

# Classic4Children: Adapting Chinese Literary Classics for Children with Large Language Model

Jiali Chen<sup>1,2</sup>, Xusen Hei<sup>1,2</sup>, Yuqi Xue<sup>1,2</sup>, Zihan Wu<sup>1,2</sup>, Jiayuan Xie<sup>3</sup>, Yi Cai<sup>1,2\*</sup>

<sup>1</sup>Key Laboratory of Big Data and Intelligent Robot

(South China University of Technology) Ministry of Education,

<sup>2</sup>School of Software Engineering, South China University of Technology,

<sup>3</sup>The Hong Kong Polytechnic University

\*Correspondence: ycai@scut.edu.cn

## Abstract

Chinese literary classics hold significant cultural and educational value, offering deep insights into morality, history, and human nature. These works often include classical Chinese and complex narratives, making them difficult for children to read. To bridge this gap, we introduce a child-friendly literary adaptation (CLA) task to adapt the Chinese literary classic into engaging and accessible text for children. However, recent large language models (LLMs) overlook children's reading preferences (*i.e.*, vivid character portrayals, concise narrative structures, and appropriate readability), which poses challenges in CLA. In this paper, we propose a method called InstructChild, which augments the LLM with these preferences for adaptation. Specifically, we first obtain the characters' personalities and narrative structure as additional information for fine-grained instruction tuning. Then, we devise a readability metric as the reward to align the LLM with the children's reading level. Finally, a lookahead decoding strategy is applied to improve the readability of the generated text during inference. To support the evaluation of CLA task, we construct the Classic4Children dataset, which comprises both the original and child-friendly versions of the Four Great Classical Novels of Chinese literature. Experimental results show that our InstructChild significantly improves automatic and human evaluation performance.

## 1 Introduction

The Chinese literary classics portray iconic characters and their stories to convey significant cultural and educational value (Hsia, 2016; Legge, 2022). These works are not only a vital part of China's rich literary heritage but also play a crucial role in shaping moral values and cultural understanding, especially for young learners. Given their importance, these classics are a fundamental part of the Chinese educational curriculum, and many children are required to engage with these texts during their



Original Text

那妖道：“你是哪里和尚？从哪里来？到哪里去？快快说明！”三藏道：“我本是唐朝僧人，奉大唐皇帝敕命，前往西方访取经，经过贵山，特来塔下谒圣，不期惊动威严，望乞恕罪。待往西方取得经回东土，永注高名也。”那妖闻言，呵呵大笑道：“我说是上邦人物，果然是你。正要吃你哩，却来的甚好！甚好！不然，却不错过了？”...

The demon asked, "Where are you from, monk? Where are you headed? Speak quickly!" Tripitaka replied, "I am a monk from the Tang Dynasty, sent by the Emperor of the Tang Empire on a mission to the West to seek sacred scriptures. Passing through your esteemed mountain, I have come to pay my respects, though I did not intend to disturb you. I beg your forgiveness. Once I have obtained the scriptures and returned to the East, your name will be forever honored." Upon hearing this, the demon laughed heartily and said, "I thought it must be someone important, and indeed, it is you. I was just planning to eat you, and you've come at the perfect time! Otherwise, I might have missed my chance!" ...



GPT-4o

那妖道：“你是哪里来的和尚？从哪里来，要到哪里去？快说清楚！”三藏答道：“我是唐朝的和尚，奉唐朝皇帝的命令，要去西天取经。路过你们这里，特意来拜见圣人，没想到打扰了您，请原谅。等我取得真经，回到唐朝，一定会把您的名声传扬出去。”妖怪听完哈哈大笑，说：“我还以为是什么厉害的人物，原来就是你！正好，我正想要吃你呢！你自己送上门来，真是太好了！否则还差点让你溜走了！”...

The demon said, "Where are you from, monk? Where are you from and where are you going? Tell me clearly!" Tripitaka replied, "I am a monk from the Tang Dynasty. I am on a journey to the West to obtain Buddhist scriptures under the order of the Tang Dynasty emperor. I passed by your place and came to see the saint. I didn't expect to disturb you. Please forgive me. When I obtain the true scriptures and return to the Tang Dynasty, I will definitely spread your reputation." The demon laughed and said, "I thought you were some powerful person, but it turned out to be you! It just so happens that I wanted to eat you! It's great that you came to my door on your own! Otherwise, I would have let you slip away!" ...



Human Writer

那个老妖恶狠狠地问道：“和尚，你从哪里来，要到哪里去？”三藏战战兢兢地回答：“我是大唐的僧人，前往西方拜佛取经，路过此地，没想到惊动了大王，还请恕罪。”老妖狰狞地哈哈大笑道：“正好我要吃你，你自己送上门来了！”...

The old demon snarled, "Monk, where do you come from, and where are you going?" Trembling with fear, Tripitaka replied, "I am a monk from the Tang Dynasty, journeying to the West to seek the Buddhist scriptures. I happened to pass through this place and didn't intend to disturb you, my lord. Please forgive me." The demon let out a sinister laugh and said, "Perfect timing! I was just thinking of eating you, and here you are, delivering yourself to me!" ...

Figure 1: The sample from our Classic4Children dataset. We use a one-shot human-written example in Table 10 to query GPT-4o. The green text emphasizes the character's personality traits.

schooling years (Luo, 2019). However, these works are frequently written in classical Chinese and involve complex plots and themes, presenting significant challenges for children to read. Traditionally, many writers make much effort to manually transform these complex literary classics into child-friendly versions (Müller, 2013; Hui, 2024), which is a time-consuming and labor-intensive process. Therefore, we propose the child-friendly literary adaptation (CLA) task, which aims to automatically make content accessible and engaging for children.

Promisingly, the recent advanced large lan-

language models (LLMs) have performed impressively across various natural language processing tasks, including text style transfer (Reif et al., 2022; Liu et al., 2024a) and text simplification (Valentini et al., 2023; Kew et al., 2023). Nevertheless, such methods simply modify specific stylistic elements (*e.g.*, sentiment, formality, author-style and lexicon) within sentences, which overlook the importance of children’s reading preferences, leading to poor performance in the CLA task. Moreover, it is well known that LLMs can capture language patterns from provided examples through in-context learning (Dong et al., 2022). We utilize a carefully crafted prompt with a one-shot human-written example to query GPT-4o (OpenAI, 2023), exploring whether children’s reading preferences can be captured by GPT-4o, as shown in Fig. 1. Surprisingly, it also fails to produce text similar to human-written content, instead generating an overly lengthy translation of classical Chinese without simplification, which is unsuitable for children. These findings underscore the limitations of current LLMs in effectively adapting literary classics for children.

Scrutinizing the adapted child-friendly text by human writers, we identify three key children’s reading preferences that are crucial for effective adaptation. **i)** Considering that the literary classic often contains numerous characters, vividly portraying each character’s personality can help children better remember and distinguish them. As shown in Fig. 1, the green-highlighted words in the adaptation emphasize the different personalities of Tripitaka and the Demon. For instance, Tripitaka presents a timid and respectful nature by his trembling response, showcasing his fear and humility in the face of danger. **ii)** Concise narrative structure, rather than overly detailed or complex plots, is effective in sustaining children’s interest in reading. As shown in Fig. 1, the adapted text simplifies the dialogue while maintaining the core storyline. Specifically, several unnecessary details, like “sent by the Emperor” and “obtained the scriptures” are removed in Tripitaka’s explanation, as they introduce background information and future outcomes that may confuse children. **iii)** Besides, the adaptation also take into account the children’s reading level to ensure appropriate readability, facilitating easier comprehension and engagement with the content (Chitez et al., 2024).

In this paper, we propose the InstructChild, a method to effectively adapt Chinese literary classics for children with the LLM. It consists of three

key techniques: First, we apply **fine-grained instruction tuning**, which incorporates personality information and narrative structure for LLM to generate text that emphasizes character traits and follows a concise narrative. Specifically, the personality is assessed based on the Big Five Personality Traits (BFPT) (Roccas et al., 2002) theory and the narrative structure is presented by the entity-relation triplets of the original text. Next, we design a Chinese readability metric to guide the **refinement** process. It ensures the LLM generates adapted text with easier-to-understand sentences for children. Finally, inspired by (Wang et al., 2023), during inference, the **lookahead decoding strategy** considers the impact of potential subsequent tokens based on the readability metric for current token selection. In addition, we collect both original and adapted versions of the Four Great Classical Novels of Chinese literature to construct the Classic4Children dataset for evaluation.

