and qualitative analysis of LLM performance in TSST tasks. In Appendix A.2, we investigate the potential overlap between different evaluation dimensions and demonstrate that evaluation models trained with data generated by different LLMs exhibit minimal bias against test models.

5.1 Setups

Alpaca-13B (Taori et al., 2023), Vicuna-13B (Chiang et al., 2023), Llama 2-Chat-13B (Touvron et al., 2023) and GPT-3.5 were employed to generate speech-style outcomes using the following setups.

Specifically, we sample 120 official-style texts from the dataset mentioned in Section 3.2 as input (40 texts for each category of News, Abstract, and Wikipedia). We use concise and enhanced instructions shown in Figure 5. The enhanced instructions emphasize that the output should be vivid and interactive to facilitate the model to improve in these two dimensions.

---

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

Instruction	Model	Stylistic Assessment score				Content Preservation score	
		FillterWord	Emotionality	Vividness	Interactivity	BERTScore	BLEU_RT(Var)
Concise	Alpaca	4.1361	7.2711	6.6480	2.6906	0.2058	0.4775 (0.0111)
	Vicuna	4.0648	5.5771	5.0936	0.7434	0.2999	0.5294 (0.0070)
	Llama 2	4.4657	8.1972	7.5537	1.9828	0.3400	0.5207 (0.0060)
	GPT-3.5	3.5568	7.9880	7.6161	0.8699	0.3893	0.5472 (0.0045)
Enhanced	Alpaca	4.2621	7.9248	7.2252	3.6030	0.1187	0.4460 (0.0062)
	Vicuna	4.9094	7.1825	6.6116	3.0335	0.2732	0.5049 (0.0053)
	Llama 2	4.7725	9.0836	8.3987	4.1680	0.2233	0.4791 (0.0052)
	GPT-3.5	4.6692	10.2208	9.8527	3.8906	0.2689	0.5097 (0.0047)

Table 4: Evaluation Results by our multi-dimension style strength evaluation models. The enhanced prompts improve style strength across all dimensions while resulting in a decrease in content preservation.

5.2 Results and Analysis

5.2.1 Different Instructions for Generating Results

As shown in Appendix D.1 and Table 4, we discover some common phenomena: when using concise instructions, the model's output has weak transfer strength but performs well in content retention. The output shows significant improvements in four stylistic dimensions when using enhanced instructions while it often omits more information. So, we believe that designing appropriate instructions is an effective way to improve the model's performance on specific dimensions in the TSST task.

5.2.2 Problems with LLMs

We analyze and summarize the characteristics and shortcomings of LLMs in performing the TSST task. More examples are shown in Appendix D.1.

- **Instruction Sensitivity** Given the same input and different expressions of instructions, Alpaca-13B exhibits a high sensitivity to those instructions. It is capable of understanding some instructions (Case 1, 2 in Figure 9), However, there are instances where it struggles to comprehend instructions accurately. In these cases, the generated text may end up being a paraphrase of the original text without any stylistic changes (Case 3 in Figure 9).

- **Uncontrolled Length Generation** As shown in Case 1 and Case 2, regardless of whether concise or enhanced instructions are used, Alpaca-13B generates very short text, which results in significant information loss. Conversely, the output of ChatGPT is always too long (Case 8, 9 in Figures 13 and 14), containing redundant and hallucinated content.

- **Weak Transfer Strength** When using the concise instruction set, the transfer strength of Llama-13B and Vicuna-13B is relatively weak (Case 4, 6 in Figures 10 and 11). Although the transfer strength of the output of ChatGPT is higher (Case 8, 9 in Figures 13 and 14), the transfer strength at the beginning and end is obviously higher than that in the main content part in the middle, which is lower or even absent. When using enhanced instructions, the transfer strength of LLMs can be improved.

- **Inconsistent Content** As mentioned former, when Alpaca-13B can understand instructions (Case 1, 2 in Figure 9), the output only has the same topic as the original text, but almost no information from the original text is retained. Vicuna-13B and Llama-13B exclude more information when using enhanced instructions (Case 5, 7 in Figure 10 and 12), and sometimes there are illusions that do not exist in the original text.

6 Conclusion

This paper introduces TSST for investigating the potential of LLMs to simulate human cognitive processes by means of text speech-style transformation. We extract four essential features of speech style through an analysis of real-world speech data, which serve as the foundation for the fine-grained ensemble evaluation models' training. Furthermore, we effectively harness the capabilities of LLMs to generate parallel data representing the speech style, in accordance with the four proposed evaluation dimensions. Our evaluation system demonstrates superior consistency with human assessment outcomes compared to LLM-based method (4×GPT-Neo-1.3B VS Llama 2-Chat-70B). The evaluation results reveal that current LLMs still exhibit certain limitations in the TSST task, which underscores the effectiveness of TSST.

7 Limitations

We outline limitations and potential enhancements of this paper as follows:

7.1 Cross-field Technology Enhancement

The Spoken Dialogue model holds significant research implications for human-computer interaction (Thórisson, 2002; Adiwardana et al., 2020; Roller et al., 2020). However, solely learning human speech patterns from the auditory modality is insufficient, as it lacks deeper semantic understanding (Nguyen et al., 2023). To address this, we intend to approach the characteristics of spoken language from a textual perspective. This approach will provide valuable insights that can enhance the performance of text-audio-based spoken dialogue generation models, bridging the gap between text and speech by incorporating richer semantic information.

