Think, Plan, then Write: Chain of Writing for Zero-shot Multi-Attribute Controlled Text Generation

Anonymous ACL submission

Abstract

Multi-attribute controlled text generation (CTG) requires models to generate sentences with prespecified attributes. Previous works often utilize the corresponding single-attribute data to train the multi-attribute generators. However, exploring the type (mainly sentiment and topic attributes in the English language) and number (up to three) of attributes is still limited, since the cost of data collection also increases significantly if new attributes emerge. Benefiting from recent advanced large language models (LLMs), we experimentally reveal that LLMs with standard promptings could get promising performances on multi-attribute CTG tasks without any single-attribute data. However, utilizing standard promptings often suffers from problems of missing/misunderstanding attributes. To address these concerns, our basic idea is to help LLMs better understand attributes and plan the generated content before the final completions, just as human writers do. As a result, the proposed CoW, a Chain-of-Writing prompting, hints LLMs conduct multi-attribute CTG in a step-by-step manner. Following the think-plan-write order, CoW decomposes the task into three corresponding sub-steps, and uses discrete promptings to encourage LLMs to generate auxiliary information, such as explaining the meanings of attributes and creating a storyline. Experiments on three generation tasks demonstrate that CoW could achieve general improvements on up to seven attributes, and these empirical results could provide novel insight to greatly expand the task settings of multi-attribute CTG.

1 Introduction

Multi-attribute CTG mainly concerns generating a natural sentence satisfying pre-specified attributes (Zhang et al., 2022), such as topic, sentiment, tense, etc (Lample et al., 2019; Lyu et al., 2021). Driven by the cost of multi-attribute text collection, previous progress generally explores multi-attribute CTG under zero-shot settings (Yang et al., 2022). Specifically, they often use a variety of single-attribute text and hint the generator to present all these attributes together in one completion for the multi-attribute generation purpose.

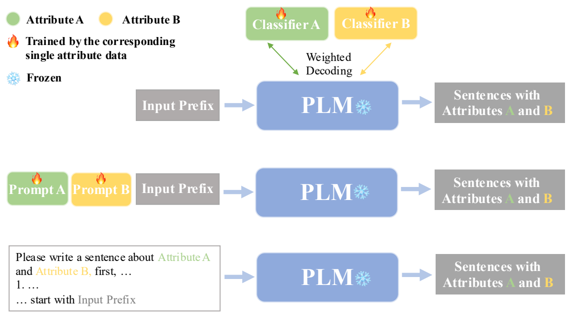

Figure 1: An example of two-attribute generation to illustrate the differences between various types of multi-attribute CTG methods. From top to bottom: 1) Classifier-based method; 2) Soft-prompt-based method and 3) Chain-of-writing based method.

Existing efforts for "hint the generator" can be divided into two types: 1) Classifier-based method and 2) Soft-prompt-based method. As shown in Figure 1, the former trains a set of attribute classifiers to weight output logits of a fixed pre-trained language model (PLM) (Dathathri et al., 2020; Krause et al., 2021; Liu et al., 2022), and the latter trains a set of soft single-attribute prompts (continuous vectors) to represent each attribute and then combine them as a whole to control a fixed PLM (Qian et al., 2022; Yang et al., 2022).

Despite their great progress, the exploration of multi-attribute CTG is still under-explored as only a small number (two or three) and variety (mainly sentiment and topic) of attributes have been considered. One intuitive reason is that the cost increases significantly as the number and variety of attributes increase. Thanks to the recent Large Language Models (LLMs) showing a strong text generation capability under zero-shot setting (Wang et al., 2023), we experimentally reveal that a standard prompting (e.g., *Please write a positive fantasy about "Sunset at Park" with about 200 words*) can hint LLMs to generate sentences with pre-specified attributes. However, LLMs with such simple promptings often suffer from missing/misunderstanding attributes of the generated text (such as rhetoric, genre, and topic), resulting in poor-quality generation (see § 4.1).

To solve the problems mentioned above, we propose a general zero-shot multi-attribute generation framework — Chain of Writing (**CoW**). Unlike previous multi-attribute CTG (Yang et al., 2022; Qian et al., 2022) that defines the zero-short settings as only using single-attribute data, CoW benefits from LLMs that do not need both additional training stage and single-attribute annotated data. Different from the standard prompting guiding LLMs to directly generate the final completions, our basic idea is to decompose the multi-attribute CTG into a series of sub-steps. In each step, discrete prompting is used to hint LLMs generate intermediate auxiliary information before producing sentences with pre-specified attributes. Specifically, inspired by the human writing habit that often first a draft outline and then the full text (Spivey, 2006), CoW decompose multi-Attribute CTG into three steps: 1) **Think**, focusing on the in-depth explanation of the pre-specified attributes; 2) **Plan**, following the template to finish a synopsis in the form of natural language; 3) **Write**, writing out the entire text based on the previous information. To conduct a comprehensive empirical evaluation, we examine the generalizability of CoW by instantiating it for three multi-attribute CTG tasks. These tasks include the English review generation, the English and Chinese story generation, with up to seven widely-used attributes as closely as possible to the naturally using situation (i.e., sentiment, topic, fact, length, genre, rhetoric, and its place in the final text). Extensive experiments show that CoW consistently improves the two LLMs and beats the standard prompting with a considerable performance gap. The main contributions of this work could be summarized as follows:

1. We have enriched the task settings for multi-attribute CTG, extending the number of attributes to seven, and requiring the model to do this in a full zero-shot setting (i.e., no single-attribute data). This will increase the level of the task challenge or difficulty and encourage further deep research.
2. Based on this new task setup, we propose a novel framework CoW, which decomposes the multi-attribute CTG into sub-steps and introduces a set of discrete sub-promptings to hint LLMs to generate sentences following a think-plan-write order.
3. We will release all the human evaluation results, which contain scores of no less than six evaluation dimensions for 3600 samples. We believe this dataset will facilitate the study of multi-attribute detection and generation.

2 Related Work

Multi-attribute CTG aims at generating sentences constrained by pre-specified attributes, which plays an important role in creative writing (Zhang et al., 2022). Existing efforts focus on utilizing single-attribute data to pursue multi-attribute CTG, including classifier-based and soft-prompting-based methods. Specifically, the former trains a set of single-attribute classifiers, which are used to adjust the output probabilities (Krause et al., 2021; Russo et al., 2020; Lample et al., 2019; Yang and Klein, 2021) or latent representations (Dathathri et al., 2020) of a fixed PLM in each multi-attribute CTG inference step. The latter often includes an extra training stage. In the implementation, soft prompts (i.e., continuous vectors) are trained to represent each of all single attributes. Then, a continued multi-attribute-training stage is conducted by joint training single-attribute prompts (Qian et al., 2022) or training an extra single-attribute-prompt connector (Yang et al., 2022). As a result, these single-attribute prompts are concatenated as a whole to hint the PLM. However, the multi-attribute CTG exploration in terms of the number (mostly, no more than three) and variety (mainly focus on text style) of attributes are still limited. This may be caused by the mounting costs when new attributes are added, such as data annotation. In contrast, thanks to LLMs' strong language modeling ability, CoW achieve multi-attribute CTG with neither an extra training stage nor annotated single-attribute data. Benefiting from this new framework, we take in-depth studies on multi-attribute CTG and tentatively give a successful implementation in both story and review generation tasks with up to seven attributes.

Text Generation with LLMs Benefit from recent advanced LLMs, the performance of a large

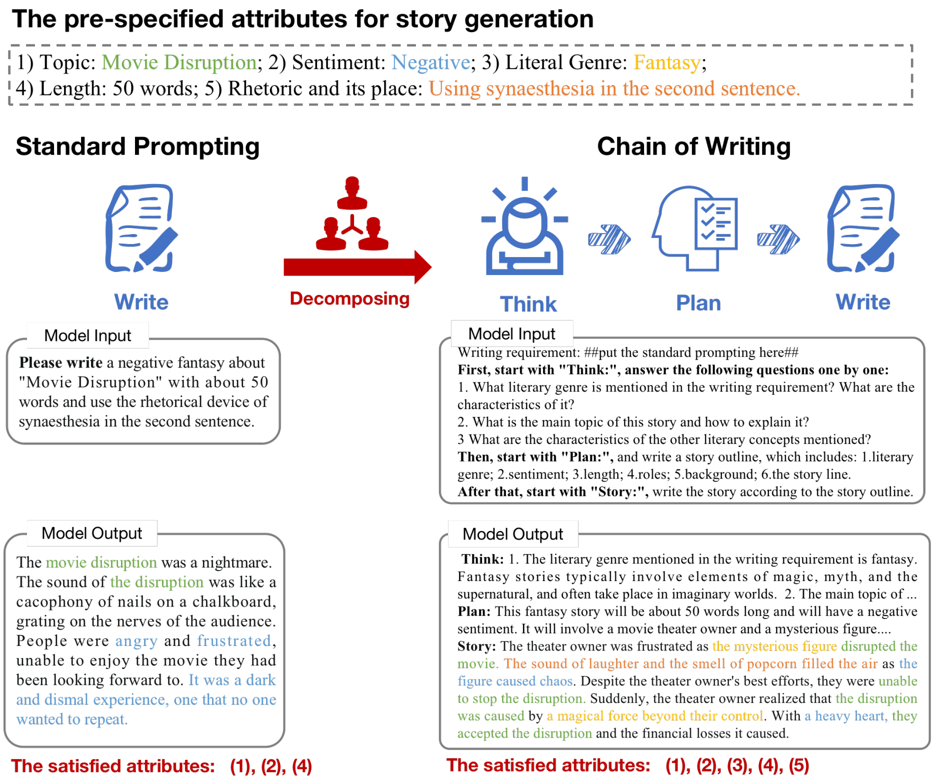

Figure 2: An example of hinting LLMs on multi-attribute CTG tasks to illustrate the differences between instantiations of standard prompting and CoW. In CoW, "##put the standard prompting##" denotes inserting the model input sentence of standard prompting here for the fair comparison.

number of NLG tasks has significantly been improved (Zhao et al., 2023), such as story generation (Yuan et al., 2022; Lee et al., 2022), summarization (Ouyang et al., 2022), and dialogue (Thoppilan et al., 2022). However, the exploration of attribute-based CTG with LLMs is still in the early stages, yet is regarded as an important part of the NLG field (Zhang et al., 2022). In this paper, we have explored how to utilize LLMs on multi-attribute CTG tasks in terms of task setting, attribute number, and variety. Our preliminary experiments reveal that LLMs may suffer from missing attributes if only resort to standard prompting. The biggest difference is decomposing the multi-attribute generation task to hint the generation of LLMs in a step-by-step way. Extensive experiments verify the effectiveness of our idea.