Our main contributions are: **i)** We highlight children’s reading preferences (*i.e.*, vivid character portrayals, concise narrative structure and appropriate readability) are essential in adapting Chinese literary classics for children (CLA). We also find that state-of-the-art LLMs struggle to capture these preferences in their adaptations. **ii)** We propose the InstructChild for the CLA task, which effectively incorporates these preferences with the LLM, using fine-grained instruction tuning, refinement with readability metric and lookahead decoding strategy. **iii)** We construct Classic4Children dataset from the Four Great Classical Novels of Chinese literature for evaluation. Our InstructChild achieves significant performance gains over the existing LLMs.

To facilitate further research, we make the code and dataset available at <https://github.com/Gary-code/Classic4Children>.

## 2 Related Work

### 2.1 Text Style Transfer

The objective of text style transfer (TST) is to endow text with a different style (*e.g.*, positive  $\rightarrow$  negative) while preserving its semantic content. The traditional paradigm explicitly divides text into content and style information and then employs a target style for desired text generation (Yuan et al., 2022; Zhu et al., 2023; Zhao et al., 2024a). Specifically, Zhu et al. (2023) address the task of author-style transfer and implement content-style disentanglement and stylization at the

discourse level. Zhao et al. (2024a) develop a multi-layer Joint Style-Content Weighed (JSCW) module along with a style consistency loss to ensure both content preservation and consistent style across generated sentences. Recently, LLMs have shown promising results on TST through fine-tuning (Dementieva et al., 2023; Mukherjee et al., 2024), in-context learning (Mai et al., 2023; Chen, 2024) and prompt-based editing (Suzgun et al., 2022; Liu et al., 2024b). However, previous methods primarily transfer the text styles related to sentiment and formality. This study adapts Chinese literary classics into a child-friendly style, emphasizing vivid character descriptions and concise narrative structure tailored to children’s reading levels.

## 2.2 Text Simplification

Text simplification aims to reduce the complexity of the text, which can be categorized into two main branches (*i.e.*, editing operations and lexical-syntactic rules). Specifically, the editing operations simplify the text through various editing techniques, such as replacing difficult words and re-ordering sentence components (Kumar et al., 2020; Mallinson et al., 2020). Moreover, some works introduce the lexical and syntactic rules for simplification (Qiang et al., 2020; Valentini et al., 2023). For instance, Valentini et al. (2023) incorporate the LLM with the lexical simplification model to simplify the children’s story. In contrast to text simplification, which aims to simplify vocabulary and syntax, Chinese literary classics often feature classical language that is challenging for children to understand. Our approach involves clearly explaining complex expressions while maintaining children’s engagement through vivid character descriptions and an age-appropriate narrative structure.

## 3 Methodology

In this section, we introduce InstructChild, a large language model (LLM) based method to facilitate the adaptation of Chinese literary classics into a child-friendly style. This method bridges the gap between the complexity of the original text and the cognitive abilities of children with fine-grained instruction tuning, refinement and a lookahead decoding strategy. Fig. 2 illustrates the proposed InstructChild, and the details of our method are elaborated in the following sections.

### 3.1 Fine-grained Instruction Tuning

Inspired by previous studies (Wang et al., 2023; Ouyang et al., 2022; Yuan et al., 2024) that successfully apply well-designed prompts as guidance to generate text with desired attributes, we develop instruction incorporating personality and narrative structure to fine-tune the LLM (*i.e.*, Qwen2-7B-Instruct (Yang et al., 2024)), aiming to adapt text with vivid character portrayals and concise narrative structure. Specifically, they are prepended to input text as the instruction for fine-tuning.

#### 3.1.1 Personality Assessment

In the field of psychology, the Big Five Personality Traits (BFPT) (Roccas et al., 2002) categorizes personality into five dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. It provides a robust framework for assessing character personality. Technically, we first identify the character names in the original text and filter out the top 50 characters by frequency of occurrence for each literature. Given that large language models (LLMs) have demonstrated the ability to effectively capture the personalities of characters (Zhao et al., 2024b; Wang et al., 2024b), we prompt GPT-4o (OpenAI, 2023) to obtain the personality dimension scores and brief descriptions of characters in the original text. The prompt is shown in the Table 7. Specifically, the score for each personality dimension ranges from 1 to 5, with higher values indicating a stronger presence of that specific dimension. Consequently, we obtain the personality information for each character based on these dimensions with the corresponding score and brief description, as shown in Fig. 2(a).

#### 3.1.2 Narrative Structure Extraction

In addition to character personality, the concise narrative structure also distinguishes adapted child-friendly style text from the original literary classic (Müller, 2013). It can be constructed from the entities and their relationships in the text (Zhao and Zhang, 2024; Xie et al., 2024). Therefore, we also employ the GPT-4o to identify entities and relationships as supplementary information for further fine-tuning. The prompt is shown in the Table 8. These triplets can be integrated into the instructions to guide the LLM in focusing on important narrative elements.

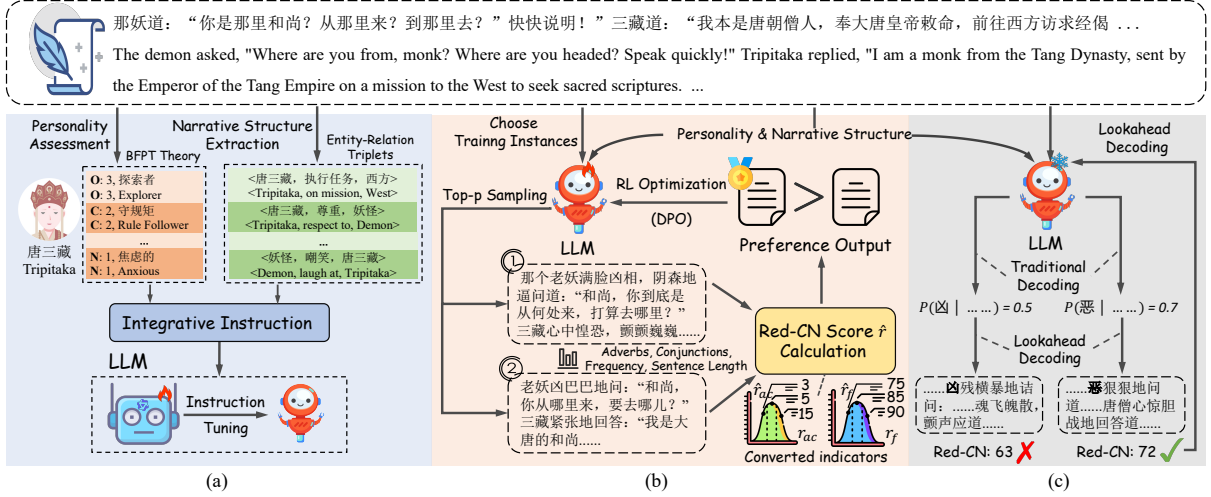


Figure 2: Overview of InstructChild. (a) fine-grained instruction tuning, which incorporates the characters’ personalities and narrative structure to fine-tune the large language model (LLM) with LoRA. (b) refinement with reinforcement learning, which considers a readability metric (*i.e.*, Red-CN) as a reward to further optimize the LLM to align with the children’s reading level. (c) lookahead decoding strategy, which extends the traditional decoding strategy with the readability score during inference.

### 3.1.3 Integrative Instruction for LLM

The integrative instruction comprises the characters’ personalities, narrative structure (*i.e.*, entity-relation triplets) and the original literary texts. Considering that each input literary text typically contains only a limited number of characters, we first identify the character names and then retrieve the corresponding personality information of the character. An exemplar instruction for fine-tuning the large language model is shown in Table 9. The integrative instruction is directly fed into the frozen large language model (*i.e.*, Qwen2-7B-Instruct) with learnable LoRA layers for child-friendly style text adaptation. The language modeling loss can be formulated as:

$$\mathcal{L}_{lan} = - \sum_{t=1}^T \log p(y_t | y_{<t}, \text{Ins}), \quad (1)$$

where  $\log p$  is the negative log-likelihood,  $\text{Ins}$  is the integrative instruction, and  $y_{<t}$  represents the words before the  $t$ -th word.