7.2 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" dimension. However, such descriptions may not align with our expectations.

Granularity Refinement As shown in Tabel 6, our emotionality evaluation lacks granularity. In future work, we aim to enhance this aspect by initially identifying the event content in the source text and its potential emotional tendencies. Subsequently, we will meticulously evaluate whether the model generates an appropriate response to specific events based on this analysis.

Effective Evaluation of Content Preservation This paper primarily focuses on assessing style strength, with limited discussion on semantic consistency. Future work will delve into evaluating the preservation of crucial information, examining the logical relationships between events and other specifics. The objective is to determine if the model overlooks essential content and exhibits confusion in logical relationships.

7.3 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.

8 Ethical Considerations

Here we discuss the primary ethical considerations of TSST

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

Content and Impact The TSST 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. However, 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

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Open AI. 2022. Introducing chatgpt.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with large language models. *CoRR*, abs/2305.16867.

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.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

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

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 955–964.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Talat, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in nlp. *ArXiv*, abs/2104.14337.

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.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *ArXiv*, abs/2307.03172.

Bill Z. Manaris. 1998. Natural language processing: A human-computer interaction perspective. *Adv. Comput.*, 47:1–66.

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*.

Lili Mou and Olga Vechtomova. 2020. Stylized text generation: Approaches and applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–22, Online. Association for Computational Linguistics.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Phil Ostheimer, Mayank Kumar Nagda, Marius Kloft, and Sophie Fellenz. 2023. A call for standardization and validation of text style transfer evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10791–10815, Toronto, Canada. Association for Computational Linguistics.

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.

Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. 2023. Diminished diversity-of-thought in a standard large language model.

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Kristinn R. Thórisson. 2002. *Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action*, pages 173–207. Springer Netherlands, Dordrecht.

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.

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

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.

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 Dataset and Experiments

A.1 Description and of Source Dataset

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[5]. 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[6]. 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[7], which is a collection of over 100 million tokens extracted from the set of verified Good and Featured articles on Wikipedia.

Category	**Corpus**	**Instances**	**Filtered Instances**
News	Fake and Real News Dataset	21,417	9,247
Abstract	arXiv Dataset	2,302,911	189,298
Wikipedia	WikiText	29,565	550

Table 5: Statistics of source data.

A.2 Ablation Study

In Section A.2.1, we investigate the potential overlap among the three dimensions utilized in the evaluation, aiming to discern whether information from each dimension can be independently captured without being influenced by other dimensions. Subsequently, in Section A.2.2, we empirically demonstrate that evaluation models trained on data generated by specific LLM exhibit minimal bias towards distinct models.

A.2.1 Cross Validation of Evaluation Models

We conducted cross-validation experiments to observe the performance of each evaluation model on the data of the other two evaluation models.

We generated 300 sets of data lists for each dimension using the method described in Section 4.1.1. Based on the consistency of the score order, we calculate the accuracy of three evaluation models on the data of each evaluation dimension. As shown in the table 6, the emotionality and vividness evaluation models can achieve good performance on each other's data sets, which shows that these two dimensions can easily generalize to each other. However, both of them perform poorly on interactivity data, indicating

Evaluation Model	**Emotionality Data**	**Vividness Data**	**Interactivity Data**
Emotionality	**93.86**	92.77	31.16
Vividness	91.33	**94.68**	40.94
Interactivity	69.31	72.24	**96.74**

Table 6: The accuracy of three evaluation models on the data of each evaluation dimension. The results indicate a potential overlap between *Emotionality* and *Vividness*, whereas the assessment of *Interactivity* appears relatively independent, capable of capturing distinct characteristics within its own dimension.

---

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

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

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

that there are obvious differences between interactivity data and others. The interactivity evaluation model is more sensitive to data.

A.2.2 Fairness of Evaluation Models

In this section, we aim to investigate potential biases in evaluation models based on the source of their training datasets. Specifically, the “better” text within a training set (comprising 4 text pieces with varying levels of style strength) may contain style features from the source model (e.g., GPT-3.5) responsible for generating this set. These features may be captured by the trained evaluation model (e.g., **EvalModel-1.3B$_{\textbf{gpt-3.5}}$**), which in turn is used to score the source model (e.g., GPT-3.5) resulting in a higher score.

Dimension	**Pearson’s ρ**
Emotionality	87.45
Vividness	87.66
Interactivity	93.19
Average	89.43

Table 7: **Pearson’s** ρ between **EvalModel-1.3B$_{\textbf{gpt-3.5}}$** and **EvalModel-1.3B$_{\textbf{llama2-13B}}$** evaluation results in each dimension.

Results and Analysis Table 7 shows the correlation between **EvalModel-1.3B$_{\textbf{gpt-3.5}}$** and **EvalModel-1.3B$_{\textbf{llama2-13B}}$** and Table 3 shows the correlation between EMs and human evaluator. The strong correlation observed in the evaluation results of **EvalModel-1.3B$_{\textbf{gpt-3.5}}$** and **EvalModel-1.3B$_{\textbf{llama2-13B}}$** across various dimensions indicates that the evaluation model, trained with data from distinct sources, exhibits a high degree of consistency in evaluation preferences and has low possibility of bias.