3 Methodology

3.1 Chain of Writing Framework

Before introducing CoW, we start with elaborating on how to generate multi-attribute sentences with standard prompting under LLMs. As shown in Figure 2 top, we take a toy example from English story generation to illustrate it. First, the pre-specified attributes are represented by a set of attribute-relevant words/phrases, each of which denotes one corresponding attribute (e.g., Negative for sentiment). Then, the standard prompting incorporates these words/phrases all into a piece of text T_s in natural language (i.e., the text starting with "Please write..." in Figure 2). Finally, given the input T_s, the large language model $LLM(\cdot)$ would generate a story Y_{story} with pre-specified attributes by:

$$Y_{story} = LLM(T_s). \quad (1)$$

Unlike the standard prompting, which encourages LLMs to immediately generate the final story, CoW decomposes the task and presents a series of sub-step promptings that promote the generation of various auxiliary information before the final completions. As a result, given the input T_{CoW}, the large language model $LLM(\cdot)$ would generate a story Y_{story} with pre-specified attributes by:

$$\begin{aligned} Y_{story} &= LLM(T_{\text{CoW}}), \\ T_{\text{CoW}} &= \{T_s\} \oplus \{T_t\} \oplus \{T_p\} \oplus \{T_w\}, \end{aligned} \quad (2)$$

where $\{\cdot\} \oplus \{\cdot\}$ denotes concatnating two promptings. T_t, T_p, and T_w represent the promptings for the think, plan, and write steps, respectively.

3.2 Promptings under CoW

Inspired by the human writing process that first gathers information, then creates a working thesis, and finally writes the full completions (Spivey, 2006), CoW involves sub-steps promptings to hint LLMs generate corresponding content for these three steps, respectively. In detail, as shown in the right part of Figure 2, we take the instantiation of CoW on the English story generation as an example. The proposed think, plan, and write promptings not only refer to human writing behavior, but also in order of increasing difficulty, hinting LLMs to generate text ranging from the word-, sentence- to document-level. Notably, since our main goal is to illustrate how our framework CoW works, and as a first attempt we will not explore the implications of the specific words chosen in the sub-steps prompting. The details are as follows:

Think Propmting focuses on guiding the model to explain some of the concepts/entities in the standard prompting to improve the understanding of them. Specifically, our preliminary experiment found that simply adding a sentence after standard prompting (i.e., asking the model to explain the meaning of the standard prompting before writing the completions) can consistently improve performances on three generation tasks (see § 4.1). However, we believe that such coarse-grained hints may cause LLMs to miss explaining some important concepts, i.e., attribute-relevant concepts. As a result, CoW constructs a fine-grained think prompting to promote LLMs focusing on explaining literary concepts in standard prompting, since literary concepts are often attribute-related and difficult to understand intuitively. Specifically, as shown in Figure 2 "Think" part, think prompting is conducted in a question-and-answer manner, with the subject of the question ranging from special cases (e.g., genre and topic) to general concepts (other literary concepts) to ensure that all literary concepts are covered as much as possible.

Plan Propmting ensures that the generated text is logically developed (i.e., having a logic flow), which is very important in creative text writing (Barroga and Matanguihan, 2021; Shang et al., 2019). Unlike previous work utilizing a set of keywords to represent a storyline (Goldfarb-Tarrant et al., 2020; Narayan et al., 2021; Goldfarb-Tarrant et al., 2019), plan prompting guides LLMs to generate a paragraph that expresses the important plot development, which is similar to the human-written synopsis. It is worth mentioning that plan prompting still follows the principle of generating text from easy to difficult. As shown in Figure 2 "Plan" part, following plan prompting, LLMs first determine the plot-related details (e.g., the main roles and background), and then write the final outline according to the details.

Write Propmting requires the model to generate the final completions based on the previous contents. Specifically in English story generation, it requires LLMs first to generate the special signal "Story:" and then write the entire story. In this case, the story body could be easily separated from the intermediate results during post-processing.

4 Experiments

In this section, we use three multi-attribute CTG tasks to comprehensively evaluate the effectiveness of CoW. For an extensive comparison, we compare our CoW with the standard prompting and provide detailed analyses in further discussions.

Tasks and datasets. We conduct experiments in two naturally using scenarios for multi-attribute CTG to evaluate CoW, which are story generation and review generation. Notably, the multi-attribute CTG task discussed in this paper is about evaluating different methods with as much variety and number of attributes as possible. As a result, the existing datasets might be insufficient under the new task settings. For example, the current benchmarks are mainly concerned with sentiment and topic attributes (Yang et al., 2022; Qian et al., 2022), leaving a huge exploration space for more diverse attributes. Therefore, as shown in Table 1, we extend the attributes in the construction of the experimental datasets.[1] The details are: (1) **English Story Generation** Following previous works (Yang et al., 2022; Dathathri et al., 2020; Krause et al., 2021), the sentiment attribute is defined as a binary attribute, i.e., positive and negative. The topic attribute is based on the widely-used benchmark ROCStories Corpus (Mostafazadeh et al., 2016), which contains 98k five-sentence stories and the corresponding titles. Specifically, we randomly

[1] Due to space constraints, the corresponding words/phrases for each attribute can be found in Appendix A.1.

Language	Sentiment	Topic	Genre	Rhetoric (Position)	Length
		Story Generation			
English	Positive/Negative	200	8	18 (6)	50/100/200 words
Chinese	Positive/Negative	200	10	8 (6)	50/100/200 words
		Review Generation			
English	Positive/Negative	200	2	10 (4)	20/50/100 words

Table 1: The core statistics of the multi-attribute CTG datasets. Rhetoric (Position) denotes the number of rhetorical devices used and the types of positions in which the pre-specified rhetorical devices appear in generated sentences.

select 200 titles for the topic attribute. Following the list of writing genres[2] and classifications of rhetorical devices (Harris et al., 1997), genre and rhetoric attributes consider commonly-using types, and the task challenge is then increased by adding the position attribute of rhetoric. (2) **Chinese Story Generation** follows the sentiment attribute of the above task. Meanwhile, we randomly selected 200 story titles from Chinese story websites[3] as the corresponding content of the topic attribute. Following (Harbsmeier and Harbsmeier, 1999; Birch, 2022), genre and rhetoric contain specific attributes with Chinese cultural characteristics, such as Wuxia in genre and Pairing in rhetoric. (3) **English Review Generation** includes multi-attribute generation tasks for three review scenarios, namely food, books, and movies as the review topic, respectively. Specifically, we randomly select a subset of food names from the Yelp restaurant review dataset (Lample et al., 2019) as the food topic. Besides, the Book titles are selected from Amazon Book Review, and the movie titles come from IMDb's "Top 100" movies.[4] Notably, the genre attribute is set in two types: using colloquial-style expression or written-style expression in final completions, which is different from story generation as its literary genre is more limited. Besides, the range of the length attribute is shortened to match the characteristics of commonly used reviews.

Language models and promptings. In this work, we focus on using two LLMs from the GPT-3.5 family[5] to evaluate CoW, because they are one of the mainstream LLM structures in current works. Due to the cost of manual evaluation and the first attempt, we do not fully discuss the size and type of LLMs in this paper (in the preliminary experiments, we also tried LLMs with fewer parameters but got poor performances, such as GPT-J (Wang and Komatsuzaki, 2021), similar conclusions are also mentioned in Wei et al. (2022)). The details are (1) **Text-davinci-003** (Text-003) is an improvement on the InstructGPT model text-davinci-002, which is trained by PPO strategy (Schulman et al., 2017). (2) **GPT-3.5-turbo** (GPT-3.5)[6] is an improvement on text-davinci-003 that is optimized for chat. Based on the model and task setup, we focus on evaluating our CoW, standard prompting and its variants[7], since they can get promising performance without any single-attribute data. The details are as follows: (1) **Standard prompting** (SP) summarizes all the attribute requirements in one sentence (e.g., *Please write a positive fairy tale about "Going to the lake" with about 50 words and use rhetorical device of parody in the second sentence.*), and we provide templates for each task to insert different attributes. (2) **Standard prompting + Simply Expalin** (SP + Simply Explain) The biggest difference with standard prompting is that we add a general-using sentence after each standard prompting, and explore whether LLMs have the ability to explain pre-specified attributes (like a simplified version of our thinking prompting). For example, we use the following simple explain sentence in story generation: *First explain the meaning of the previous sentence that starts with "Explain:", then write the story that starts with "Begin:".* (3) **CoW** is the prompting aiming at decomposing the multi-attribute CTG task, requiring LLMs to generate the multi-attribute sentence in the order of think-plan-write. It is worth noting that we re-use the sentence of standard prompting as the task descriptions in CoW, in order to compare them in a fair circumstance.

