From Style to Story: A Curriculum Learning Approach for Imitative Novel Generation

Anonymous ACL submission

Abstract

Great novels create immersive worlds with rich character arcs, well-structured plots, and nuanced writing styles. However, current novel generation methods often rely on brief, simplistic story outlines and generate details using plain, generic language. To bridge this gap, we introduce the task of *Imitative Novel Generation*, which requires the generated novels to imitate the distinctive features of the original work, including understanding character profiles and world views, predicting plausible plot developments, and writing concrete details using vivid, expressive language. To achieve this, we propose WriterAgent, a novel generation system designed to master the core aspects of literary imitative. WriterAgent is trained through a curriculum learning paradigm, progressing from low-level stylistic mastery to high-level narrative coherence. Its key tasks include language style learning, character modeling, plot planning, and stylish writing, ensuring comprehensive narrative control. To support this, WriterAgent leverages the WriterLoRA framework, an extension of LoRA with hierarchical and cumulative task-specific modules, each specializing in a different narrative aspect. We evaluate WriterAgent on multilingual classics like *Harry Potter* and *Dream of the Red Chamber*, demonstrating its superiority over baselines in capturing the target author's settings, character dynamics, and writing style to produce coherent, faithful narratives. We hope this work inspires literary creativity in NLP: WriteAgent.

1 Introduction

Novels create rich, immersive worlds with intricate plots and distinct styles, captivating readers through complex storytelling (Bai et al., 2024). A significant amount of research (Ammanabrolu et al., 2020; Yao et al., 2019) has proposed new model architectures to improve story generation. With the emergence of LLMs, recent efforts have shifted towards improved prompt-based techniques (Wang et al.; Han et al., 2024). For example, Ma et al. (2024) proposed modular premise synthesis, providing concrete information such as background, persona, and theme to guide the generation process. While these methods have enhanced novel generation performance (Hu et al., 2024), they fall short in capturing the irreplaceable qualities of real-world literary classics: engaging plots, vivid characters, and distinctive language that immerse readers in complex and authentic storytelling.

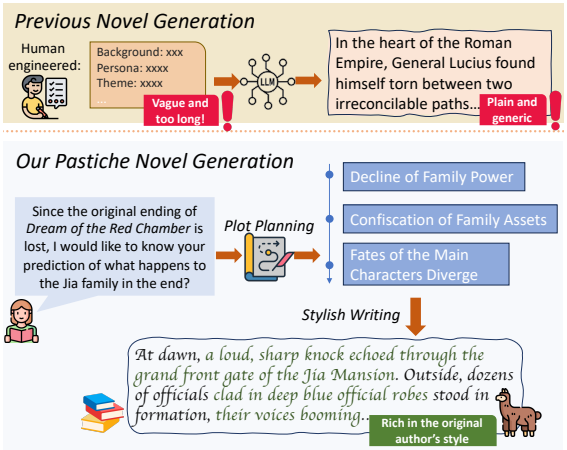

Figure 1: Comparison of traditional story generation and our imitative novel generation, showcasing enhanced narrative depth, character development, and stylistic fidelity.

Hence, in this work, we propose the *Imitative Novel Generation* task, which aims to generate novels that faithfully emulate the original author's style and narrative depth. This task presents two key challenges: (1) *plot planning* that aligns with the novel's established worldview and character dynamics, and (2) *stylish writing* that produces narrative text reflecting the target author's personalized writing style. As illustrated in Figure 1, for instance, given the rich context of Dream of the Red Chamber, with its intricate interpersonal conflicts

and lavish lifestyles, the model should predict significant plot point outcomes, such as the eventual downfall of the Jia family. Additionally, the model must accurately reproduce the linguistic and stylistic features of the original text, including evocative phrases like "rolled up their sleeves" and "please issue the decree", which reflect the author's unique writing style.

To address these challenges, we propose WriterAgent, a novel generation model designed to emulate a target author's writing style. The model is trained in sequence on four core tasks to master key narrative and stylistic elements of the author's work. This sequential training follows the natural writing process, from conceptual elements like language style and world-building to high-level plotting and fine details: 1) *Language Style Learning*: Teaching the LLM to capture the author's distinctive writing style through tasks like next-word prediction, ensuring consistent character voices. 2) *World Building*: Guiding the model to introduce characters and define their relationships, constructing an interconnected world aligned with the original narrative. 3) *Plot Planning*: Enabling the model to generate coherent plotlines that evolve character arcs in line with the story's structure. 4) *Stylish Writing*: Enhancing descriptive details of settings, character interactions, and events, ensuring immersive storytelling that reflects the author's tone and depth. To train WriterAgent efficiently, we propose WriterLoRA, an extension of the Low-Rank Adaptation (LoRA) (Hu et al., 2021). The original LoRA A and B matrices act as a general expert, preserving the original text's style as a reference. We augment this with specialized B matrices dedicated to specific aspects such as world-building, plot, and detail writing, trained sequentially in a curriculum-based approach from simple to complex tasks.

We evaluated WriterAgent on the English *Harry Potter series* and the Chinese *Dream of the Red Chamber*. Given the absence of prior work on the personalized novel generation task, we developed a set of automatic metrics to assess writing style, including language style, expression methods, and sentence complexity, and plot development, covering the story mainline, character behavior, and emotions. Both automatic evaluations and human annotations demonstrate that WriterAgent effectively captures the target author's style and constructs coherent, engaging narratives.

Our main contributions are as follows: (1) We introduce the Imitative Novel Generation task, which aims to generate novels that mimic a target author's style, narrative structure, and character development. (2) We develop WriterAgent, an LLM with a WriterLoRA structure, trained to learn character profiles, predict future plotlines, and reconstruct full stories, enabling consistent and context-aware storytelling. (3) We demonstrate that WriterAgent outperforms baseline models in mimicking multilingual classics such as Harry Potter, which highlights new possibilities for literary creativity and personalized storytelling.

2 Related Work

Story Generation. The task of long-form story generation has received significant attention in recent years due to advancements in LLMs. Early approaches primarily focused on developing new modules to enhance narrative coherence and consistency. For example, Ammanabrolu et al. (2020); Fan et al. (2019, 2018) leveraged graph structures to organize events more effectively and improve narrative consistency. Another example is Peng et al. (2018), which introduced an interface for human-computer interaction to generate personalized stories and applied it to RNN-based models for controlling story endings and storylines. However, these methods often struggled to maintain coherence and consistency over extended sequences. More recently, prompt engineering techniques have been adopted to tap into the generative power of LLMs (Giray, 2023). For instance, Han et al. (2024) proposed a director-actor agent collaboration framework for controllable and interactive drama script generation, while Huang et al. (2023) explored dynamic beam sizing and affective reranking to generate engaging narratives.

Despite these advancements, existing methods lack a framework to emulate complex narratives and distinct authorial styles, essential for real-world long-form novel writing.

Parameter-Efficient Fine-Tuning. Parameter-efficient fine-tuning (PEFT) (He et al.) reduces the computational costs of fine-tuning LLMs by introducing additional modules, avoiding direct updates to the large-scale pretrained weights. Adapters (Houlsby et al., 2019) insert extra feature transformations between model blocks, while prefix tuning (Li and Liang, 2021) optimizes parameters through learnable prefixed embeddings without modifying the pretrained weights. More recently, LoRA (Hu et al., 2021) introduces low-rank

matrix decomposition to efficiently fine-tune models by updating small, trainable parameter matrices. Extensions of LoRA have further explored task-specific adapters, enabling specialization in distinct aspects of text generation (Zhou et al., 2023; Li et al., 2024; Luo et al., 2024; Chen et al., 2024). Unlike these works, our WriterLoRA combines advantages of a mixture-of-LoRA architecture but is specifically designed to handle the challenges of imitative novel generation.

Stylish Novel Generation Stylish Novel Generation is a task that focuses on generating narratives with coherent plots, consistent characters, and distinctive stylistic elements reflective of specific authors or genres. With the emergence of LLMs, recent research has shifted towards improving prompt-based techniques (Luo et al., 2023; Ye et al., 2022; Zhu et al., 2020; Tao et al., 2024; Ye et al., 2022). Ye et al. (2022) proposed using style examples as prompts to guide LLMs in generating text in specific styles. PLoRA (Zhang et al., 2024) (Personalized LoRA) is a plug-and-play framework for human-computer interactive text understanding, enabling efficient adaptation to users' language styles and effectively capturing personalized writing. MultiLoRA (Wang et al., 2023b) addresses parameter conflicts and knowledge transfer problems between different tasks in multi-task learning.