We believe that employing sentence-level data alleviates the speech-style traits of the source model and ameliorates the bias of the evaluation model. Intriguingly, our human evaluation shows that, in the interactivity dimension, GPT-3.5 frequently utilizes phrases like "*please imagine*", "*picture this*" and "*my friends*" whereas Llama 2-Chat often employs expressions such as "*I know you might think this way* " to convey interaction with the audience. The sentence-based fine-grained evaluation method utilized in this paper helps the evaluation model to mitigate the risk of capturing such characteristics to a certain extent, promoting fairness.

B Training Details

For each evaluation dimension, we fine-tuned GPT-Neo-1.3B (Gao et al., 2020) using Lora (Hu et al., 2021) as the evaluation model. The training parameters are detailed in Table 8.

Optimizer	**LR_Scheduler**	**Learning Rate**	**Batch Size**	**Epochs**	**Weight Decay**
AdamW	Warmup=0.03 Decay="cosine"	2×10^{-5}	16	6	0.001

Lora r	**Lora α**	**Lora Dropout**	**Lora Target Module**
8	32	0.1	q_proj v_proj

Table 8: Details of evaluation modeling.

C Prompts

C.1 Prompt for LLMs to Evaluate Speech-Style Strength

Figure 6 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.

Output:[Candidate Text]
The above is a output of a formal-to-speech style transfer task, a good speech 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 speech 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.