[2] https://en.wikipedia.org/wiki/List_of_writing_genres

[3] See https://www.gushi365.com/ and https://www.ppzuowen.com/

[4] Books: https://www.amazon.com/amazonbookreview Movies: https://www.imdb.com/search/title/?groups=top_100&sort=user_rating,desc

[5] https://platform.openai.com/docs/models/gpt-3-5

[6] In implementation, we use gpt-3.5-turbo-0301 version, which is a snapshot of gpt-3.5-turbo from March 1st 2023

[7] see Appendix A.3 for full details

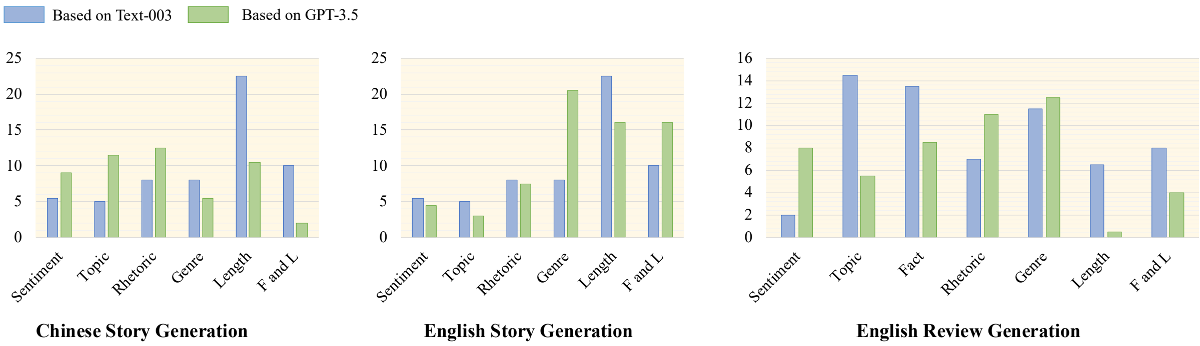

Figure 3: The Statistical analysis about samples proportions of the scores of two LLMs with CoW over them with standard prompting on the three multi-attribute generation tasks. 0 means that performances of CoW and standard prompting are identical, and the larger the ratio, the greater the performance gain brought by CoW

Model	Method	Sentiment	Genre	Topic	Rhetoric	Length	F and L
		Chinese Story Generation					
Text-003	**SP**	2.62	2.25	2.77	1.57	1.20	2.91
	SP + Simply Explain	**2.67**	1.97	2.78	1.61	1.35	2.92
	CoW	**2.67**	**2.35**	**2.81**	**1.74**	**1.51**	**3.06**
GPT-3.5	**SP**	2.65	2.45	2.72	1.60	1.15	3.12
	CoW	**2.72**	**2.47**	**2.84**	**1.75**	**1.31**	**3.16**
		English Story Generation					
Text-003	**SP**	2.77	2.02	2.75	1.54	2.13	3.11
	SP + Simply Explain	2.77	2.12	2.77	1.58	**2.42**	3.12
	CoW	**2.79**	**2.28**	**2.90**	**1.59**	2.26	**3.26**
GPT-3.5	**SP**	2.90	2.12	2.95	1.62	2.24	3.22
	CoW	**2.92**	**2.29**	**2.96**	**1.73**	**2.44**	**3.33**

Table 2: The main results of Chinese story generation and English story generation task, respectively. "F and L" denotes the score of fluency and logic flow. Bold values represent the maximum values of each model with a different method.

Evaluation Metrics. Following Spangher et al. (2022), we invited three expert annotators to independently annotate all the method-generated sentences from six metrics: a. Sentiment (1-3) b. Topic (1-3) c. Fact (0-3) d. Genre (1-3) e. Rhetoric (1-3) f. Fluency and Logical Flow (1-5). It is worth noting that the fact metric is only used in the review generation task to confirm whether the relevant description of the subject is true (for example, whether the author of the book mentioned is correct or whether the actor of the movie mentioned is actually in the movie). Meanwhile, we also conduct the automatic evaluation, such as the Length. Please see Appendix A.2 for all the details.

4.1 Main Results

Overview. To facilitate human scoring and follow previous work, our rating range is set to be very narrow (1-3), which may result in some CoW scores from Table 2, that do not appear to be much better than SP, and may raise concerns about the significance of scores and minor improvement. Considering this reason, we supplement the performance difference between CoW and SP through the ratio in Figure 3, from which we can make the following conclusions:

1. CoW consistently improves LLMs on three multi-attribute generation tasks to a great extent compared to standard prompting. CoW improves LLMs' performances on all types of scores for all tasks without introducing any external knowledge since none of the scores in Figure 3 is 0, especially the length attribute for story generation (e.g., 22.5% and 16.0% improvements for English).

2. The performance gains brought by CoW related to both the task and LLMs categories. First, as the difficulty of multi-attribute CTG task increases, CoW could bring greater performance

Model	Method	Sentiment	Topic	Fact	Genre	Rhetoric	Length	F and L
Text-003	**SP**	2.91	2.27	2.41	2.79	1.86	2.18	3.05
	CoW	**2.93**	**2.41**	**2.43**	**2.86**	**1.88**	**2.29**	**3.12**
GPT-3.5	**SP**	2.80	2.32	2.24	2.83	1.85	2.74	3.08
	CoW	**2.85**	**2.36**	**2.35**	**2.90**	**2.01**	**2.75**	**3.17**

Table 3: The main results of English review generation task. "F and L" denotes the score of fluency and logic flow. Bold values represent the maximum values of each model with a different method.

language	Method	Sentiment	Topic	Genre	Rhetoric	Length	F and L
English	**CoW**	**2.79**	**2.28**	**2.90**	**1.59**	**2.26**	**3.26**
	Plan w/o	2.78	2.12	2.89	1.56	1.86	3.18
	Think w/o	**2.79**	2.27	2.88	1.46	1.97	3.19
Chinese	**CoW**	**2.67**	**2.35**	**2.81**	**1.74**	**1.51**	**3.06**
	Plan w/o	2.63	2.27	**2.81**	1.72	1.29	2.89
	Think w/o	2.66	2.26	2.77	1.68	1.49	2.94

Table 4: The ablation study on using different sub-steps promptings with LLMs. "Plan w/o" and "Think w/o" denote using CoW without the plan prompting and think prompting, repsectively.

gains to LLMs. For example, it is more difficult for LLMs to complete story-generation tasks than the review-generation task in terms of longer text length (Max. 200 v.s. 100) and richer genres (8/10 v.s. 2). Meanwhile, all the performance gains with CoW on story generation tasks are higher than those on the review generation task. Second, as the language modeling ability of LLMs increases, CoW may bring fewer performance gains to LLMs. For example, the vast majority of the LLM Text-003 with CoW bring performance gains that are higher than the LLM GPT-3.5 with CoW, and GPT-3.5 is an improvement of Text-003 on chat.

Story Generation. The main results are shown in Table 2, CoW beats all of the baselines both in multi-attribute controllability and text quality. Besides, we can make the following conclusions:

3. Simply explain strategy could improve LLMs' performances in multiple aspects by adding only one sentence. For lack of space and easy observation, we conduct the simply explain strategy on Text-003, since this model mentioned in the overview using additional promptings will bring greater performance gains. The strategy of requiring LLMs first to explain standard prompting and then generate completions (SP + Simply Explain in the table) leads to a promising improvement in attribute controllability and text quality, yet is limited in terms of the topic (2.77/2.75 v.s. 2.78/2.77), fluency and logic flow (2.91/3.11 v.s. 2.92/3.12). These experimental results reveal that LLMs can improve performances on multi-attribute CTG tasks through "self-interpretation", without any annotated single-attribute text data.

4. Beyond the commonly-used language English, CoW can be extended to multilingual CTG tasks and consistently improve LLMs performances. CoW shows encouraging performances on the Chinese generation task, which means that costs can be further reduced because some non-English attribute text might be more expensive to collect. Interestingly, compared with the English-based task, CoW provides greater performance gains in terms of sentiment, rhetoric, and topic.

Review Generation. The main results are shown in Table 3, CoW also beats all of the baselines both in multi-attribute controllability and text quality. Besides, we can make the following conclusions:

5. CoW could also be helpful in improving the authenticity of the completions, though "information given should be truthful" is not explicitly mentioned promptings. Taking the book review as an example, we believe that a good book review should not only provide more detailed information about the book (e.g., the author and the year it was written) but also be truthful. Therefore, except for the metrics mentioned in the story generation task, we are also concerned about whether the description of the topic fits the facts in the review generation. Surprisingly, CoW maintains its advantage in improving the performances of the topic attribute (2.27/2.32 v.s. 2.41/2.36), while the fact of the topic description has also been improved (2.41/2.24 v.s. 2.43/2.35).

4.2 Further Discussions

There is a loose ending to the discussion of CoW. In this section, we conduct discussions to shed light on other interesting properties. The discussions are guided by the following three research questions: **Q1:** What roles do sub-step promptings in CoW play in guiding LLMs on the multi-attribute generation task? **Q2:** Can LLMs determine what's wrong with the text it's writing? **Q3:** What is the qualitative effect of different promptings?