3 Problem Formulation

We begin by introducing the notations and key concepts for the task of personalized long-form novel generation. Formally, the training dataset for a novel is hierarchically structured into character profiles, plots, and words. Each character profile C_i consists of the i-th character's name and a detailed description, including key traits and relationships with other characters. The narrative text is divided into segments of words $\{x_1^i, x_2^i, \ldots, x_n^i\}$, each corresponding to an individual plot summary P_j, where j represents the plot index within the chapter. These plots $\{P_j\}$ capture the key story developments within their respective text segments.

The task involves two core subtasks: (1) *plot prediction*, where the model predicts the next plot point $\hat{P}_t$ based on previous plot points $\{P_1, \ldots, P_{t-1}\}$ and ensures logical consistency by comparing it to the ground truth plot point P_t; (2) *stylish writing*, where the model generates a word sequence $\hat{\mathcal{Y}}_t = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$ from the predicted plot point $\hat{P}_t$ and ensures coherence and stylistic alignment by comparing it to the ground truth text $\mathcal{Y}_t = \{x_1, x_2, \ldots, x_n\}$.

4 Method

In this section, we first introduce the vanilla LoRA, then present our adapted WriterLoRA built on it, along with the overall WriterAgent framework, as shown in Figure 2.

4.1 Preliminaries

LoRA fine-tunes LLMs efficiently by adding trainable low-rank matrices instead of updating all parameters, reducing computational cost. Concretely, a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ is updated using a low-rank decomposition $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with $r \ll \min(d, k)$. Here, B and A are the trainable low-rank matrices, and r represents the rank of the decomposition. During training, W_0 is frozen and does not receive gradient updates, while A and B are optimized. Given an input $x \in \mathbb{R}^k$, the forward pass through the modified weight matrix is:

$$h' = W_0 x + \Delta W x = W_0 x + BAx$$

Here, A serves as an encoder-like transformation, mapping $x \in \mathbb{R}^k$ into a lower-dimensional representation $\mathbf{z} \in \mathbb{R}^r$, while B acts as a decoder-like transformation, projecting $\mathbf{z}$ back into output space.

4.2 Curriculum Learning Tasks

Our training tasks are inspired by real-world observations of how authors create novels. Typically, an author begins by determining the work's style, defining key characters and their traits, outlining interactions, and ultimately developing a complete narrative based on these plots. Following this natural progression, our model training tasks are designed to emulate this process. First, the model is pretrained on a next-word prediction task for learning the writing style. Then, we design three downstream fine-tuning tasks:

World-Building Learning In this task, the model maps character names, represented as $\mathcal{N} = \{N_1, N_2, \ldots, N_n\}$, to their corresponding profiles $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$. For each character name N_i, the model generates a detailed profile C_i, including the character's attributes, relationships, and role in the narrative. This process is formulated as:

$$\hat{C}_i = f_{\text{world}}(N_i).$$

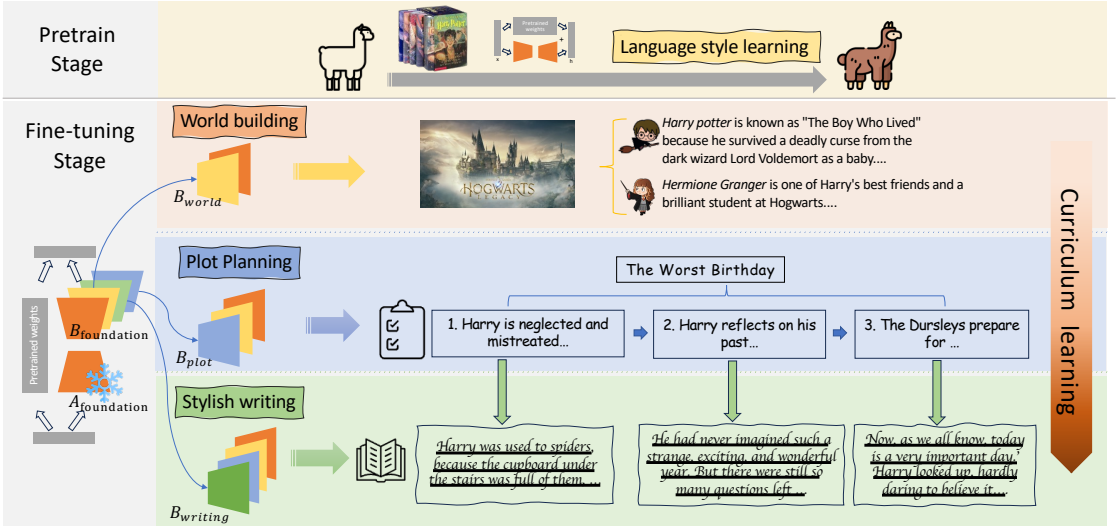

Figure 2: The entire training process can be divided into two parts: the pretraining phase and the fine-tuning phase. During the fine-tuning phase, tasks are divided into three stages of increasing complexity: world-building learning, plot structure learning, and stylish writing learning. These stages are integrated using curriculum learning.

This task ensures that the model develops a comprehensive understanding of the story world, forming the foundation for subsequent tasks.

Plot Structure Learning Building on the character profiles established in the previous task, this step focuses on predicting narrative progression. The model is trained to generate the next plot point $\hat{P}_t$ based on the most recent N_p predicted plot points:

$$\hat{P}_t = f_{plot}(P_{t-N_p}, ..., P_{t-1}).$$

Using the N_p most recent plots, the model ensures narrative continuity and coherence within the story's timeline.

Stylish Writing Following plot prediction, the Stylish Writing task involves generating the narrative text based on the predicted plot $\hat{P}_t$.The model produces a sequence of words $\hat{\mathcal{Y}}_t = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$ as: $\hat{\mathcal{Y}}_t = f_{\text{writing}}(\hat{P}_t)$. The generated text is compared with the ground truth $\mathcal{Y}_t = \{x_1, x_2, \ldots, x_n\}$ for plot coherence and stylistic alignment.

4.3 WriterLoRA

A straightforward approach to train an LLM involves sequentially training it on the above tasks in a parameter-efficient manner using LoRA. However, this method may fail to optimize task-specific performance while maintaining cross-task synergy. To address this limitation, we propose WriterLoRA, a structured multi-task learning framework that maximizes efficiency through shared components while ensuring task-specific specialization.

Shared Foundation Initially, the model undergoes next-word prediction training on the entire corpus, using a pair of LoRA matrices, $A_{Fdn.}$ and $B_{Fdn.}$. Here, $A_{Fdn.}$ serves as the shared matrix across all tasks, while $B_{Fdn.}$ collaborates with task-specific B-matrices. The motivation for sharing the A-matrix is to enhance learning efficiency and task synergy. As a compression matrix, $A_{Fdn.}$ extracts core representations for language understanding and generation, ensuring consistency across tasks while reducing redundancy and improving cross-task transfer learning. After pretraining, $A_{Fdn.}$ is fixed to maintain stable representations and prevent catastrophic forgetting.

Task-Specific Adaptations The weight update process originally defined in Equation 4.1 is extended to account for task-specific requirements, beginning with language style learning and progressing through subsequent tasks. The weight update process defined in Equation 4.1 is extended to meet task-specific requirements, starting with language style learning and progressing through world-building, plot structure, and stylish writing tasks. During language style learning, the model uses $B_{Fdn.}$ and $A_{Fdn.}$ for next-word prediction, capturing linguistic and stylistic nuances. For world-building, the model maps character names to profiles using B_{world} alongside $B_{Fdn.}$ to define characters and relationships. In plot structure learn-

ing, both B_{plot} and B_{world} are used, where B_{plot} controls narrative flow and B_{world} ensures consistency. For stylish writing, $B_{writing}$ integrates with $B_{Fdn.}$, B_{world}, and B_{plot} to generate stylistic text.

The generalized weight update is:

$$h' = W_0 x + \Sigma_t \alpha_t B_t A_{Fdn.} x$$

The task-specific weights α_t are computed using:

$$\alpha_t = \frac{\exp(w_t)}{\sum_{t'} \exp(w_{t'})},$$

where $w_t = 1$ for the active task and $w_{t'} = 0$ for others. This ensures the active task's matrix B_t dominates, while others provide auxiliary contributions, enabling efficient task adaptation and knowledge sharing.

5 Experimental Setup

5.1 Dataset

We selected two renowned literary works for our dataset: the classic Chinese novel Dream of the Red Chamberand the Harry Potter series, chosen for their rich narratives and literary significance. For Dream of the Red Chamber, the first 80 chapters were used for training and the last 40 for testing. Similarly, the first six Harry Potter books were used for training, with the final book for testing.

In addition to the primary text, our dataset incorporates supplementary information to enhance its utility. First, we collected detailed, human-written *introductions for the main characters*. These character profiles provide valuable context for tasks such as role-playing and characterization. Specifically, profiles for Dream of the Red Chamber were sourced from Sohu website[1], while those for Harry Potter were obtained from Wikipedia[2]. Secondly, we used GPT-4 to segment the text into sections and generate concise *plot summaries* for each section. A section is smaller than a chapter but longer than a paragraph, with the division based on self-contained and relatively complete narrative events. These summaries offer structured descriptions of key narrative developments, supporting tasks such as plot-aware content generation.