## 3.2 Refinement with Readability Metric

After fine-grained instruction fine-tuning, the LLM can effectively emphasize character personalities and the narration of a concise storyline. However, the instruction tuning phase does not include an explicit mechanism to verify whether the generated text aligns with the children’s reading level. Instead, it relies solely on token-wise gradient updates through teacher forcing based on reference

texts. Therefore, we design the readability metric tailored for Chinese child-adapted literary classics, namely Red-CN, to assist LLM in generating text with the desired readability. Existing studies (Ouyang et al., 2022; Xu et al., 2023), have demonstrated that supplementing initial instruction tuning with a subsequent reinforcement learning phase can be beneficial, where the model is further refined by the reward function. For our implementation, we utilize the designed readability metric Red-CN as a reward for refinement.

### 3.2.1 Readability Metric as Reward

We design the readability metric Red-CN, which takes into account the complexity of Chinese sentences. Inspired by (Valentini et al., 2023; Chitez et al., 2024), we first assess the suitability of Chinese characters and syntax. Specifically, the metric is determined by the proportion of adverbs and conjunctions  $r_{ac}$  and the frequency statistics of the character  $r_f$  based on the collection (Da, 2004) in each sentence. We analyze expert-adapted child-friendly literary classics and find that these two indicators typically cluster around values of 5 and 85 per sentence, respectively. It suggests that excessively low or high values of these indicators are not conducive to optimal reading levels for children. We assign the maximum value of 1.0 when any indicator matches the desired value when each indicator matches its corresponding target value previously mentioned, with the reward decreasing

exponentially toward 0 as it deviates from the target values. Thus, we convert these two indicators with the normalized Gaussian distribution  $\mathcal{F}(\cdot)$  centered at their target values:

$$\mathcal{F} = \mathcal{N}(\bar{r}, \delta^2), \quad (2)$$

$$\hat{r}_{ac} = \mathcal{F}(r_{ac}), \hat{r}_f = \mathcal{F}(r_f), \quad (3)$$

where  $\bar{r}$  is the corresponding target value,  $\delta$  is the standard deviation.  $\hat{r}_{ac}$  and  $\hat{r}_f$  are the converted indicators. The rationale behind introducing the Gaussian distribution is to ensure that the reward decreases nonlinearly, in the sense that minor deviations from the target readability lead to slight reductions, whereas more significant deviations result in increasingly larger penalties. More details are shown in the Appendix B.1.1 and B.1.2.

Furthermore, Xu et al. (2016) suggest that reducing the number of characters contributes to more effective adaptations of literary classics for children. We use the token length of the adapted text as an additional indicator to prevent generating overly verbose text. This indicator is also normalized to the range of [0, 1], using its proportional value relative to the original input text, denoted as  $\hat{r}_t$ , where higher values represent fewer tokens:

$$\hat{r}_t = \max\left(0, 1 - \frac{\text{output\_len}}{\text{input\_len}}\right), \quad (4)$$

where `input_len` and `output_len` represent the length of the original text and the generated text, respectively. Finally, following the emphasis by Da (2004) on the strong correlation between character frequency and readability in Chinese, we assign the specific weights to obtain the overall readability metric (*i.e.*, Red-CN), as the reward score  $\hat{r}$ :

$$\hat{r} = 0.3\hat{r}_{ac} + 0.4\hat{r}_f + 0.3\hat{r}_t. \quad (5)$$

We also report correlation coefficients in Table 6 of Appendix B.2, showing good correlations between Red-CN and human evaluation metrics.

### 3.2.2 DPO

The reinforcement learning with the readability metric is used to ensure the LLM generates text aligned with the children’s reading level. Specifically, we randomly choose 1,000 instances from the Classic4Children training data and use the top sampling method to produce  $K$  outputs for each sample, represented as  $\{\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_K\}$ . Next, we form them as rank pairs based on the readability

metric score, ignoring those with a score difference less than 3. The direct preference optimization (DPO) (Rafailov et al., 2023) is employed as a reinforcement learning policy to further optimize the LLM with learnable LoRA layers.

### 3.3 Lookahead Decoding

Getting inspiration from (Wang et al., 2023), we devise a lookahead decoding strategy during inference to maximize the LLM’s ability to generate text with high readability scores, which extends the traditional strategy with the readability metric, as illustrated in Fig. 2(c). The core idea involves forecasting potential subsequent tokens and then adjusting the selection process toward higher readability scores. Specifically, the LLM first generates  $L$  candidate samples at the  $t$ -th decoding step. Each sample starts at the  $(t - 1)$ -th step and continually generates  $n$  subsequent tokens, which can be denoted as  $\hat{y}_{<t-1+n}$ . Then, we calculate the readability scores of these candidate samples. The calculation of each sample can be formulated as:

$$\mathcal{G}(\hat{y}_{<t-1+n}) = \hat{r}, \quad (6)$$

where  $\mathcal{G}(\cdot)$  denotes an supplementary guidance function during decoding. Consequently, the selection criterion for the  $t$ -th token can be mathematically expressed as:

$$f(y_t) = \log p(y_t | y_{<t}, \text{Ins}) + \lambda \max_L \mathcal{G}(\hat{y}_{<t-1+n}), \quad (7)$$

where  $\lambda$  is the hyperparameter that adjusts the influence of readability on the token selection process.

## 4 Experiment

### 4.1 Classic4Children Dataset

In this paper, we construct Classic4Children, a Chinese child-friendly literary adaptation dataset based on the Four Great Classical Novels of Chinese literature (*i.e.*, Journey to the West, Romance of the Three Kingdoms, Water Margin, Dream of the Red Chamber). Initially, we collect widely recognized versions of the original texts and their corresponding child-adapted editions<sup>12</sup>. Following (Zhu et al., 2023), since there are too many tokens in each chapter of the literary classic, we manually annotate the corresponding paragraph fragments as training samples, with each chapter being divided into multiple

<sup>1</sup><https://5000yan.com/>

<sup>2</sup>[http://www.bph.com.cn/shaoer.html?\\_isa=1](http://www.bph.com.cn/shaoer.html?_isa=1)

fragments. Additionally, these fragments generally span several paragraphs. Finally, we manually match these fragments from both versions as training pairs. The Classic4Children dataset consists of 2,686 and 300 samples for training and testing, respectively. More details of the dataset construction are described in the Appendix A.

## 4.2 Implementation Details

We train our InstructChild on a Tesla A40 48GB GPU card and it is initialized with Qwen2-7B-Instruct<sup>3</sup>. We insert the LoRA layers into each self-attention layer, setting the rank to 8 for optimization. Specifically, the batch size is 24 and we train 3 epochs during the fine-grained instruction tuning. The desired values (*i.e.*, the expectation) of the Gaussian distribution in Eq. 2 are 5 and 85 for  $\hat{r}_{ac}$  and  $\hat{r}_f$ , respectively. After experimenting with different values for the standard deviation  $\delta$ , we set it at half the expectation (*i.e.*, 2.5 and 42.5, respectively). This choice allows the mapping values to be concentrated as closely as possible around the target values. Before the refinement, we randomly sample 1,000 instances from the training set and employ top-p sampling strategy, where the temperature is set to 0.8 and  $p=0.9$  to generate four candidate samples for each input. We maintain those with a score difference greater than 3 for direct preference optimization. For the lookahead decoding strategy, the LLM generates  $L=5$  candidate samples with  $n=20$  subsequent tokens at each decoding step. The hyperparameter  $\lambda$  for the token selection in Eq. 7 is 1.