Now you need to score the output based on the three 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 "
}
```

Figure 6: Prompt used for Llama 2-Chat-70B to score the speech-style strength in three dimensions.

C.2 Prompt for LLMs to Generate Sentence-Level Data to Train Evaluation Models

Figure 7 shows the prompts utilized in Section 4.1.1 to generate training data for training the evaluation model in vividness and interactivity dimensions. The subsequent prompts are directly employed for GPT-3.5. For Llama 2-Chat-113B, we only modified instruction to *You only need to reply in the following format, and do not reply with anything else* to ensure the model to return dictionary format for easy processing.

C.3 Prompt for LLMs to Perform TSST

We design two prompt types for LLMs to conduct Text Speech-Style Transfer (TSST): a concise prompt and an enhanced prompt, illustrated in Figures 8. 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 Case Study

one can experience certain speech-style text in audio version here.

D.1 Bad Cases of LLMs in TSST

In Figures 9, 10, 11, 12, 13, 14, we provide specific cases of challenges encountered by LLMs during the execution of TSST, as discussed in Section 5.2.2. These cases encompass issues denoted as Instruction Sensitivity, Uncontrolled Length Generation, Weak Transfer Strength and Inconsistent Content.

As shown in these cases, the red part of the input represents the part that is lost during transformation. The red part in the output represents the illusion produced by the model that is not found in the input text, and the green part indicates that the stylization is good.

Prompt for Vividness:
The input is a sentence from a speech. Please transform this sentence into four sentences with increasing degrees of vividness, while retaining the semantics of the input sentence. Incorporate vivid language and rhetoric to create a more vivid and captivating sentence. Four sentences should be close in length.
Please generate the following JSON formatted output and nothing else:
{
    “1”: “[Sentence without vividness.]”,
    “2”: “[Sentence with a little vividness]”,
    “3”: “[Sentence with medium vividness.]”,
    “4”: “[Sentence with a lot of vividness]”
}

Prompt for Interactivity:
The input is a sentence from a speech. Please transform this sentence into four sentences with increasing degree of interactivity, while retaining the semantics of the input sentence. Incorporate elements of interaction with the audience, such as asking questions, getting the audience to think, asking them to raise their hands, etc., to enhance audience participation and connection. Four sentences should be close in length.
Please generate the following JSON formatted output and nothing else:
{
    “1”: “[Sentence without interactivity.]”,
    “2”: “[Sentence with slight interactivity.]”,
    “3”: “[Sentence with medium interactivity.]”,
    “4”: “[Sentence with a lot of interactivity.]”
}

Figure 7: Prompt used for LLMs to generate four sentences with increasing degree of vividness and interactivity to train evaluation model in the two dimension respectively.

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

Inconsistency and insensitivity to style strength of LLMs as speech-style strength refer 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 15.

D.3 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 16. 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.

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. Imagine you are addressing a crowd as you transform this passage into a speech.

4. Re-contextualize the following passage into a speech format that is engaging and persuasive.

5. Reframe the following passage as a compelling speech that would resonate with your audience.

6. Transform the tone and style of the text to make it more suitable for a attractive speech.

Enhanced prompts:

1. Please transform the following passage into a captivating speech that would grab the audience's attention. Throughout the speech, incorporate engaging storytelling techniques, use vivid language to paint images in the minds of the audience, and keep them hooked by building suspense and excitement. Remember to maintain a conversational tone, as if you were directly addressing the audience, making them feel involved.

2. Convert the tone of the following text to be more suitable for delivering a speech. Take moments to pause and add emphasis to key points, allowing the audience to absorb and reflect on your words. You can use humor or anecdotes to break the ice and foster a warm rapport with your listeners. By doing so, you'll maintain their attention and leave a memorable impression.

3. Imagine you are addressing a crowd as you transform this passage into a speech. Tailor your speech to resonate with them on a personal level. You could consider asking the audience questions or requesting a show of hands to encourage active participation. Intersperse your speech with inspiring quotes or funny sentences, ensuring that your message reaches not just their ears but also their hearts.

4. Re-contextualize the following passage into a speech format that is engaging and persuasive. Use vivid language to make the content resonate with the audience. Incorporate rhetorical techniques, such as repetition, alliteration, or parallelism, to enhance the speech's persuasiveness and memorability. Interact with the audience through questions or reflections that encourage active engagement.

5. Reframe the following passage as a compelling speech that would resonate with your audience. Use persuasive language, emotional appeals, and relatable anecdotes to make the speech memorable and thought-provoking. Incorporate elements of audience interaction, such as asking for their input, sharing personal reflections, or creating opportunities for discussion.

6. Transform the tone and style of the text to make it more suitable for an attractive speech. Use dynamic language, engaging metaphors, and evocative imagery to create a vivid and captivating experience for the audience and stimulate audience involvement. Seek opportunities for audience interaction, such as brief discussions, sharing personal experiences, or inviting questions, to foster a sense of participation and connection.

Figure 8: Concise and enhanced instructions to help LLMs perform TSST. The purple indicates specific guidance.

Case 1
instruction: Please transform the following passage into a speech that would captivate an audience.
input:
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.
output:
Good evening, everyone! Today, I am here to talk to you about a vision for the future of Chile. A future where we are not only a nation among many, but a nation that stands out on its own. A nation that is admired and respected by the world. A nation that is a model for development and prosperity. A nation that can provide a better life for all its citizens. This future is within our reach if we work together. If we unite and strive for a common goal, we can make Chile a developed nation by 2025. We need to grow our economy, create jobs, and increase wages. We must ensure that our people have access to quality education and healthcare. We must make sure that our businesses can compete and our workers are protected. We must make sure that our country is a place where everyone can live in peace and security. We have the potential to make Chile a beacon of progress and success. But this will only happen if we have the courage and vision to make it a reality. Let us work together to create a future that all Chileans can be proud of. Thank you.

Case 2