5.2 Comparison Methods

We selected ChatGLM and Qwen as backbone models for evaluating Chinese language performance, and Llama3 for English tasks. The key backbones include: (1) *Qwen2-7B-Instruct* (Bai et al., 2023): An instruction-tuned variant of Qwen2-7B, optimized for handling complex queries and interactive tasks. (2) *ChatGLM2-6B* (Zhang et al., 2023): A bilingual LLM optimized for both English and Chinese languages. (3) *Llama3-8B* (Grattafiori et al., 2024): Excels in reasoning, creative writing, and coding tasks. We implemented four representative approaches across the three open-source LLMs above: (1) Prompt-based method (Luo et al., 2023), which guides pre-trained language models to generate text in specific styles using carefully crafted prompts. (2) Pre-trained Language Model (PLM), representing the traditional full-parameter fine-tuning technique, where we directly fine-tune the pre-trained language model using works from the target author. (3) PLoRA (Zhang et al., 2024), a parameter-efficient fine-tuning framework that learns specialized low-rank adapters for each target author, which can be dynamically loaded into the base model as plugins. (4) MultiLoRA (Wang et al., 2023b), which proposes a democratized LoRA framework designed to improve parameter sharing and task specialization in multi-task learning.

5.3 Implementation Details

We implemented our experiments using PyTorch and conducted them on an NVIDIA A100 GPU. All models were configured with a maximum sequence length and cutoff length of 2048 tokens. We trained the model for 3 epochs using the AdamW optimizer and BF16 precision for efficiency. The rank r was set to 8 for both PLoRA and each LoRA module within Multi-LoRA. To simulate a larger batch size, we set the batch size to 1 and applied gradient accumulation over 8 steps. More implementation details are in Technical Appendix.

5.4 Evaluation Metrics

To assess the quality of generated novels, we employ both traditional evaluation methods and advanced aspect-based metrics. Our evaluation framework consists of two components: the assessment of the final novel generation and the evaluation of its intermediate subtask, plot planning.

Novel Generation Evaluation: Traditional metrics like ROUGE-N and ROUGE-L (Lin, 2004) assess content coverage, fluency, and structural alignment. ROUGE-N measures n-gram overlap, while ROUGE-L evaluates the longest common subse-

[1] https://www.sohu.com/a/773246248_121948389

[2] https://www.wikiwand.com/en/List_of_Harry_Potter_characters

Dataset	Model	Traditional Metrics			Advanced Metrics					
		ROUGE-1	ROUGE-2	ROUGE-L	LS	EX	SLC	SM	CBM	EM
Dream of the Red Chamber	Prompt-based(Qwen)	17.43	3.70	17.57	2.80	2.72	2.96	2.04	2.28	2.24
	PLM(Qwen)	20.98	5.18	18.83	2.79	2.68	3.07	2.10	2.57	2.53
	PLoRA(Qwen)	31.30	9.55	25.58	3.14	2.80	2.78	2.14	2.21	2.36
	Muti-LoRA(Qwen)	23.54	6.06	16.30	3.00	2.98	3.02	2.20	2.54	2.58
	WriterAgent(Qwen)	**34.81**	**9.73**	**24.11**	**3.58**	**3.25**	**3.27**	**3.19**	**3.15**	**3.12**
	Prompt-based(ChatGLM)	15.87	2.32	13.18	2.08	1.96	1.96	2.19	1.87	1.79
	PLM(ChatGLM)	15.32	2.28	13.46	2.37	2.58	2.43	2.23	2.12	2.05
	PLoRA(ChatGLM)	14.58	2.21	12.53	2.54	2.60	2.87	2.26	2.30	2.23
	Multi-LoRA(ChatGLM)	15.89	2.98	16.00	2.20	2.13	2.31	2.46	1.95	1.81
	WriterAgent(ChatGLM)	**16.29**	**3.21**	**16.01**	**3.03**	**2.96**	**2.94**	**2.59**	**2.41**	**2.40**
Harry Potter	Prompt-based(Llama)	25.17	3.29	16.92	2.55	2.64	3.03	3.03	3.04	2.85
	PLM(Llama)	26.30	4.29	18.71	2.42	2.32	2.95	2.20	2.31	2.26
	PLora(Llama)	26.18	6.35	18.97	2.56	2.53	2.95	2.53	2.46	2.45
	Multi-LoRA(Llama)	25.43	7.93	20.23	2.68	2.58	3.15	3.23	2.95	2.53
	WriterAgent(Llama)	**29.24**	**8.79**	**23.91**	**3.05**	**2.86**	**3.29**	**3.54**	**3.03**	2.75

Table 1: Performance comparison of baseline models and our WriterAgent on two classic literary datasets, Harry Potter (English) and Dream of the Red Chamber (Chinese), evaluated for personalized long-form novel generation. The scores are represented as follows: best and second. Numbers in **bold** mean that the improvement to the best baseline is statistically significant (a two-tailed paired t-test with p-value <0.01).

Dataset	Model	Advanced Metrics		
		SM	CBM	EM
Dream of the Red Chamber	Prompt-based(Qwen)	1.89	2.17	2.18
	PLM(Qwen)	1.94	2.13	2.26
	PLoRA(Qwen)	1.97	2.14	2.29
	Multi-LoRA(Qwen)	2.01	2.23	2.35
	WriterAgent(Qwen)	**2.14**	**2.54**	**2.78**
	Prompt-based(ChatGLM)	1.78	1.56	1.43
	PLM(ChatGLM)	1.86	1.75	1.53
	PLoRA(ChatGLM)	1.80	1.79	1.60
	Milti-LoRA(ChatGLM)	1.88	1.83	1.72
	WriterAgent(ChatGLM)	**1.86**	**1.91**	**1.82**
Harry Potter	Prompt-based(Llama)	1.97	1.86	1.49
	PLM(Llama)	2.06	1.95	1.56
	PLoRA(Llama)	2.03	2.01	1.76
	Multi-LoRA(Llama)	2.08	2.11	1.85
	WriterAgent(Llama)	**2.10**	**2.13**	**1.87**

Table 2: Scores for plot prediction ability, with best highlighting and **bold** denoting statistically significant improvement (p-value <0.01).

quence to capture broader structural coherence.

Beyond these traditional measures, we introduce six advanced aspect-based metrics to evaluate both ***writing style*** and ***plot development***. These aspects are derived from human observations of a randomly selected set of 100 samples. Writing style evaluation includes: *Language Style (LS)*, which assesses similarity in vocabulary choices, sentence structures, tone, and voice; *Expression Methods (EX)*, which evaluates consistency in emotional expression, descriptive techniques, and use of metaphors; *Sentence Length and Complexity (SLC)*, which compares sentence structures, their complexity, and overall paragraph organization. Plot development evaluation includes: *Story Mainline (SM)*, which measures the alignment and coherence of the central plotline; *Character Behavior and Motivation (CBM)*, which examines whether character actions and motivations are logically consistent with the story's progression; *Emotions (EM)*, which evaluates the emotional flow and conflict dynamics within the text. Given the high correlation of GPT-4o with human judgments, particularly for creative NLG tasks (Wang et al., 2023a), we adopt both GPT-4o-based scoring and human evaluation. GPT-4o rates each generated text from 1 to 5, based on comparison with the ground truth. We used a few-shot prompt to evaluate the output results. Detailed prompts can be found in Technical Appendix.

Plot Planning Evaluation: As an intermediate step crucial to creating a well-structured and engaging narrative, plot planning is evaluated using a subset of advanced aspect-based metrics centered on plot coherence. Specifically, *SM*, *CBM*, and *EPM* are applied. The evaluation follows the same GPT-4o-based scoring setup as described earlier.

6 Experimental Results

6.1 Main Results

Table 1 and Table 2 present the experimental results of Prompt-based, PLM, Plora, Multi-LoRA, and WriterAgent methods across two datasets.