## 4.3 Baselines and Ablation Models

### 4.3.1 Baseline

To verify the superiority of our InstructChild, we compare it with three types of baselines. **i)** Closed-source large language models (LLMs) with large-scale parameters (**API-based**): GPT-4o (OpenAI, 2023), GLM-4 (GLM et al., 2024), QWen2.5 (Yang et al., 2024), and EAPMT (Wang et al., 2024a). We access these models via API interfaces, which exhibit strong reasoning capabilities across various natural language processing tasks. Specifically, we also ask them to focus on the characters personality and concise narrative structure with a one-shot example. In particular, EAPMT first generates a detailed explanation for each input sentence, encompassing both the literal content and deeper mean-

ings. Then it paraphrases sentences into a child-friendly style format based on the explanation. The prompts for these LLMs is shown in Table 10 and Table 11. **ii)** Text style transfer (TST) models (**TST-based**): P&R (Suzgun et al., 2022), StoryTrans (Zhu et al., 2023), and ParaGuide (Horvitz et al., 2024). Specifically, P&R is a training-free few-shot TST model and we utilize the Qwen2 as the backbone. We also expand the Chinese vocabulary for diffusion-based Paraguide and re-trained it for our task. **iii)** Open-source large language models with relatively moderate parameters (**FT-based**): Llama2-13B (Touvron et al., 2023), Llama3-8B (Dubey et al., 2024), Qwen2-7B (Yang et al., 2024) and GLM4-9B (GLM et al., 2024). We utilize LoRA (Hu et al., 2022) technique to fine-tune these LLMs. Notably, Llama models have been extended with the Chinese vocabulary and further pre-trained on Chinese instruction (Cui et al., 2023).

### 4.3.2 Ablation Models

We conduct ablation experiments to evaluate the effectiveness of various components within our InstructChild method. **InstructChild w/o Per** and **InstructChild w/o Nat**: InstructChild without personality information and narrative structure, respectively. **InstructChild w/o Ref**: InstructChild without refinement stage by reinforcement learning. It does not implement the lookahead decoding strategy for generation, as it is closely tied to the Red-CN metric used during refinement. **InstructChild w/o Look**: InstructChild without lookahead decoding strategy during inference.

## 4.4 Evaluation Metric

### 4.4.1 Automatic Evaluation Metrics

Following previous studies (Zhu et al., 2023), we evaluate the performance with BLEU-(1 to 2) (Papineni et al., 2002) and BERTScore (Zhang et al., 2020) metrics. Specifically, we report the recall (BS-R), precision (BS-P) and F1 score (BS-F1) of the BERTScore. They are commonly used for measuring the lexical and semantic similarity in text generation. Moreover, we also use the readability scores in Eq. 5 as an additional metric (*i.e.*, Red-CN) to assess whether the generated text aligns with children’s reading levels (higher is better).

### 4.4.2 Human Evaluation Criteria

We also conduct human evaluation for baselines (*i.e.*, GPT-4o, StoryTrans and GLM4-9B) and our InstructChild. Specifically, we randomly select 200

<sup>3</sup><https://huggingface.co/Qwen/Qwen2-7B-Instruct>

samples from the test set for evaluation and invite five volunteers with good educational backgrounds in Chinese to assess the following criteria: Fluency (**Flu**) mainly reflects the grammatical correctness and fluency of the generated text. Content Preservation (**CP**) refers to the conservation of the core storyline and coherence of the narrative. Character Clarity (**CC**) measures the clarity and emphasis of character descriptions. Narrative Efficiency (**NE**) indicates the conciseness of the story narration with less redundant plots. Each criterion is scored on a scale from 0 to 2, with higher scores indicating better performance.

## 4.5 Results and Analysis

### 4.5.1 Performance Comparison

Table 1 shows the results of baselines and our InstructChild method on Classic4Children dataset. We find that: **i**) The closed-source LLMs, despite their strong performance in many NLP tasks, do not perform well on the child-friendly literary adaptation (CLA) task. Specifically, our InstructChild scores higher than these LLMs under all evaluation metrics, e.g. “+2.17” and “+4.93” on BS-F1 and Red-CN, respectively, compared to GPT-4o, which has over 10 times more parameters. The reason is that these LLMs tend to generate detailed explanations of the original texts and produce excessively lengthy outputs (Chen et al., 2023; Hu et al., 2024; Chen et al., 2024), resulting in suboptimal performance. **ii**) There is a significant performance gap between existing text style transfer (TST) models and our InstructChild, which indicates that the CLA task is not simply a matter of text style transfer but also requires the consideration of children’s reading preference for effective adaptation. These TST models cannot capture the child-friendly style for adaptation and often include words that are too complex for children. By integrating characters’ personalities and narrative structure, our InstructChild effectively produces text that is both accessible and engaging for children. **iii**) Despite our efforts to continually pre-training the Llama models on Chinese instruction datasets, their performance remains considerably distant from the intended reading level. Specifically, Llama3-8B and Llama2-13B score significantly lower than other models on the Red-CN metric. Notably, they tend to generate specific phrases that closely match the reference text, leading to a higher BS-P value. Since they do not fully cover all the content in the

reference text, the BS-R value is low. Furthermore, QWen2-7B and GLM4-9B also perform worse than our InstructChild, indicating that straightforward LoRA fine-tuning alone for the LLM is insufficient.

### 4.5.2 Ablation Study

The results of ablation experiments are shown in Table 1. We observe that: **i**) By removing personality information and narrative structure from the integrative instruction, there is a decline in performance on all metrics. It demonstrates that this additional information is crucial for the LLM to produce text adaptation more similar to those written by humans. Moreover, although InstructChild w/o Per and InstructChild w/o Nat employ refinement, the drop in readability scores (*i.e.*, Red-CN) indicates that initial instruction tuning can encourage the LLM to further generate sentences that align with the children’s reading level. **ii**) Then we investigate the impact of the refinement with Red-CN metric. Comparing the result of InstructChild and InstructChild w/o Ref, the refinement stage improves performance, particularly on the Red-CN metric. Simultaneously, most traditional natural language generation metrics also improve. It demonstrates that our Red-CN metric is a reasonable measure of children’s reading ability in Chinese literary classics. **iii**) Finally, we notice a slight performance improvement when applying the lookahead decoding strategy. It suggests that this strategy can further encourage the LLM to generate child-friendly style text with appropriate readability during inference.

### 4.5.3 Human Evaluation Results

To validate the reliability of our human evaluation, we separately calculate standard deviations of each human evaluation metric, as shown in Table 2. The statistical analysis confirms the faithfulness of our evaluation results. Table 2 also presents the results of the human evaluation. We find that: **i**) Our InstructChild achieves competitive results on the **Flu** metric with the much larger LLM (*i.e.*, GPT-4o) to generate fluent sentences. For the **CP** metric, our integrative instruction guides the LLM to simplify or omit certain intricate plots. In contrast, GPT-4o often attempts a more comprehensive paraphrase, which results in higher scores. **ii**) The **CC** metric indicates that our model is more likely to generate adapted texts with rich character descriptions. This outcome is attributed to the inclusion of character personality during instruction tuning,

Method	Type	BLEU-1	BLEU-2	BS-P	BS-R	BS-F1	Red-CN
GPT-4o (OpenAI, 2023)	API-based	20.91	10.42	70.59	65.96	68.04	68.19
GLM-4 (GLM et al., 2024)		20.57	10.54	71.44	65.04	67.97	61.95
QWen2.5 (Yang et al., 2024)		20.30	9.30	69.84	64.92	67.17	63.09
EAPMT (Wang et al., 2024a)		18.26	8.91	71.42	63.09	66.89	57.47
P&R (Suzgun et al., 2022)	TST-based	8.60	4.12	57.94	59.32	58.62	56.73
StoryTrans (Zhu et al., 2023)		7.59	4.41	59.19	60.32	59.65	69.74
Paraguide (Horvitz et al., 2024)		13.88	4.64	60.43	61.24	60.72	71.87
Llama2-13B (Touvron et al., 2023)	FT-based	13.78	8.47	<b>73.86</b>	60.72	66.51	58.89
Llama3-8B (Dubey et al., 2024)		8.97	5.45	73.66	58.60	65.09	50.06
QWen2-7B (Yang et al., 2024)		17.63	8.09	66.04	68.12	67.07	67.31
GLM4-9B (GLM et al., 2024)		16.87	7.16	65.80	67.42	66.44	67.54
InstructChild w/o Per		20.07	9.63	67.87	68.32	68.10	71.85
InstructChild w/o Nat	-	21.46	10.93	67.96	69.03	68.21	71.56
InstructChild w/o Ref		21.37	11.98	69.55	70.26	69.36	67.88
InstructChild w/o Look		22.45	12.89	69.72	70.16	69.94	72.97
InstructChild		<b>22.93</b>	<b>13.09</b>	70.01	<b>70.42</b>	<b>70.21</b>	<b>73.12</b>

Table 1: Main results of baselines and our method. **Bold**: the maximum value in the column.