instruction: Please transform the following passage into a captivating speech that would grab the audience's attention. Throughout the speech, incorporate engaging storytelling techniques, use vivid language to paint images in the minds of the audience, and keep them hooked by building suspense and excitement. Remember to maintain a conversational tone, as if you were directly addressing the audience, making them feel involved.
input:
Same as case 1.
output:
Good morning everyone! Today, I am here to talk to you about a transformative vision for the future of Chile - a vision that will create a country that is not just economically successful, but one that is prosperous and just. I am committed to making Chile the first Latin American nation to achieve developed nation status by 2025. To do this, I will double our current growth rate, increase investment, and eventually cut corporate taxes. This will bring more wealth to all Chileans, not just the few. I will also ensure that this wealth is shared fairly, so that everyone can benefit from our nation's prosperity. This is a bold vision, but I am confident that together we can make it a reality. Thank you!

Case 3
instruction: Convert the tone of the text to be more suitable for delivering a speech.
input:
Same as case 1.
output:
Chile's conservative presidential candidate Sebastian Pinera on Wednesday promised to double economic growth and make the country the first in Latin American to achieve developed nation status by 2025. He pledged to 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 nations. 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.
….(Omit 238 words.)

Figure 9: Cases of Alpaca-13B performing the TSST task with concise and enhanced instructions.

Case 4
instruction: Please transform the following passage into a speech that would captivate an audience.
input:
A compromise Republican tax bill released late Friday does not eliminate a $7,500 electric vehicle tax credit as Republicans in the U.S. House of Representatives had previously proposed. The measure follows the lead of the Senate version approved last month that did not eliminate the credit. Killing the credit could have hurt automakers like General Motors Co (GM.N), Volkswagen AG (VOWG_p.DE), Tesla Inc (TSLA.O) and Nissan Motor Co (7201.T). Consumers under current law are eligible for a $7,500 tax credit to defray the cost of plug-in electric vehicles. The electric vehicle tax credit starts to phase out after a manufacturer sells 200,000 plug-in vehicles. After an automaker hits that point, the $7,500 tax credit is still available for at least three more months before phasing out. Consumers are currently allowed to take the credit on vehicles until the manufacturer hits 200,000 plug-in vehicles sold. Electric vehicles have expensive batteries that make them pricier than gasoline-powered vehicles. The Electric Drive Transportation Association said in a statement late Friday it was pleased the credit would remain in law. “The credit supports innovation and job creation while helping drivers access advanced vehicle technology,” the group said. More than 50 automakers and other companies and groups released a letter earlier this week urging Congress to retain the credit, including Ford Motor Co (F.N), BMW AG (BMWG.DE), GM, Uber Technologies Inc. Former President Barack Obama repeatedly proposed hiking the tax credit for electric vehicles to $10,000 and converting it to a point-of-sale rebate, but Congress did not approve the measure. Automakers face mandates from California and a dozen other states to produce a rising number of zero-emission vehicles and have said the credits are essential to meeting requirements.
output:
Ladies and gentlemen, I have some good news to share with you today. A compromise Republican tax bill has been released, and it does not eliminate the $7,500 electric vehicle tax credit. This is a relief for automakers like General Motors, Volkswagen, Tesla, and Nissan, as well as for consumers who are looking to purchase plug-in electric vehicles.
…(Omit 132 words.)
It's worth noting that former President Barack Obama proposed hiking the tax credit for electric vehicles to $10,000 and converting it to a point-of-sale rebate, but Congress did not approve the measure. However, automakers face mandates from California and a dozen other states to produce a rising number of zero-emission vehicles, and they have said that the credits are essential to meeting these requirements.
In conclusion, I'm pleased to announce that the electric vehicle tax credit will remain in law, and I'm confident that it will continue to support innovation and job creation in the automotive industry. Thank you.

Case 5
instruction: Please transform the following passage into a captivating speech that would grab the audience's attention. Throughout the speech, incorporate engaging storytelling techniques, use vivid language to paint images in the minds of the audience, and keep them hooked by building suspense and excitement. Remember to maintain a conversational tone, as if you were directly addressing the audience, making them feel involved.
input:
Same as case 4.
output:
Ladies and gentlemen, have you ever imagined a world where driving a car could be more affordable, environmentally-friendly, and innovative? Well, that world is closer than you think, and it all starts with a simple yet impactful decision made by our lawmakers.
Picture this: You're driving down the highway, enjoying the scenic view, when suddenly you notice a sleek, shiny car passing by. It's not just any car; it's an electric vehicle – a symbol of the future of transportation. And you can't help but feel a sense of awe and admiration for the technology that made it possible.
…(Omit 84 words.)
So, what does this mean for you and me? It means that we have the opportunity to take advantage of a $7,500 tax credit to defray the cost of plug-in electric vehicles. And the best part? This credit doesn't disappear the moment an automaker sells 200,000 plug-in vehicles. It's still available for at least three more months before phasing out.
Now, imagine the impact this can have on our environment. Electric vehicles are zero-emission, which means they significantly reduce air pollution and our carbon footprint. And with more than 50 automakers and other companies and groups urging Congress to retain the credit, we can be sure that innovation and job creation will continue to thrive.
But it's not just about the cars; it's about our future. Automakers face mandates from California and a dozen other states to produce a rising number of zero-emission vehicles. And they've said that the credits are essential to meeting these requirements.
So, let's take a moment to appreciate the power of our government's decision. By keeping the electric vehicle tax credit in law, we're not only supporting innovation and job creation, but we're also giving ourselves a chance to drive into a greener, cleaner future.
And if you're still not convinced, just think about it this way: Wouldn't it be amazing to be part of a movement that's shaping the world for the better? To look back and say, "I was there when we started driving electric"?
So, let's embrace this change, and let's make our future brighter than ever before. Thank you!

Figure 10: Cases of Vicuna-13B performing the TSST task with concise and enhanced instructions.

Case 6

instruction: Please transform the following passage into a speech that would captivate an audience.

input:

