Towards A “Novel” Benchmark: Evaluating Literary Fiction with Large Language Models

Wenqing Wang[1,2], **Mingqi Gao**[1], **Xinyu Hu**[1], **Xiaojun Wan**[1]
[1] Wangxuan Institute of Computer Technology, Peking University
[2] School of Software & Microelectronics, Peking University
{wangwenqing}@stu.pku.edu.cn
{gaomingqi,huxinyu,wanxiaojun}@pku.edu.cn

Abstract

Current exploration on creative generation focuses mainly on short stories, poetry, and scripts. With the expansion of Large Language Models (LLMs) context windows, "novel"[1] avenues emerge. This study aims to extend the boundaries of Natural Language Generation (NLG) evaluation by exploring LLMs' capabilities in more challenging long-form fiction. We propose a new multi-level evaluation framework that incorporates ten metrics across the Macro, Meso, and Micro levels. An annotated fiction dataset, sourced from human authors, LLMs, and human-AI collaborations in both English and Chinese is then constructed. Human evaluation reveals notable disparities between LLM-generated and human-authored fictions, particularly the “high-starting, low-ending” pattern in LLM outputs. We further probe ten high-performing LLMs through different prompt templates, achieving moderate correlations by strategically utilizing diverse LLMs tailored to different levels, as an initial step towards better automatic fiction evaluation. Finally, we offer a fine-grained analysis of LLMs capabilities through six issues, providing promising insights for future advancements. Our dataset and code are publicly available[2].

1 Introduction

Fiction is widely recognized as a cornerstone of literary expression (Song and Liu, 2024), which plays a crucial role in fostering empathy (Hale, 2020), moral reflection (Eliot, 2010), and cultural engagement in an increasingly globalized world (Grancher, 2023). Recent advances has demonstrated the potential of Large Language Models (LLMs) in various Natural Language Generation (NLG) tasks, ranging from traditional tasks like automatic summarization (Chang et al., 2024; Patel et al., 2024) to more creative domains such as storytelling (Tian et al., 2024; Subbiah et al., 2024), yet their potential within the long-form literary fiction remains largely unexplored.

With the recent orders-of-magnitude growth in LLMs context window sizes, enabling them to process inputs over 100K tokens (An et al., 2024; Fu et al., 2024) and generate outputs of up to 10K tokens (Pham et al., 2024; Bai et al., 2024b), exploring their potential in literary fiction has become more attainable. Current efforts have incorporated long-form fiction texts to facilitate tasks in quotation attribution (Vishnubhotla et al., 2022), and evaluate LLMs capabilities in comprehension and reasoning (Yu et al., 2024b; Karpinska et al., 2024), yet none have specifically targeted the fiction evaluation task. This paper aims to fill this gap by exploring automatic fiction evaluation using long-context LLMs with context windows of at least 128K, thereby providing a more comprehensive understanding of LLMs' evaluation abilities.

The overall pipeline of our work is illustrated in Figure 1. Specifically, we constructed a diverse fiction dataset in both English and Chinese, encompassing award-winning and ordinary works by human authors, generated fictions from two LLMs under different prompts, and human-AI collaborative works (Zhang et al., 2024), spanning seven mainstream genres: Fantasy, Sci-Fi, Mystery, Martial Arts, Romance, Coming-of-Age, and Realism. Our focus is on entertainment-oriented short fiction (3K-20K[3]), targeted at adult audiences. Given that literary fiction poses increased challenges for NLG Evaluation (Vaezi and Rezaei, 2019), with complex plot arcs, diverse character ensembles, sophisticated narrative techniques, profound thematic depth, and high demands for originality and linguis-

[1] To clarify, the 3K–20K word fiction we study isn't novel-length. We used “novel” in quotes to convey both its “fiction” and “new” meanings.

[2] https://github.com/wenqing-01/Fiction_Eval

[3] Please note that throughout this paper, length is measured in words (or characters for Chinese) rather than tokens.

Findings of the Association for Computational Linguistics: ACL 2025, pages 21648–21673
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

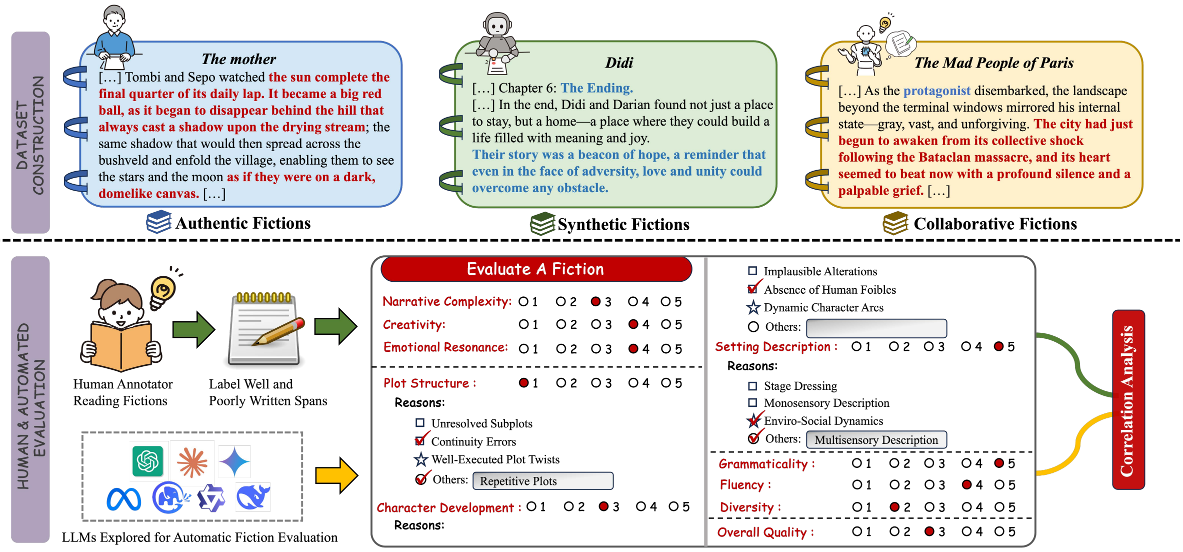

DATASET CONSTRUCTION
The mother
[…] Tombi and Sepo watched the sun complete the final quarter of its daily lap. It became a big red ball, as it began to disappear behind the hill that always cast a shadow upon the drying stream; the same shadow that would then spread across the bushveld and enfold the village, enabling them to see the stars and the moon as if they were on a dark, domelike canvas. […]
Authentic Fictions
Didi
[…] Chapter 6: The Ending.
[…] In the end, Didi and Darian found not just a place to stay, but a home—a place where they could build a life filled with meaning and joy.
Their story was a beacon of hope, a reminder that even in the face of adversity, love and unity could overcome any obstacle.
Synthetic Fictions
The Mad People of Paris
[…] As the protagonist disembarked, the landscape beyond the terminal windows mirrored his internal state—gray, vast, and unforgiving. The city had just begun to awaken from its collective shock following the Bataclan massacre, and its heart seemed to beat now with a profound silence and a palpable grief. […]
Collaborative Fictions
HUMAN & AUTOMATED EVALUATION
Human Annotator Reading Fictions
Label Well and Poorly Written Spans
LLMs Explored for Automatic Fiction Evaluation
Evaluate A Fiction
Narrative Complexity: ○1 ○2 ●3 ○4 ○5
Creativity: ○1 ○2 ○3 ●4 ○5
Emotional Resonance: ○1 ○2 ○3 ●4 ○5
Plot Structure : ●1 ○2 ○3 ○4 ○5
Reasons:
□ Unresolved Subplots
☑ Continuity Errors
☆ Well-Executed Plot Twists
○ Others: Repetitive Plots
Character Development : ○1 ○2 ●3 ○4 ○5
Reasons:
□ Implausible Alterations
☑ Absence of Human Foibles
☆ Dynamic Character Arcs
○ Others:
Setting Description : ○1 ○2 ○3 ○4 ●5
Reasons:
□ Stage Dressing
□ Monosensory Description
☆ Enviro-Social Dynamics
○ Others: Multisensory Description
Grammaticality : ○1 ○2 ○3 ○4 ●5
Fluency : ○1 ○2 ○3 ●4 ○5
Diversity : ○1 ●2 ○3 ○4 ○5
Overall Quality : ○1 ○2 ●3 ○4 ○5
Correlation Analysis

Figure 1: Overview of the fiction dataset construction and evaluation pipeline. Top: Excerpts from human writers, LLMs, and human-LLM collaboration, with **Red** highlighting well-written spans and **Blue** for poor ones. Bottom: Human annotation and and automatic evaluations within a unified framework. LLMs' evaluation capability is assessed by its correlation with human annotations.

tic skill, across extended contexts and varied genres, we designed a robust fiction evaluation framework featuring 10 metrics across 3 levels, with specific exceptions tailored for certain genres. We affirm its validity both theoretically through domain experts and empirically through consistent human evaluation, achieving a Krippendorff's α of 0.74. Building upon the raw dataset and framework for literary fiction, rigorous human assessments and automated evaluation involving 5 openly available and 5 closed-source LLMs were systematically conducted, along with different prompt strategies. We conducted an in-depth analysis of LLMs capabilities, centering on three primary issues. Based on the analysis, we found that LLMs produce strong writing in the first 40%-60% of the text, but decline as the fiction progresses. Additionally, we highlight their potential in evaluation tasks with a moderate correlation, their inherent biases, and a strong correlation between their writing and evaluation capabilities within the same dimensions. To summarize, our contributions are:

- We are the first to explore LLMs in long-form fiction evaluation and propose a multi-level framework for quantitative analysis;

- We release an annotated fiction dataset with various sources in both English and Chinese, along with our refined guidelines, to benchmark literary fiction evaluation;

- We evaluate ten top LLMs, initially enhancing automated fiction evaluation through utilizing different LLMs with diverse prompt strategies, and provide valuable insights to enhance their capabilities based on a fine-grained analysis.

2 Fiction Evaluation Framework

The intricate nature of literary fiction can cause ambiguities in setting evaluation criteria. To address this, we developed a comprehensive evaluation framework through expert consultation and iterative refinement. It is a dual-part framework that assesses fiction utilizing 10 universal metrics encompassing cognitive and emotional factors, while also accommodating Genre-Specific Exceptions, such as intentional violations of physics in Fantasy genre[4]. Each criterion is rated on a 5-point Likert scale, with higher scores indicating better quality.

2.1 Macro-level Evaluation

Macro-level focuses on the overall structure of the fiction, considering its narrative structure, creativity, and ability to resonate with the reader.

Narrative Complexity (Somasundaran et al., 2018): Strategic utilization of literary skills—like non-linear narration, plot twists, and double perspectivation—to enrich plot development.

[4] Limited space permits us to present only the universal metrics here. The second part of our framework are Genre-Specific Exceptions, which are also incorporated in the human and automated evaluation process. Please refer to §A.

Creativity (Sternberg, 2006): Degree of originality exhibited in the core premises, narrative perspectives, and structural design.

Emotional Resonance (Schrock et al., 2004): Effectiveness to evoke empathy in readers, fostering immersive identification with characters and cognitive involvement with the narrated experiences.

2.2 📄Meso-level Evaluation

Meso-level is centered on the core elements—plot, characters, and setting. Building upon the MQM (Burchardt, 2013) framework, we implemented a point deduction policy to prevent score inflation, complemented by bonus categories to distinguish outstanding work, as illustrated in Table 1.

Plot Structure: Events arrangement in the fiction, evaluated for its logical coherence, conflict development, and resolution effectiveness.

Character Development: Portrayal and growth of the protagonist within the fiction, evaluated for authenticity, depth, and evolution.

Setting Description: Depiction of the physical, temporal, and cultural environment, evaluated for its contribution to atmosphere building, character development, and plot advancement.

2.3 ≡Micro-level Evaluation

Micro-level evaluation targets sentence-level grammar, fluency, and diversity to ensure linguistic clarity and readability, in contrast to the broader narrative focus of Macro and Meso levels.

Grammaticality: Adherence to rules of grammar, syntax, and sentence structure, as noted in Chiang and Lee (2023a); Li et al. (2023b); Hong et al. (2024). Capitalization is not factored in.

Fluency: Smoothness and natural flow of the language, utilized in Sottana et al. (2023); Razumovskaia et al. (2024); Migal et al. (2024).

Diversity: Variety in linguistic expression, including rich vocabulary and varied sentence structures, is especially crucial in fiction to avoid monotony over long contexts (Hong et al., 2024).

2.4 Overall Quality Evaluation

Finally, we measure the **Overall Quality** of the fiction, considering all assessment aspects across the aforementioned levels and granularities.

3 Raw Fiction Dataset

We constructed a bilingual fiction dataset from three sources: existing literary works, LLM generation, and human-AI collaboration, referred to as **Authentic Fictions**, **Synthetic Fictions**, and **Collaborative Fictions**, respectively. The context length is restricted to 3K–12K words for English fictions and 6K–20K characters for Chinese[5].

3.1 Authentic Fictions

The Authentic fictions corpus consists of two categories under the Realism genre: **Award-winning** works from the *O. Henry Award* for English and the *Lu Xun Literary Prize* for Chinese, along with ordinary works collected through random **Web-scraping.**[6] The raw corpus underwent strict preprocessing, involving manual removal of metadata like authorship and publication details, and filtering out works outside predefined length range. Web-scraped fictions created by award-winning authors and translated versions from other languages were excluded. We examined the genre of each fiction and discarded the ones that were not Realism genre.