Method	Flu	CP	CC	NE
GPT-4o	<b>1.95</b> /0.25	<b>1.85</b> /0.22	1.52/0.18	0.95/0.12
StoryTrans	0.86/0.23	1.12/0.21	0.95/0.15	0.89/0.17
GLM4-9B	1.72/0.28	1.35/0.16	1.56/0.19	1.25/0.09
InstructChild	1.92/0.19	1.69/0.17	<b>1.83</b> /0.21	<b>1.78</b> /0.14

Table 2: Human evaluation results. Each value is presented as  $\tau/\rho$ , where  $\tau$  is the metric value and  $\rho$  is the standard deviation. **Bold**: the maximum value.

making the adapted content more relatable and engaging for children. **iii**) The **NE** metric indicates that our model significantly outperforms baselines by producing a more concise narrative structure aligned with the children’s reading level. Upon examining the generated outputs, we observe that GPT-4o often provides overly detailed explanations for certain terms, resulting in longer text than the original. This verbosity and inclusion do not align with children’s reading preferences.

#### 4.6 Case Study

Fig. 3 shows the adapted text generated by GPT-4o and our InstructChild. Specifically, InstructChild produces a more concise narrative that emphasizes key characters’ traits. It describes Cao Cao’s determination and urgency while simplifying his inner monologue. In contrast, GPT-4o solely rewrites the original text, resulting in overly lengthy content with complex words. Moreover, the output of GPT-4o is even longer than the original text, further reducing its readability for children. Although InstructChild performs well on the CLA task, it focuses on the key narrative where Cao Cao’s sword

Original Text:
次日，曹操佩着宝刀，来至相府，问：“丞相何在？”从人云：“在小阁中。”操径入。见董卓坐于床上，吕布侍立于侧。卓曰：“孟德来何迟？”操曰：“ <b>马羸行迟耳。</b> ”卓顾谓布曰：“ <b>吾有西凉进来好马，奉先可亲去拣一骑赐与孟德。</b> ”布领令而出。操暗付曰：“ <b>此贼合死！</b> ” <b>即欲拔刀刺之，惧卓力大，未敢轻动。卓胖大不耐久坐，遂倒身而卧，转面向内。</b> 操又思曰：“ <b>此贼当休矣！</b> ” <b>急掣宝刀在手，恰待要刺，不想董卓仰面看衣镜中，照见曹操在背后拔刀，急回身问曰：“孟德何为？”</b> 时吕布已牵马至阁外。操 <b>惶遽</b> ，乃持刀跪下曰：“ <b>操有宝刀一口，献上恩相。</b> ”卓接视之，见其刀长尺余，七宝嵌饰，极其锋利，果宝刀也；遂递与吕布收了。 <b>操解鞍布衣，卓引操出阁看马，操谢曰：“愿借试一骑。”卓就教与鞍马。</b> 操牵马出相府，加鞭东南。
GT:
第二天，曹操来到董卓家，将吕布支开去给曹操选马，曹操趁董卓躺在床上时候要刺杀他。董卓在镜子中发现曹操在身后举起刀，急忙回身问道：“孟德做什么？”这时吕布也回来了，曹操 <b>连忙</b> 说：“我有一口宝刀，想要献给丞相。”董卓接过刀来，见这宝刀长约一尺多，七种宝石镶嵌其上，锋利无比，便让吕布收了。曹操拜谢后骑马向东南离去。
GPT-4:
第二天，曹操拿着一把宝刀来到相府，问：“丞相在哪里？”有人告诉他：“在小阁里。”曹操径直走进来，看到董卓坐在床上，吕布站在旁边。董卓问：“孟德为什么这么晚才来？”曹操回答说：“马走得慢，所以来晚了。”董卓让吕布带他去挑一匹西凉进贡的好马送给曹操。曹操心想：“ <b>这个坏人应该死！</b> ” <b>想要拔刀刺他，但又害怕董卓力气大，不敢轻举妄动。董卓因为太胖坐久了不舒服，于是躺下来休息，脸朝里面。</b> 曹操又想：“ <b>这个坏人应该休息了！</b> ”他急忙拔出宝刀，准备刺他，却没想到董卓仰头看着镜子，发现曹操正要刺他，赶紧转身问：“孟德你要干什么？”这时吕布已经把马牵到了阁外。曹操慌乱之下，跪在地上说：“我有一把宝刀，献给您。”董卓接过一看，发现刀长又锋利，果然是宝刀，便交给吕布收起。曹操拔出刀交给吕布。 <b>董卓带着曹操出去看马，曹操谢过后说：“我想试一下这四匹马。”董卓让人给他备好鞍马。</b> 曹操牵着马离开相府，加快速度向东南方向离去。
InstructChild:
曹操带剑来到相府，正碰见董卓和吕布在一起。董卓让吕布去挑一匹好马给曹操，吕布领命去了。曹操 <b>心中暗想：“这董卓是该死的时候了！”</b> 他刚拔出宝刀，董卓却突然转过身，看见曹操在后面拔刀，急忙问道：“孟德为什么这样做？”原来董卓正在镜子前看自己，看到了曹操的举动，连忙转身问。这时候吕布已经牵着马到了门外，曹操 <b>急忙把宝刀递给吕布</b> ，说：“我有一口宝刀，献给恩相。”董卓接过一看，见这把刀长达一尺多，镶嵌着宝石，十分锋利，就收下了。曹操拜谢后牵着马出了相府，加鞭向东南方向去了。

Figure 3: A case study. The **green text** denotes words that emphasize the characters’ personalities. The **red text** are redundant plots and words with low readability for children. The **blue text** may cause misunderstanding.

ultimately reaches Lv Bu. However, such simplifications can sometimes lead to misunderstandings. We believe that the balance between simplifying character relationships and ensuring the clarity of key details is a crucial direction for future research.



## 5 Conclusion

In this paper, we introduce the child-friendly literary adaptation (CLA) task to automatically adapt Chinese literary classics into engaging and age-appropriate text for children. Moreover, we pinpoint three key children’s reading preferences (*i.e.*, vivid character portrayals, concise narrative structures, and appropriate readability) for CLA, even state-of-the-art LLMs struggle to capture these preferences for adaptation. Our proposed InstructChild explicitly leverages these preferences to guide the LLM in generating child-friendly text for children. Additionally, we construct the Classic4Children dataset for a comprehensive evaluation. Experimental results show that our InstructChild significantly outperforms the existing LLMs.

## 6 Limitations

In this paper, we propose InstructChild to adapt Chinese literary classics into a child-friendly style. Although our method significantly enhances the readability of adapted texts for children, the look-ahead decoding strategy incurs substantial computational overhead during inference. Consequently, both the number of subsequent tokens  $n$  and candidate samples  $L$  must be limited. To mitigate these computational costs, a potential solution is to apply distillation techniques to improve decoding efficiency. Additionally, our CLA task currently focuses on paragraph fragments as training data due to limitations on model input length. In the future, adapting the model to handle the full chapter sequences should be considered. Moreover, we also believe that the balance between simplifying character relationships and ensuring the clarity of key details is a crucial direction for future research.

## Acknowledgments

This research is supported by Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010), the National Natural Science Foundation of China (62076100, 62476097), the Fundamental Research Funds for the Central Universities, South China University of Technology (x2rjD2240100), Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078).