Venezuela s opposition refused on Wednesday to swear in newly-elected governors before a pro-government legislative superbody it deems unconstitutional, as a small protest broke out in southern Bolivar state over fraud allegations. The pro-government electoral council announced in the middle of the night that the ruling socialists had won the Bolivar governorship, meaning President Nicolas Maduro s government took 18 of 23 states in Sunday s vote. Polls had put the opposition far ahead, and anti-Maduro politicians have alleged a litany of dirty tricks including switching electoral centers to dangerous areas at the eleventh hour and gross abuse of state resources. However, they have failed to give evidence of ballot-tampering, and some opposition candidates have conceded they lost due to high abstention in their demoralized ranks. Still, the disparate opposition coalition said its five winning candidates would not be sworn in by the controversial legislative superbody known as the Constituent Assembly. The governors-elect will only be sworn in as established in the constitution and the laws of the Republic, the Democratic Unity coalition said in a statement on Wednesday. Leftist Maduro has previously said that governors not sworn in by the pro-government legislative body will not be allowed to take their posts in a country reeling from widespread food and medicine shortages, a collapsing currency and soaring inflation. He described Venezuela s electoral system as the world s most secure and slammed U.S. President Donald Trump and other foreign leaders who questioned the veracity of the vote. Bolivar became a flashpoint after the electoral council briefly showed the opposition winning on its web site Sunday night before proclaiming the Socialist Party candidate as winner in the early hours of Wednesday. Opposition candidate Andres Velasquez accused the electoral council of invalidating some ballots cast for him. Pockets of his supporters protested outside the electoral board s offices in state capital Ciudad Bolivar, with some clashes breaking out on Monday and Tuesday. Some 50 people rallied on Wednesday, though the protest quickly fizzled. I am going to demonstrate to the world that this electoral process is fraudulent, Velasquez told local radio in Bolivar, home to many of Venezuela s gold and diamond mines. However, nationwide protest like those that rocked Venezuela for four months earlier this year are not expected, given fatigue and disappointment among demonstrators. The perennially divided opposition is in disarray after Sunday s election, with some leaders calling fraud and others conceding defeat, often in uncoordinated press statements. Sunday s election has left the socialists more confident of winning a presidential vote expected in late 2018.

output:

My fellow Venezuelans,

Today, we stand at a crossroads. The recent elections have left us with a choice: do we accept the results and move forward, or do we stand up for our rights and demand justice? The answer is clear. We will not be silenced by fraudulent elections and the tyranny of the ruling socialist party.

Yesterday, we saw the true colors of the electoral council, as they brazenly tampered with the results and stole our democratic victory. They Switched electoral centers to dangerous areas at the eleventh hour, and grossly abused state resources to secure their power. But we will not be deterred.

We, the opposition, have refused to swear in the newly-elected governors before this pro-government legislative superbody, which we deem unconstitutional. We will only be sworn in as established in the constitution and the laws of the Republic. We will not be bullied by President Maduro and his regime.

They may have the power and the weapons, but we have the truth and the people on our side. We will not be silenced by their fraudulent tactics and their abuse of power. We will fight for our rights and for the future of our country.

Bolivar has become a flashpoint in this struggle. The electoral council briefly showed the opposition winning on its website Sunday night before proclaiming the Socialist Party candidate as the winner in the early hours of Wednesday. Opposition candidate Andres Velasquez accused the electoral council of invalidating some ballots cast for him. Pockets of his supporters protested outside the electoral board's offices in state capital Ciudad Bolivar, with some clashes breaking out on Monday and Tuesday. But we will not be deterred. We will continue to fight for our rights and for the truth.

I, Andres Velasquez, will demonstrate to the world that this electoral process is fraudulent. We will not be silenced by the regime's tactics. We will continue to protest and fight for our rights, and we will not back down.

My fellow Venezuelans, the time for change is now. We cannot continue to live under this tyranny any longer. We must stand together and fight for our freedom and our future. We will not be silenced. We will not be oppressed. We will rise up and take back our country.

Join us in this fight for democracy and justice. Join us in this fight for our rights. We will not be silenced. We will not be defeated. We will rise up and take back our country.

Venezuela, join us in this struggle for freedom and democracy. We will not be silenced. We will not be defeated. We will rise up and take back our country.

¡Venceremos! We will triumph!

Figure 11: Case of Llama 2-Chat-13B performing the TSST task with concise instruction.

Case 7
instruction: Please transform the following passage into a captivating speech that would grab the audience's attention. Throughout the speech, incorporate engaging storytelling techniques, use vivid language to paint images in the minds of the audience, and keep them hooked by building suspense and excitement. Remember to maintain a conversational tone, as if you were directly addressing the audience, making them feel involved.
input:
Venezuela s opposition refused on Wednesday to swear in newly-elected governors before a pro-government legislative superbody it deems unconstitutional, as a small protest broke out in southern Bolivar state over fraud allegations. The pro-government electoral council announced in the middle of the night that the ruling socialists had won the Bolivar governorship, meaning President Nicolas Maduro s government took 18 of 23 states in Sunday s vote. Polls had put the opposition far ahead, and anti-Maduro politicians have alleged a litany of dirty tricks including switching electoral centers to dangerous areas at the eleventh hour and gross abuse of state resources. However, they have failed to give evidence of ballot-tampering, and some opposition candidates have conceded they lost due to high abstention in their demoralized ranks. Still, the disparate opposition coalition said its five winning candidates would not be sworn in by the controversial legislative superbody known as the Constituent Assembly. The governors-elect will only be sworn in as established in the constitution and the laws of the Republic, the Democratic Unity coalition said in a statement on Wednesday. Leftist Maduro has previously said that governors not sworn in by the pro-government legislative body will not be allowed to take their posts in a country reeling from widespread food and medicine shortages, a collapsing currency and soaring inflation. He described Venezuela s electoral system as the world s most secure and slammed U.S. President Donald Trump and other foreign leaders who questioned the veracity of the vote. Bolivar became a flashpoint after the electoral council briefly showed the opposition winning on its web site Sunday night before proclaiming the Socialist Party candidate as winner in the early hours of Wednesday. Opposition candidate Andres Velasquez accused the electoral council of invalidating some ballots cast for him. Pockets of his supporters protested outside the electoral board s offices in state capital Ciudad Bolivar, with some clashes breaking out on Monday and Tuesday. Some 50 people rallied on Wednesday, though the protest quickly fizzled. I am going to demonstrate to the world that this electoral process is fraudulent, Velasquez told local radio in Bolivar, home to many of Venezuela s gold and diamond mines. However, nationwide protest like those that rocked Venezuela for four months earlier this year are not expected, given fatigue and disappointment among demonstrators. The perennially divided opposition is in disarray after Sunday s election, with some leaders calling fraud and others conceding defeat, often in uncoordinated press statements. Sunday s election has left the socialists more confident of winning a presidential vote expected in late 2018.