3.2 Synthetic Fictions

We employed two LLMs for generating Synthetic fictions: the closed-source GPT-4o (OpenAI, 2024), referred to as **GPT-writer** and open-weight LongWriter (Bai et al., 2024b), which is optimized for long-context writing based on GLM-4-9B (GLM et al., 2024), named **GLM-writer**. Given GPT-4o's limit of producing outputs of over 2,000 words in a single run (Bai et al., 2024b), we implemented a plan-then-write pipeline (Seredina, 2024) involving several steps: First, generate a high-level outline of the fiction's structure, including the beginning, setup, climax, resolution, and conclusion. Next, break each chapter into multiple paragraphs, detailing content and word count requirements. The LLM then writes each chapter while ensuring logical coherence by referencing previously written chapters. Finally, compile and verify the chapters for length, with instructions to regenerate if the word count is insufficient. In contrast, GLM-writer produces full-length fiction directly from initial instructions. Fiction-writing prompts are designed for two scenarios: using human-written titles randomly selected from the Authentic subset and limited to the Realism genre for alignment; and

[5] This gap in length corresponds with our observations of bilingual fiction translations, where Chinese texts are nearly twice as long as their English counterparts, a similar pattern also noted in Bai et al. (2024a).

[6] English fictions were obtained from `https://novel.tingroom.com/`, and Chinese from `https://www.chinawriter.com.cn/`, both offering short fiction panels, ensuring access to our target corpus.

Plot Structure	Character Development	Setting Description
👎 **Unresolved Subplots** (Ackerman, 2022): Secondary storylines remain incomplete or abandoned, leaving unresolved elements that disrupt the cohesion and impact of the main plot. 👎 **Continuity Errors** (Ackerman, 2022): Discrepancies within the narrative, arising from internal contradictions or external factual errors, undermining plot cohesion and credibility.	👎 **Implausible Alterations** (Corbett, 2013): Unexplained shifts in a character's actions, beliefs, or personality across scenes, undermining their consistency and believability. 👎 **Absence of Human Foibles** (Jenkins, 2024): Characters are portrayed as one-dimensional paragons, devoid of redeemable human flaws, resulting in unrealistic or overly idealized portrayals.	👎 **Stage Dressing** (Ackerman, 2022): Treating the setting as a mere decorative backdrop rather than an integral element that enhances the narrative or contributes to character development. 👎 **Monosensory Description** (Ackerman, 2022): Relying exclusively on a single sensory modality, typically visual, to describe settings, leading to a flat ambiance and limiting immersion.
👍 **Well-Executed Plot Twists** (Truby, 2008): The incorporation of unexpected yet plausible plot twists that enhance narrative complexity and engagement, surprising readers while preserving logical consistency.	👍 **Dynamic Character Arcs** (McKee, 1997): Depicting a character's gradual evolution throughout the narrative, marked by significant personal growth or a meaningful transformation in beliefs, behavior, or emotional depth.	👍 **Enviro-Social Dynamics** (Malewitz, 2021): Skillfully integrating natural and social environments with multisensory details to enhance the emotional tone and reflect the era and relationships driving the plot.
• **Others**: Additional plot-related merits or demerits to be enumerated.	• **Others**: Additional character-related merits or demerits to be enumerated.	• **Others**: Additional setting-related merits or demerits to be enumerated.

Table 1: Definition of 👎Deductions and 👍Bonus reasons in Meso-level Evaluation.

employing LLM-generated titles under the other six genres outlined in §1, combined with three narrative settings and two writing constraints. Full prompts can be found in §B.1. Furthermore, an originality check was also conducted to eliminate any plagiarized content, as detailed in §B.2.

3.3 Collaborative Fictions

Inspired by Zhang et al. (2024), we additionally constructed a **Collaborative**[7] fiction subset that integrates both human- and AI-written content to explore the potential for human-machine collaboration in fiction writing. Specifically, we randomly selected 100 fictions from Authentic subset, with an equal split between Chinese and English, and provided them to GPT-4o for initial comprehension and summarization into outlines. The original fictions were then discarded, and GPT-4o was employed to rewrite the content solely based on the generated outlines (See the full prompt in §B.3).

3.4 Data Statistics

Ultimately, we obtained a diverse fiction dataset with 500 entries spanning 2 languages, 3 sources featuring 5 categories, and 7 genres. Each language accounts for half of the dataset, with each category contributing 50 entries per language. On average, English fictions consist of 6,338.68 words, while Chinese contains 10,734.33 characters, with detailed statistics available in §B.4 Table 8.

4 Human Annotation

The assessment of literary fiction differs from gold-standard evaluations as it cannot rely on references. Therefore, both human annotation and LLM's automatic evaluation are reference-free.

4.1 Annotation Protocol

We recruited eight annotators from the Faculty of Arts, all highly proficient in both Chinese and English, with extensive experience in reading literary works. Each annotator first completed pre-annotation training and quality testing, with a strict quality control mechanism applied during the annotation process. Detailed annotation procedure and full guidelines are publicly available in §C.

In the annotation interface, fictions—whether human-authored or LLM-generated[8]—are displayed anonymously on the left to prevent bias, while the scoring rubric is presented on the right. The procedure consists of four steps: Annotators first label the well and poorly written spans on the fiction text. They then answer a GPT-4o-generated Attention Check Question based on original text details. After fully understanding the content, they rate each of the ten evaluation criteria on a 5-point Likert scale. During the Meso-level evaluation, annotators are required to label predefined deductions and bonuses as the rationale, but these do not need

[7] Notably, although referred to as "Collaborative" in accordance with Zhang et al. (2024), the construction does not involve real-time human-AI interaction; the term "Semi-synthetic" may be more accurate.

[8] Special tokens, such as "###", "***", or other redundant texts (e.g. "Title:..."), have been manually removed prior to annotation to avoid interference.

to exactly align with the score to avoid bias (Guan et al., 2021). Finally, they determine whether the fiction was human-authored or LLM-generated.

Each sample was annotated by two distinct annotators, with the final score determined by averaging their ratings. For binary authorship classification, a third annotator made the final decision in case of disagreement. Inter-Annotator Agreement (IAA) was evaluated using Krippendorff's α (Krippendorff, 2011) for the ten evaluation criteria and Cohen's Kappa (Cohen, 1960) for authorship determination. The results showed a substantial IAA, with Krippendorff's $\alpha = 0.735$ and $\kappa = 0.743$ in English and $\alpha = 0.754$ and $\kappa = 0.770$ in Chinese. Table 9 in §C.1 displays the IAA for each metric and genre, demonstrating the high consistency of human annotation and the framework's applicability across literary fiction with various genres.

4.2 Annotation Result

Table 2 presents the average annotation scores for each fiction category. In this section, we conducted a detailed analysis of the annotation results by answering the following three questions.

1) Can LLMs compete with human novelists, both ordinary and award-winning? The answer is No. Both Table 2 and Figure 14 in §E.2 shows that LLMs' fiction-writing capabilities are obviously inferior to web-scraped works and far behind top novelists. Their writing style is distinguishable, featuring a homogeneously positive tone, fragmented narratives, and AI-like subheadings, as illustrated in Figure 1 **Blue** spans. These style patterns are also observed in shorter storytelling (Tian et al., 2024). However, a more distinguished flaw in LLM-generated fictions is repetitive narration, as discussed in §E.1. This is also supported by Table 2 where LLMs are competitive with human authors in terms of grammaticality and fluency at the Micro level, but fall short in linguistic diversity. Notably, Collaborative fictions outperform purely Synthetic ones, highlighting the potential of human-AI collaboration to enhance LLMs' capabilities.

2) Are LLMs biased towards certain languages, genres, or prompts in fiction creating? The answer is Yes. Table 2 highlights this score disparity between the two languages. For more precise comparison, we calculated the intra-individual annotation score differences between each Collaborative fiction and its corresponding Authentic fiction, as they are paired, allowing a direct comparison of the human-LLM gap between the two languages. The results in Figure 2 show that the GPT-4o's gap with human authors is consistently larger in Chinese (average score difference of -0.90) than in English (-0.56). Regarding genres and prompt templates, the spider plot of the average scores for each setup in §E.2 Figure 13, indicates that LLMs shows no dependence on prompt templates, but exhibits a preference for genres, with Sci-Fi and Fantasy genres achieving strong performance while Coming-of-age consistently lags in all dimensions.

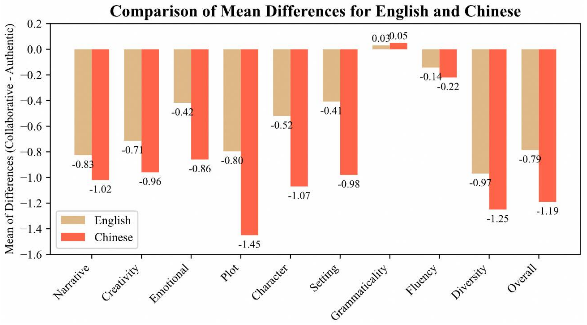


Figure 2: LLM-Human performance gap in English and Chinese. Negative scores indicate that Collaborative fictions are rated lower than Authentic ones.

3) Is length a factor in LLM performance for fiction creation? The answer is Yes. Figure 3 and 15 illustrate that the fictions generated by the two LLMs follows a "high-starting, low-ending" pattern in both English and Chinese, with strong performance primarily concentrated in the initial 40%-60% of the text, unlike the consistently high quality of Authentic fiction. This explains LLMs' high performance in shorter creative tasks (Gómez-Rodríguez and Williams, 2023; Franceschelli and Musolesi, 2024; Marco et al., 2025), which typically fit within their optimal performance window with 20%-50% of the fiction length. One reason for the quality decline in later chapters is its excessive overlap with earlier content, which can be inferred from the lower diversity and the deductions of repetitive narrative, as analyzed in Question 1).

5 Experiments

Due to the limited training data, our goal in this section is to investigate the LLMs' current capabilities in fiction evaluation through prompting rather than developing an advanced system through training.

Long-context LLMs. We evaluated ten high-performing long-context LLMs, including 5 proprietary models, namely GPT-4-TURBO, GPT-4O, O1-PREVIEW, CLAUDE-3.5-SONNET-V2, and GEMINI-2.0-FLASH, along with 5 open-source

		English Fiction					Chinese Fiction				
		Authentic		Synthetic		Collaborative	Authentic		Synthetic		Collaborative
		Award	Web	GPT	GLM		Award	Web	GPT	GLM	
📕Macro-level	Narrative	4.32	3.78	3.05	2.81	3.40	4.35	3.85	2.26	2.03	3.12
	Creativity	4.16	4.19	2.94	2.65	3.51	4.31	3.77	2.37	2.06	3.14
	Emotional	4.63	3.89	2.91	2.83	3.91	4.59	4.10	2.25	1.89	3.46
🗎Meso-level	Plot	4.40	4.09	3.11	2.71	3.52	4.61	4.11	2.12	1.68	2.96
	Character	4.56	4.16	3.60	3.06	3.87	4.54	4.13	2.34	2.01	3.26
	Setting	4.63	4.19	3.78	3.55	4.08	4.61	4.01	2.50	2.19	3.37
☰Micro-level	Grammaticality	4.90	4.83	4.83	4.73	4.86	4.71	4.43	4.57	4.51	4.62
	Fluency	4.57	4.20	4.39	4.10	4.35	4.50	4.18	3.75	3.67	4.17
	Diversity	4.46	4.18	2.78	2.61	3.39	4.45	3.93	2.21	1.86	3.04
	Overall Quality	4.60	4.08	2.91	2.50	3.65	4.69	4.03	2.24	1.76	3.19
	Avg.	4.52	4.16	3.43	3.16	3.85	4.54	4.05	2.66	2.37	3.43
Authorship Determination (%Acc.)		84.00	76.00	56.00	72.00	36.00	80.00	76.00	76.00	80.00	48.00

Table 2: Human evaluation average scores and author identification accuracy across three fiction sources in two languages. As the recognition for authorship may improve along the evaluation process (Marco et al., 2024a), we calculated determination accuracy only relied on the annotation results from the latter half.

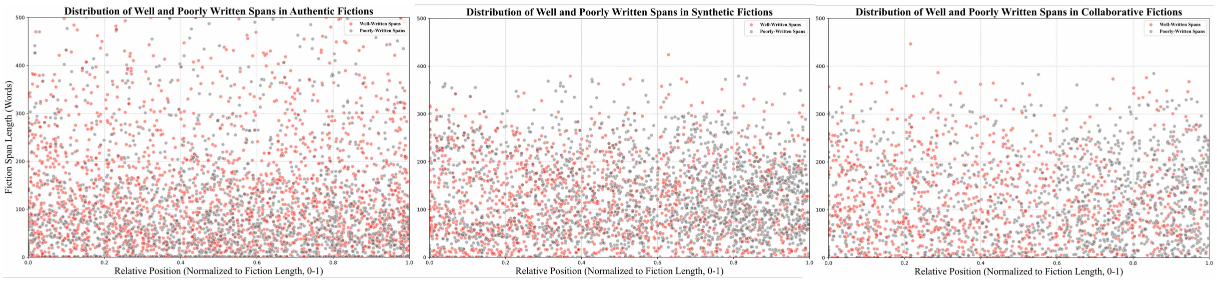


Figure 3: Scatter plot of highlighted spans from three fiction sources in English (Chinese in Figure 15), with Red dots indicating well-crafted spans and Gray denoting poor ones. The x-axis represents the span's relative position in the fiction, and the y-axis denotes the span's length, combining results from two annotators for each LLM-generated fiction. Collaborative fictions exhibit sparser distributions due to containing half the data volume of other types.

ones: LLAMA3.1-8B, LLAMA3.1-70B, GLM-4-9B, QWEN2.5-72B, and DEEPSEEK-V3. We set the output temperature T of 0.5, and results were averaged over three runs following Wang et al. (2022). Please see §D.1 for detailed model cards.

Prompting Strategies. We prompted each model with the full fiction text and identical evaluation instructions based on human assessments in a zero-shot manner[9]. We first conducted a pilot study on 50 fictions, comparing three prompt formats (provided in §D.2): *Rate-only*, *Rate-then-explanation*, and *Explanation-then-rate*. We selected the best-performing *Explanation-then-rate* format, in which the model first explains its reasoning before assigning a final score, aligning with the Chain-of-Thought prompting (Wei et al., 2023).

Evaluation Metrics. We assessed model-human rating alignment using Pearson (r) and Spearman (ρ) correlation coefficients.