Firstly, *evaluation results reveal that different metrics capture complementary aspects of writing quality.* ROUGE favors models like PLoRA for content coverage and structural alignment, while

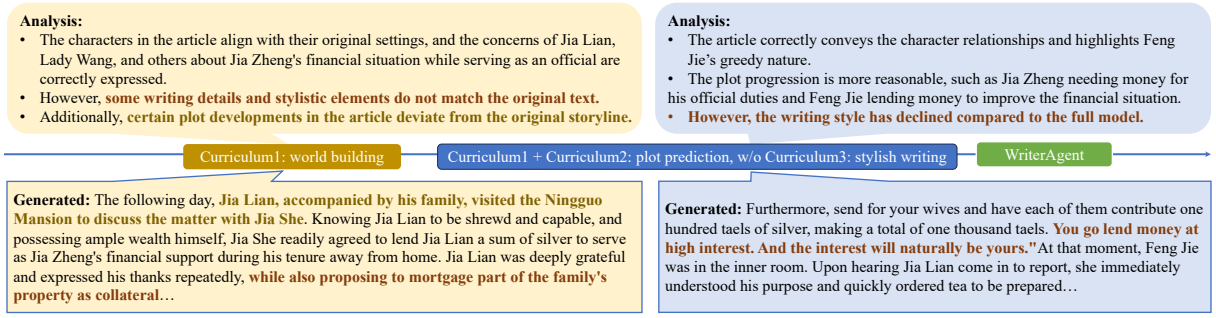

Figure 3: Demonstration of the model's stepwise learning on Dream of the Red Chamber: from curriculum 1 to curriculum 1 & 2. The text colors indicate the corresponding problems. Chinese version is in Technical Appendix.

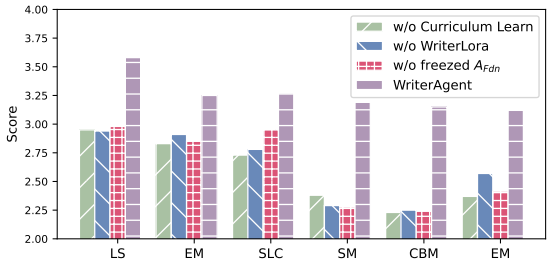

Figure 4: Ablation performance of WriterAgent.

aspect-based metrics highlight Pretrain and Multi-LoRA's strengths in plot coherence and emotional tone. This underscores the need for diverse evaluation frameworks to fully assess model performance. Secondly, Multi-LoRA, the best-performing baseline, benefits from specialized multi-task training, producing coherent plots and well-aligned emotional tones, but *its language style is not well-preserved due to the loss of pre-trained textual style during multi-task learning*. Finally, our proposed *WriterAgent consistently outperforms all baselines across datasets and metrics*, achieving significant improvements in plot coherence, character development, emotional depth, and overall narrative quality. Additionally, it addresses Multi-LoRA's language style limitation by freezing the matrix A, thereby preserving the pre-trained writing style and ensuring stylistic consistency in the generated narratives.

6.2 Ablation Study

We conduct an ablation study on Dream of the Red Chamber as shown in Figure 4. Removing the curriculum learning strategy led to a significant performance drop across all metrics, showing that learning all tasks simultaneously without a structured progression reduces the model's effectiveness. Similarly, removing WriterLoRA with only one LoRA weakened task-specific adaptation, reducing the model's ability to handle diverse constraints. Finetuning the foundation matrix A_{fdn}. instead of keeping it fixed led to performance degradation, showing the importance of a stable foundation for consistent improvements. We further analyze the curriculum learning sequence in Technical Appendix to validate our method.

6.3 Analysis of Curriculum Learning

Our curriculum learning consists of three stages: world building, plot prediction, and stylish writing. Figure 3 presents a case study showing the model's output after each stage. After curriculum 1 (world building), the generated text demonstrates characters and traits that closely align with the original work's character settings. However, the plot structure and textual details remain significantly different from the original. After curriculum 2 (plot prediction), the output retains the vivid character traits learned in the first stage. For instance, Wang Xifeng is depicted as highly fond of wealth and enthusiastic about managing the household, consistent with her original portrayal. However, since the plot construction relies on modern vernacular organization, the model's ability to capture writing details and mimic the original literary style is insufficient. Finally, after incorporating the final step of curriculum learning, the overall performance improves significantly in terms of word choice, plot structure, and style, as shown in Figure 5. We present Chinese output in Technical Appendix.

6.4 Human Evaluation

In addition to automatic evaluation, we conducted a human evaluation with two PhD annotators on the Dream of the Red Chamber dataset, both native speakers with strong literary backgrounds. We conducted a survey with 100 randomly selected test questions, each featuring a plot summary and five options from baseline models and WriterA-

Figure 5: Comparison of reference and generated texts from the baseline and our WriterAgent. We highlight the weaknesses of the baseline model and the strengths of our approach. Some linguistic nuances may be lost in translation; see the Appendix (Chinese) for accuracy.

gent. Experts chose the option most likely from the original text. We averaged their choices to compute hit rates per model, shown in a pie chart in Technical Appendix. The results indicate that our model was preferred in 73.1% of cases, achieving a Kappa score of 0.78, which signifies substantial agreement.

In Figure 5, we present a case study of the generated novel text and categorize common errors in baseline models. First, ***character relationship*** errors frequently occur, leading to misinterpretations of key familial roles. For instance, baselines such as Qwen and LoRA misidentify Lady Wang as Jia Lian's mother, distorting the intricate family dynamics in Dream of the Red Chamber. Second, ***plot inconsistencies*** arise, where baselines shift core narrative elements. For example, PLoRA and Multi-LoRA prioritize financial matters over official duties, altering the intended thematic focus and disrupting the novel's structured storytelling. Third, ***stylistic deviations are prevalent***, as baselines modernize language, losing conciseness and tradition; for example, Qwen adds "hope of the Jia family", disrupting the reserved tone. Additionally, errors in emotional tone make dialogues overly sentimental and expressive, diverging from the novel's subtle and restrained emotional depth. In contrast, our model maintains plot consistency, character dynamics, classical language, and restrained emotion, aligning with Lady Wang's focus on political risks and family reputation. More analysis is in Technical Appendix.

7 Conclusion

We introduce WriterAgent, an LLM designed for novel generation by learning character profiles, predicting plotlines, and reconstructing full stories. At its core, WriterLoRA employs LoRA modules and curriculum learning to progressively master narrative elements, enhancing coherence and stylistic consistency. Experiments on two datasets in different languages show that WriterAgent outperforms baselines in capturing complex settings and character dynamics. In future work, we will explore incorporating reader feedback and interactive story generation, enabling dynamic adaptation to audience preferences.

Limitations

Despite the effectiveness of WriterAgent in generating imitative novels, several limitations remain. First, while determining the plot before drafting is a natural part of the writing process, plots typically exist in the author's mind rather than as explicit, structured data. As a result, we need to manually construct plot datasets, which introduces potential biases and may not fully capture the organic, evolving nature of storytelling. Second, automatic evaluation heavily relies on LLMs' ability to understand an author's style, yet current LLMs do not possess comprehensive knowledge of all literary nuances. To mitigate this, we incorporated few-shot prompting in our evaluation, allowing the model to refine its understanding of specific authors. However, this approach is still limited and cannot fully replace human judgment, as LLM-based evaluation may overlook deeper stylistic elements and narrative coherence that human readers naturally perceive. These limitations highlight the challenges in literary imitative generation and evaluation, underscoring the need for future research on more refined plot modeling and improved evaluation frameworks that better align with human literary perception.

Ethical Considerations The development of WriterAgent, aimed at generating long-form literary imitatives that emulate the styles, themes, and narrative elements of iconic works such as Harry Potter and Dream of the Red Chamber, raises important ethical questions. While our goal is to explore stylistic modeling for literary creativity and assistive generation, we acknowledge that the close imitation of authorial voices, character behaviors, and narrative arcs requires thoughtful reflection on authenticity, ownership, and appropriate use.

Potential ethical risks associated with imitative novel generation include: (1) Authorship Confusion and Misattribution: Generated content that closely mimics the tone and style of a well-known author may be mistaken for original material, leading to misattribution in educational, fan, or commercial contexts if not clearly identified as AI-generated. (2) Intellectual Property and Copyright Violation: Although some works used for training are in the public domain, others—like Harry Potter—remain under copyright. High-fidelity imitation of such works may risk infringing on intellectual property rights, particularly if outputs are used beyond research or educational purposes. (3) Cultural and Literary Misrepresentation: Iconic texts often embody complex cultural, historical, or political meanings. Inaccurate, superficial, or decontextualized reproduction of such elements may trivialize or distort the source material, misrepresenting its original intent or cultural significance. (4) Overreliance and Creativity Dilution: If widely adopted in writing or education contexts, systems like WriterAgent could encourage reliance on AI-generated content at the expense of human creativity, critical interpretation, and original authorship.