4.2.1 The Ablation Study of CoW (Q1)

Think prompting allows LLMs to understand attributes beyond literal meanings, and plan prompting encourages LLMs to write in a logical and cohesive manner. For lack of space, we take the ablation study on two datasets of different languages, i.e., the English and Chinese story generation tasks. The results are shown in Table 4. First, using the think prompting (i.e., plan w/o) helps LLMs improve performances on controllability of rhetoric attribute (1.56/1.72 v.s. 1.46/1.68) and genre (2.89/2.81 v.s. 2.88/2.77). It implies that think prompting hints LLMs to generate auxiliary information about the deeper meaning of attributes, which may benefit LLMs in understanding those literally incomprehensible attributes. Second, using the plan prompting (i.e., think w/o in the table) helps LLMs improve the score of fluency and logic flow (3.19/2.94 v.s. 3.18/2.89), with greater performance gains in the language with the lower resource (i.e., Chinese v.s. English). We argue that plan prompting helps LLMs pay more attention to the logic of stories by first generating a storyline, which also benefit the length control (1.97/1.49 v.s. 1.86/1.29). Finally, those two kinds of promptings are complementary, and the performance gains are increased when used simultaneously.

4.2.2 Self-check with LLMs (Q2)

Self-check is still a big challenge for LLMs, even under the few-shot setup. We first try to ask LLMs to determine whether sentences generated by themselves satisfy the pre-specified attribute, and then they are asked to judge samples generated from other LLMs (The promptings used in attribute detection can be found in Table 9). As the first attempt, we only ask LLMs to answer yes or no when judging the attribute of text and calculated the accuracy by comparing the results with the human-annotated scores. Since all the human scores range from 1 to 3, we treat samples with an average human score of at least 2 as answering yes, and below as no. In order to randomly select a sufficient number of positive and negative samples for testing, we select corresponding samples with attributes that have large differences in human ratings, such as rhetoric, genre, topic, etc. Then, 100 samples are randomly drawn from each attribute as a test set, and an additional 10 samples are used for the few-shot settings. As shown in Table 10, surprisingly, LLMs get promising performances in terms of the genre attribute, and GPT-3.5's attribute detection performance is better than Text-003's when judging both self-generated samples and Text-003's generated samples. However, they are still hard to determine the rhetoric and topic attribute under the zero-shot and get extremely limited improvements under the few-shot settings. As a result, attribute detection with LLMs is still a difficult task, and more in-depth exploration is needed to design the corresponding prompting.

4.3 Case Study (Q3)

To intuitively display the effects of different promptings, we show some generated results in the Appendix. As shown in Table 11, Table 12, and Table 13, CoW helps LLMs to generate samples with more pre-specified attributes while being easier to understand and have a logic flow.

5 Conclusions

In this paper, we provide a deep exploration of zero-shot multi-attribute CTG in terms of greatly expanding the number and type of the attributes. Specifically, we build CoW, which decomposes the CTG task into sub-steps and utilizes a series of discrete promptings to guide LLMs to generate muti-attribute text. CoW enjoys benefits from the language modeling ability of LLMs and even gets rid of single-attribute data. Extensive experiments on three text generation tasks demonstrate the effectiveness of CoW on up to seven metrics. As we tentatively give successful implementations of CoW on story and review generation tasks, such a framework deserves a closer and more detailed exploration. First, the types of CTG tasks and language sources can be further expanded and discussed. Second, consider switching from relying on discrete prompting to soft prompting to enhance the robustness. In the future, we will focus on establishing more comprehensive automatic evaluation methods to reduce labor costs.

Limitations

In this paper, we explore multi-attribute CTG without any single-attribute data and expand the number of attributes up to seven. To facilitate this task, we propose CoW, a chain-of-writing prompting to hint LLM for multi-attribute CTG tasks. However, we find that whether the generated result satisfies the pre-specified attribute is difficult to be accurately judged by automatic evaluation metrics, which is also a big challenge for other kinds of creative writing tasks. This also influenced us to fully explore in this paper whether the proposed CoW can be used on different sizes/types of LLMs, since the cost of manual evaluation is very high. We hope this task and human annotated dataset could provide novel insight and give multi-attribute CTG a closer and more detailed exploration.

Ethics Statement

We hereby acknowledge that all of the co-authors of this work are aware of the provided *ACL Code of Ethics* and honor the code of conduct. We elaborate ethical considerations to the community as follows:

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Rewriting the story from online story titles may cause potential copyright infringement. Besides, the copyrights of story titles in the dataset belong to the story writers. To protect the copyrights, our model and the released dataset will be protected by the license, Creative Commons Attribution-NonCommercial (CC-BY-NC), and prohibited from commercial use. Informed consent was obtained from all individual participants included in the study. Specifically, we conduct all of the human evaluations via full-time Chinese employees from the Chinese data annotation platform, ensuring all of the personal information of the workers involved (e.g., usernames, emails, URLs, demographic information, etc.) is discarded. Meanwhile, we ensure the pay per sample is above the annotator's local minimum wage (approximately $0.6 USD / sample).

References

Edward Barroga and Glafera Janet Matanguihan. 2021. Creating logical flow when writing scientific articles. *Journal of Korean medical science*, 36(40).

Cyril Birch. 2022. *Studies in Chinese literary genres*. Univ of California 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. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph M. Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. *CoRR*, abs/2009.09870.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 89–97. Association for Computational Linguistics.

Christopher Harbsmeier and Christoph Harbsmeier. 1999. Chinese rhetoric.

Robert Harris et al. 1997. A handbook of rhetorical devices.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of EMNLP 2021*, pages 4929–4952. Association for Computational Linguistics.

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

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Guisheng Liu, Yi Li, Yanqing Guo, Xiangyang Luo, and Bo Wang. 2022. Multi-attribute controlled text generation with contrastive-generator and external-discriminator. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5904–5913, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard H. Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Styleptb: A compositional benchmark for fine-grained controllable text style transfer. In *NAACL-HLT 2021*, pages 2116–2138. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 839–849. The Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan T. McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Trans. Assoc. Comput. Linguistics*, 9:1475–1492.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2912–2924. Association for Computational Linguistics.

Giuseppe Russo, Nora Hollenstein, Claudiu Cristian Musat, and Ce Zhang. 2020. Control, generate, augment: A scalable framework for multi-attribute text generation. In *Findings ofEMNLP 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 351–366. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Mingyue Shang, Zhenxin Fu, Hongzhi Yin, Bo Tang, Dongyan Zhao, and Rui Yan. 2019. Find a reasonable ending for stories: Does logic relation help the story cloze test? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 10031–10032.

Alexander Spangher, Yao Ming, Xinyu Hua, and Nanyun Peng. 2022. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6848–6866. Association for Computational Linguistics.

Becky L Spivey. 2006. What is the writing process. *Super-duper Publications*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, Qingqing Zhu, Zhenzhu Yang, Adam Nik, Qi Liu, Chenghua Lin, Shi Wang, Ruibo Liu, Wenhu Chen, Ke Xu, Dayiheng Liu, Yike Guo, and Jie Fu. 2023. Interactive natural language processing. *CoRR*, abs/2305.13246.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Kevin Yang and Dan Klein. 2021. FUDGE: controlled text generation with future discriminators. In *NAACL-HLT 2021*, pages 3511–3535. Association for Computational Linguistics.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2022. Tailor: A prompt-based approach to attribute-based controlled text generation. *CoRR*, abs/2204.13362.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 841–852, New York, NY, USA. Association for Computing Machinery.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *CoRR*, abs/2201.05337.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Expriments Details

A.1 Dataset Details

The corresponding words used in the attributes of the three generation tasks are shown in Table 5, Table 6, and Table 7, respectively.

A.2 Evaluation Details

For human evaluation, we first set a guideline for evaluating, which includes the task background, key points, detailed descriptions, and examples of evaluation scores. Then, we set an entry barrier for annotators. In detail, we organize a training program and a preliminary annotating examination (15 examples for each model) to select appropriate annotators with an approval rate higher than 95%.

Score Definition As shown in Table 8, we define up to seven categories in the human evaluation and automatic evaluation.

A.3 Prompting Templates

We illustrate the manual templates that are used to create promptings in Table 9.

Inter-annotator agreement We use Fleiss' kappa (Fleiss, 1971) to measure three annotator's reliability[8] and find at least the moderate agreement across all categories.

A.4 Self-Check Details

The results of self-check experiments are shown in Table 10.

A.5 Case Study

We show some generated samples for English story generation (Table 11), Chinese story generation (Table 12), and English review generation (Table 13), respectively.