6 Results and Analysis

This section aims at analyzing LLMs' capabilities in fiction evaluation through three questions.

1) Are LLMs effective at evaluating fictions? The answer is: They Show Promise. As illustrated in Table 3 and Table 13 in §F.1, LLMs achieve at best moderate correlations (> 0.45), but weak at the Meso-level. No single LLM stands out across all levels: CLAUDE-3.5-SONNET-V2 excelled at Macro-level and Overall quality, potentially benefiting from its 200K context window, while GPT-4O dominated Micro-level evaluations, aligning with its performance in fiction writing. Nevertheless, none of the LLMs showed consistently strong performance at the Meso-level, with particularly weak correlations (< 0.3) for Synthetic fiction. Annotation analysis in §E.1 found frequent undefined deduction/bonus reasons (i.e., "Others") in LLM-generated fictions. Motivated by this, we further developed an **Extended-Criteria Prompt** that incorporates the top-three sub-reasons from the "Others"

[9] Due to the extensive length of fictions, few-shot prompting was not utilized.

LLMs for Evaluation	Macro-Eval				Meso-Eval				Micro-Eval				Overall-Eval			
	Aut	Syn	Col	Avg.	Aut	Syn	Col	Avg.	Aut	Syn	Col	Avg.	Aut	Syn	Col	Avg.
English Fiction																
GPT-4-TURBO	0.26	0.22	0.21	0.23	0.20	0.24	0.27	0.24	0.26	0.22	0.21	0.23	0.21	0.17	0.18	0.19
GPT-4O	0.33	0.44	0.24	0.34	0.38	0.23	**0.34**	0.31	**0.49**	**0.41**	**0.54**	**0.48**	0.30	0.29	0.22	0.27
O1-PREVIEW	0.42	0.29	0.17	0.29	0.38	0.16	0.21	0.25	0.43	0.36	0.42	0.40	0.23	0.24	0.26	0.24
CLAUDE-3.5-SONNET-V2	0.49	**0.49**	**0.51**	**0.50**	0.39	0.26	0.28	0.31	0.47	0.37	0.52	0.45	**0.45**	0.46	**0.51**	**0.47**
GEMINI-2.0-FLASH†	**0.52**	0.34	0.39	0.42	**0.42**	0.16	0.26	0.28	0.41	0.37	0.33	0.37	0.34	0.41	0.32	0.36
LLAMA3.1-8B†	0.11	0.19	0.15	0.15	0.10	0.13	0.16	0.13	-0.05	0.05	0.15	0.05	-0.01	0.15	0.11	0.08
LLAMA3.1-70B	0.25	0.18	0.26	0.23	0.28	0.23	0.28	0.26	0.26	0.26	0.12	0.22	0.24	0.36	0.13	0.24
GLM-4-9B†	0.20	0.16	0.15	0.17	0.19	0.05	0.19	0.14	0.23	0.12	0.13	0.16	0.18	0.16	0.10	0.15
QWEN2.5-72B	0.18	0.21	0.18	0.19	0.26	0.17	0.20	0.21	0.26	0.25	0.32	0.28	0.23	0.21	0.20	0.21
DEEPSEEK-V3	0.37	0.36	0.41	0.38	0.38	**0.29**	**0.34**	**0.33**	0.32	0.36	0.50	0.39	0.34	**0.48**	0.45	0.42
Chinese Fiction																
GPT-4-TURBO	0.24	0.21	0.24	0.23	0.27	0.10	0.20	0.19	0.22	0.27	0.23	0.24	0.22	0.27	0.20	0.23
GPT-4O	0.26	0.22	0.37	0.28	0.26	0.20	0.40	0.29	0.40	**0.42**	**0.42**	**0.41**	0.33	0.21	0.24	0.26
O1-PREVIEW	0.18	0.16	0.25	0.20	0.26	0.13	0.29	0.23	0.30	0.34	0.30	0.31	0.25	0.13	0.30	0.23
CLAUDE-3.5-SONNET-V2	**0.52**	**0.44**	**0.47**	**0.48**	0.33	**0.26**	0.34	0.31	0.33	0.38	0.34	0.35	**0.43**	**0.47**	**0.53**	**0.48**
GEMINI-2.0-FLASH†	0.34	0.23	0.35	0.31	0.20	0.16	**0.42**	0.26	0.21	0.20	0.40	0.27	0.23	0.26	0.50	0.33
LLAMA3.1-8B†	0.22	0.27	-0.04	0.15	0.10	0.07	0.10	0.09	0.10	0.10	-0.06	0.05	0.04	0.24	-0.03	0.08
LLAMA3.1-70B	0.38	0.30	0.11	0.26	0.24	0.17	0.12	0.18	0.31	0.24	0.14	0.23	0.20	0.35	0.26	0.27
GLM-4-9B†	0.17	0.15	0.03	0.12	0.03	0.13	0.04	0.07	0.10	0.10	0.22	0.14	0.11	0.13	0.01	0.08
QWEN2.5-72B	0.29	0.11	0.10	0.17	0.14	0.12	0.11	0.12	0.16	0.19	0.07	0.14	0.20	0.25	0.13	0.20
DEEPSEEK-V3	0.47	0.41	0.26	0.38	**0.36**	0.23	**0.42**	**0.34**	**0.44**	0.38	0.27	0.36	0.33	0.46	0.47	0.42

Table 3: Pearson correlation between 5 Proprietary and 5 Open-source models with human ratings on Authentic fictions (Aut), Synthetic fictions (Syn), Collaborative fictions (Col), and the average score (Avg.), comparing English (top) and Chinese fictions (bottom). † indicates that the model's results are not computed on full datasets due to missing expected outputs (§D.3). The corresponding Spearman coefficients can be found in §F.1 Table 13.

LLMs w/ Prompt Strategies	Authentic				Synthetic				Collaborative				Average			
	EN		CH		EN		CH		EN		CH		EN		CH	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
GPT-4O	0.38	0.29	0.26	0.19	0.23	0.24	0.20	0.20	0.34	0.35	0.40	0.38	0.31	0.29	0.29	0.26
+ Extended	0.25	0.24	0.27	0.16	0.49	0.44	0.44	0.48	**0.47**	**0.46**	0.43	**0.47**	0.40	0.38	0.38	0.37
- Direct	0.41	0.44	0.40	0.41	0.42	0.44	0.36	0.43	0.41	0.45	0.46	0.44	0.41	0.44	0.41	0.43
CLAUDE-3.5-SONNET-V2	0.39	0.33	0.33	0.32	0.26	0.26	0.26	0.30	0.28	0.29	0.34	0.43	0.31	0.30	0.31	0.35
+ Extended	0.33	0.38	0.28	0.17	0.48	0.51	0.46	0.49	0.40	0.42	0.32	0.33	0.40	0.43	0.35	0.33
- Direct	0.46	0.44	0.45	0.48	0.48	0.47	**0.49**	0.49	0.42	0.43	0.41	0.34	0.45	0.45	**0.45**	0.43
GEMINI-2.0-FLASH	0.42	0.35	0.20	0.17	0.16	0.15	0.16	0.22	0.26	0.28	0.42	0.43	0.28	0.26	0.26	0.27
+ Extended	0.37	0.36	0.18	0.20	0.40	0.37	0.43	0.46	0.34	0.35	0.48	0.45	0.37	0.36	0.36	0.37
- Direct	**0.48**	0.46	0.30	0.26	0.39	0.36	0.36	0.36	0.37	0.36	0.38	0.42	0.41	0.39	0.34	0.34
DEEPSEEK-V3	0.38	0.32	0.36	0.35	0.29	0.29	0.23	0.21	0.34	0.34	0.42	0.39	0.33	0.32	0.34	0.32
+ Extended	0.28	0.23	0.36	0.37	**0.50**	**0.52**	0.47	**0.55**	0.41	0.38	0.46	0.39	0.40	0.38	0.43	**0.44**
- Direct	0.47	**0.50**	**0.50**	**0.51**	0.48	0.46	0.36	0.38	0.46	0.44	**0.49**	0.42	**0.47**	**0.47**	**0.45**	**0.44**

Table 4: Pearson (r) and Spearman (ρ) coefficients for the top 4 performing models at Meso-level evaluations. For each model: Row 1 presents the original prompt results, while Rows 2-3 show results from two optimized prompts. Underperforming scores compared to the baseline in Row 1 are marked in Gray.

category in §E.1, requiring LLMs to identify and explain the six predefined deduction/bonus types before scoring. We also designed a **Direct-Scoring Prompt** that aligns with the other two levels by omitting explicit identification. As shown in Table 4, both prompts enhanced performance by approximately 10% and achieved a moderate correlation. Interestingly, the Direct-Scoring Prompt outperformed the two identification-based methods, and Extended-Criteria boosted Synthetic fiction consistency by 23.3% but reduced Authentic fiction by 8.5%. This divergence suggests differing deduction/bonus patterns between AI and human writing, emphasizing the need for caution in future work to evaluate both types together in beneficial ways. Our exploration demonstrates the potential of LLMs for effective fiction evaluation by strategically using different models suited to each level.

2) Do LLMs enjoy their self-generated fictions more? The answer is Yes. We calculated the score discrepancies across all metrics between human evaluations and two LLMs, GPT-4o and GLM-4, which participated in both the creation and evaluation tasks. The delta distributions are shown in Figure 4. Both models displayed positive discrepancies for all LLM-generated fictions, with a noticeable preference for their self-generated ones, especially in the case of GPT-4o. Moreover, GPT-4o showed a significantly lower preference for Collaborative fiction compared to its entirely self-written

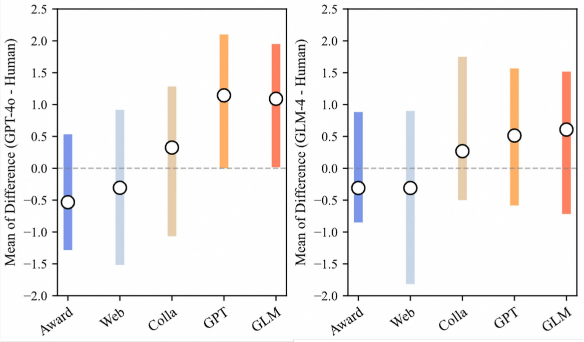


Figure 4: Dumbbell plot of the average difference between LLMs and human ratings on five fiction categories, with GPT-4o on the left and GLM-4 on the right.

works, underscoring a clear distinction between Collaborative and purely Synthetic fiction.

3) Is there a correlation between LLM fiction writing and evaluation abilities? The answer is

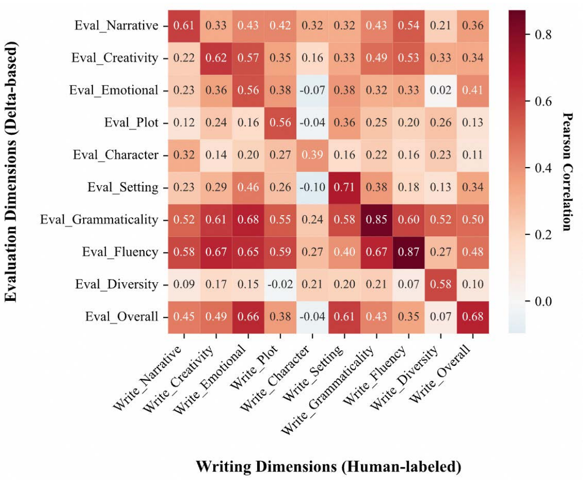


Figure 5: Heatmap of Pearson correlation between GPT-4o writing and evaluation capabilities across different dimensions. GLM-4 is shown in §F.2 Figure 16.

Yes. We assess LLM's fiction writing ability for each dimension (i) using human-annotated scores, denoted as write_dim(i) = human_dim(i), where write_dim(i) represents the LLM's writing capability for the i-th dimension, and human_dim(i) is the corresponding human score. LLM's evaluation ability is quantified as: eval_dim(i) $= 1 - \frac{|\text{LLM\_dim}(i) - \text{human\_dim}(i)|}{4}$, with eval_dim$(i)$ representing the LLM's evaluation capability and LLM_dim(i) being the LLM-rated score. The denominator of 4 accounts for the maximum score discrepancy. More detailed formula can be found in §F.3. Pearson correlation is calculated using all samples for each dimension, and Figures 5 and 16 reveal a strong correlation between LLMs' evaluation performance and their inherent writing abilities with the same dimensions. GPT-4o, in particular, demonstrates a high Micro-level correlation, likely due to its outstanding Micro-level alignment in writing (Table 2) and evaluation (Table 3) tasks. This indicates that future work could boost LLMs' evaluation abilities by improving their generation capabilities.

7 Related Work

Creative Generation by LLMs. The advent of LLMs marks a significant shift in the landscape of creative generation (Zhao et al., 2023; Harel-Canada et al., 2024), displaying comparable performance in a range of automatic generation tasks, such as short storytelling (Huang et al., 2023; Xie and Riedl, 2024), poetry (Chen et al., 2024; Yu et al., 2024a) and creative writing (Ippolito et al., 2022; Gómez-Rodríguez and Williams, 2023). Nevertheless, the expansion of LLMs' context windows size has prompted us to explore more intricate tasks. In this study, we take a preliminary step toward exploring literary fiction and utilize long-context LLMs to generate long-form fiction.

Natural Language Generation Evaluation. Recent studies in NLG evaluation have been predominantly driven by LLM-based methods (Liu et al., 2023; Chiang and Lee, 2023b; Hu et al., 2024; Kim et al., 2024) and applied to various tasks (Tian et al., 2024; Chakrabarty et al., 2024; Chhun et al., 2024; Marco et al., 2024a; Walsh et al., 2024). The most relevant work to ours is Vaezi and Rezaei (2019), who introduced a fiction rubric. Nevertheless, their focus was solely on the creation of literary criteria, without considering LLM-based evaluation, thereby leaving an absence in the NLG evaluation domain for literary fiction. In response to Gao et al. (2024), who advocates for exploring new task scenarios across different languages, our work bridges this gap in literary fiction and provide a more comprehensive understanding of LLMs' capabilities.