To mitigate these risks, we implement the following ethical safeguards: (1) Clear Disclosure of AI-Generated Content: All outputs produced by WriterAgent are explicitly labeled as machine-generated. This ensures transparency and prevents confusion between AI outputs and authentic literary works. (2) Restricted Use Scope: We intend WriterAgent primarily for research, literary analysis, and creative exploration in controlled settings. We discourage any use that may lead to plagiarism, unauthorized derivative works, or the commercial exploitation of copyrighted styles or content. (3) Ethical Data Use and Copyright Awareness: Public domain texts are used for training when possible. In cases involving copyrighted works, we follow fair use guidelines strictly for research purposes, and outputs are not released for public or commercial redistribution. (4) Preservation of Literary Integrity: We emphasize careful design of generation tasks and evaluation criteria to avoid reductive or disrespectful reinterpretations of culturally significant literature. Where applicable, we incorporate expert feedback and cross-lingual validation to maintain fidelity to the original spirit and context.

Our ethical stance is informed by prior research on AI-generated persona simulation (Shanahan et al., 2023), and we align with general best practices in role-based LLMs. We adhere to three key principles: respect for the original author, transparency of AI generation, and restriction of use within appropriate boundaries.

References

Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark O. Riedl. 2020. Bringing stories alive: Generating interactive fiction worlds. *CoRR*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *Preprint*, arXiv:2408.07055.

Xiuying Chen, Mingzhe Li, Shen Gao, Xin Cheng, Qingqing Zhu, Rui Yan, Xin Gao, and Xiangliang Zhang. 2024. Flexible and adaptable summarization via expertise separation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2018–2027.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proc. of ACL*.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2019. Strategies for structuring story generation. *CoRR*.

Louie Giray. 2023. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, pages 2629–2633.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 93 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation. In *Proc. of ACL*, pages 1607–1619.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *Proc. of ICLR*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proc. of ICML*, pages 2790–2799.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *Proc. of ICLR*.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *Preprint*, arXiv:2410.14255.

Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen, and Snigdha Chaturvedi. 2023. Affective and dynamic beam search for story generation. In *Proc. of EMNLP Findings*, pages 11792–11806.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhengmao Ye, Zhiyuan Cheng, Yinghao Tang, Yan Zhang, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts. *Preprint*, arXiv:2404.15159.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *ACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Guoqing Luo, Yu Tong Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. *Preprint*, arXiv:2301.11997.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *Preprint*, arXiv:2402.12851.

Yan Ma, Yu Qiao, and Pengfei Liu. 2024. Mops: Modular story premise synthesis for open-ended automatic story generation. *ACL*.

Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, pages 493–498.

Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. 2024. Cat-llm: Prompting large language models with text style definition for chinese article-style transfer. *Preprint*, arXiv:2401.05707.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *Preprint*, arXiv:2303.04048.

Yichen Wang, Kevin Yang, Xiaoming Liu, and Dan Klein. Improving pacing in long-form story planning. In *Proc. of EMNLP*.

Yiming Wang, Yu Lin, Xiaodong Zeng, and Guannan Zhang. 2023b. Multilora: Democratizing lora for better multi-task learning. *Preprint*, arXiv:2311.11501.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proc. of AAAI*, pages 7378–7385.

Jiacheng Ye, Jiahui Gao, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Progen: Progressive zero-shot dataset generation via in-context feedback. *Preprint*, arXiv:2210.12329.

Xiyuan Zhang, Xinyue Zhang, and Ying Yu. 2023. Chatglm-6b fine-tuning for cultural and creative products advertising words. In *2023 International Conference on Culture-Oriented Science and Technology (CoST)*, pages 291–295.

You Zhang, Jiaxin Wang, Lung-Cheng Yu, Dongkuan Xu, and Xiang Zhang. 2024. Personalized lora for human-centered text understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19588–19596.

Juexiao Zhou, Xiuying Chen, and Xin Gao. 2023. Path to medical agi: Unify domain-specific medical llms with the lowest cost. *medRxiv*, pages 2023–06.

Anna Zhu, Xiongbo Lu, Xiang Bai, Seiichi Uchida, Brian Kenji Iwana, and Shengwu Xiong. 2020. Few-shot text style transfer via deep feature similarity. *IEEE Transactions on Image Processing*, 29:6932–6946.

A Implementation Details

We implemented our experiments using PyTorch and conducted them on an NVIDIA A100 GPU. All models were configured with a maximum sequence length and cutoff length of 2048 tokens. We trained the model for 3 epochs using the AdamW optimizer and BF16 precision for efficiency. The learning rate was set to 1.0e-4 and scheduled using a cosine decay strategy with a warmup ratio of 0.1. A dropout rate of 0.05 was applied to prevent overfitting. The rank r was set to 8 for both LoRA and each LoRA module within Multi-LoRA.

The training progress was monitored every 10 steps, with checkpoints saved every 500 steps to enable recovery and evaluation. Loss curves were plotted to track convergence, with the output directory overwritten during updates to maintain consistency. To improve efficiency, data preprocessing utilized 8 parallel workers. During training, we set the batch size to 1 and used gradient accumulation over 8 steps to simulate a larger batch size. The number of samples was limited to 1000 to control training time. The training dataset followed the Alpaca format, and for the multi-task model, a *task_id* field was added to the data to classify tasks during training.

We used the AdamW (Loshchilov, 2017) optimizer for training, which decouples weight decay from gradient updates, improving both training stability and generalization. This makes it particularly effective for large-scale models, such as Transformers, by avoiding overfitting and enhancing optimization efficiency.

The version of the Transformers library was chosen to match the architectures of the models being trained. For the ChatGLM2-6B model, which adopts a GLM (General Language Model) architecture with bidirectional and autoregressive training, version 4.30.2 of Transformers was used. However, for the Llama 3-8B and Qwen2-7B-Instruct models, the recommended version was 4.45.0 to ensure compatibility and optimal performance.

B Case Study of Harry Potter

In this case study, we evaluate different models on their ability to generate coherent plots and novels and maintain consistency with the world and character personalities of the Harry Potter series.

As shown in Figure 6, the baseline models and our WriterAgent are tested on their performance in predicting plot progressions that align with prior story context. The results reveal significant differences in coherence, character consistency, and adherence to the magical world's logic. The Base Llama model frequently generates plots disconnected from previous events, with characters behaving inconsistently with their established traits. For example, Harry's decisions often lack continuity and conflict with his determined, courageous nature in the novels. The PLoRA model improves slightly but suffers from repetitive content and omits key world-building knowledge, such as the importance of destroying Horcruxes to defeat Voldemort. By contrast, WriterAgent generates coherent plot progressions where Harry's actions align with his character and successfully weaves in critical elements of the magical world, like his mission to destroy Horcruxes.

Figure 7 further highlights differences in the generated text. The base Llama model produces incorrect content, including the claim that "Voldemort's plan was to use Harry's blood to gain immortality", contradicting the established magical rule that his blood was only used for resurrection. The Llama model also introduces vague phrases like "leaving Harry to continue his journey", failing to capture the tense, high-stakes atmosphere of the final battle. The PLoRA model depicts the Dursley family as comforting Harry, which contradicts their antagonistic relationship in the novels, highlighting its limited understanding of character dynamics. The Mutil-LoRA model generates overly extended content, introducing numerous ghost characters at Hogwarts and shifting the scene to a lakeside setting. While this adds richness, it strays from the original narrative's focus. In contrast, WriterAgent accurately identifies Harry's parents as the key figures providing emotional support, portraying their encouragement in a way that is faithful to the original story. The generated text reflects the original tone and concise writing style while capturing Harry's inner strength. By producing contextually appropriate, emotionally resonant, and character-consistent content, WriterAgent demonstrates a clear advantage in simulating deep personas and generating coherent narratives.

C Case Study of Red Chamber

In Figure 8, we show a case of *generated plot*, and Chinese version in Figure 9. The QWen model continues the story directly, overlooking potential objections from the Jia family about Xichun becom-

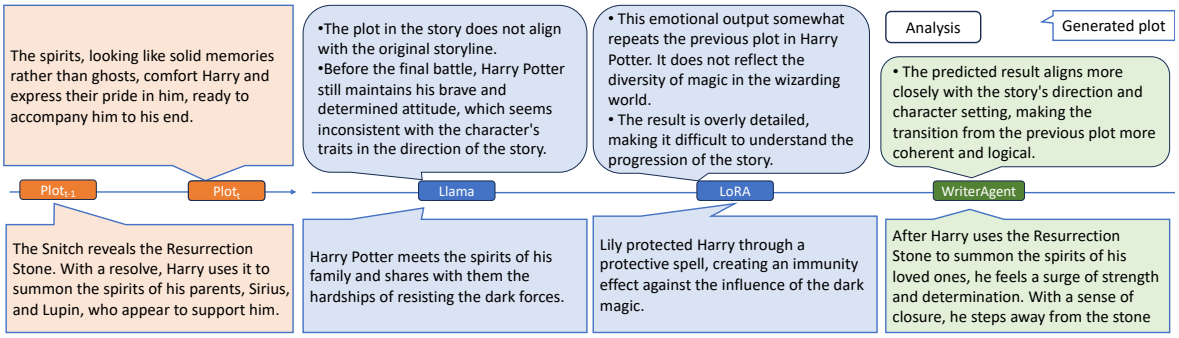

Figure 6: Comparison of generated plot continuations for Harry Potter's use of the Resurrection Stone.

ing a nun, as well as the family's power dynamics. Jia Mu, as the family matriarch, would likely not agree easily, making the model's output feel oversimplified and lacking in emotional depth. The LoRA model generates a classical-style continuation but focuses on a specific detail, with Xichun's tone becoming harsh and defiant. While emotional, this feels abrupt and lacks narrative coherence. In contrast, our WriterAgent model shows the Jia family's complex reactions—shock, reluctance, and eventual acceptance—better fitting the worldview of Dream of the Red Chamber.