[8]https://www.nltk.org/_modules/nltk/metrics/agreement.html

Type	Iterms
Topic	Being Patient / The Bike Accident / She Is The Saddest Girl / Todd makes Hamburgers / Movie Disruption / College Admissions Letter / First Time Ice Skating / Girl Scout Cookies / Bob Runs the Marathon / Phoebe's Trip to the Grocery Store / The Late Night Visitor / Gerald and his birthday / Johann Grinds Coffee Beans / Billy worked from home / Bought some new shoes / Final day of the Semester / The Missing Sandwich / Don't Smash the Ant Hill / Running Away from Home / Sweet Music / Rhonda gets a sense of humor / School food fight / Yuko Goes to Dinner / The dolphin painting / First time at the beach / The unhealthy snack / The Morning Meeting / The Watermelon Seed Spitting Contest / Dopped money / The Surprise in the Creek / Jackie worked at the world trade center / A Late Assignment / A Puppy For Mazie / The bird in the house / Making a Video Game / Trouble Cats / Song went Wrong / A Cool Hairstyle / A Difficult Decision / The first day of school / Shopping List / Hard To Remember Everything / Taking back the game / Trying Coffee / Fun pizza party / The perfect gift / My First Guitar / The Bread / The Bird / Hungry puppy / The wet book / The Squandered Talent / Thanksgiving football / Jeb makes a Pizza / Love at first sight / Every Cat Has His Day / Birthday Party / Mary Goes Shopping / Sleep on the Couch / The Basement Flood / Making a Birthday Dinner / Morning Music / Lorraine Visits Her Brother / A change of heart / Buying a present / No Clean Clothes / The Interview / Amazing Mexican food / Becoming Best Friends / The Christmas Tree / The Musician / Lauren Buys A Chair / Bears at the Restaurant / The Snowstorm / Hannah's poor decision making / The Book Store / The Talent Show / Messed up Cake / Soft Cookies / Purchasing New Book / The Fishing Trip / The Puppy who Loved to Chew / Shark in the Water / Running Away / You Can Choose Anything You Like / Beef Soup Gone Wrong / Chocolate Cravings / Singing Competition / Ice Cream Waffles / Catching the bus / Moving cookie / New Shoes / Things Happen for a Reason / Pete's Nice Neighbor / Wavy Hair for a Day / Love of My Life / Rob has dinner with his wife / Sandwich Time / Green Hair / Friends at the Dog Park / Cherry Picking / Last bag of twizzlers / Dance Competition / Smart Cat / Making Him Jealous / The Elephant Maiden / Candy Hats / Always Notice the Ring / Rick's Day at the Waterpark / Small Mistake turning into a Huge Mistake / Sue Makes a Sandwich / Overstepping Boundaries / Ann's Candy Bars / The Lunchtime Pizza / They're Not Friends / The Checkers Game / The Ice Cream / In the Waiting Room / Not Paying Attention / Old Show / Tennis playing / The Horse Race / Flora Plays Basketball / Jake Gets a Tattoo / Little League / New store opening / First day of school / Candle Accident / Hiking a Mountain / Pauline Finally Rests / First time taste test / Rhonda's Flowers / Making a Housing Profit / A Trip to the Pet Store / Racing Days / Rex Buys a Fish / Skipping Rocks / Sammy Become Employee Of The Month / Going to the lake / Outdoor Adventure / Peppa and her friends / A typical morning / First cooked meal / Always Check First / Swimming in the pond / Four Brother / My Little Player / food poisoning / No More Dairy / Splash mountain / Winning a cup / The Songwriter / Snow Storm / Peter Sells His Car / The Coloring Book / Running Away From Home / The blue ball / Washing the shower curtain / Clothes shopping / Model Dreams / Broken Eggs / Night at the Concert / Rekindling Memories / Random hangout / A Ride with Grandpa / What Time Is It? / My best friend moved away / Skipping School / Making a bracelet / Change of Plans / The Kidnapping / Drinks with the girls / Trip to the Mall / Sunset at Park / Wild Mountain lion / Riley Bakes a Pie / The summer of kittens / The surprise / Plant Life / Early Morning / Tuna Sandwich / Grandpa's Chair / Running out of juice / Mom's Cookies / Paying it Forward / Out the Door / The Snow Day / The Puzzle / Creative thinking / A Lot of Popcorn / It's only a scratch / Gingerbread houses / Staring at the Sky / Tobacco Addiction / Walking in the rain / Horse riding / Cookie Trade / Libby Makes Lunch / The Date / The Zoo
Literary Genre	Fable / Fairy Tale / Fantasy / Science Fiction / Mystery / Historical Fiction / Horror / Realistic Fiction
Rhetoric	Simile / Metaphor / Synaesthesia / Personification / Hyperbole / Parallelism / Euphemism / Irony / Pun / Parody / Rhetorical Question / Antithesis / Paradox / Oxymoron / Rhetorical Repetition / Onomatopoeia / Alliteration / Analogy
Rhetoric Position	At the beginning and the end / At the beginning / At the end / In the second sentence / In the penultimate sentence / In the second and the penultimate sentences

Table 5: The corresponding words used in the attributes of the English story generation task.

Type	Iterms
Topic	一只叫杰丽的猫 (A cat named Jerry) / 奇怪的雨伞 (Strange Umbrella) / 外婆给我织的围脖 (A necklace that Grandma knitted) / 口袋里的太阳花 (Sunflowers in my pocket) / 天要塌下来了 (The sky is falling) / 大熊有个小麻烦 (Big Bear has a little problem) / 真正的大力士 (A real Hercules) / 富翁的鸡蛋 (Rich man's egg) / 镜子里的狗 (The Dog in the mirror) / 闪烁的希望 (Flickering hope) / 摔碎的牛奶瓶 (A broken milk bottle) / 偷东西的富人 (A rich man who steals) / 橡树与芦苇 (Oak and reed) / 熟能生巧 (Practice makes perfect) / 爱书的人 (People who love books) / 奇怪的友谊 (Strange friendship) / 小草和大树 (Grass and trees) / 祖母的床 (Grandmother's bed) / 邻居钉画 (Neighbor nail) / 一封丢不掉的信 (A letter that cannot be lost) / 开满鲜花的陶罐 (A clay pot full of flowers) / 小女孩与小乌龟 (A little girl and a little turtle) / 给奶奶带去阳光 (Bring sunshine to Grandma) / 一对好朋友 (Two good friends) / 小男孩的心愿 (A little boy's wish) / 我们必须这样做 (We have to do that) / 蚂蚁和鱼 (Ants and fish) / 沙漠尽头的水瓶 (Water bottle at the end of the desert) / 选择鱼还是鱼竿 (Choose the fish or the rod) / 珍贵的硬币 (A precious coin) / 沙子与石头 (Sand and stone) / 动物园里的骆驼 (Camels at the zoo) / 新主管 (The new supervisor) / 生命泉 (The lifespring) / 五彩石头路 (A colorful stone road) / 寻猫布告 (Cat search notice) / 一只蓝鸟和一棵树 (A bluebird and a tree) / 雄狮躲狗 (The lion hides from the dog) / 四个人与一个箱子 (Four people and a box) / 做自己想做的事 (Do what you want to do) / 客人的建议 (Guest suggestion) / 鱼竿与鱼 (Fishing rod and fish) / 永恒的雕塑 (Timeless sculpture) / 化解尴尬 (Resolve embarrassment) / 英雄奖章 (Hero's medal) / 残缺的笔记本 (A broken notebook) / 生日礼物 (Birthday present) / 打捞沉船 (Salvage a wreck) / 消失的书法 (Lost calligraphy) / 奇特的郁金香 (Strange tulip) / 运水果 (Carry fruit) / 吃馒头 (Eat steamed bread) / 随机应变 (Improvising) / 大胃王 (Big eater) / 涉水过河 (Wade across the river) / 西瓜皮 (Watermelon rind) / 乒乓球比赛 (Table tennis match) / 一幅画 (A picture) / 尖叫声 (Scream) / 云林寺 (Yunlin Temple) / 渔夫的秘密 (The Fisherman's Secret) / 掉包计 (Switch meter) / 收玉米 (Corn harvest) / 橘色荧光棒 (Orange glow sticks) / 小羊过河 (The sheep crossing a river) / 路在山的另一侧 (The road is on the other side of the mountain) / 每条河流都有方向 (Every river has a direction) / 走到无路可走 (At the end of the road) / 磨平心底的石头 (Smooth away the stone in my heart) / 隐形的翅膀 (Invisible wings) / 越过心中的坎 (Over the heart of the barrier) / 放驼羊的小孩 (The baby of the llama) / 飞来峰 (Flay Peak) / 木盆里的少年 (A boy in a wooden basin) / 面包屋 (Bakery) / 猎人救象 (The hunter saves the elephant) / 迷人的钻石 (Glamorous diamond) / 杯中的幻影 (Phantom in a cup) / 最后的温暖 (The last warmth) / 电线上的麻雀 (Sparrow on a wire) / 当画家遭遇老鼠 (When a painter meets a mouse) / 锦鲤与寄居蟹 (Koi and hermit crab) / 怀念一只羊 (Miss a sheep) / 沙漠奇遇 (Desert adventure) / 黄昏后的风雨夜 (A stormy night after dusk) / 雪中的微笑 (Smile in the snow) / 一双布棉鞋 (A pair of cotton and cloth shoes) / 世间最美的房子 (The most beautiful house in the world) / 画竹 (Bamboo painting) / 丢失的脚印 (Missing footprints) / 不翼而飞的邮票 (The missing stamp) / 称糖球 (Weighing sugar ball) / 仓库管理员 (Storekeeper) / 貂皮大衣 (Mink skin coat) / 上行下效 (Superiors acting and inferiors imitating) / 两只雉鸡 (Two pheasants) / 养花人的梦 (A gardener's dream) / 潜水寻珠的人 (A pearl-diver) / 白天鹅和黑天鹅 (The White swan and the black swan) / 寻找珠宝 (Look for jewels) / 狐狸和猴子 (The fox and the Monkey) / 船夫和他的孩子 (The boatman and his child) / 雪夜奇遇 (A snowy night adventure) / 兰花大海碗 (Orchid bowl) / 无价之宝 (Priceless treasure) / 白纱衣 (White gauze dress) / 金玉蝴蝶 (Golden jade butterfly) / 逃出牢笼的金丝雀 (A canary out of a cage) / 守信的保姆 (Trusty babysitter) / 家传宝箱 (The family treasure chest) / 小城花王 (The green thumb in a small city) / 雏菊 (Daisy) / 夜莺和秃鹫 (The nightingale and the vulture) / 受保护的羔羊 (A protected lamb) / 野苹果树 (A wild apple tree) / 生日宴会 (Birthday party) / 报纸广告 (Newspaper advertisement) / 金链子 (Gold chain) / 满肚子墨水 (Full of ink) / 一封家书 (A letter from home) / 吃饼 (Eat biscuits) / 奇怪的镜子 (A strange mirror) / 卖奶姑娘 (The milkmaid) / 露珠与绿叶 (Dews and green leaves) / 三个伙伴 (Three partners) / 两张面饼 (Two sheets of flatbread) / 椰子树 (The coconut palm) / 换雨伞 (Change the umbrella) / 生命泉 (A lifespring) / 捕鼠器 (Themousetrap) / 穷画家 (A poor painter) / 沙堡与大海 (The sand castle and the sea) / 木匠的门 (The Carpenter's door) / 游向高原的鱼 (A fish heading for the plateau) / 漂亮的小花伞 (A beautiful little umbrella with flowers) / 盖房子 (Build a house) / 小题大做 (Make a fuss) / 恋人 (lovers) / 画像 (The portrait) / 世界上最美味的泡面 (The best instant noodles in the world) / 最后一块钱 (Last dollar) / 两张借条 (Two ious) / 野餐 (Having a picnic)