8 Conclusion

This work aims to enhance the understanding of LLMs' capabilities by incorporating challenging literary fiction tasks. We developed a multi-level evaluation framework and constructed the first fiction dataset through LLM-assisted writing with quality annotations in both Chinese and English. Through human evaluation, we identified key distinctions between human and LLM-generated capabilities in fiction creation. We also evaluated ten LLMs and, based on our analysis, proposed a promising evaluation strategy by applying various models to different levels, establishing a benchmark for future study. We believe that our efforts lays the groundwork for a more systematic exploration of literary fiction and hope that the framework, dataset and proposed perspectives will attract broader academic interest to advance literary fiction evaluation.

Limitations

Evaluating literary fiction is a challenging task. This paper acts as a preliminary exploration in expanding NLG evaluation scope by examining literary fiction from a macro viewpoint, which inherently limits the study's depth. Finer granularity, such as the internal consistency of long-form fiction, require further in-depth investigation. Due to the current limitations on the output windows of LLMs, this study is confined to short fictions within 20K words or characters, leaving the analysis of medium and long-form fictions for future exploration. Additionally, only English and Chinese languages are evaluated, necessitating the investigation of lower-resource languages to achieve a more comprehensive understanding of LLM capabilities. Nonetheless, we would anticipate an even greater disparity between LLMs and human writers in other languages that have fewer online resources.

Ethics Statement

We recruit annotators from a college campus, and their participation in our annotation process is entirely voluntary. All our annotators read these fictions with copyright and give informed consent for the use of their evaluations in LLM assessment. The payment is $9 per hour, which is higher than the local minimum wage. No personal information is included in our collected dataset, and any identifying information is removed after annotation. Additionally, LLM-generated fictions might contain toxic language, which could cause discomfort for the annotators. We reviewed the data prior to annotation and found no problematic samples. We check the licenses of the artifacts used in this study and do not find conflicts. The license of the dataset we will release is CC BY-NC 4.0.

Due to copyright restrictions, the Authentic fictions corpus will not be publicly distributed, but metadata including titles, authors, and annotation results will be made available for individual acquisition of the subset. Furthermore, the Synthetic and Collaborative fictions, which aren't limited by copyright, will be released in full.

Acknowledgements

This work was supported by Beijing Science and Technology Program (Z231100007423011) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank Professor Ehud Reiter for providing constructive suggestions. We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Angela Ackerman. 2022. Setting description mistakes that weaken a story.

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. 2024. Training-free long-context scaling of large language models. *Preprint*, arXiv:2402.17463.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024a. Longbench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508.

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

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Gioele Cadamuro and Marco Gruppo. 2023. A distribution-based threshold for determining sentence similarity. *Preprint*, arXiv:2311.16675.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. *Preprint*, arXiv:2309.14556.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. *Preprint*, arXiv:2310.00785.

Yanran Chen, Hannes Gröner, Sina Zarrieß, and Steffen Eger. 2024. Evaluating diversity in automatic poetry generation. *Preprint*, arXiv:2406.15267.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Preprint*, arXiv:2405.13769.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

David Corbett. 2013. *The art of character: Creating memorable characters for fiction, film, and TV*. Penguin.

George Eliot. 2010. *Silly novels by lady novelists*. Penguin UK.

Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *Preprint*, arXiv:2304.00008.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. *Preprint*, arXiv:2402.10171.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *Preprint*, arXiv:2402.01383.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Romain Grancher. 2023. The politics of fiction: The significance of fiction as a medium to advance dialogue and advocacy around the human condition in politically challenging times. *World Sustainability Series*, pages 181–199.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics.

Dorothy J Hale. 2020. *The novel and the new ethics*. Stanford University Press.

Fabrice Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. Measuring psychological depth in language models. *Preprint*, arXiv:2406.12680.

Xudong Hong, Asad Sayeed, and Vera Demberg. 2024. Summary of the visually grounded story generation challenge. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 39–46, Tokyo, Japan. Association for Computational Linguistics.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: A reference-free nlg evaluation language model with flexibility and interpretability. *Preprint*, arXiv:2406.18365.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11792–11806, Singapore. Association for Computational Linguistics.

Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *Preprint*, arXiv:2211.05030.

Jerry Jenkins. 2024. 8 types of characters in fiction and how to use them.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. Tigerscore: Towards building explainable metric for all text generation tasks. *Preprint*, arXiv:2310.00752.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. *Preprint*, arXiv:2406.16264.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and

Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *Preprint*, arXiv:2402.14848.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023a. Generative judge for evaluating alignment. *Preprint*, arXiv:2310.05470.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023b. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *Preprint*, arXiv:2310.19740.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Raymond Malewitz. 2021. What is a setting?

Guillermo Marco, Julio Gonzalo, Ramón del Castillo, and María Teresa Mateo Girona. 2024a. Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing? *Preprint*, arXiv:2407.01119.

Guillermo Marco, Julio Gonzalo, M.Teresa Mateo-Girona, and Ramón Del Castillo Santos. 2024b. Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19654–19670, Miami, Florida, USA. Association for Computational Linguistics.

Guillermo Marco, Luz Rello, and Julio Gonzalo. 2025. Small language models can outperform humans in short creative writing: A study comparing slms with humans and llms. *Preprint*, arXiv:2409.11547.

Robert McKee. 1997. Substance, structure, style, and the principles of screenwriting. *Alba Editorial*.

Aleksandr Migal, Daria Seredina, Ludmila Telnina, Nikita Nazarov, Anastasia Kolmogorova, and Nikolay Mikhaylovskiy. 2024. Overview of long story generation challenge (LSGC) at INLG 2024. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 47–53, Tokyo, Japan. Association for Computational Linguistics.

OpenAI. 2024. Openai: Hello gpt-4o.

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

Kevin Patel, Suraj Agrawal, and Ayush Kumar. 2024. Tweak to trust: Assessing the reliability of summarization metrics in contact centers via perturbed summaries. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 172–186, Mexico City, Mexico. Association for Computational Linguistics.

Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following for long-form text generation. *Preprint*, arXiv:2406.19371.

Evgeniia Razumovskaia, Joshua Maynez, Annie Louis, Mirella Lapata, and Shashi Narayan. 2024. Little red riding hood goes around the globe: Crosslingual story planning and generation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10616–10631, Torino, Italia. ELRA and ICCL.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. Defining and detecting vulnerability in human evaluation guidelines: A preliminary study towards reliable nlg evaluation. *Preprint*, arXiv:2406.07935.

Douglas Schrock, Daphne Holden, and Lori Reid. 2004. Creating emotional resonance: Interpersonal emotion work and motivational framing in a transgender community. *Social Problems*, 51(1):61–81.

Daria Seredina. 2024. A report on LSG 2024: LLM fine-tuning for fictional stories generation. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 123–127, Tokyo, Japan. Association for Computational Linguistics.

Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

Li Song and Ying Liu. 2024. Approaches and challenges for resolving different representations of fictional characters for Chinese novels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1408–1421, Torino, Italia. ELRA and ICCL.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

Robert J Sternberg. 2006. The nature of creativity. *Creativity research journal*, 18(1):87.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024. Reading subtext: Evaluating large language models on short story summarization with writers. *Preprint*, arXiv:2403.01061.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? *Preprint*, arXiv:2407.13248.

John Truby. 2008. *The anatomy of story: 22 steps to becoming a master storyteller*. Farrar, Straus and Giroux.

Maryam Vaezi and Saeed Rezaei. 2019. Development of a rubric for evaluating creative writing: a multi-phase research. *New Writing*, 16(3):303–317.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848, Marseille, France. European Language Resources Association.

Melanie Walsh, Anna Preus, and Maria Antoniak. 2024. Sonnet or not, bot? poetry evaluation for large models and datasets. *Preprint*, arXiv:2406.18906.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Kaige Xie and Mark Riedl. 2024. Creating suspenseful stories: Iterative planning with large language models. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. 2024a. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–325, Bangkok, Thailand. Association for Computational Linguistics.

Linhao Yu, Qun Liu, and Deyi Xiong. 2024b. LFED: A literary fiction evaluation dataset for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10466–10475, Torino, Italia. ELRA and ICCL.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436, Mexico City, Mexico. Association for Computational Linguistics.

Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre L. S. Filipowicz. 2023. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, CC '23, page 368–370, New York, NY, USA. Association for Computing Machinery.

A Genre-Specific Exemptions

Based on ten universal criteria, we additionally define exceptions for certain genres to acknowledge conventions that might seem flawed by universal standards but are essential within their particular traditions, ensuring more equitable assessments. The Genre-Specific Exemptions is the second part of our fiction evaluation framework, which is also incorporated in both human and automatic evaluations procedures.

🪄 **Fantasy and Sci-Fi Genres**: Evaluation should exempt *Plot Structure* and *Character Development* criteria. Specifically, magic, supernatural events, or futuristic technology should not be dismissed illogical simply for defying real-world physics, provided they maintain internal consistency. Similarly, characters with extraordinary abilities can remain credible if their core motivations and emotional struggles are compelling.

🔨 **Martial Arts Genre**: Evaluation should exempt *Grammaticality*, *Fluency*, and *Diversity* criteria, i.e., the use of dialects, archaic expressions, or other non-standard grammar are acceptable in martial arts narratives if it serves a deliberate literary purpose rather than stemming from accidental errors.

🔍 **Mystery Genre**: Evaluation should exempt *Creativity* and *Emotional Resonance* criteria. Classic detective archetypes (e.g., private investigators, police detectives) and familiar scenario setups (e.g., locked-room mysteries, serial killers) are inherent to the genre and should not be dismissed as overused clichés. Furthermore, exploring the ethical and societal foundations of crime should not be penalized for its dark themes.

☕ **Coming-of-Age, Romance, and Realism Genres**: Evaluation should exempt *Narrative Complexity* and *Creativity* criteria. Fictions centered around personal growth, self-reflection, or everyday life should not be regarded as plain narrative. Their depth often arises from emotional nuance, societal insights, or well-developed character arcs rather than intricate narrative techniques.

B Dataset Construction Details

This section provides additional details on the construction of the fiction dataset, including prompts designed to instruct the model in generating fiction, originality verification for the generated fictions, and data statistics.

B.1 Prompts for Synthetic Fiction

This section provides detailed prompts for generating Synthetic fictions. Figures 6 illustrate the plan-then-write pipeline utilized by GPT-wirter: first, generating a high-level outline of the fiction's chapters; second, breaking down each chapter into multiple paragraphs with specified content and word count requirements; and third, expanding each chapter into a fully developed narrative based on the planned structure. Figure 7 presents the prompt of GLM-writer, which directly generates fictions from an initial instruction without prior planning.

The fiction writing instructions are further divided into two main categories: model-generated titles and given human-authored titles (Title-Based Prompt), following Marco et al. (2024b), who found that LLMs produce higher-quality writing when using human-written titles rather than self-generated ones. We randomly selected 14 titles from Authentic fictions and instructed the LLMs to generate content of Realism genre based solely on these titles. To prevent data contamination and ensure reliance of parametric knowledge, we measured the similarity[10] between the generated and original Authentic content, replacing the title and regenerating the fiction when a high similarity was detected. Details of the similarity computation are provided in the §B.2. For model-generated titles, there are 36 combinations encompassing 6 genres: Fantasy, Martial Arts, Coming-of-Age, Sci-Fi, Mystery, and Romance; 3 narrative settings: Basic Prompt, which only defines the genre; Theme-Oriented Prompt, which further refines the genre with a specific theme; and Character-Centric Prompt, which designates the protagonist's identity; and 2 writing scenarios with or without quality requirements, as detailed in Table 5.

The only difference between English and Chinese prompts lies in the word limit, with the Chinese prompt allowing up to 12,000 characters. The word count specifications for both English and Chinese prompts are based on the average length of Authentic fiction collected.

B.2 Originality Check for Synthetic Fiction

To ensure originality and prevent plagiarism in LLMs writing based on Title-Based Prompt, we calculated the similarity between the generated and source Authentic fictions utilizing Sentence-BERT (Reimers and Gurevych, 2019). This involves two challenges: calculating similarity over long contexts and determining an appropriate threshold.

Given BERT's 512-token limit, we split both the generated and original texts into 300-word or character chunks, computing embeddings for each. Pairwise similarity scores were then calculated for all embedding combinations, with the highest score serving as the final similarity, as illustrated in Figure 8. Determining similarity thresholds proves challenging, as noted by Cadamuro and Gruppo (2023), who set thresholds by analyzing similarity score distributions. Silimarly, we first computed pairwise similarities for the 50 award-winning fictions in each language and plotted the similarity distribution for the C(50, 2) = 1,225 combinations, as shown in Figure 9. The majority of similarity scores clustered in 0.4-0.5 range. Considering the potential keyword overlap due to identical titles, we set the threshold within right-side range at 0.5. Table 7 presents the BERT-Sentence similarity scores for all the Title-Based Synthetic fictions in Chinese and English from the existing dataset, along with their BLEU (Papineni et al., 2002) scores computed in the same approach.

Additionally, we also utilized an online plagiarism detection tool following Harel-Canada et al. (2024) to further ensure the originality of the LLM-generated content, both Synthetic and Collaborative fictions showed a low probability of plagiarism, with an average of 8% and a maximum of 34%.

[10] `https://github.com/UKPLab/sentence-transformers`

Narrative Settings	**Basic Prompt:** Please compose a {Genre} short fiction with a compelling title, comprising approximately *6,000 words*. **Theme-Oriented Prompt:** Please compose a {Genre} shot fiction with the theme: {Theme}. The fiction should have a compelling title and be approximately *6,000 words* in length. **Character-Centric Prompt:** Please compose a {Genre} short fiction centered around {Description of the Protagonist}. The fiction should have a compelling title and be approximately *6,000 words* in length.
Writing Constraints	**Prompt w/ Constraints** {Narrative Setting Prompt}. The fiction should feature engaging narrative techniques, creative originality, emotional depth, a coherent plot, compelling character development, immersive settings, and grammatically sound, fluent, diverse language to ensure overall literary excellence. **Raw Prompt** {Narrative Setting Prompt}.
Human-authored Title	**Title-Based Prompt** Please compose a Realism genre short fiction titled "{Title}" with approximately *6,000 words*. The fiction should feature engaging narrative techniques, creative originality, emotional depth, a coherent plot, compelling character development, immersive settings, and grammatically sound, fluent, diverse language to ensure overall literary excellence.