D Ablation Study

Our curriculum learning sequence is based on two core principles: (1) the logical writing process of human authors, and (2) a progressive increase in task difficulty. While different authors may follow different writing patterns, we adopt a generalized progression informed by real-world observations and practical experience. We also conduct ablation studies by removing each of the three stages to demonstrate the necessity of each.

From the results, it can be seen that the sequence arrangement of our curriculum learning is reasonable.

E Analysis of Human Annotation Results

In addition to automatic evaluation, we conducted human evaluation with two PhD annotators, both native speakers with strong literary backgrounds. They assessed the generated text by comparing it with the ground truth and selecting the best-performing model on the Dream of the Red Chamber dataset.

To analyze the results, we calculated the average hit rate for each model's output and visualized it as a pie chart, as shown in Figure 12. The results indicate that our model was preferred in 73.1% of cases. The probability of experts selecting the base model's output was 0%, the pre-trained model's output 3.8%, the fine-tuned model's output 15.4%, and the Multi-LoRA trained model's output 7.7%.

From the results, we observe that the hit rate of PLoRA is slightly higher than that of Multi-LoRA. This is because the model trained with the Multi-LoRA method exhibits improved instruction-following capability. However, in some cases, the Multi-LoRA-trained model did not effectively expand upon the given abstract, leading to more rigid or limited responses. In contrast, our model demonstrates a better balance between instruction adherence and contextual expansion, resulting in more comprehensive and coherent outputs that align closely with human expectations.

F Evaluation Prompts

Here, we provide the prompt for evaluating model output using ChatGPT-4o.First, here is the prompt for the English dataset.

Evaluation Criteria:
Stylistic Similarity:
Language Style: Analyze the similarity in vocabulary choice, sentence structure, tone, and overall mood to determine whether the imitated text aligns with the writing style of the Harry Potter series.
Expression Techniques: Examine whether emotional expressions, descriptive language, and the use of metaphors are consistent between the two texts.
Sentence Length and Complexity: Compare sentence length, structural complexity, and paragraph organization in both texts.
Plot Similarity:
Main Storyline: Assess whether the core plotlines are similar and whether they present the same themes or narrative progression.
Character Behavior and Motivation: Analyze whether the characters in the given text align with those in Harry Potter in terms of personality, actions, and motivations, ensuring they fit the story's logical development.
Emotions and Conflict Analysis: Evaluate whether the relationships between characters, emotional dynamics, and

Figure 7: Comparison of reference and generated texts from the baseline and our WriterAgent on the Harry Potter dataset. We highlight the weaknesses of the baseline model and the strengths of our approach.

plot developments are consistent with those in Harry Potter.
The following are two example evaluations:
Example 1:
original text: The two men appeared out of nowhere, a few yards apart in the narrow, moonlit lane. For a second they stood quite still, wands directed at each other's chests; then, recognizing each other, they stowed their wands beneath their cloaks and started walking briskly in the same direction. "News?" asked the taller of the two. "The best," replied Severus Snape. The lane was bordered on the left by wild, low-growing brambles, on the right by a high, neatly manicured hedge. The men's long cloaks flapped around their ankles as they marched. "Thought I might be late," said Yaxley, his blunt features sliding in and out of sight as the branches of overhanging trees broke the moonlight. "It was a little trickier than I expected. But I hope he will be satisfied. You sound confident that your reception will be good?" Snape nodded, but did not elaborate. They turned right, into a wide driveway that led off the lane. The high hedge curved into them, running off into the distance beyond the pair of imposing wrought-iron gates barring the mens way. Neither of them broke step: In silence both raised their left arms in a kind of salute and passed straight through, as though the dark metal was smoke.
The yew hedges muffled the sound of the mens footsteps. There was a rustle somewhere to their right: Yaxley drew his wand again pointing it over his companions head, but the source of the noise proved to be nothing more than a pure-white peacock, strutting majestically along the top of the hedge.
He always did himself well, Lucius. Peacocks Yaxley thrust his wand back under his cloak with a snort.
output text:As they walked, their footsteps echoed off the stone walls. The air was heavy with the scent of blooming flowers. A faint mist hung in the air, and the moon cast long shadows behind them.
"The Dark Lord's plans are moving ahead, Yaxley," Snape said, his voice low and even. "I've received a message from him. He will be returning to Hogwarts soon."
Yaxley's eyes gleamed in the moonlight. "A wise move, sir. The time is ripe for the Dark Lord to assert his dominance over that...establishment." He paused, glancing up at Snape with a mixture of curiosity and wariness. "What do you make of it, Potter's Chosen One and all?" Snape's expression remained unreadable. "I think it wise to keep a low profile for the time being. The atmosphere at Hogwarts is...tense." He glanced around, as if ensuring they were alone. "Rumors are circulating about the Order of

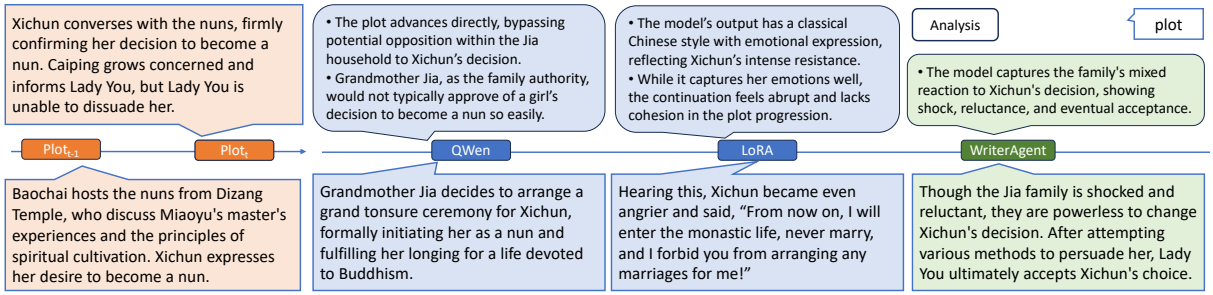

Figure 8: Comparison of generated plot continuations for Xichun's decision to become a nun using different models.

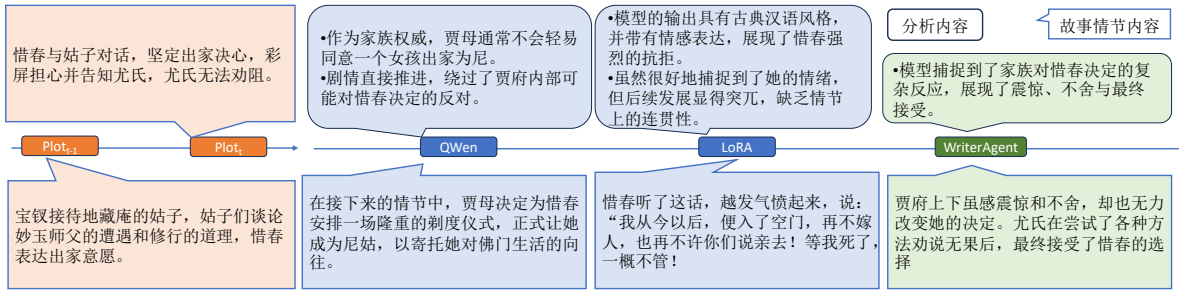

Figure 9: Comparison of generated plot continuations for Xichun's decision to become a nun in Chinese using different models.

the Phoenix. I suspect Lord Voldemort will be looking to eliminate any threats to his power." As they approached Malfoy Manor's gates, Snape raised his left arm and Yaxley followed suit. The gates swung open with a soft creak, revealing a path lined with peacocks strutting along manicured hedges. The pair made their way past the birds and hedges, the sound of their footsteps muffled by the thick grass. Ahead, the Manor's imposing façade loomed, its windows like empty eyes staring back at them.
Scoring Results: "Language Style Analysis": 1, "Expression Techniques Analysis": 1, "Sentence Length and Complexity Analysis": 2, "Main Storyline Analysis": 3, "Character Behavior and Motivation Analysis": 2, "Emotions and Conflict Analysis": 1
Example2:
original text:The two men appeared out of nowhere, a few yards apart in the narrow, moonlit lane. For a second they stood quite still, wands directed at each other's chests; then, recognizing each other, they stowed their wands beneath their cloaks and started walking briskly in the same direction. "News?" asked the taller of the two. "The best," replied Severus Snape. The lane was bordered on the left by wild, low-growing brambles, on the right by a high, neatly manicured hedge. The men's long cloaks flapped around their ankles as they marched. "Thought I might be late," said Yaxley, his blunt features sliding in and out of sight as the branches of overhanging trees broke the moonlight. "It was a little trickier than I expected. But I hope he will be satisfied. You sound confident that your reception will be good?" Snape nodded, but did not elaborate. They turned right, into a wide driveway that led off the lane. The high hedge curved into them, running off into the distance beyond the pair of imposing wrought-iron gates barring the mens way. Neither of them broke step: In silence both raised their left arms in a kind of salute and passed straight through, as though the dark metal was smoke.