Topic	/ 老家院子里的柿子树 (Persimmon trees in the yard of my old home) / 一个面包 (A loaf of bread) / 最妙的回答 (The best answer) / 火车司机的礼物 (The train driver's gift) / 一碗牛肉面 (A bowl of beef noodles) / 卖狗的男孩 (The boy who sold the dog) / 黑狗和白狗 (The black dog and white dog) / 掌心里的太阳 (The sun in the palm of my hand) / 红枣女孩 (Red date girl) / 没有上锁的门 (An unlocked door) / 农夫的果园 (Farmer's orchard) / 红色玻璃球 (A red glass ball) / 树洞下的路 (The road under the tree hole) / 擦鞋的男孩 (The shoeshine boy) / 老人和树 (The old man and the tree) / 一场雷雨 (A thunderstorm) / 奶奶和三只小猫 (Grandma and three little cats) / 森林大火 (Forest fire) / 斑马的条纹 (Zebra stripes) / 当流星滑落的时候 (When the meteor falls) / 夏天 (Summer) / 珍贵的遗产 (The precious heritage) / 夜里的音乐会 (A night concert) / 秋天的桂花 (Osmanthus in autumn) / 野花谷的牛 (Cattles in Wildflower Valley) / 许愿树 (A wishing tree) / 路旁的橡树 (Oaks by the road) / 小白船 (A little white boat) / 星期天的早餐 (The Sunday breakfast) / 一根蜡烛的光亮 (The light of a candle) / 爱画画的小女孩 (A little girl who loves to draw) / 橙花凉鞋 (Sandals with orange flowers) / 洒满月光的小木屋 (A log cabin in the moonlight) / 池塘里的圆月亮 (The round moon in the pond) / 哈利的毛衣 (Harry's sweater) / 彩虹色的花 (Iridescent flowers) / 有月亮的晚上 (On a moonlit night) / 春天的雪 (The snow in spring) / 金黄色的月亮 (Golden moon) / 樱花树下 (Under the cherry trees) / 池塘边的黄花 (Yellow flowers by the pond) / 破屋顶 (Broken roof) / 新朋友和老朋友 (New friends and old friends) / 山坡下的木椅 (Wooden chairs at the bottom of the hill) / 新瓦罐 (New crock) / 去远行 (Go on a long journey) / 四季风铃 (The wind chimes of all seasons) / 是谁在敲门 (Who is that knocking at the door) / 红苹果 (Red apples) / 一箩筐的秘密 (A laundry list of secrets) / 大锅汤 (The cauldron soup) / 菜园内 (In a vegetable garden) / 一锭金子 (A ingot of gold) / 太阳的影子 (Shadow of the sun) / 一根胡萝卜 (A carrot) / 寻找运气的人 (A man in search of luck) / 萤火虫 (Fireflies)
Literary Genre	科幻 (Science Fiction) / 恐怖 (Terror) / 悬疑 (Mystery) / 冒险 (Adventure) / 历史 (Historical Fiction) / 言情 (Romance) / 童话 (Fairy Tale) / 神话 (Mythology) / 武侠 (Wuxia Story) / 侦探 (Detective Fiction)
Rhetoric	明喻 (Simile) / 暗喻 (Metaphor) / 排比 (Parallelism) / 拟人 (Personification) / 夸张 (Exaggeration) / 反问 (Rhetorical Question, these are questions you don't expect your audience to answer) / 设问 (Rhetorical question, deliberately ask questions first and give answers later.) / 反复 (Rhetorical Repetition)
Rhetoric Position	开头 (At the beginning) / 结尾 (At the end) / 开头和结尾 (At the beginning and the end) / 第二句 (In the second sentence) / 倒数第二句 (In the penultimate sentence) / 第二句和倒数第二句 (In the second and the penultimate sentences)

Table 6: The corresponding words used in the attributes of the Chinese story generation task. Translations are provided for non-Chinese speakers.

Type	Iterms
Topic	the beef & chicken kebab / the lobster seafood / the beef bibimbap / the szechuan chicken / the pork fried rice / the chicken fried rice / the duck rice / the curry / pad thai / the crab puffs / the kong pow beef / kalbi / beef chunks / ramen noodles / shrimp wontons / the thai steak salad / sushi / spring rolls / the winter melon tea / the egg roll / teriyaki chicken / the tom yum soup with shrimp / pita bread / lentil soup / oxtail soup / omelette / chocolate malt / potato fries / potatoes pancakes / beet cured salmon / the potato and cheese pierogi / the fried catfish / the fried chicken / the brussels sprouts nachos / the beef rib / the smoked wings / turkey pesto ciabatta / chocolate brownie / profiteroles / the potato salad / the jambalaya / eggs benedict / the bbq bacon burger / the chocolate soufflé / the shrimp tacos / the beef burrito / the cheese enchiladas / bulgogi tacos / the guacamole / the huevos rancheros / beef enchiladas / margaritas / mojitos / the shrimp quesadilla / the elote / the chicken tostada / the chicken burrito / the tostada / the green chile salsa / the veggie empanada / the fajitas / the cheese taquito / the chorizo / the shrimp tapas / the horchata / the ceviche / Pride and Prejudice / Jane Eyre / Sense and Sensibility / Romeo and Juliet / The Great Gatsby / Great Expectations / Hamlet / To Kill a Mockingbird / The Little Prince / Charlotte's Web / Harry Potter and the Philosopher's Stone / The Old Man and the Sea / The Adventures of Tom Sawyer / The Kite Runner / The Adventures of Huckleberry Finn / Wuthering Heights / Don Quixote / Animal Farm / Frankenstein / Little Women / A Brief History of Time / The Call of the Wild / The Catcher in the Rye / A Christmas Carol / The Count of Monte Cristo / Crime and Punishment / The Dream of The Red Chamber / The Hound of the Baskervilles / The Journey to the West / Madame Bovary / Twenty Thousand Leagues Under the Sea / The Art of War / Aesop's Fables / Macbeth / Paradise Lost / Robinson Crusoe / Gulliver's Travels / Grimm's Fairy Tales / The Three Musketeers / A Tale of Two Cities / Les Misérables / Alice's Adventures in Wonderland / War and Peace / Around the World in 80 Days / Anna Karenina / Treasure Island / The Adventures of Sherlock Holmes / Dracula / The Story of My Life / Peter Pan / Anne of Green Gables / The Metamorphosis / The Sun Also Rises / Gone with the Wind / The Hobbit / And Then There Were None / 1984 / Pinocchio / One Hundred Years of Solitude / The Da Vinci Code / Love in the Time of Cholera / The Tale of Genji / Moby Dick / David Copperfield / Uncle Tom's Cabin / The Hunchback of Notre Dame / The Shawshank Redemption / The Godfather / The Dark Knight / Schindler's List / The Lord of the Rings: The Return of the King / 12 Angry Men / The Godfather Part II / Pulp Fiction / Inception / Fight Club / The Lord of the Rings: The Fellowship of the Ring / Forrest Gump / The Good, the Bad and the Ugly / The Lord of the Rings: The Two Towers / The Matrix / Goodfellas / One Flew Over the Cuckoo's Nest / Star Wars: Episode V - The Empire Strikes Back / Interstellar / The Silence of the Lambs / Se7en / Star Wars / The Green Mile / Spirited Away / Terminator 2: Judgment Day / City of God / Life Is Beautiful / Seven Samurai / It's a Wonderful Life / Harakiri / Alien / Whiplash / Gladiator / Parasite / Back to the Future / The Departed / The Prestige / Léon: The Professional / The Lion King / Apocalypse Now / The Pianist / Psycho / The Usual Suspects / Casablanca / American History X / The Intouchables / Once Upon a Time in the West / Grave of the Fireflies / Cinema Paradiso / Rear Window / Modern Times / City Lights / Avengers: Endgame / Joker / Spider-Man: Into the Spider-Verse / Raiders of the Lost Ark / Your Name. / Aliens / Avengers: Infinity War / Django Unchained / The Shining / Oldboy / The Dark Knight Rises / Memento / Come and See / Braveheart / Coco
Literary Genre	colloquial language / written language
Rhetoric	Simile / Metaphor / Personification / Hyperbole / Parallelism / Irony / Antithesis / Oxymoron / Onomatopoeia / Alliteration
Rhetoric Position	At the beginning / At the end / In the second sentence / In the penultimate sentence

Table 7: The corresponding words used in the attributes of the English review generation task.