Table 5: Fiction writing instructions utilized to generate Synthetic fictions for both GPT-writer and GLM-writer. The themes and protagonist settings in purple are detailed in Table 6. The top two rows represent model-generated title prompts, while the bottom row provides a human-authored title.

Genre	Theme	Protagonist
Fantasy	Courage and friendship	A young magic apprentice
Martial Arts	Loyalty and betrayal	A young swordsman
Coming-of-Age	Self-identity and the pursuit of dreams	A group of high school friends
Mystery	Conflict between justice and revenge	A retired detective
Romance	Loneliness and connection in modern interpersonal relationships	A corporate elite
Sci-Fi	Ethical dilemmas and societal transformations faced by human society in coexistence with Artificial Intelligence	A genius scientist

Table 6: The distinct themes and protagonist settings tailored to each genre that utilized in the Basic Prompt, Theme-Oriented Prompt, and Character-Centric Prompt to generate a diverse range of Synthetic fictions. These settings are derived from GPT-4o's brainstorming.

B.3 Prompts for Collaborative Fiction

For Collaborative fiction, we adopted the rewrite method from Zhang et al. (2024). The difference from Synthetic fictions lies in Step I, where it requires reading and analyzing the provided fiction content first. The generation of the high-level outline is based on extraction and summarization of Authentic fiction rather than self-composing, as shown in Figure 10. The subsequent two steps align with Synthetic fictions, which can be referenced in Figure 6.

B.4 Detailed Dataset statistics

Table 8 provides detailed data statistics for each fiction source in both Chinese and English. The Synthetic fiction encompasses six genres with model-generated titles, as well as Realism genre generated under Title-Based Prompts, making a total of seven categories. Collaborative fiction involves rewrites of Authentic fiction and therefore keeps aligned with it in genre.

C Fiction Annotation

C.1 Annotation Details

This section outlines the detailed procedure for carrying out the annotation work. Initially, each annotator was provided with an annotation guideline, which is publicly available in §C.2 and underwent a training process to ensure a thorough understanding of our evaluation framework. Before formal annotation, we conducted a qualification test with 10 fiction samples and provided one-on-one feedback on their initial annotation results, along with a chance for revision. The feedback are reminders about overly strict or lenient evaluations, treating each metric independently, insufficient span highlighting, etc., without directly providing scores. Only those whose revised annotations achieved a high agreement (Krippendorff's $\alpha > 0.8$) with our

TITLE	BERT-Sentence	BLEU	TITLE	BERT-Sentence	BLEU
Roy	0.39	0.10	《青玻璃》 *Green Glass*	0.49	0.17
The Soccer Balls of Mr. Kurz	0.42	0.18	《葛生于野》 *Vines Growing in the Wild*	0.38	0.08
Orphans	0.43	0.16	《耳朵还有什么用》 *What's the Use of Ears*	0.48	0.22
The Home Visit	0.37	0.06	《废墟》 *The Ruins*	0.39	0.00
The Import	0.34	0.11	《盼望羊羔儿》 *Longing for the Lamb*	0.35	0.02
Didi	0.41	0.14	《川流不入海》 *Rivers That Do Not Flow into the Sea*	0.49	0.16
Serranos	0.45	0.26	《我梦见过七月》 *I Have Dreamed of July*	0.46	0.31
Hiding Spot	0.38	0.08	《去有光的地方》 *To the Place of Light*	0.40	0.12
My Good Friend	0.49	0.28	《许多树》 *Many Trees*	0.42	0.16
The Castle of Rose Tellin	0.36	0.04	《折嘴鹦鹉》 *The Parrot with a Broken Beak*	0.42	0.14
Rain	0.46	0.24	《叶子》 *Leaves*	0.39	0.10
Marital Problems	0.42	0.17	《我奶奶的故事及其他》 *My Grandmother's Story and Others*	0.47	0.26
The Last Grownup	0.35	0.02	《做梦发财三题》 *Three Dreams of Getting Rich*	0.34	0.20
The Honor of Your Presence	0.47	0.22	《椿舍里》 *In the House of Spring*	0.44	0.22

Table 7: Titles and similarity scores of Synthetic fictions generated from Title-Based Prompts with English fictions on the left and Chinese on the right.

pre-annotated results were considered qualified to continue. Furthermore, considering that some evaluation metrics are influenced by the annotators' personal knowledge base, like Narrative Complexity, we strongly encourage referring to Wikipedia or other reliable sources for support during the annotation process.

We divided the dataset into batches and assigned two sets of daily tasks to each annotator. Upon receiving the daily annotations, we reviewed the results and rejected them if the Attention Check Question was answered incorrectly or the award-winning fiction's Overall Quality scored below 3. Through the quality control mechanism, we filtered about 27% pieces of the dataset, which was then re-annotated until they met the required standards.

Table 9 displays the Inter-Annotator Agreement calculated for each metric and genre. The Macro-level includes more subjective feelings, making it tough to reach consensus. The strong agreement at the Meso-level might be attributed to our deductions and bonus policy, which breaks evaluation into more fine-grained tasks of identifying weaknesses and strengths. Overall, high annotation consistency validates the robustness of our framework for literary assessment across different genres.

C.2 Annotation Guideline

We developed a comprehensive guideline through multiple revisions, integrating feedback from the authors of this work, expert consultants, and our annotators, following the guideline principles (Ruan et al., 2024). The finalized guideline is publicly available in Table 10 to contribute to future research in the community.

D Experimental Details

This section provides additional details about our evaluation experiments.

D.1 Long-context LLMs

Table 11 showcases the five proprietary and five open-source models used for our assessment. We utilized the provider's API to access the models in zero-shot manners. Although there are many high-performance evaluation LLMs (Jiang et al., 2024; Li et al., 2023a; Hu et al., 2024) designed specifically for NLG tasks, they are unsuitable for our fiction evaluation task due to context window limitations. We have not conducted additional tests on DEEPSEEK-R1 after its release. Given that it is a reasoning model, and based on our results

	English Fiction					Chinese Fiction				
	Award-winning	Web-scraping	GPT-writer	GLM-writer	Colla-borative	Award-winning	Web-scraping	GPT-writer	GLM-writer	Colla-borative
#Fictions	50	50	50	50	50	50	50	50	50	50
#Genre(s)	1	1	7	7	1	1	1	7	7	1
#Author(s)	48	-	1	1	1	28	45	1	1	1
Avg. Len	6,207.04	6,345.22	6,636.48	6,215.88	6,288.80	11,227.18	11,319.48	10,881.28	10,989.84	9,253.86
Stdev. Len	2,169.97	2,320.75	945.97	1,217.04	10,420	2,693.54	2,885.90	1,937.48	2,505.11	2,334.21
Max. Len	11,728	11,665	9,181	11,436	1,865	19,846	19,400	15,465	17,748	13,954
Min. Len	3,581	3,187	5,234	4,295	3,296	6,526	6,770	7,037	6,869	6,041

Table 8: Data statistics of constructed fiction dataset. The number of authors for the English Web-scraped fictions remains unknown as several authors are not publicly accessible on the website.

	Macro-level		Meso-level		Micro-level		Overall Quality	
	EN	CH	EN	CH	EN	CH	EN	CH
Criterion-based IAA								
Metric1	0.73	0.76	0.77	0.78	0.69	0.66	0.72	0.74
Metric2	0.67	0.71	0.75	0.79	0.71	0.73	-	-
Metric3	0.62	0.66	0.75	0.82	0.71	0.71	-	-
Macro-Avg	0.67	0.71	0.76	0.80	0.70	0.70	-	-
Micro-Avg	0.69	0.70	0.77	0.80	0.78	0.77	-	-
Genre-based IAA								
Fantasy	0.67	0.64	0.78	0.67	0.73	0.83	0.70	0.71
Coming-of-Age	0.58	0.62	0.78	0.70	0.70	0.80	0.69	0.72
Martial Arts	0.65	0.66	0.82	0.80	0.92	0.78	0.76	0.75
Mystery	0.71	0.77	0.71	0.82	0.75	0.73	0.74	0.81
Romance	0.62	0.76	0.70	0.81	0.78	0.62	0.72	0.68
Sci-Fi	0.75	0.72	0.83	0.71	0.86	0.87	0.77	0.75
Realism	0.64	0.61	0.66	0.74	0.67	0.79	0.67	0.71

Table 9: IAA of Krippendorff's α, calculated separately for each metric and genre. Metrics 1, 2, and 3 follow the order outlined in §2 across different evaluation levels. Macro-Avg represents the average of the three metrics, while Micro-Avg calculates the average by first combining all the individual samples at each level.

of O1-PREVIEW, reasoning models do not seem to demonstrate its advantage in fiction evaluation tasks.

D.2 Fiction Evaluation Prompts

In a pilot study, we conducted initial tests on three distinct prompt types using a subset that made up one-tenth of the dataset, since testing on the entire dataset would be too costly. The full prompts are illustrated in Figure 11: *Rate-only:* prompts the model to provide only the rating. *Rate-then-explanation:* prompts the model to give a rating first, followed by an analysis for each criterion. *Explanation-then-rate:* prompts the model for the analysis of each criterion fist, followed by the rating. The Pearson (r) correlation results for each prompt on the subset are shown in Table 12. Despite the minor differences between the *Answer-then-explanation* and *Explanation-then-answer* prompts, we selected the most promising *Explanation-then-answer* prompt for the final experiments.

D.3 Label extraction

We automatically extract ratings from the generated answers using regular expressions. Despite employing the *Explanation-then-rate* prompt template, we observed that models occasionally provide answers before the analysis, consistent with the findings of Levy et al. (2024). Additionally, models particularly GLM-4-9B and LLAMA3.1-8B find it challenging to follow the given template, which exhibited error rates of 42.4% and 34.8%, respectively. Additionally, GEMINI-2.0-FLASH have failed to return some results probably due to copyright issues or rate limits, despite multiple attempts. For outputs lacking rating scores, we conducted up to three retries. If the required results were still absent, the instance was marked as a failure, and only the data pieces with valid rating is utilized for evaluation.

E Supplementary Annotation Results

In this section we provide additional details about our annotation results.

E.1 Frequency of Deductions and Bonus Reasons

The frequency in Figure 12 reveals that model-generated fiction exhibit a higher-frequency pattern in the “Others” category (encompassing both strengths and weaknesses), compared to Authentic fiction’s adherence to predefined categories. This highlights the content disparities between human-authored and LLM-generated fictions. We further analyzed the top-ranked subcategories in the “Others” across three metrics through manual grouping. The top three labels were Repetition (109), Overly Optimistic Plotlines (102), and Clichés (63) in Plot Structure; Flat Characters (87), Vivid Characters (55), and Contradictory Characterization (52) in Character Development; and Irrelevant Details (48), Multi-sensory Descriptions (41), and Over-Description (22) in Setting Description. This indicates that repetitive and overly optimistic narratives frequently occur in LLMs-generated texts, setting it apart from authentic fictions crafted by human writers. Chakrabarty et al. (2024) also highlight the deficiencies in the writing capabilities of models, noting that while LLMs like GPT-4 and Claude can generate fluent narratives, they frequently contain plot holes or repetitive themes that are not well-received by human critics.

E.2 Comparison of Genres and Prompt Templates

The spider plot in Figure 13 compares the scores of Synthetic fiction under different prompt settings, which reveals that LLMs show no significant differences with various prompt templates, contradicting Marco et al. (2024a)’s observation that prompts play a crucial role in the creativity of the generated text. Conversely, the model displays a preference for certain genres, yet no single genre consistently excels in all aspects, and it notably underperforms in the Coming-of-Age genre. We did not include a comparison with the Realism genre to prevent unfairness, as it involves given human titles.

E.3 Annotation Score Distribution

Figure 14 illustrates the score distribution of Chinese and English fictions annotated across different evaluation levels. It is evident that Authentic fictions maintain a relatively stable distribution of scores across all levels, followed by Collaborative fictions, while Synthetic fictions exhibit significant variability. This suggests an uneven quality distribution for Synthetic fictions, excelling at the Micro level, but declining at broader Macro and Meso levels. This discrepancy also validates the discriminative capacity of our evaluation framework, confirming its ability to differentiate textual quality across hierarchical assessment dimensions.

E.4 Highlighted Spans in Chinese Fiction

Figure 15 presents highlighted spans in Chinese fictions from the annotation results, revealing similar patterns to those observed in English fictions: well- and poorly-written segments in Authentic fiction are relatively evenly distributed throughout the text, whereas model-generated fiction shows a decline in quality towards the end. The relative proportions of strong and weak segments closely align with their respective human evaluation scores. The main difference between Chinese and English fictions lies in the vertical axis range, with English spans being relatively longer, likely due to language-driven differences in writing styles. Additionally, the decline in quality for Chinese synthetic fiction occurs earlier than in English, possibly due to the higher word count and lower quality of Chinese fictions compared to English ones, as discussed in §4.2.

F Additional Experiment Results

F.1 Spearman’s Correlation Result

Table 13 shows the results of the Spearman’s correlation for the 10 models, aligning with the weak-to-moderate Pearson correlations observed in Table 3.

F.2 Writing and Evaluation Correlation of GLM-4

We present the correlation analysis between GLM-4’s performance of fiction generation and evaluation, as shown in Figure 16, showing moderate to high correlations between the two, consistent with the conclusion in §6. Compared to GPT-4o’s preference for Micro-level performance, GLM-4 shows more balanced performance across all levels.