The yew hedges muffled the sound of the mens footsteps. There was a rustle somewhere to their right: Yaxley drew his wand again pointing it over his companions head, but the source of the noise proved to be nothing more than a pure-white peacock, strutting majestically along the top of the hedge.
He always did himself well, Lucius. Peacocks Yaxley thrust his wand back under his cloak with a snort.
output text: The lane was deserted and moonlit. Snape's pale eyes flicked over Yaxley's unshaven face, lingering upon the red scars that had been carved into his cheek.
'I trust you are here for the same reason as I, Pucey?' said Snape, his voice low and even.
Yaxley, who had been examining Snape's pointed black hat, looked up. 'What reason would that be?' he said, his tone rather less smooth than Snape's.
'I have reason to believe that the Ministry is going to send a group of Aurors to take my - to take the Dark Lord into custody,' said Snape. 'I am on my way to warn him. 'Yaxley nodded. 'I've heard something similar. I've been sent to tell him the same thing.'
He and Snape looked at each other. They raised their left arms simultaneously and strode off along the lane. The peacocks on the grounds of Malfoy Manor looked up at the sound of their footsteps. They passed beneath the lantern that hung outside the gates and proceeded up the drive, the manicured hedges glinting silver in the moonlight. They did not speak again until they reached the front door, which swung open with a slight creak. Snape led the way across the flagged entrance hall, his black robes sweeping across the stone floor. Yaxley followed him into the great hall.
Scoring Results: "Language Style Analysis": 4, "Expression Techniques Analysis": 5, "Sentence Length and Complexity Analysis": 3, "Main Storyline Analysis": 5, "Character Behavior and Motivation Analysis": 4, "Emotions and Conflict Analysis": 3

Model	Traditional Metrics			Advanced Metrics					
	ROUGE-1	ROUGE-2	ROUGE-L	LS	EX	SLC	SM	CBM	EM
A1	21.43	4.67	17.75	3.02	2.89	3.04	2.96	3.07	3.03
A2	24.32	5.84	20.32	3.18	3.05	3.13	3.19	2.98	3.05
WriterAgent(Qwen)	**34.81**	**9.73**	**24.11**	**3.58**	**3.25**	**3.27**	**3.19**	**3.15**	**3.12**

Table 3: A1 starts with the writing detail task as the first stage, and A2 places it second. The scores are represented as follows: best and second. Numbers in **bold** mean that the improvement to the best result is statistically significant (a two-tailed paired t-test with p-value <0.01).

Based on the following two texts, evaluate their similarity in writing style and plot. Provide scores for each aspect (1 to 5, where 1 is completely different and 5 is identical). The original text is from Harry Potter, and the output text is a rewritten version. My goal is to assess whether the rewritten text aligns with the original's writing style and plot."
'Please do not provide any analytical information. Output results in the following JSON format: "Language Style Analysis":x,"Expression Analysis":x,"Sentence Length and Complexity Analysis":x,"Main Storyline Analysis":x,"Character Behavior and Motivation Analysis":x,"Emotion and Conflict Analysis":x'

Below is the prompt for the Chinese dataset.

评估要求：文风相似性：语言风格：分析两者在词汇选择、句式结构、语气和语调等方面的相似性，判断仿写文本是否与红楼梦文风相似。表达方式：分析两者的情感表达、描述性语言、比喻使用等是否一致。句子长度与复杂度：比较两者的句子长度、句型的复杂性、段落的组织方式等。
情节相似性：故事主线：比较两者的主要情节是否相似，是否呈现相同的主题或事件进展。人物行为与动机：分析文本中的任务与红楼梦中的人物角色是否保持一致，以及主要人物是否在行为和动机上保持一致，是否符合故事情节的推展。情感与冲突分析：分析情节中角色之间的人物关系是否与红楼梦中的设定保持一致，是否有相同的情感流动和情节变化。
以下是两个示例评分：实例1：原文本：话说金桂听了，将脖项一扭，嘴唇一撇，鼻孔里哧哧两声，冷笑道："菱角花开，谁见香来？若是菱角香了，正经那些香花放在那里？可是不通之极！"香菱道："不独菱花香，就连荷叶、莲蓬，都是有一般清香的。但他原不是花香可比，若静日静夜或清早半夜细领略了去，那一股清香比是花都好闻呢。就连菱角、鸡头、苇叶、芦根得了风露，那一股清香也是令人心神爽快的。"金桂道："依你说，这兰花、桂花倒香的不好了？"香菱说到热闹头上，忘了忌讳，便接口道："兰花、桂花的香，又非别的香可比。"一句未完，金桂的丫鬟名唤宝蟾的，忙指着香菱的脸说道："你可要死，你怎么叫起姑娘的名字来？"香菱猛省了，反不好意思，忙陪笑说："一时顺了嘴，奶奶别计较。"金桂笑道："这有什么，你也太小心了。但只是我想这个'香'字到底不妥，意思要换一个字，不知你服不服？"香菱笑道："奶奶说那里话，此刻连我一身一体是奶奶的。何得换一个名字反问我服不服，叫我如何当得起。奶奶说那一个字好，就用那一个。"金桂冷笑道："你虽说得是，只怕姑娘多心。"香菱笑道："奶奶原来不知，当日买了我时，原是老太太使唤的，故此姑娘起了这个名字。后来伏侍了爷，就与姑娘无涉了。如今又有了奶奶，越发不与姑娘相干。且姑娘又是极明白的人，如何恼得这些呢？"金桂道："既这样说，'香'字竟不如'秋'字妥当。菱角菱花皆盛于秋，岂不比香字有来历些？"香菱笑道："就依奶奶这样罢了。"自此后遂改了"秋"字。宝钗亦不在意。输出文本：话说那日中秋佳节，大观园内众人齐聚，赏月饮酒，谈笑风生。其中，薛宝钗之妹薛宝琴与林黛玉、贾宝玉等人一同畅饮，欢声笑语不绝于耳。忽有一阵微风吹过，带来一阵清香，众人皆为之一怔，寻香而至，只见院落中一株古树下，立着两位女子，正是薛宝琴与香菱。香菱自幼随母入贾府，虽非嫡系，却也颇得贾府上下喜爱。其人聪慧异常，且颇有才情，尤善诗文。此时，她正与薛宝琴谈论着园中菱花的香气，两人各抒己见，争论不休。金桂即薛蟠之妻，因性情泼辣，常在贾府中引起是非。她今日亦在此，听闻香菱与宝琴争执，便插嘴道："菱花之香，自是天然，何须争论？然观其名，'香菱'二字，过于俗气，不如改作'秋菱'如何？"香菱闻言，略一沉吟，随即笑道："金姑姑所言极是，菱花虽香，却不及秋水长天，更添几分清雅之意。改作'秋菱'，既符合此花之态，又显我之品格，甚好。"众人大赞金桂之智慧，皆认为此名更为合适，于是香菱便改名为"秋菱"。自此之后，"香菱"之名不再，取而代之的是"秋菱"，其名字之变，不仅象征着她内心对美好事物的追求与向往，亦反映出贾府中人物性格的多样性和复杂性。这一小插曲，虽不似大观园中其他情节般波澜壮阔，却也以细微之处见精神，展现了人物性格的微妙变化，以及家族生活的细腻情感。此番改名之事，虽只是一花一叶之小事，却也如同大观园中的一颗明珠，熠熠生辉，映照出人物命运的起伏与家族兴衰的缩影。而在这片繁华的世界里，每个人都在以自己的方式，书写着属于自己的故事，或悲或喜，或聚或散，皆是人生百态，尽在其中。" 打分结果：
"语言风格分析":1,"表达方式分析":1,"句子长度与复杂度分析":1,"故事主线分析":4,"人物行为与动机分析":1,"情感与冲突分析":1
实例2：原文本：话说贾琏到了王夫人那边，一一的说了。次日，到了部里，打点停妥，回来又到王夫人那边将打点吏部之事告知王夫人。王夫人便道："打听准了么？果然这样，老爷也愿意，合家也放心。那外任何尝是做得的？不是这样回来，只怕叫那些混账东西把老爷的性命都坑了呢。"贾琏道："太太怎么知道？"王夫人道："自从你二叔放了外任，并没有一个钱拿回来，把家里的倒掏摸了好些去了。你瞧那些跟老爷去的人，他男人在外头不多几时，那些小老婆子们都金头银面的妆扮起来了，可不是在外头瞒着老爷弄钱？你叔叔就由着他们闹去。要弄出事来，不但自己的官做不成，只怕连祖上的官也要抹掉了呢。"贾琏道："太太说的很是。方才我听见参了，吓的了不得，直等打听明白才放心。也愿意老爷做个京官，安安逸逸的做几年，才保得住一辈子的声名。就是老太太知道了，倒也是放心的。只要太太说的宽缓些。"王夫人道："我知道，你到底再去打听打听。"输出文本：话说贾琏至王夫人处，禀道："近日闻得老爷在外任上，颇有亏空之弊，我恐将来回朝，不