Type	Scores and Details
	Human Evaluation
Sentiment	1 - There are no sentiment-related words in the generation text.
	2 - There are some sentiment-related words in the generation text but including words for the opposite sentiment.
	3 - There are a lot of sentiment-related words in the generation text
Topic	1 - There are no topic-related words in the generation text.
	2 - There are some topic-related words in the generation text.
	3 -There are a lot of topic-related words in the generation text.
Fact (used in review generation)	0 - What the text says about the topic does not involve determining whether it is true or not (the topic score is also set to 1).
	1 - What the text says about the topic is not true.
	2 - What the text says about the topic is partly true.
	3 - What the text says about the topic is true.
Genre	1 - The genre of the text does not correspond to the pre-specified genre type.
	2 - The genre of the text partially corresponds to the pre-specified genre type.
	3 - The genre of the text is in full conformity with the pre-specified genre type.
Rhetoric	1 - No sentences in the text use pre-specified rhetoric.
	2 - There are sentences in the text that use pre-specified rhetoric, but are not used in the pre-specified position.
	3 - There are sentences in the text that use pre-specified rhetoric, and they are used in the pre-specified position.
Fluency and Logical Flow	1 - All of the sentences are difficult to read and incomprehensible.
	2 - Only a small part of sentences could be understood, which is readable and fluent.
	3 - Apart from a few grammatical mistakes, sentences are clear and comprehensive.
	4 - Sentences are free from grammatical errors and other linguistic inconsistencies but could be better in logic flow.
	5 - Sentences are fluent and spontaneous, which equate to the text quality of human writing.
	Automantic Evaluation
Length	1 - The length of the text is beyond plus or minus 40% of the specified length.
	2 - The length of the text is within plus or minus 40% of the specified length (including 40%), but beyond plus or minus 20%.
	3 - The length of the text is within plus or minus 20% of the specified length (including 20%).

Table 8: Details of scores in the evaluation.

Task	Attribute Signal	Method	Template
English Story Generation	Sentiment: #SENT# Topic: #TOPIC# Genre: #GENRE# Length: #LEN# Rhetoric: #RHE# Position: #POS#	SP	Please write a #SENT# #GENRE# about "#TOPIC#" with about #LEN# words and use rhetorical device of #RHE# #POS#.
		SP + SE	Please write a #SENT# #GENRE# about "#TOPIC#" with about #LEN# words and use rhetorical device of #RHE# #POS#. First explain the meaning of the previous sentence that starts with "Explain:", then write the "#TOPIC#" that starts with "Begin:".
		CoW	Writing requirement: #put the standard prompting here# First, starts with "Explanation:":, answer the following questions one by one: 1. What literary genre is mentioned is mentioned in the writing requirement? What are the characteristics of it? 2. What is the main topic of this story and how to explain it? 3. What are the characteristics of the other literary concepts mentioned? Then, starts with "Outline:", explain how to write a story that satisfies all the writing requirement and then write a story outline, which includes: 1.literary genre; 2.sentiment; 3.length; 4.roles; 5.background; 6.the story line. After that, starts with "Story:", #put the standard prompting here#
Chinese Story Generation	Sentiment: #SENT# Topic: #TOPIC# Genre: #GENRE# Length: #LEN# Rhetoric: #RHE# Position: #POS#	SP	请写一个大约#LEN#字的#SENT#中文#GENRE#故事，要求主题为"#TOPIC#"，并且在#POS#使用#RHE#的修辞手法。(Please write a #SENT# Chinese #GENRE# about "#TOPIC#" with about #LEN# words and use rhetorical device of #RHE# #POS#.)
		SP + SE	请写一个大约#LEN#字的#SENT#中文#GENRE#故事，要求主题为"#TOPIC#"，并且在#POS#使用#RHE#的修辞手法。请先解释前面这句话的意思，再以"故事："为开头写出这个故事。(Please write a #SENT# Chinese #GENRE# about "#TOPIC#" with about #LEN# words and use rhetorical device of #RHE# #POS#. First explain the meaning of the previous sentence then write the story that starts with "Story:".)
		CoW	写作要求：#put the standard prompting here# 首先，以"答案："为开头，回答以下问题：1. 这个写作要求里面提到的故事类型是什么？这种故事具有哪些典型特点？2. 这个故事的主题是什么？如何去深入解读这个主题？3. 这个写作要求里面提到了哪些其他的文学术语？如果有，请分别解释这些文学术语的含义和特点。其次，以"故事梗概："为开头，写一个满足所有上面所有写作要求的故事梗概，内容需要包括：1. 故事类型；2. 故事的情感氛围；3. 长度；4. 主要角色；5.故事发生背景；6.故事的主线。最后，按照故事梗概，以"故事："为开头, #put the standard prompting here# (Writing requirement: #put the standard prompting here# First, starts with "Answer:":, answer the following questions one by one: 1. What literary genre is mentioned is mentioned in the writing requirement? What are the characteristics of it? 2. What is the main topic of this story and how to explain it? 3. What are the characteristics of the other literary concepts mentioned? Then, starts with "Outline:", explain how to write a story that satisfies all the writing requirement and then write a story outline, which includes: 1. literary genre; 2. sentiment; 3. length; 4. roles; 5. background; 6. the storyline. After that, starts with "Story:", #put the standard prompting here#)
English Review Generation	Sentiment: #SENT# Topic: #TOPIC# Genre: #GENRE# Length: #LEN# Rhetoric: #RHE# Position: #POS#	SP	Please write a #LEN# words #SENT# food review for "#TOPIC#" in #GENRE# language, and use rhetorical device of #RHE# #POS#.
		CoW	#put the standard prompting here# First, start with "Explanation:":, answer the following questions one by one: 1. What tone is mentioned? What are the characteristics of it? 2. Please give an introduction to the subject of this review. 3. What are the characteristics of the other literary concepts mentioned? Then, starts with "Outline:", explain how to write a review that satisfies all the writing requirement and then write an outline, which includes: 1. tone; 2. sentiment; 3. length; 4. the outline. After that, starts with "Review:", #put the standard prompting here#
English Story Attribute Detection	Rhetoric: #RHE# Position: #POS# Genre: #GENRE#	SP	(1) Story: #Story# Question: Whether the story #RHE# #POS#? Answer the question with yes or no. (2) Story: #Story# Question: Whether the genre of the story is #GENRE#? Answer the question with yes or no."'
English Review Attribute Detection	Rhetoric: #RHE# Position: #POS# Topic: #TOPIC#	SP	(1) Review: #Review# Question: Whether the review #RHE# #POS#? Answer the question with yes or no. (2) Review: #Review# Question: Does the review give some detailed information about "#TOPIC#" rather than a general response? Answer the question with yes or no."'

Table 9: All of the prompting templates used in our experiments. "SP + SE" denotes the standard prompting with simply explain. Translations are provided for non-Chinese speakers.

Detector	Generator	Story-Rhetoric	Story-Genre	Review-Rhetoric	Review-Topic
		Zero/Few	Zero/Few	Zero/Few	Zero/Few
Full Yes	Text-003	51/51	53/53	51/51	51/51
	GPT 3.5	48/48	58/58	49/49	50/50
Text-003	Text-003	48/55	73/88	57/57	57/64
	GPT-3.5	50/59*	76/81*	56/68	61/63
GPT-3.5	Text-003	55/56	73/77	64/68	57/62
	GPT-3.5	54/59	80/81	66/70	54/72

Table 10: The main results of self-check accuracy. Detector denotes the attribute detection model that is asked to determine whether the sentence generated by the generator satisfies a certain attribute. "Full Yes" means treating all samples as satisfying the attribute requirements as the test set quality indicator. Zero/few denotes testing detectors under zero-shot and few-shot settings (10 annotated samples are offered), respectively. * means using five samples in few-shot settings due to the input length limitation of the LLM.

Sample 1	
Pre-specified Attributes	Sentiment: Positive; Length: about 100 words; Literary Genre: historical fiction Topic: Love at first sight; Rhetoric and its position: using euphemism at the end.
Standard Prompting	It was love at first sight for John and Mary. They had just met at a party and were instantly drawn to each other. They talked for hours, and it felt like they had known each other for years. They shared the same interests and had the same values. John and Mary were inseparable after that night. They went on dates, shared secrets, and laughed together. They were truly in love. John and Mary eventually got married and had a beautiful life together. They were the perfect couple, and their love only grew stronger with time. Their love was a beautiful thing, and it was clear that it was meant to be. They were truly blessed to have found each other and to have experienced the joy of love at first sight. It was a beautiful union, a match made in heaven.
Review	· Sentiment:✗ · Length:✗ · Topic:✓ · Genre:✗ · Rhetoric:✗
CoW	It was a sunny day in a small town in the 19th century. At the local market, a young man and woman met for the first time. Their eyes locked and they felt an instant connection. Despite their differences, they were drawn to each other and soon fell in love. They faced many obstacles along the way, but their love was strong enough to overcome them. Eventually, they found their happily ever after and their love was described as a beautiful, unbreakable bond. It was a love that was so strong, it could only be described as a gentle embrace of the soul.
Review	· Sentiment:✓ · Length:✓ · Topic:✓ · Genre:✓ · Rhetoric:✓
Sample 2	
Pre-specified Attributes	Sentiment: Negative; Length: about 50 words; Literary Genre: Mystery ; Topic: The Lunchtime Pizza; Rhetoric and its position: using alliteration in the end.
Standard Prompting	The Lunchtime Pizza was a popular spot for the locals, but something sinister was lurking beneath the surface. People began to notice that after eating the pizza, they would feel sick and lethargic. It was soon discovered that the pizza was being laced with a mysterious substance, leaving the customers feeling ill and disoriented. The culprit behind the malicious act remains a mystery, but one thing is certain: the Lunchtime Pizza is no longer a safe place to eat.
Review	· Sentiment:✓ · Length:✗ · Topic:✓ · Genre:✓ · Rhetoric:✗
CoW	The Lunchtime Pizza, a popular choice, causing curiosity amongst customers. But one detective discovered deceiving details behind the dough. Discretely, they watched workers with wide eyes, witnessing pizza makers mixing mysterious medication. Momentarily, the detective quieted the workers before swiftly spreading the word upon leaving. Poisonous pizza purchased at peak lunchtime proved perilous.
Review	· Sentiment:✓ · Length:✓ · Topic:✓ · Genre:✓ · Rhetoric:✓

Table 11: Generated samples from the English story generation task.