## References

- Jiali Chen, Zhenjun Guo, Jiayuan Xie, Yi Cai, and Qing Li. 2023. Deconfounded visual question generation with causal inference. In *Proc. of ACM MM*, pages 5132–5142. ACM.
- Jiali Chen, Xusen Hei, Yuqi Xue, Yuancheng Wei, Jiayuan Xie, Yi Cai, and Qing Li. 2024. Learning to correction: Explainable feedback generation for visual commonsense reasoning distractor. In *Proc. of ACM MM*, pages 8209–8218. ACM.
- Jianlin Chen. 2024. Lmstyle benchmark: Evaluating text style transfer for chatbots. *CoRR*, abs/2403.08943.
- Madalina Chitez, Mihai Dascalu, Aura Cristina Udrea, Cosmin Striletschi, Karla Csürös, Roxana Rogobete, and Alexandru Oravitan. 2024. Towards building the LEMI readability platform for children’s literature in the romanian language. In *Proc. of COLING*, pages 16450–16456. ELRA and ICCL.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *CoRR*, abs/2304.08177.
- Jun Da. 2004. A corpus-based study of character and bigram frequencies in chinese e-texts and its implications for chinese language instruction. In *Proceedings of the fourth international conference on new technologies in teaching and learning Chinese*, pages 501–511. Citeseer.
- Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proc. of IJCNLP*, pages 1083–1101. Association for Computational Linguistics.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen R. McKeown. 2024. Paraguide: Guided diffusion paraphrasers for plug-and-play textual style transfer. In *Proc. of AAAI*, pages 18216–18224. AAAI Press.
- Chih-tsing Hsia. 2016. *The classic Chinese novel: A critical introduction*. The Chinese University of Hong Kong Press.

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proc. of ICLR*. OpenReview.net.
- Zhiyuan Hu, Yuliang Liu, Jinman Zhao, Suyuchen Wang, Yan Wang, Wei Shen, Qing Gu, Anh Tuan Luu, See-Kiong Ng, Zhiwei Jiang, and Bryan Hooi. 2024. Longrecipe: Recipe for efficient long context generalization in large language models. *CoRR*, abs/2409.00509.
- Haifeng Hui. 2024. Children’s literature adaptation studies: Interdisciplinary theoretical resources with a special focus for education. *History of Education and Children’s Literature: HECL: XIX, 1, 2024*, pages 629–647.
- Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: benchmarking large language models on sentence simplification. In *Proc. of EMNLP*, pages 13291–13309. Association for Computational Linguistics.
- Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proc. of ACL*, pages 7918–7928. Association for Computational Linguistics.
- James Legge. 2022. *The Chinese Classics: Vol. I*. BoD–Books on Demand.
- Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024a. Step-by-step: Controlling arbitrary style in text with large language models. In *Proc. of COLING*, pages 15285–15295. ELRA and ICCL.
- Qingyi Liu, Jinghui Qin, Wenxuan Ye, Hao Mou, Yuxuan He, and Keze Wang. 2024b. Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. In *Proc. of AACL*, pages 18689–18697. AACL Press.
- Zhao Luo. 2019. Research on the teaching of chinese language and literature in modern educational thoughts. *Review of Educational Theory*, 2(2):29–33.
- Huiyu Mai, Wenhao Jiang, and Zhi-Hong Deng. 2023. Prefix-tuning based unsupervised text style transfer. In *Proc. of EMNLP Findings*, pages 14847–14856. Association for Computational Linguistics.
- Jonathan Mallinson, Aliaksei Severyn, Eric Malmi, and Guillermo Garrido. 2020. FELIX: flexible text editing through tagging and insertion. In *Proc. of EMNLP Findings*, volume EMNLP 2020 of *Findings of ACL*, pages 1244–1255. Association for Computational Linguistics.
- Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. 2024. Are large language models actually good at text style transfer? In *Proc. of INLG*, pages 523–539. Association for Computational Linguistics.
- Anja Müller. 2013. *Adapting Canonical Texts in Children’s Literature*. A&C Black.
- OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proc. of NeurIPS*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.
- Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lsbert: A simple framework for lexical simplification. *CoRR*, abs/2006.14939.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proc. of NeurIPS*.
- Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proc. of ACL*, pages 837–848. Association for Computational Linguistics.
- Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801.
- Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proc. of EMNLP*, pages 2195–2222. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu 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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.
- Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children’s stories. In *Proc. of EMNLP*, pages 3588–3598. Association for Computational Linguistics.
- Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: self-consistent chain-of-thought distillation. In *Proc. of ACL*, pages 5546–5558. Association for Computational Linguistics.
- Shanshan Wang, Derek F. Wong, Jingming Yao, and Lidia S. Chao. 2024a. What is the best way for chatgpt to translate poetry? *CoRR*, abs/2406.03450.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proc. of ACL*, pages 1840–1873. Association for Computational Linguistics.
- Jiayuan Xie, Yi Cai, Jiali Chen, Ruohang Xu, Jiexin Wang, and Qing Li. 2024. Knowledge-augmented visual question answering with natural language explanation. *IEEE Trans. Image Process.*, 33:2652–2664.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Trans. Assoc. Comput. Linguistics*, 4:401–415.
- Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. In *Proc. of EMNLP*, pages 5967–5994. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Li Yuan, Yi Cai, Haopeng Ren, and Jiexin Wang. 2024. A logical pattern memory pre-trained model for entailment tree generation. In *Proc. of COLING*, pages 759–772.
- Li Yuan, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2022. Hierarchical template transformer for fine-grained sentiment controllable generation. *Information Processing & Management*, 59(5):103048.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proc. of ICLR*. OpenReview.net.
- Jie Zhao, Ziyu Guan, Cai Xu, Wei Zhao, and Yue Jiang. 2024a. SC2: towards enhancing content preservation and style consistency in long text style transfer. *CoRR*, abs/2406.04578.
- Jinman Zhao, Zifan Qian, Linbo Cao, Yining Wang, and Yitian Ding. 2024b. Bias and toxicity in role-play reasoning. *CoRR*, abs/2409.13979.
- Jinman Zhao and Xueyan Zhang. 2024. Large language model is not a (multilingual) compositional relation reasoner. In *First Conference on Language Modeling*.
- Xuekai Zhu, Jian Guan, Minlie Huang, and Juan Liu. 2023. Storytrans: Non-parallel story author-style transfer with discourse representations and content enhancing. In *Proc. of ACL*, pages 14803–14819. Association for Computational Linguistics.

## A Dataset Construction

We describe the data collection process based on the Four Great Classical Novels of Chinese literature (*i.e.*, Journey to the West, Romance of the Three Kingdoms, Water Margin, Dream of the Red Chamber). Specifically, the original texts are collected from the publicly available website<sup>4</sup>. For the children’s adapted versions, we choose the widely acclaimed children’s reading series<sup>5</sup>. Following (Zhu et al., 2023), since there are too many tokens in each chapter of the literary classic, we use paragraph fragments as training samples, with each chapter being divided into multiple fragments. Considering that a single chapter in the children’s version may correspond to multiple chapters in the original version, we first manually align the chapter IDs of the children’s edition with those of the original work. This approach effectively narrows down the scope for subsequent annotation, ensuring more accurate correspondence between the versions. Subsequently, using the text of each paragraph from the children’s version as a reference, we manually search for the corresponding text in the original version and constructed them into pairs. Moreover, we filter out paragraphs that lacked a matching counterpart in the original text. This process ensures that the data accurately reflects the intended alignment between the two versions. After this meticulous process, we obtained 819, 744, 742, and 681 data samples for “Journey to the West”, “Romance of the Three Kingdoms”, “Water Margin” and “Dream of the Red Chamber,” respectively. Finally, the collected data are divided into training and testing samples, comprising 2,686 samples for training and 300 samples for testing, with 75 samples from each literary work are randomly selected for testing.

## B More Experimental Details

### B.1 Indicators of Readability Metric

#### B.1.1 Adverbs and Conjunctions

Inspired by (Chitez et al., 2024), the appropriate proportion of adverbs and conjunctions in sentences plays a crucial role in the readability of Chinese texts. Therefore, we utilize publicly available code<sup>6</sup> for the calculation.

<sup>4</sup><https://5000yan.com/>

<sup>5</sup>[http://www.bph.com.cn/shaoer.html?\\_isa=1](http://www.bph.com.cn/shaoer.html?_isa=1)

<sup>6</sup><https://pypi.org/project/cntext/>

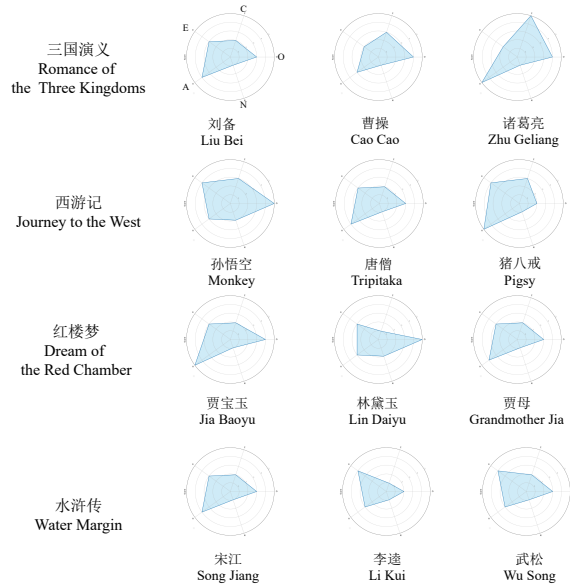


Figure 4: Some cases of personality assessment.