output:
Ladies and gentlemen, today I want to talk to you about a story that's not just about politics, but about the fight for democracy, freedom, and the future of a nation. It's a story that's been unfolding in Venezuela, where the people have been facing unimaginable challenges and struggling to make their voices heard.
On Sunday, Venezuelans went to the polls to elect their governors, hoping for a change that would bring them closer to a better life. But what they got was a farce, a sham of an election that has left them feeling betrayed and disillusioned.
The ruling socialist party, led by President Nicolas Maduro, has been accused of using every dirty trick in the book to maintain their grip on power. They've been accused of switching electoral centers to dangerous areas at the eleventh hour, of gross abuse of state resources, and of tampering with the ballots. But despite these allegations, the opposition has failed to provide concrete evidence of these claims.
But here's the thing, my friends. It doesn't matter if there's concrete evidence or not. What matters is that the people of Venezuela have been crying out for change, and they were promised that change in this election. They were promised that their voices would be heard, that their votes would count. And now, it seems that those promises have been broken.
The opposition has refused to swear in the newly-elected governors before a pro-government legislative superbody that they deem unconstitutional. And why should they? Because that body is not a legitimate representative of the Venezuelan people. It's a body that has been stacked with Maduro's loyalists, a body that has been used to silence the opposition and maintain the status quo.
But the people of Venezuela are not giving up. They are fighting back, they are protesting in the streets, they are demanding to be heard. And I stand with them, because I believe that the voices of the people must be heard, no matter what the cost.
In Bolivar state, a small protest broke out over fraud allegations, and it's a sign of things to come. The people of Venezuela are not going to let this injustice stand. They are going to continue to fight for their rights, for their freedom, and for their future.
And so, my friends, I ask you to stand with the people of Venezuela. Stand with them in their fight for democracy, for freedom, and for a better life. Because the struggle of the Venezuelan people is not just their struggle, it's our struggle too. It's the struggle of all of us who believe in the power of the people, who believe in the power of democracy, and who believe in the power of freedom.
So let us stand together, my friends, and let us stand with the people of Venezuela. Let us show them that we hear their cries, that we see their struggles, and that we will not let them fight alone. Let us show them that we are all in this together, and that together, we can make a difference.
Thank you.

Figure 12: Case of Llama 2-Chat-13B performing the TSST task with enhanced instruction.

Case 8
instruction: Please transform the following passage into a speech that would captivate an audience.
input:
A smartphone is a portable computer device that combines mobile telephone functions and computing functions into one unit. They are distinguished from older-design feature phones by their more advanced hardware capabilities and extensive mobile operating systems, which facilitate wider software, access to the internet (including web browsing over mobile broadband), and multimedia functionality (including music, video, cameras, and gaming), alongside core phone functions such as voice calls and text messaging. Smartphones typically contain a number of metal–oxide–semiconductor (MOS) integrated circuit (IC) chips, include various sensors that can be leveraged by pre-installed and third-party software (such as a magnetometer, a proximity sensor, a barometer, a gyroscope, an accelerometer, and more), and support wireless communication protocols (such as Bluetooth, Wi-Fi, or satellite navigation). More recently, smartphone manufacturers have begun to integrate satellite messaging connectivity and satellite emergency services into devices for use in remote regions where there is no reliable cellular network.
Following the rising popularity of the iPhone in the late 2000s, the majority of smartphones have featured thin, slate-like form factors with large, capacitive screens with support for multi-touch gestures rather than physical keyboards and have offered the ability for users to download or purchase additional applications from a centralized store and use cloud storage and synchronization, virtual assistants, as well as mobile payment services. Smartphones have largely replaced personal digital assistant (PDA) devices, handheld/palm-sized PCs, portable media players (PMP), and, to a lesser extent, handheld video game consoles.
Improved hardware and faster wireless communication (due to standards such as LTE) have bolstered the growth of the smartphone industry. In 2022, 1.43 billion smartphone units were shipped worldwide. 75.05 percent of the world population were smartphone users as of 2020.
output:
Ladies and gentlemen, distinguished guests, and fellow technology enthusiasts, I stand before you today to take you on a captivating journey into the world of smartphones. This remarkable invention has revolutionized the way we communicate, connect, and experience the world around us. So, fasten your seatbelts and prepare to be amazed!
Imagine a device that combines the power of a computer with the convenience of a mobile phone—a device that fits in the palm of your hand and opens up a universe of possibilities. Enter the smartphone, a portable marvel that has reshaped our lives in ways we could never have imagined.
Gone are the days of clunky feature phones. Smartphones have emerged as the pinnacle of technological advancement, boasting cutting-edge hardware capabilities and sophisticated operating systems. These miniature powerhouses grant us access to a vast array of software, bring the entire internet to our fingertips, and provide us with a plethora of multimedia functionalities. Music, videos, cameras, gaming—the possibilities are endless.
But a smartphone is more than just a collection of circuits and sensors. It is a gateway to a world of seamless communication and boundless exploration. With features like voice calls, text messaging, and the ability to browse the web over mobile broadband, smartphones have become indispensable companions in our daily lives.
Think about the incredible technology packed into these devices. Smartphones are equipped with metal-oxide-semiconductor integrated circuit chips, which fuel their lightning-fast performance. They house an array of sensors, including magnetometers, proximity sensors, barometers, gyroscopes, and accelerometers. These sensors, combined with pre-installed and third-party software, enable a host of innovative applications and experiences.
Communication knows no bounds for these remarkable devices. With support for wireless communication protocols like Bluetooth, Wi-Fi, and satellite navigation, smartphones have truly connected us in ways we never thought possible. And now, even in the most remote regions, where reliable cellular networks are scarce, manufacturers have integrated satellite messaging connectivity and emergency services into these incredible devices.