F.3 Detail Formula For Heatmap

Assume there are K quality dimensions (e.g., Narrative Complexity), and an LLM generates N fiction samples. The human ratings for these N fictions on dimension k are represented as an N-dimensional vector $h^k = (h_i^k)_{i=1}^N$, where h_i^k denotes the human evaluation score for the i-th fiction on that dimension. Similarly, the LLM's ratings for these N fictions on dimension k are represented as $q^k = (q_i^k)_{i=1}^N$. The LLM's fiction writing ability on dimension k, denoted as w^k, is represented by the human evaluation scores: $w^k = h^k$. The LLM's evaluation ability on dimension k, denoted as e^k, is measured by the difference between its ratings and the human evaluation scores: $e^k = \left(1 - \frac{|q_i^k - h_i^k|}{4}\right)_{i=1}^N$. The denominator of 4 accounts for the maximum score discrepancy. We compute the Pearson correlation coefficient $\{r(w^i, e^j)\}_{i=1,j=1}^{K,K}$ between any pair of w^i and e^j.

STEP I : High-Level Outline

You are an exceptional novelist. Please analyze the following Fiction Writing Instructions, then create a compelling title and generate a high-level outline of the fiction's chapters, covering the beginning, setup, climax, resolution, and conclusion.

\n Fiction Writing Instructions:
\n {Write-Prompt}
\n Please structure the fiction in the following format, output the title first, then list each chapter with description on a separate line:
\n Title: [Fiction Title]
\n Chapter 1 – Beginning: [Description of the chapter]
\n Chapter 2 – Setup: [Description of the chapter]
\n …
\n Ensure that the fiction framework is logically coherent and forms a cohesive narrative. Do not output any other content.

STEP II : Detailed Description

You are an exceptional novelist. Please analyze the original Fiction Writing Instructions and the provided Writing Outline, then break down each chapter into multiple paragraphs, specifying the main content of each paragraph and corresponding word count requirements.

\n Fiction Writing Instructions:
\n {Write-Prompt}
\n Fiction Title and Writing Outline:
\n {Generated result in STEP I}
\n Please break down the chapter into the following format, output the chapter description first, then list each paragraph on a separate line:
\n Chapter 1 – Beginning: [Description of the chapter]
\n Paragraph 1 – Main Content: [Detailed description of the paragraph's content] - Word Count Requirement: [e.g., 300 words]
\n Paragraph 2 – Main Content: [Detailed description of the paragraph's content] - Word Count Requirement: [e.g., 500 words]
\n …
\n Ensure that each paragraph is clear and specific, and that all paragraphs adhere to the overall fiction writing instructions. Do not divide paragraphs too finely; each paragraph should be no less than 200 words and no more than 1000 words. Do not output any other content.

STEP III : Incremental Writing

You are an exceptional novelist. I will provide you with the original Fiction Writing Instructions and the Writing Outline with detailed descriptions for each paragraph. Additionally, I will supply you with the chapter that have already written. Please continue to write the next chapter of the fiction with the pre-planned paragraph descriptions.

\n Fiction Writing Instructions:
\n {Write-Prompt}
\n Writing Plan with descriptions for each paragraph :
\n {Generated result in STEP II }
\n Already Written Fiction Chapter:
\n {Previous generated chapter}

\n Please integrate the original Fiction Writing Instructions, Writing Outline with descriptions for each paragraph, and the already Already Written Fiction Chapter, and based on {The outline of the i-th chapter}, continue writing. If necessary, you may add a chapter subtitle at the beginning. Please output only the chapter you have written and do not repeat the already written content.

Figure 6: Prompt of plan-then-write for GPT-writer.

LongWriter Prompt w/o Planning

You are an exceptional novelist. {Write-Prompt}.
\n Please output in the following format:
\n Title: [Fiction Title]
\n Fiction Content: [Fiction Content]

Figure 7: Prompt for GLM-writer to generate fictions directly. {Write-Prompt} can be found in Table 5.

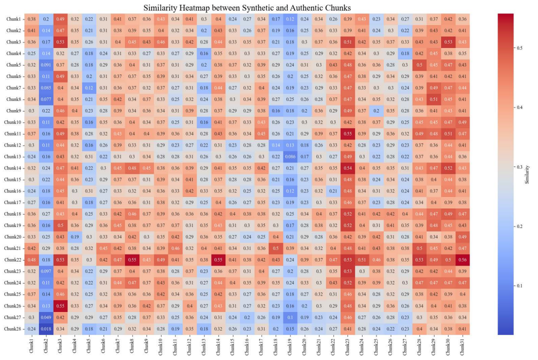


Figure 8: A similarity calculation example: The O. Henry Award-winning work *Orphans* (x-axis) and its corresponding Synthetic fiction (y-axis) were each split into multiple chunks. The highest pairwise score of 0.56 was taken as the final similarity, and the example was filtered out.

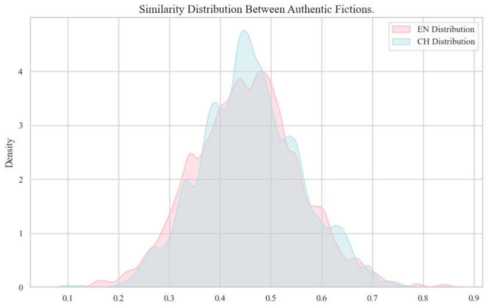


Figure 9: Similarity distribution of all Award-winning fiction combinations in both Chinese and English.

STEP I : Read and Summarize the High-Level Outline

Please read and analyze the provided fiction content, then extract and summarize a high-level outline of the fiction's chapters, covering the beginning, setup, climax, resolution, and conclusion.

\n Fiction Content
\n {The input fiction content}
\n Please summarize the outline in the following format, output the fiction title first, then list each chapter with description on a separate line:
\n Title: [Fiction Title]
\n Chapter 1 – Beginning: [Description of the chapter]
\n Chapter 2 – Setup: [Description of the chapter]
\n …
\n Ensure that the fiction framework is logically coherent and forms a cohesive narrative. Do not output any other content.

Figure 10: Initial step for Collaborative fiction generation, keeping the last two steps aligned with Figure 6.

📚Annotation Guideline for Fiction Evaluation

1. Task Overview

Thank you for participating in this task! We are conducting a project aimed at evaluating human-authored and LLM-generated fictions in both Chinese and English through large language models (LLMs). The average lengths are 10,734 characters for Chinese and 6,339 words for English. Our entertainment-oriented fictions are intended for adult readers which typically explore profound themes and complex narratives, spanning seven mainstream genres including Fantasy, Sci-Fi, Mystery, Martial Arts, Romance, Coming-of-Age, and Realism. You will be presented with a random set of fictions, either extracted from real fictions or generated by LLMs.

Your task is to rate each fiction based on the provided evaluation framework, which includes Narrative Complexity, Creativity and Emotional Resonance at the Macro level, Plot Structure, Character Development, and Setting Description at the Meso level, and Grammaticality, Fluency, Diversity at the Micro level, along with Overall Quality that integrates all these aspects, each using a 5-point Likert scale. Please note that certain criteria have set exemptions depending on the genre; you can find these specifics in Section 3.

Participation in the research project is completely voluntary. Your annotations will contribute to improving automation of fiction evaluation.

2. Fiction Evaluation Aspects

2.1 Macro-level Evaluation

Macro-level focuses on the overall structure of the fiction, considering its narrative structure, creativity, and ability to resonate with the reader.

Ø **Narrative Complexity:** Strategic utilization of literary skills—like non-linear narration, plot twists, and double perspectivation—to enrich plot development.

Ø **Creativity**: Degree of originality exhibited in the core premises, narrative perspectives, and structural design.

Ø **Emotional Resonance:** Effectiveness to evoke empathy in readers, fostering immersive identification with characters and cognitive involvement with the narrated experiences.

2.2 Meso-level Evaluation

Meso-level evaluation is centered on the core elements of fiction—the plot, characters (specifically denotes the protagonist in the fiction), and setting.

We have implemented a point deduction and bonus policy to prevent score inflation and highlight outstanding work. Specifically, a fiction should get punishment in corresponding points if it contains any of the following 👎flaws, and will be awarded if it includes the following 👍highlights. Accordingly, you are encouraged to highlight segments of the fiction that you find particularly well-written (e.g., engaging plot twists, intense conflicts, effectively rendered atmosphere) or poorly executed (e.g., logical inconsistencies, bland storylines, superficial settings). You don't need to precisely align your ratings with the points that are deducted or awarded to prevent bias.

Ø **Plot Structure:** Events arrangement in the fiction, evaluated for its logical coherence, conflict development, and resolution effectiveness.

👎Unresolved Subplots: Secondary storylines remain incomplete or abandoned, with unresolved elements that detract from the main plot.

👎Continuity Errors: Inconsistencies in the narrative, caused either by internal contradictions with established principals or by incorrect external factual errors, undermining plot cohesion and credibility.

👍Well-Executed Plot Twists: The incorporation of unexpected yet plausible plot twists that enhance narrative complexity and engagement, surprising readers while preserving logical coherence.

- Others: Additional plot-related merits and demerits to be enumerated.

Ø **Character Development:** Portrayal and growth of the protagonist within the fiction, evaluated for authenticity, depth, and evolution.

👎Character InconsistencyUnexplained alterations in a character's actions, beliefs, or personality between scenes, undermining believability.

👎Absence of Human FoiblesCharacters lacking redeemable human flaws, resulting in unrealistic or overly idealized portrayals.

👍Dynamic Character ArcsThe evolution of a character throughout the journey, demonstrating personal growth or a meaningful transformation in beliefs, behavior, or emotions.

- Others: Additional character-related merits and demerits to be enumerated.

Ø **Setting Description:** Depiction of the physical, temporal, and cultural environment, evaluated for its contribution to atmosphere building, character development, and plot advancement.

👎Stage DressingTreating the setting merely as a backdrop or decoration, without leveraging it meaningfully to enrich the narrative or develop characters.

👎Monosensory DescriptionRelying solely on a single sensory modality, typically visual, to describe settings, thereby limiting immersion and making the setting feel flat.

👍Enviro-Social DynamicsSkillfully integrating descriptions of natural and social environments, along with multisensory details, to enhance the emotional atmosphere and reflect the era, social dynamics, and relationships driving the plot.

- Others: Additional setting-related merits and demerits to be enumerated.

2.3 Micro-level Evaluation

Micro-level evaluation concentrates on sentence-level grammar, fluency, and linguistic precision, ensuring clarity and readability, in contrast to the Macro and Meso levels, which assess broader narrative structures. Similarly, we invite you to point out any sentences you consider exemplary (e.g., rich vocabulary, smooth flow) or problematic (e.g., grammatical errors or repetitive phrasing).

Ø **Grammaticality:** Adherence to rules of grammar, syntax, and sentence structure. Capitalization is not factored in.

Ø **Fluency:** Smoothness and natural flow of the language.

(Continued from previous page)

Ø **Diversity:** Variety in linguistic expression, including rich vocabulary and varied sentence structures, is especially crucial in fiction to avoid monotony over long contexts.

2.4 Overall Quality Evaluation

Finally, we measure the **Overall Quality** of the fiction, considering all assessment aspects across the aforementioned levels and granularities.

3. Genre-Specific Exemptions

Based on the ten universal criteria, exceptions are defined for certain genres to clarify specific conventions that might be viewed as flaws under general standards but are acceptable within their particular traditions, ensuring more equitable assessments.

3.1 Fantasy and Sci-Fi Genres

- Exemption criteria: Plot Structure, Character Development

- Guidance: Magic, supernatural events, or futuristic technology should not be dismissed illogical simply for defying real-world physics, provided they maintain internal consistency. Similarly, characters with extraordinary abilities can remain credible if their core motivations and emotional struggles are compelling.

3.2 Martial Arts Genre

- Exemption criteria: Grammaticality, Fluency, Diversity

- Guidance: The use of dialects, archaic expressions, or other non-standard grammar are acceptable in martial arts narratives if it serves a deliberate literary purpose rather than stemming from accidental errors.

3.3 Mystery Genre

- Exemption criteria: Creativity, Emotional Resonance

- Guidance: Classic detective archetypes (e.g., private investigators, police detectives) and familiar scenario setups (e.g., locked-room mysteries, serial killers) are inherent to the genre and should not be dismissed as overused clichés. Furthermore, exploring the ethical and societal foundations of crime should not be penalized for its dark themes.

3.4 Coming-of-Age, Romance, and Realism Genres

- Exemption criteria: Narrative Complexity, Creativity

- Guidance: Fictions centered around personal growth, self-reflection, or everyday life should not be regarded as plain narrative. Their depth often arises from emotional nuance, societal insights, or well-developed character arcs rather than intricate narrative techniques.

4. Annotation Procedure

1. Reading Comprehension: Thoroughly read the entire text of the fiction to gain a full understanding of its content. We have set an Attention Check Questions for each fiction. Please answer these questions based on the original text before assigning any evaluation scores.

2. Segment Highlighting: Identify and highlight both well-executed and poorly executed components, encompassing Meso-level segments (e.g., engaging plot twists or logical inconsistencies) and Micro-level sentences (e.g., smooth flow or grammatical errors).

3. Evaluation and Rating: Score each of the ten evaluation aspects on a scale of 1 to 5, where 1 represents the lowest quality and 5 represents the highest. In evaluating each metric, it's necessary to refer to Section 3 to determine whether any exemptions need to be considered based on the current genre. For Meso-level aspects (Plot Structure, Character Development, and Setting Description), please label predefined bonuses or penalties as the rationale for the score.

4. Authorship Determination: After evaluating the above criteria, determine whether the fiction is human-authored or LLM-generated based on your analysis of its style, structure, and language.

5. Review and Submit: Repeat the evaluation process until the entire fiction has been thoroughly reviewed and all aspects have been carefully scored. Submit the annotation results on time.

5. Emphasis and Caution

- **Daily Annotation Requirement:** Fictions for this task will be provided in batches. You are required to annotate two sets, totaling 10 fictions, each day. Please submit your daily annotations before 24:00 (midnight) on the same day.