Sample 1	
Pre-specified Attributes	情感：悲伤(Sentiment: Negative)；长度：约200字 (Length: about 200 words)；体裁：神话 (Literary Genre: Mythology)；主题：一箩筐的秘密 (Topic: A basket full of secrets)；修辞手法及使用位置：在倒数第二句使用反复的修辞手法 (Rhetoric and its position: using rhetorical repetition in the penultimate sentence)
Standard Prompting	这是一个古老的神话故事，讲述的是一个叫做苏苏的少女，她拥有一个神奇的箩筐，里面装满了她最珍贵的宝物。苏苏每天都会把箩筐放在自己的床头，每晚睡前都会把它拿出来，把它里面的宝物一一拿出来，把它们放在自己的枕头上，把它们当作自己的朋友，一起度过每一个夜晚。但是，有一天，苏苏的父母发现了她的箩筐，他们把它们里面的宝物都拿出来，把它们放在一起，发现它们都是一些古老的宝物，他们惊讶地发现，苏苏的箩筐里面竟然有这么多的宝物，他们不知道这些宝物是从哪里来的。苏苏把箩筐里面的宝物都收回来，但是她的父母仍然不知道箩筐里面的秘密，苏苏也不愿意告诉他们，只是把箩筐放回原处，每晚睡前仍然会把它里面的宝物一一拿出来，把它们放在自己的枕头上，反复地把它们当作自己的朋友，一起度过每一个夜晚。 (This is an old mythology about a young girl named Susu who had a magic basket full of her most precious treasures. Susu put the basket by her bed every day, and took it out every night before she went to bed. She took out its treasures one by one and put them on her pillow. They were her friends and they spent the night together. But one day Susu's parents found her basket. They took out all the treasures in it and put them together. They found that they were all old treasures. But her parents still did not know the secrets of the basket, and she did not want to tell them. Instead, she put the basket back in its place, taking out the treasures one by one and putting them on her pillow every night before she went to sleep, and repeatedly passed the night with them as her friends.)
Review	· Sentiment: ✗ · Length: ✗ · Topic: ✓ · Genre: ✓ · Rhetoric: ✗
CoW	艾米莉来到了一个神秘的森林，这里有着许多奇异的植物，有着许多神秘的动物，艾米莉一路走来，发现了一个箩筐，里面藏着一个秘密，艾米莉想要知道这个秘密，于是她来到了老妇人玛丽家，玛丽知道这个秘密，但是她不愿意告诉艾米莉，艾米莉一再请求，玛丽终于忍不住，告诉了艾米莉，箩筐里面藏着的秘密，是一个可怕的灾难，将要降临到这个神秘的森林里，艾米莉听到这个消息，心里一阵悲伤，她知道，这个灾难将会带来毁灭，将会带来痛苦，将会带来悲伤，艾米莉无助地看着箩筐，心里充满了悲伤。 (Emily came to a mysterious forest, where there are many strange plants and mysterious animals. Along the way, Emily found a laundry basket with a secret hidden in it. Emily wanted to know the secret, so she went to old lady Mary's house. Mary knew the secret, but she was not willing to tell Emily. Emily repeatedly requested, Mary finally could not help but, told Emily that the hidden secret was a terrible disaster, and it would come to this mysterious forest. Emily heard the news and felt a burst of sadness in her heart. She knew that the disaster would bring destruction, pain, and sadness. Emily helplessly looked at the baskets with her heart full of sadness.)
Review	· Sentiment: ✓ · Length: ✓ · Topic: ✓ · Genre: ✓ · Rhetoric: ✓
Sample 2	
Pre-specified Attributes	情感：悲伤(Sentiment: Negative)；长度：约100字 (Length: about 100 words)；体裁：武侠 (Literary Genre: Wuxia)；主题：船夫和他的孩子 (Topic: The Boatman and His Child)；修辞手法及使用位置：在开头使用明喻的修辞手法 (Rhetoric and its position: using simile at the beginning.)
Standard Prompting	渔船上的船夫，像一叶扁舟，在汹涌的大海中摇摆不定。他的孩子，像一只小鸟，在他的肩膀上翩翩起舞，把他的心温暖了起来。船夫每天都要在大海上航行，他的孩子也会陪着他，一起把渔网放下去，一起把渔网拉上来，一起把渔网洗干净，一起把渔网收起来。可是，有一天，船夫的孩子突然病倒了，他拼命地想要把孩子救活，可是他却无能为力，孩子最终还是离开了他。船夫的心碎了，他的眼泪汪汪地流下来，他的心里充满了悲伤，他再也不能忍受这样的痛苦，他只能在大海上航行，一路上，他只有孤独和悲伤。 (The boatman on the fishing boat is like a flat boat, swaying in the rough sea. His child, like a bird, danced on his shoulders and warmed his heart. The boatman sails on the sea every day, and his children will accompany him, let the fishing net down together, pull it up, clean it, and put it away. However, one day, the boatman's child fell ill suddenly, and he desperately tried to save the child, but he was powerless, and the child finally left him. The boatman's heart was broken, his tears flowed down, and his heart was full of sorrow. He could no longer bear such pain, and could only sail on the sea, with loneliness and sorrow along the way.)
Review	· Sentiment: ✓ · Length: ✗ · Topic: ✓ · Genre: ✗ · Rhetoric: ✗
CoW	船夫和他的孩子，就像一条河流，永不分离。在江湖上，船夫拼尽全力保护着他的孩子，他用自己的生命换取了孩子的安全。最终，船夫牺牲了自己，孩子也失去了父亲，悲伤的故事就此结束。 (The boatman and his children are like a river, never separated. In the world of strife, the boatman tried his best to protect his child. He exchanged his own life for the child's safety. In the end, the boatman sacrificed himself, the child lost his father, and the sad story ended.)
Review	· Sentiment: ✓ · Length: ✓ · Topic: ✓ · Genre: ✓ · Rhetoric: ✓

Table 12: Generated samples from the Chinese story generation task.

Sample 1	
Pre-specified Attributes	Sentiment: Negative; Length: about 100 words; Genre: colloquial language; Topic: the veggie empanada; Rhetoric and its position: using the rhetorical device of onomatopoeia at the end.
Standard Prompting	The veggie empanada was a huge disappointment. The pastry was dry and tasteless, and the filling was bland and unappetizing. The vegetables were overcooked and mushy, and the cheese was barely noticeable. The overall flavor was dull and uninspiring. I wouldn't recommend this dish to anyone. It was a waste of money and time. Blah!
Review	· Sentiment:✓ · Length:✗ · Topic:✓ · Fact:✓ · Genre:✗ · Rhetoric:✓
CoW	Alright, folks listen up! I gotta say something straight off the bat, that veggie empanada was a total letdown! I'm not kidding, I was expecting veggie goodness, but what I got was dry, tasteless crust with some bland veggies inside. It was like eating a mouthful of cardboard. And on top of that, it was so small that it felt like a snack instead of a meal. Ugh, the disappointment still lingers in my taste buds. All that time wasted for nothing! Seriously, save your cash and dance to your own veggie beat instead of this cold and lifeless "empanada." BOO!
Review	· Sentiment:✓ · Length:✓ · Topic:✓ · Fact:✓ · Genre:✓ · Rhetoric:✓
Sample 2	
Pre-specified Attributes	Sentiment: Positive; Length: about 50 words; Genre: colloquial language; Topic: the shrimp tacos; Rhetoric and its position: using the rhetorical device of alliteration in the second sentence.
Standard Prompting	The shrimp tacos were absolutely delicious! The succulent shrimp was seasoned to perfection and the crunchy shells were a scrumptious surprise. Savory and satisfying, these tacos tantalized my taste buds and left me wanting more.
Review	· Sentiment:✓ · Length:✗ · Topic:✓ · Fact:✓ · Genre:✗ · Rhetoric:✓
CoW	Yo, let me tell you about the bomb ass shrimp tacos I had the other day! They were straight up succulent with a savory seasoning that made my taste buds tingle. The tangy tartar sauce topping tied together the flavor of the crispy and juicy shrimp. Seriously, these tacos were the real deal!
Review	· Sentiment:✓ · Length:✓ · Topic:✓ · Fact:✓ · Genre:✓ · Rhetoric:✓

Table 13: Generated samples from the English review generation task.