$L$	BLEU-2	BS-F1	Red-CN
2	12.86	69.98	73.01
5	13.09	70.21	73.12
8	<b>13.27</b>	<b>70.67</b>	<b>73.15</b>

Table 3: Experimental results of different  $L$  for token selection. **Bold**: the maximum value in the column.

### B.1.2 Chinese Character Frequency

In our analysis of Chinese novel and their texts, we encounter difficulties in calculating word frequency. It arises primarily due to the prevalence of character names and proprietary nouns, which are seldom included in common word lists. As a result, these terms often exhibit anomalously low frequency, even they are simple for children to read. Consequently, we calculate the frequency of individual characters for each sentence based on the collection (Da, 2004). Specifically, we use only the top 5000 most frequent Chinese characters from the collection and recalibrate their statistical frequencies for calculation.

### B.1.3 Sentence Length

Furthermore, Xu et al. (2016) suggest that reducing the number of characters contributes to more effective adaptations of literary classics for children. We utilize the sentence length of the adapted text as an additional indicator to prevent generating overly verbose text. This indicator  $r_t$  is also normalized to the range of [0, 1], using its proportional value relative to the original input text, where higher values

$n$	<b>BLEU-2</b>	<b>BS-F1</b>	<b>Red-CN</b>
10	12.94	69.91	72.80
20	13.09	70.21	73.12
50	<b>13.17</b>	<b>70.46</b>	<b>73.64</b>

Table 4: Experimental results of different  $n$  for token selection. **Bold**: the maximum value in the column.

$\lambda$	<b>BLEU-2</b>	<b>BS-F1</b>	<b>Red-CN</b>
0.5	<b>13.17</b>	70.10	72.42
1	13.09	<b>70.21</b>	<b>73.12</b>
2	12.89	69.87	72.61

Table 5: Experimental results of different  $\lambda$  for the token selection. **Bold**: the maximum value in the column.

represent fewer tokens:

$$r_t = \max\left(0, 1 - \frac{\text{output\_len}}{\text{input\_len}}\right), \quad (8)$$

where `input_len` and `output_len` represent the length of the original text and the generated text, respectively.

## B.2 Correlation Comparison

Following (Xu et al., 2023), we report Pearson’s correlation coefficients between the automatic metrics (*i.e.*, BERTScore and reward calculation Red-CN) and human evaluation metrics (*i.e.*, Flu, CP, CC and NE, as defined in the Human Evaluation Criteria section) to assess the effectiveness of our reward calculation. The Core results are shown in Table 6, and the p-values of our results are lower than 0.05, which indicates statistical significance. Specifically, we sample 100 generated instances from the model and invite volunteers with good educational backgrounds in Chinese to provide human evaluation results. The experimental results show that all human evaluation metrics exhibit statistically significant correlations with Red-CN, demonstrating the effectiveness of our reward calculation.

<b>Metric</b>	<b>Flu</b>	<b>CP</b>	<b>CC</b>	<b>NE</b>
BERTScore	23.25	15.87	24.66	18.36
Red-CN	<b>41.64</b>	<b>34.93</b>	<b>43.39</b>	<b>37.95</b>

Table 6: Pearson correlation on data samples from Classic4Children dataset.

## B.3 Experiment on Hyperparameter

We conduct experiments to investigate the sensitivity of hyperparameters during inference, including the number of candidate samples  $L$ , the number of subsequent tokens  $n$  and the weight  $\lambda$  for the token selection. Specifically, we explore different parameter  $\lambda$  settings, including ( $\lambda = 0.5, 1, 2$ ) in Table 5, and find that the fluctuations of the results are minimal. Table 3 and Table 4 show the results of different  $L$  and  $n$ , respectively. We observe that the model performance has a slight improvement as  $L$  and  $n$  increase. However, larger values of  $L$  and  $n$  are computationally costly, as they require the generation of additional future tokens at each step during inference. Therefore, our InstructChild model employs the parameter setting of  $L = 5$  and  $n = 20$  to reduce computational overhead.

---

**Prompt:**

请根据大五人格特质分析角色：开放性、尽责性、外向性、亲和性、神经质。对西游记中孙悟空的性格进行分析，每个特质分配一个分数（1-5），并为每个分数提供简短的解释（最多10个字）。分数越高，表示该特质的存在感越强。请遵循以下格式：

经验开放性：2，好奇但保守。

尽责性：1，轻浮不认真。

外向性：4，社交活跃。

亲和性：3，有些自私。

神经质：1，易怒冲动。

Please analyze the character based on the Big Five personality traits (BFPT): openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Analysis of Monkey's character in the Journey to the West. Then assign a score (1-5) for each trait based on the role and provide a brief explanation (max 10 words) for each score. A higher score indicates a stronger presence of that trait. You should use the format as follow:

Openness: 2, curious but reserved.

Conscientiousness: 1, frivolous and not serious.

Extraversion: 4, sociable.

Agreeableness: 3, somewhat selfish.

Neuroticism or emotional stability: 1, irritable and impulsive.

---

**GPT-4o:**

经验开放性：5，创新冒险。

尽责性：3，随性自由。

外向性：4，活泼好动。

亲和性：3，慷慨侠义。

神经质：2，冲动易怒。

Openness: 5, innovative and adventurous.

Conscientiousness: 3, casual and free.

Extraversion: 4, lively and active.

Agreeableness: 3, generous and chivalrous.

Neuroticism: 2, irritable and impulsive.

---

Table 7: The exemplar prompt is used for GPT-4o to assess personality dimension scores of characters based on the Big Five Personality Traits (BFPT) theory.

---

**Prompt:**

请根据下面文本，抽取故事叙事过程当中的实体和关系，用格式<A, R, B>的形式来进行输出，其中A和B分别是头实体和尾实体，R代表实体之间的关系。输入文本：那妖道：“你是那里和尚？从那里来？到那里去？”快快说明！”三藏道：“我本是唐朝僧人，奉大唐皇帝救命，前往西方访求经偈...”

The demon asked, "Where are you from, monk? Where are you headed? Speak quickly!" Tripitaka replied, "I am a monk from the Tang Dynasty, sent by the Emperor of the Tang Empire on a mission to the West to seek sacred scriptures. ..."

---

**GPT-4o:**

<妖道, 询问, 和尚>, <和尚, 来自, 唐朝>, <和尚, 奉命, 大唐皇帝>, <和尚, 前往, 西方>, <和尚, 目的, 访求经偈> ...

<The demon priest, inquires of, the monk>, <The monk, comes from, the Tang Dynasty>, <The monk, is ordered by, the Emperor of the Great Tang>, <The monk, is heading to, the West>, <The monk, aims to, seek scriptures> ...

---

Table 8: The exemplar prompt is used for GPT-4o to extract the entity-relation triplets as the narrative structure.

---

**Instruction:**

请将以下中国经典名著原文内容改编为适合儿童阅读的版本。你需要依据大五人格理论对角色评估的分数，突出人物的性格特征。并利用实体-关系三元组来构建故事的叙事框架，适当简化或省略一些复杂的叙事情节。

**# 人物性格:**

孙悟空: [经验开放性: 5, 创新冒险。尽责性: 3, 随性自由。外向性: 4, 活泼好动。亲和性: 3, 慷慨侠义。神经质: 2, 冲动易怒] ...

**# 实体-关系三元组:**

<道士, 在, 滚油锅里>, ..., <国王, 扑在, 御案>。

**# 原文内容:**

三藏、八戒、沙僧立在殿前, 见那道士在滚油锅里打挣...

Please adapt the following Chinese classics into a version suitable for children. You need to highlight the character's personality traits based on the evaluation scores. Use entity-relationship triples to construct the story's narrative framework, and appropriately simplify or omit some complex narrative plots.