- **Quality Assurance:** If the answers to the two verification questions for a fiction are incorrect, or if the award-winning fiction receives an overall quality score below 3, the HIT will be rejected. Additionally, we will conduct daily inspections of annotated data. Please ensure high-quality annotations.

- **Support and Reference:** If you encounter any confusion regarding professional knowledge or context, feel free to reach out for clarification. You may also refer to Wikipedia or other reliable sources for further understanding, particularly when evaluating the Narrative Complexity and Creativity metric, which may be influenced by individual knowledge.

- **Feedback Mechanism:** A discussion board is available on the annotation interface for recording queries, concerns, or suggestions. We encourage sharing insights for learning and problem-solving.

6. Confidentiality and Anonymity

Raw data and participant identities will remain confidential and will not be disclosed to anyone outside the research team. The data you provide may be analyzed and used in publications, dissertations, reports, or presentations related to the research project, but your identity will not be revealed. Personal information of participants must not be included in the document, and we will verify the absence of participant identities when aggregating submission data.

Table 10: Full instructions given to annotators of the fiction evaluation task.

MODEL	CONTEXT WINDOW	CHECKPOINTS	AVAIL.	#PARAM
GPT-4-TURBO	128K	gpt-4-turbo-2024-04-09	🔒	-
GPT-4O	128K	gpt-4o-2024-11-20	🔒	-
O1-PREVIEW	128K	o1-preview-2024-09-12	🔒	-
CLAUDE-3.5-SONNET-V2	200K	claude-3-5-sonnet-20241022-v2	🔒	-
GEMINI-2.0-FLASH	1,000K	gemini-2.0-flash-exp	🔒	-
LLAMA3.1-8B	128K	Meta-Llama-3.1-8B-Instruct	🔓	8B
LLAMA3.1-70B	128K	Meta-Llama-3.1-70B-Instruct	🔓	70B
GLM-4-9B	128K	glm-4-9b-chat	🔓	9B
QWEN2.5-72B	128K	Qwen2.5-72B-Instruct	🔓	72B
DEEPSEEK-V3	128K	deepseek-chat	🔓	671B

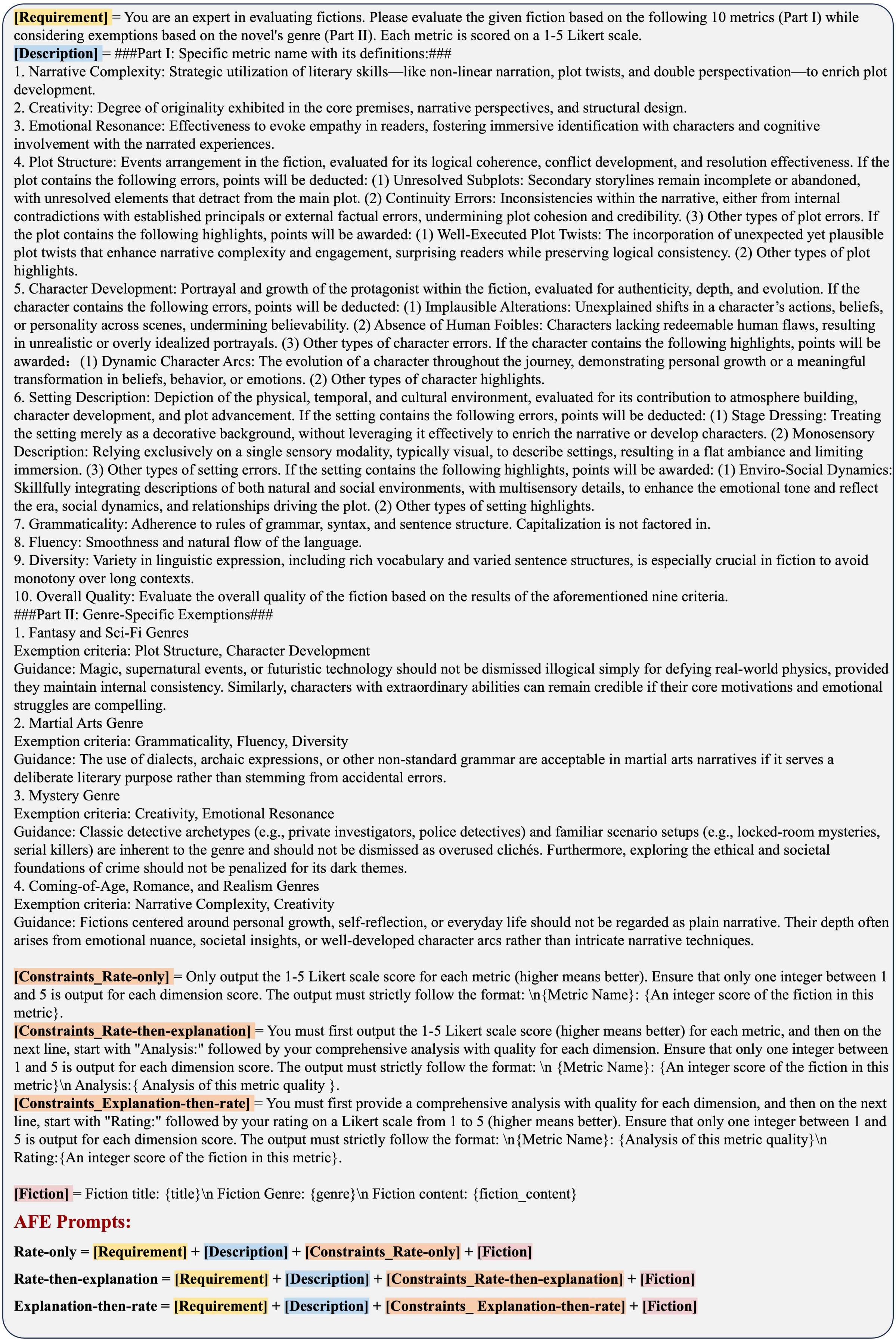
Table 11: Model cards utilized for the experiments.

	Rate-only		**Rate-then-explanation**		**Explanation-then-rate**	
	EN	CH	EN	CH	EN	CH
GPT-4-TURBO	0.18	0.11	0.25	0.14	0.24	0.19
GPT-4O	0.25	**0.26**	**0.35**	0.26	0.35	0.28
O1-PREVIEW	0.21	0.13	0.29	0.18	0.27	0.20
CLAUDE-3.5-SONNET-V2	0.30	0.25	**0.35**	**0.36**	**0.41**	**0.33**
GEMINI-2.0-FLASH	0.29	0.30	0.29	0.31	0.33	0.30
LLAMA3.1-8B	0.01	-0.12	0.00	-0.05	0.04	-0.03
LLAMA3.1-70B	0.15	0.05	0.16	0.08	0.21	0.13
GLM-4-9B	0.05	0.09	0.09	0.09	0.14	0.13
QWEN2.5-72B	0.11	0.09	0.14	0.13	0.17	0.12
DEEPSEEK-V3	**0.31**	0.25	0.34	0.30	0.33	**0.33**

Table 12: The Pearson (r) correlation results for each prompt template on the subset with 50 fiction pieces indicate that the *Explanation-then-rate* performed the best, followed by *Rate-then-explanation*, with *Rate-only* performing the worst.

	📕Macro-Eval				📄Meso-Eval				≡Micro-Eval				Overall-Eval			
	Aut	Syn	Col	Avg.	Aut	Syn	Col	Avg.	Aut	Syn	Col	Avg.	Aut	Syn	Col	Avg.
English Fiction																
GPT-4-TURBO	0.26	0.21	0.19	0.22	0.23	0.25	0.29	0.26	0.29	0.23	0.26	0.26	0.29	0.24	0.17	0.23
GPT-4O	0.37	0.43	0.29	0.36	0.29	0.24	**0.35**	0.29	0.39	**0.41**	0.46	0.42	0.32	0.33	0.24	0.30
O1-PREVIEW	0.46	0.30	0.22	0.33	0.32	0.14	0.22	0.23	0.39	0.30	0.38	0.36	0.22	0.23	0.22	0.22
CLAUDE-3.5-SONNET-V2	**0.53**	**0.46**	**0.57**	**0.52**	0.33	0.26	0.29	0.30	**0.44**	0.40	0.46	**0.43**	**0.44**	0.43	0.47	**0.45**
GEMINI-2.0-FLASH†	0.51	0.31	0.32	0.38	**0.35**	0.15	0.28	0.26	0.37	0.36	0.39	0.37	0.32	0.32	0.31	0.32
LLAMA3.1-8B†	0.17	0.18	0.15	0.16	0.08	0.11	0.16	0.12	-0.01	0.05	0.15	0.06	0.05	0.20	0.07	0.11
LLAMA3.1-70B	0.29	0.18	0.29	0.25	0.23	0.25	0.28	0.25	0.19	0.25	0.13	0.19	0.23	0.36	0.18	0.26
GLM-4-9B†	0.25	0.13	0.19	0.19	0.16	0.05	0.22	0.14	0.27	0.13	0.18	0.19	0.21	0.16	0.09	0.15
QWEN2.5-72B	0.22	0.21	0.19	0.20	0.21	0.13	0.23	0.19	0.24	0.23	0.27	0.25	0.22	0.21	0.22	0.22
DEEPSEEK-V3	0.41	0.36	0.40	0.39	0.32	**0.29**	0.34	**0.32**	0.29	0.37	**0.48**	0.38	0.35	**0.49**	**0.48**	0.44
Chinese Fiction																
GPT-4-TURBO	0.22	0.17	0.26	0.22	0.27	0.13	0.21	0.20	0.24	0.22	0.26	0.24	0.13	0.28	0.19	0.20
GPT-4O	0.22	0.22	0.35	0.26	0.19	0.20	0.38	0.26	0.39	**0.44**	0.40	**0.41**	0.26	0.21	0.22	0.23
O1-PREVIEW	0.15	0.17	0.23	0.18	0.24	0.15	0.30	0.23	0.28	0.36	0.28	0.31	0.18	0.16	0.27	0.20
CLAUDE-3.5-SONNET-V2	**0.54**	**0.45**	**0.48**	**0.49**	0.32	**0.30**	**0.43**	**0.35**	0.36	0.41	0.43	0.40	**0.41**	0.40	**0.56**	**0.46**
GEMINI-2.0-FLASH†	0.35	0.28	0.36	0.33	0.17	0.22	**0.43**	0.27	0.33	0.21	**0.46**	0.33	0.27	0.35	0.45	0.36
LLAMA3.1-8B†	0.18	0.26	-0.05	0.13	0.01	0.08	0.12	0.07	0.04	0.11	-0.04	0.04	-0.06	0.24	0.02	0.06
LLAMA3.1-70B	0.34	0.30	0.12	0.26	0.18	0.18	0.14	0.17	0.28	0.25	0.15	0.23	0.17	0.39	0.23	0.26
GLM-4-9B†	0.15	0.17	0.02	0.11	-0.01	0.15	0.07	0.07	0.06	0.08	0.21	0.12	0.00	0.17	0.02	0.06
QWEN2.5-72B	0.30	0.14	0.08	0.17	0.10	0.19	0.12	0.14	0.27	0.20	0.08	0.18	0.22	0.34	0.10	0.22
DEEPSEEK-V3	0.44	0.42	0.26	0.37	**0.35**	0.21	0.39	0.32	**0.40**	0.40	0.27	0.36	0.33	**0.46**	0.53	0.44

Table 13: Spearman correlation between LLMs and human ratings for both Chinese and English fictions. GEMINI-2.0-FLASH† is calculated based on 436 data pieces, LLAMA3.1-8B† on 288, GLM-4-9B† on 326 and the remaining models are calculated on the full 500-fiction dataset.



[Requirement] = You are an expert in evaluating fictions. Please evaluate the given fiction based on the following 10 metrics (Part I) while considering exemptions based on the novel's genre (Part II). Each metric is scored on a 1-5 Likert scale.

[Description] = ###Part I: Specific metric name with its definitions:###

1. Narrative Complexity: Strategic utilization of literary skills—like non-linear narration, plot twists, and double perspectivation—to enrich plot development.
2. Creativity: Degree of originality exhibited in the core premises, narrative perspectives, and structural design.
3. Emotional Resonance: Effectiveness to evoke empathy in readers, fostering immersive identification with characters and cognitive involvement with the narrated experiences.
4. Plot Structure: Events arrangement in the fiction, evaluated for its logical coherence, conflict development, and resolution effectiveness. If the plot contains the following errors, points will be deducted: (1) Unresolved Subplots: Secondary storylines remain incomplete or abandoned, with unresolved elements that detract from the main plot. (2) Continuity Errors: Inconsistencies within the narrative, either from internal contradictions with established principals or external factual errors, undermining plot cohesion and credibility. (3) Other types of plot errors. If the plot contains the following highlights, points will be awarded: (1) Well-Executed Plot Twists: The incorporation of unexpected yet plausible plot twists that enhance narrative complexity and engagement, surprising readers while preserving logical consistency. (2) Other types of plot highlights.
5. Character Development: Portrayal and growth of the protagonist within the fiction, evaluated for authenticity, depth, and evolution. If the character contains the following errors, points will be deducted: (1) Implausible Alterations: Unexplained shifts in a character's actions, beliefs, or personality across scenes, undermining believability. (2) Absence of Human Foibles: Characters lacking redeemable human flaws, resulting in unrealistic or overly idealized portrayals. (3) Other types of character errors. If the character contains the following highlights, points will be awarded：(1) Dynamic Character Arcs: The evolution of a character throughout the journey, demonstrating personal growth or a meaningful transformation in beliefs, behavior, or emotions. (2) Other types of character highlights.
6. Setting Description: Depiction of the physical, temporal, and cultural environment, evaluated for its contribution to atmosphere building, character development, and plot advancement. If the setting contains the following errors, points will be deducted: (1) Stage Dressing: Treating the setting merely as a decorative background, without leveraging it effectively to enrich the narrative or develop characters. (2) Monosensory Description: Relying exclusively on a single sensory modality, typically visual, to describe settings, resulting in a flat ambiance and limiting immersion. (3) Other types of setting errors. If the setting contains the following highlights, points will be awarded: (1) Enviro-Social Dynamics: Skillfully integrating descriptions of both natural and social environments, with multisensory details, to enhance the emotional tone and reflect the era, social dynamics, and relationships driving the plot. (2) Other types of setting highlights.
7. Grammaticality: Adherence to rules of grammar, syntax, and sentence structure. Capitalization is not factored in.
8. Fluency: Smoothness and natural flow of the language.
9. Diversity: Variety in linguistic expression, including rich vocabulary and varied sentence structures, is especially crucial in fiction to avoid monotony over long contexts.
10. Overall Quality: Evaluate the overall quality of the fiction based on the results of the aforementioned nine criteria.