…(Omit 78 words.)
The rise of the smartphone has been nothing short of extraordinary. Improved hardware and faster wireless communication standards, such as LTE, have propelled the industry to unprecedented heights. In 2022 alone, a staggering 1.43 billion smartphone units were shipped worldwide. And as of 2020, a staggering 75.05 percent of the global population were smartphone users—highlighting the widespread impact these devices have had on our lives.
As we look to the future, we can only anticipate further advancements in smartphone technology. Artificial intelligence, augmented reality, and 5G connectivity are poised to unleash a new wave of innovation, promising even more immersive experiences and endless possibilities.
So, my friends, let us embrace the marvels of the smartphone. Let us celebrate the way it has transformed our lives, bridging gaps, and connecting us across borders. Let us marvel at the power we hold in our hands—a power that has forever changed the way we communicate, work, and play.
In the palm of your hand lies a gateway to a world of endless possibilities. It's time to unlock the full potential of the smartphone and embark on a journey of discovery, innovation, and connection.
Thank you.

Figure 13: Case of GPT-3.5 performing the TSST task with concise instruction.

Case 9

instruction: Please transform the following passage into a captivating speech that would grab the audience's attention. Throughout the speech, incorporate engaging storytelling techniques, use vivid language to paint images in the minds of the audience, and keep them hooked by building suspense and excitement. Remember to maintain a conversational tone, as if you were directly addressing the audience, making them feel involved.

input:

A smartphone is a portable computer device that combines mobile telephone functions and computing functions into one unit. They are distinguished from older-design feature phones by their more advanced hardware capabilities and extensive mobile operating systems, which facilitate wider software, access to the internet (including web browsing over mobile broadband), and multimedia functionality (including music, video, cameras, and gaming), alongside core phone functions such as voice calls and text messaging. Smartphones typically contain a number of metal–oxide–semiconductor (MOS) integrated circuit (IC) chips, include various sensors that can be leveraged by pre-installed and third-party software (such as a magnetometer, a proximity sensor, a barometer, a gyroscope, an accelerometer, and more), and support wireless communication protocols (such as Bluetooth, Wi-Fi, or satellite navigation). More recently, smartphone manufacturers have begun to integrate satellite messaging connectivity and satellite emergency services into devices for use in remote regions where there is no reliable cellular network.

Following the rising popularity of the iPhone in the late 2000s, the majority of smartphones have featured thin, slate-like form factors with large, capacitive screens with support for multi-touch gestures rather than physical keyboards and have offered the ability for users to download or purchase additional applications from a centralized store and use cloud storage and synchronization, virtual assistants, as well as mobile payment services. Smartphones have largely replaced personal digital assistant (PDA) devices, handheld/palm-sized PCs, portable media players (PMP), and, to a lesser extent, handheld video game consoles.

Improved hardware and faster wireless communication (due to standards such as LTE) have bolstered the growth of the smartphone industry. In 2022, 1.43 billion smartphone units were shipped worldwide. 75.05 percent of the world population were smartphone users as of 2020.

output:

Ladies and gentlemen, imagine a world where communication, entertainment, and endless possibilities fit right in the palm of your hand. A world where a single device combines the power of a computer and the convenience of a mobile phone. It's a world where the future has arrived, and it comes in the form of a captivating device known as the smartphone.

But what exactly is a smartphone? Well, it's not just any ordinary gadget. No, my friends, it's a marvel of technology that weaves together the realms of imagination and practicality. Picture this: a sleek and slender device, exuding elegance, with a vibrant, capacitive screen that responds to the gentlest touch of your finger. Gone are the days of clunky physical keyboards, for now, you can effortlessly glide your fingers across the screen, engaging in a symphony of multi-touch gestures that unlock a world of possibilities.

These incredible devices, my friends, are more than just phones. They are companions, confidants, and gateways to a universe of knowledge and entertainment. With the tap of a finger, you can browse the vast expanse of the internet, exploring its depths and uncovering treasures of information. No longer bound by wires and cables, you can connect with friends and loved ones across the globe, sharing laughter, tears, and moments of pure joy. The smartphone has become the heartbeat of our digital age, pulsating with life and energy.

But it doesn't stop there, my friends. These ingenious gadgets are equipped with a symphony of sensors, hidden beneath their sleek exteriors. A magnetometer, a proximity sensor, a barometer, a gyroscope, an accelerometer, and more—these sensors work harmoniously, like the gears of a finely tuned machine, to enhance our experiences. They bring a new dimension to gaming, immersing us in virtual worlds that defy imagination. They empower us to capture the beauty of life with stunning cameras, preserving memories that will forever be etched in our hearts.

Now, let me take you on a journey through time. Cast your minds back to the late 2000s when a revolution was brewing. It was the rise of a device that would forever change the landscape of technology—the iPhone. With its sleek design, large screen, and support for multi-touch gestures, it set the stage for a new era of smartphones. And just like a wildfire, the iPhone sparked a revolution that engulfed the world, captivating the hearts and minds of millions.

…(Omit 236 words.)

In 2022 alone, a staggering 1.43 billion smartphone units were shipped worldwide, a testament to their irresistible allure. And as of 2020, a whopping 75.05 percent of the world's population had embraced the smartphone revolution, their lives forever intertwined with these magnificent devices.

So, my friends, as we stand here today, let us marvel at the wonders that surround us. Let us recognize the incredible journey that brought us here, from the humble beginnings of the feature phone to the extraordinary realm of smartphones. And let us embrace the future with open arms, for the smartphone is not just a tool but a gateway to infinite possibilities. It is a symbol of human ingenuity and our unwavering spirit to conquer the unknown.

Together, let us embark on this journey, where technology meets imagination, where dreams become reality, and where the smartphone remains at the heart of it all. Thank you, and may the wonders of the smartphone forever inspire us to reach for the stars.

Figure 14: Case of GPT-3.5 performing the TSST task with enhanced instruction.

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
EM(gpt_based): GPT-3.5 > Llama 2-Chat-13B
EM(llama_based): GPT-3.5 > Llama 2-Chat-13B
Human: GPT-3.5 > Llama 2-Chat-13B

Figure 15: An instance of the evaluation failure in the interactivity dimension using **Llama 2-Chat-70B**. The provided prompt is shown in Figure 6 . 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 16: Prompts and results of GPT-4 evaluating the semantic consistency of two texts. The length of Text 2 is about 60% less than Text 1.