**# Character:**

Monkey: [Openness: 5, innovative and adventurous. Conscientiousness: 3, casual and free. Extraversion: 4, lively and active. Agreeableness: 3, generous and chivalrous. Neuroticism: 2, impulsive and irritable.] ...

**# Entity-relationship triples:**

<Taoist priest, in, boiling oil pot>, ..., <King, fell on, the imperial case>.

**# Original content:**

Tripitaka, Pigsy, and Friar Sand stood in front of the hall and saw the Taoist priest struggling in the boiling oil pot ...

---

Table 9: The exemplar integrative instruction.

---

**Prompt:**

请将以下中国经典名著的原文改写为儿童易懂的版本, 参考以下示例格式:

**# 示例:****## 原文内容:**

孙悟空在旁闻讲, 喜得他抓耳挠腮, 眉花眼笑。忍不住手之舞之, 足之蹈之。忽被祖师看见, 叫孙悟空道: “你在班中, 怎么颠狂跃舞, 不听我讲?” 悟空道: “弟子诚心听讲, 听到老师父妙音处, 喜不自胜, 故不觉作此踊跃之状。望师父恕罪!” 祖师道: “你既识妙音, 我且问你, 你到洞中多少时了?” 悟空道: “弟子本来懵懂, 不知多少时节。只记得灶下无火, 常去山后打柴, 见一山好桃树, 我在那里吃了七次饱桃矣。” 祖师道: “那山唤名烂桃山。你既吃七次, 想是七年了。”

**## 输出:**

悟空听到精彩的地方, 高兴得抓耳挠腮。祖师很生气, 问他: “你怎么在下面疯疯癫癫的, 不听我讲?” 悟空说: “弟子在认真地听, 听到精妙的地方, 喜不自禁, 希望师父原谅!” 祖师说: “你到洞中有多长时间了?” 悟空说: “弟子不知, 只记得经常去山后打柴, 在那里吃了七次饱桃了。” 祖师说: “那应当是七年了。”

**## 原文内容:**

{原文内容}

Please rewrite the original text of the following Chinese classics into a version that is easy for children to understand, referring to the following sample format:

**# Example:****## Original content:**

Monkey was listening to the lecture and was so happy that he scratched his ears and cheeks, smiling. He couldn't help dancing with his hands and feet. Suddenly, the master saw him and called Monkey, saying, "Why are you dancing wildly in the class instead of listening to me?" Monkey said, "I listened to the lecture sincerely. I was overwhelmed with joy when I heard the wonderful sound of the master, so I jumped up and down unconsciously. I hope the master will forgive me!" The master said, "Since you know the wonderful sound, let me ask you, how long have you been in the cave?" Monkey said, "I was ignorant and didn't know how long it had been. I only remember that there was no fire under the stove, so I often went to the back of the mountain to collect firewood. I saw a mountain of beautiful peach trees, and I ate peaches there seven times." The master said, "That mountain is called Rotten Peach Mountain. Since you have eaten seven times, it must be seven years."

**## Output:**

Monkey was so happy to hear the wonderful part that he scratched his cheeks. The master was very angry and asked him, "Why are you acting crazy down there and not listening to me?" Monkey said, "I am listening carefully. I am so happy to hear the wonderful part. I hope you will forgive me!" The master asked, "How long have you been in the cave?" Monkey said, "I don't know. I just remember that I often went to the back of the mountain to collect firewood and ate peaches seven times there." The master said, "That should be seven years."

**## Original content:**

{Original content}

---

Table 10: The prompt template for GPT-4o, GLM-4, Qwen2.5 baselines.

---

**Prompt:**

请对以下输入内容生成详细的帮助儿童理解的解释，并结合解释请将以下中国经典名著的原文改写为儿童易懂的版本，参考以下示例格式：

# 示例：

## 原文内容：

孙悟空在旁闻讲，喜得他抓耳挠腮，眉花眼笑。忍不住手之舞之，足之蹈之。忽被祖师看见，叫孙悟空道：“你在班中，怎么颠狂跃舞，不听我讲？”悟空道：“弟子诚心听讲，听到老师父妙音处，喜不自胜，故不觉作此踊跃之状。望师父恕罪！”祖师道：“你既识妙音，我且问你，你到洞中多少时了？”悟空道：“弟子本来懵懂，不知多少时节。只记得灶下无火，常去山后打柴，见一山好桃树，我在那里吃了七次饱桃矣。”祖师道：“那山唤名烂桃山。你既吃七次，想是七年了。”

## 解释：

故事里，孙悟空表现出对学习的喜爱，他听课时因为兴奋而不自觉地跳起来，表现了他很喜欢学习，并且心中充满了好奇和快乐。师父也通过询问，指出了他在学习的进步，暗示了孙悟空在师父的教导下已经学到了很多。这个情节让我们看到了孙悟空的活泼和热爱学习的精神，同时也反映了师徒之间的有趣互动。

## 输出：

孙悟空在一旁听着师父讲课，听得特别开心，忍不住搔搔头、笑得眯起了眼睛，心里美滋滋的。高兴得他开始手舞足蹈，开心地跳了起来。可是他正跳着呢，突然被师父看到了。师父说：“孙悟空，你在课堂上怎么这样乱跳乱舞，不认真听课呢？”孙悟空连忙解释说：“师父，我真的很认真听呢！只是因为您的讲课太精彩了，让我忍不住开心地跳了起来。希望师父不要怪我！”师父听了，微笑着说：“既然你听得这么认真，那我来问你一个问题。你记得自己在这里待了多久了吗？”孙悟空挠挠头说：“师父，我不知道具体的时间。只记得这里的厨房没火，我经常去山后砍柴。那儿有一棵很好的桃树，我在那里吃了七次桃子，都吃得饱饱的。”师父笑了笑，说：“那座山叫‘烂桃山’。你已经吃了七次桃子，说明你在这里学习了七年啦！”

## 原文内容：

{原文内容}

Please generate detailed explanations for the following input content to help children understand it, and based on the explanations, please rewrite the original text of the following Chinese classics into a version that is easy for children to understand:

# Example:

## Original content:

Monkey was listening to the lecture and was so happy that he scratched his ears and cheeks, smiling. He couldn't help dancing with his hands and feet. Suddenly, the master saw him and called Monkey, saying, "Why are you dancing wildly in the class instead of listening to me?" Monkey said, "I listened to the lecture sincerely. I was overwhelmed with joy when I heard the wonderful sound of the master, so I jumped up and down unconsciously. I hope the master will forgive me!" The master said, "Since you know the wonderful sound, let me ask you, how long have you been in the cave?" Monkey said, "I was ignorant and didn't know how long it had been. I only remember that there was no fire under the stove, so I often went to the back of the mountain to collect firewood. I saw a mountain of beautiful peach trees, and I ate peaches there seven times." The master said, "That mountain is called Rotten Peach Mountain. Since you have eaten seven times, it must be seven years."

## Explanation:

In the story, Monkey showed his love for learning. He jumped up unconsciously because of excitement during the class, which showed that he liked learning very much and was full of curiosity and joy. The master also pointed out his progress in learning through questioning, implying that Monkey had learned a lot under the guidance of the master. This plot allows us to see Monkey's liveliness and love of learning, and also reflects the interesting interaction between the master and the apprentice.

## Output:

Sun Wukong was listening to the master's lecture. He was so happy that he couldn't help scratching his head and smiling until his eyes narrowed. He was so happy that he began to dance and jump up happily. But when he was jumping, he was suddenly seen by the master. The master said: "Sun Wukong, why are you dancing around in class and not listening carefully?" Sun Wukong quickly explained: "Master, I am really listening carefully! It's just that your lecture is so wonderful that I can't help but jump up happily. I hope Master won't blame me!" The master smiled and said: "Since you listen so carefully, let me ask you a question. Do you remember how long you have been here?" Sun Wukong scratched his head and said: "Master, I don't know the exact time. I only remember that there is no fire in the kitchen here, so I often go to the back of the mountain to chop firewood. There is a very good peach tree there. I ate peaches there seven times and was full every time." The master smiled and said: "That mountain is called 'Rotten Peach Mountain'. You have eaten peaches seven times, which means you have been studying here for seven years!"

## Original content:

{Original content}

---

Table 11: The prompt template for EAPMT baseline.