###Part II: Genre-Specific Exemptions###

1. Fantasy and Sci-Fi Genres
Exemption criteria: Plot Structure, Character Development
Guidance: Magic, supernatural events, or futuristic technology should not be dismissed illogical simply for defying real-world physics, provided they maintain internal consistency. Similarly, characters with extraordinary abilities can remain credible if their core motivations and emotional struggles are compelling.
2. Martial Arts Genre
Exemption criteria: Grammaticality, Fluency, Diversity
Guidance: The use of dialects, archaic expressions, or other non-standard grammar are acceptable in martial arts narratives if it serves a deliberate literary purpose rather than stemming from accidental errors.
3. Mystery Genre
Exemption criteria: Creativity, Emotional Resonance
Guidance: Classic detective archetypes (e.g., private investigators, police detectives) and familiar scenario setups (e.g., locked-room mysteries, serial killers) are inherent to the genre and should not be dismissed as overused clichés. Furthermore, exploring the ethical and societal foundations of crime should not be penalized for its dark themes.
4. Coming-of-Age, Romance, and Realism Genres
Exemption criteria: Narrative Complexity, Creativity
Guidance: Fictions centered around personal growth, self-reflection, or everyday life should not be regarded as plain narrative. Their depth often arises from emotional nuance, societal insights, or well-developed character arcs rather than intricate narrative techniques.

[Constraints_Rate-only] = Only output the 1-5 Likert scale score for each metric (higher means better). Ensure that only one integer between 1 and 5 is output for each dimension score. The output must strictly follow the format: \n{Metric Name}: {An integer score of the fiction in this metric}.

[Constraints_Rate-then-explanation] = You must first output the 1-5 Likert scale score (higher means better) for each metric, and then on the next line, start with "Analysis:" followed by your comprehensive analysis with quality for each dimension. Ensure that only one integer between 1 and 5 is output for each dimension score. The output must strictly follow the format: \n {Metric Name}: {An integer score of the fiction in this metric}\n Analysis:{ Analysis of this metric quality }.

[Constraints_Explanation-then-rate] = You must first provide a comprehensive analysis with quality for each dimension, and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). Ensure that only one integer between 1 and 5 is output for each dimension score. The output must strictly follow the format: \n{Metric Name}: {Analysis of this metric quality}\n Rating:{An integer score of the fiction in this metric}.

[Fiction] = Fiction title: {title}\n Fiction Genre: {genre}\n Fiction content: {fiction_content}

AFE Prompts:

Rate-only = [Requirement] + [Description] + [Constraints_Rate-only] + [Fiction]

Rate-then-explanation = [Requirement] + [Description] + [Constraints_Rate-then-explanation] + [Fiction]

Explanation-then-rate = [Requirement] + [Description] + [Constraints_ Explanation-then-rate] + [Fiction]

Figure 11: Three prompts for LLMs on fiction evaluation, with Explanation-then-rate as the final prompt for the experiment.

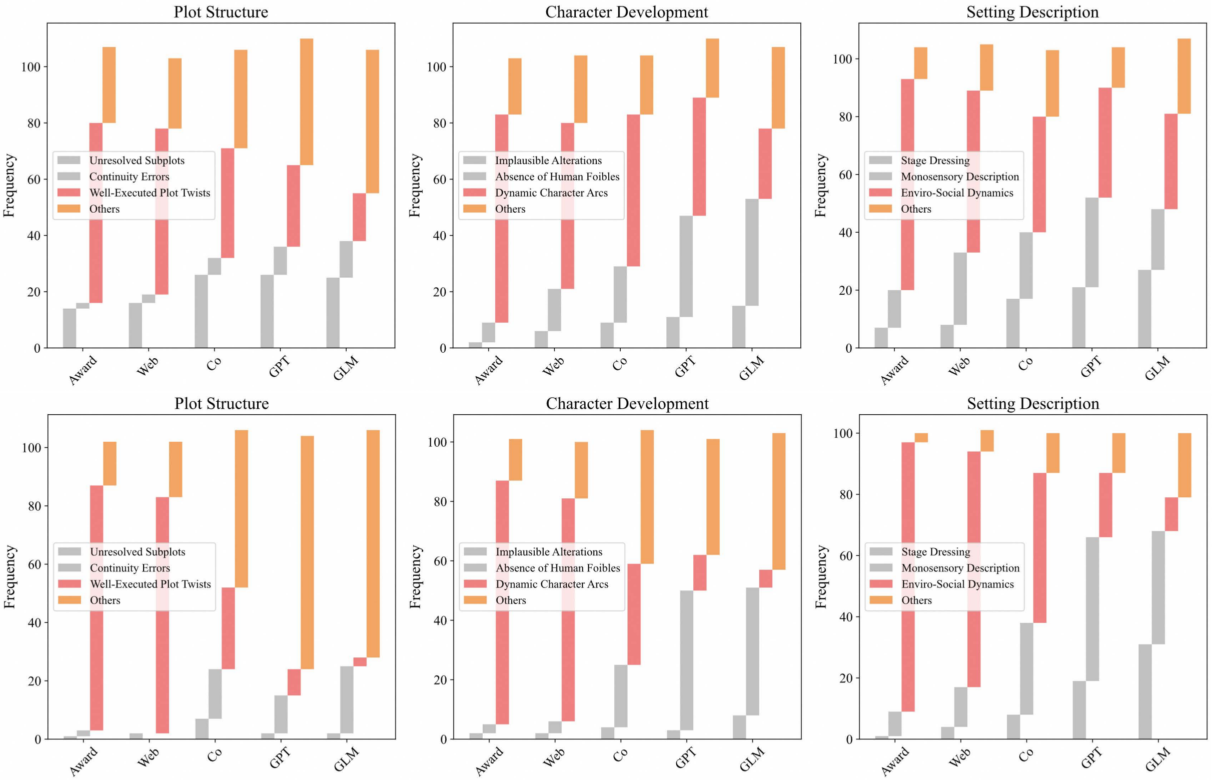


Figure 12: Frequency of Deductions, Bonus and Others for scoring reasons of Award-winning (Award), Web-scraping (Web), Collaborative (Col), GPT-writer (GPT), and GLM-writer (GLM) fictions, in both English (Top) and Chinese (Bottom).

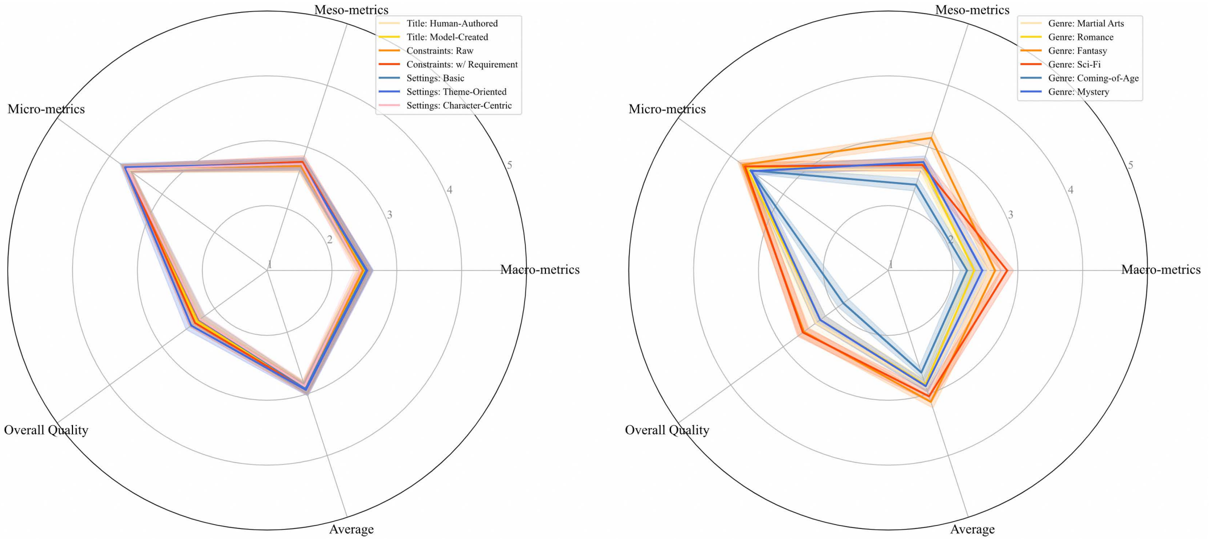


Figure 13: Spider plot comparing LLMs performance based on various prompt templates (left) and genres (right). Values are averages of the Likert scores received for Chinese and English fictions under each setup.

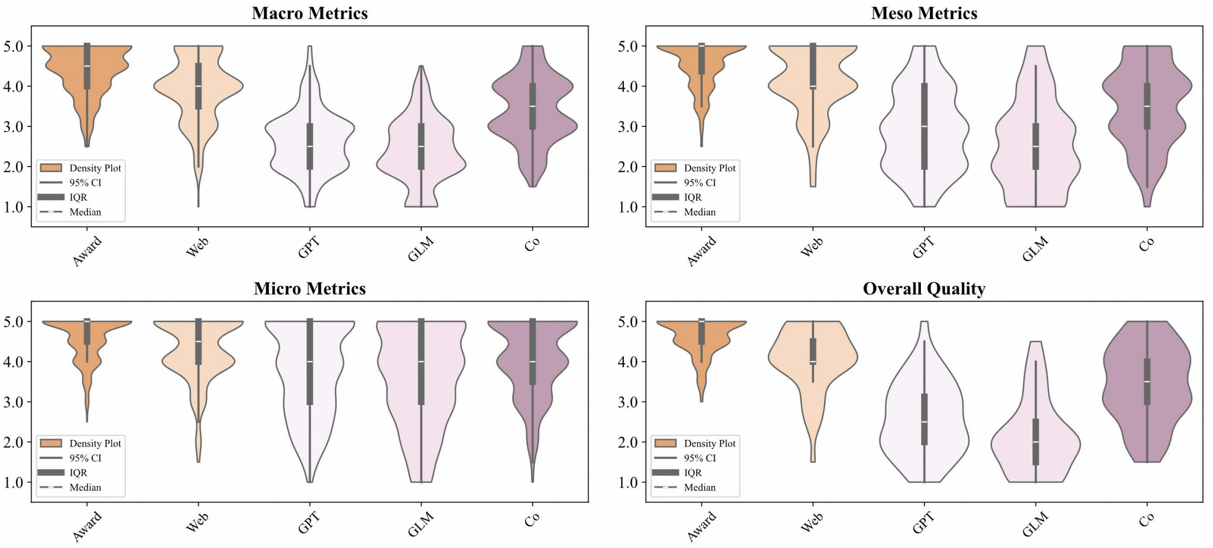


Figure 14: Violin plots showing the score distribution across three evaluation levels and overall quality for Award-winning (Award), Web-scraping (Web), Collaborative (Col), GPT-writer (GPT), and GLM-writer (GLM) fictions, with Chinese and English scores combined.

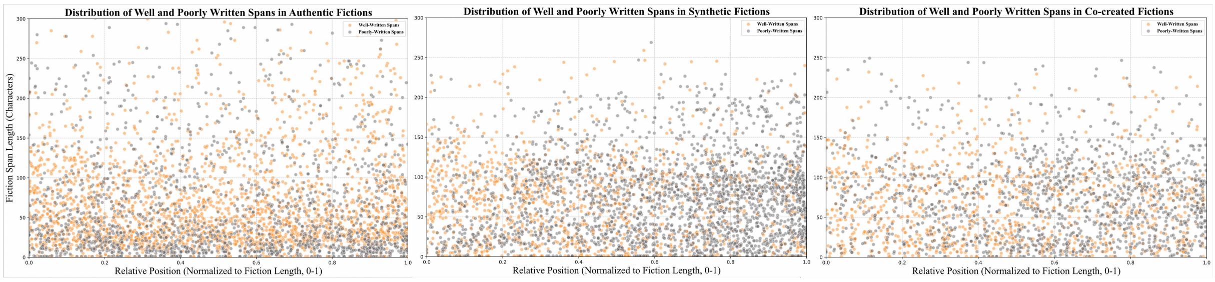


Figure 15: Scatter plot of highlighted spans from three Chinese fiction sources, uses Yellow dots for well-crafted spans and Gray for poor ones.

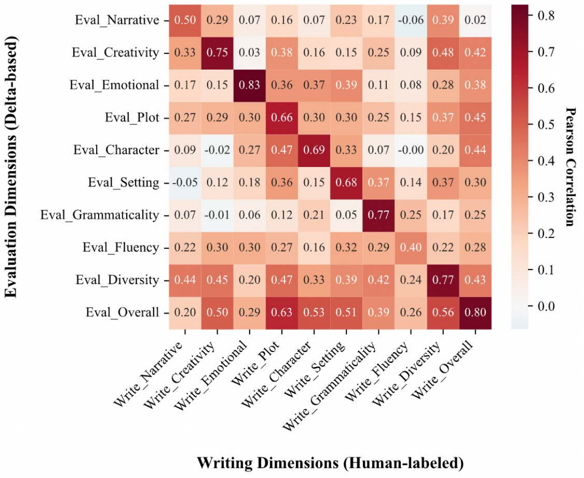


Figure 16: Heatmap displaying the Pearson correlation coefficients between GLM-4's writing and evaluation abilities across different dimensions.