能善后，因此特来求太太。老爷在京做官，一则可以省却许多烦难，二则亦可保全声名。” 王夫人听了，不觉一怔，随即笑道：“你这孩子，怎么又想出这些事来？难道老爷在外，就该贪污不成？” 贾琏忙笑道：“太太何出此言！不过为的是体面，才想出这个主意来。再者，老爷在外，也须要有人照应才是。” 王夫人点头道：“这话倒是。只是如今谁可托付呢？” 贾琏笑道：“太太只管放心，我自有安排。” 打分结果：’"语言风格分析":5,"表达方式分析":5,"句子长度与复杂度分析":4,"故事主线分析":4,"人物行为与动机分析":4,"情感与冲突分析":4’ 请根据下面两个文本，对它们的文风相似性和情节相似性进行打分。请分别给出文风相似性和情节相似性的评分（1到5分，1为完全不同，5为完全相同）。其中原文本为红楼梦中的原文文本，输出文本为我微调大模型输出的仿写文本，我的目的在于让你给出仿写文本是否贴近原著文风和情节的打分。 请不要给我输出任何分析信息，请按以下json格式输出结果："语言风格分析":x,"表达方式分析":x,"句子长度与复杂度分析":x,"故事主线分析":x,"人物行为与动机分析":x,"情感与冲突分析":x’

964

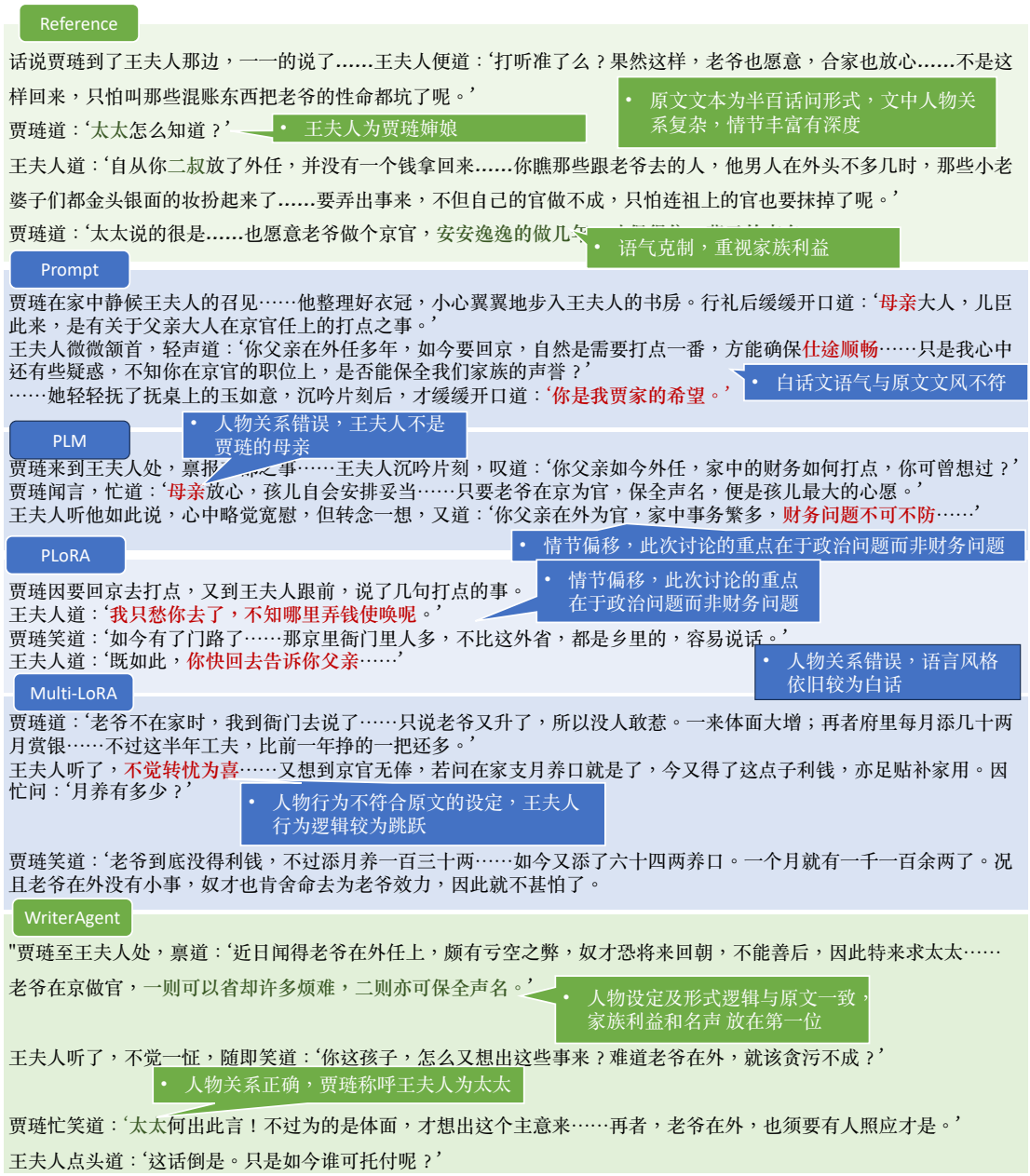

Figure 10: Comparison of reference and generated texts from the baseline and our WriterAgent on Dream of the Red Chamber in Chinese. We highlight the weaknesses of the baseline model and the strengths of our approach.

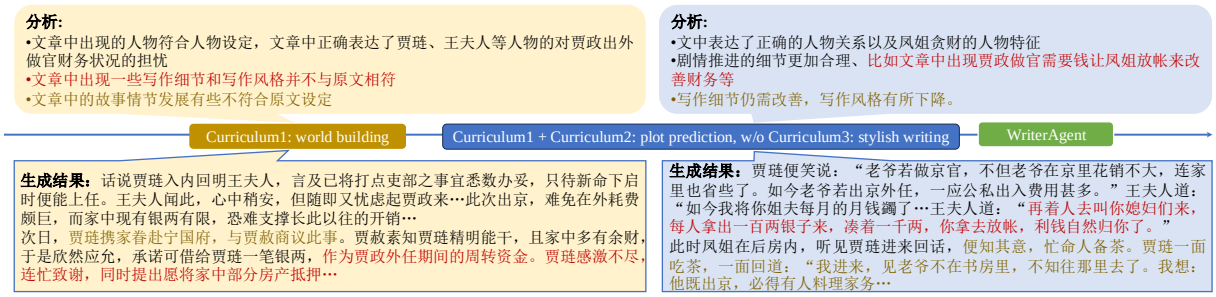

Figure 11: Demonstration of the model's stepwise learning on Dream of the Red Chamber in Chinese: from curriculum 1 to curriculum 1 & 2. The text colors indicate the corresponding problems.

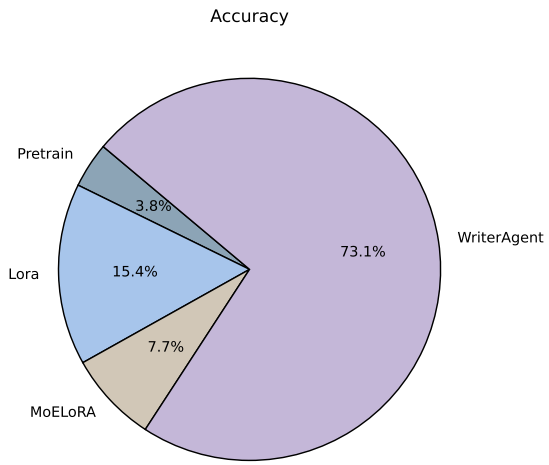

Figure 12: Human Annotation Hit Rate.