ChangJuan: A Comprehensive Benchmark for Book-Length Chinese Story Evaluation

Dingyi Yang[1*], **Mingshuo Wang**[2*], **Qin Jin**[2†]
[1]Nanyang Technological University
[2]Renmin University of China
dingyi.yang@ntu.edu.sg, {wangmingshuo2004,qjin}@ruc.edu.cn

Abstract

Automatic evaluation of book-length stories remains underexplored, particularly for non-English literature. We introduce ChangJuan, the first benchmark for *book-length Chinese story evaluation*, comprising 300 novels with metadata, human ratings, and large-scale user reviews. To mitigate the subjectivity of raw reviews, we propose a distillation method to aggregate them into generally agreed viewpoints (pros and cons) across key evaluation aspects such as plot and character. We conduct systematic experiments to benchmark current LLMs, analyze aspect importance, and examine genre differences. For book-length story evaluation, we propose an enhanced summary-based method that leverages length-detail balanced summaries and representative excerpts, generates aspect-specific reviews, and considers genre-aware aspect weighting to assign a final score. Using this framework and our distilled viewpoints, we fine-tune an 8B model, CLEM, which outperforms open-source baselines and raises Qwen3's Kendall's tau correlation with human judgments from 24.8 to 34.1. Our datasets and codes are available at https://github.com/DingyiYang/ChangJuan.

1 Introduction

Recent advances in Large Language Models (LLMs) have significantly enhanced the capacity of Automatic Story Evaluation (Chhun et al., 2024). Leveraging their superior semantic comprehension, LLM-based methods correlate much better with human judgments than traditional metrics such as BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020).

Despite this progress on short narratives, evaluating book-length stories remains a formidable challenge. LongStoryEval (Yang and Jin, 2025) made an early attempt, but three key questions remain open: (1) Do benchmarking results remain consistent across *cultures*? For instance, do readers from different cultural backgrounds prioritize the same evaluation aspects? (2) While training on user reviews can enhance models' book evaluation capabilities, such reviews are *inherently subjective*. How can such subjective reviews be better utilized? (3) Summary-based evaluation (using book summaries and selected excerpts) is efficient but inevitably *discards important details*. Are there more effective ways to "compress" lengthy stories while preserving essential elements?

To address these questions, we introduce **ChangJuan**, a comprehensive benchmark dataset for book-length Chinese story evaluation. It comprises 300 long-form fictions published within the past 5 years, with newly published works as the test set to mitigate data contamination problem. For each book, we collect metadata, average user ratings, and multiple user reviews from Douban, the largest Chinese book review platform[1]. To transform inherently *subjective* user reviews into more *objective and general* evaluation signals, we design a **novel pipeline for general viewpoints extraction** (Figure 1): we first extract diverse viewpoints across different evaluation aspects (e.g., plot, world-building; see Figure 3), then aggregate them into shared viewpoints mentioned by multiple readers, ultimately producing the *general pros and cons* of each book.

With ChangJuan, we conduct systematic experiments to benchmark current models, analyzing overall performance, aspect-level effectiveness, and cross-genres differences (see Section 5). To address the limitations of prior summary-based evaluation, we propose **an enhanced summary-based framework** (Figure 2) that integrates: (1) summaries balancing length and plot detail cover-

*Equal contribution.
†Corresponding author.

[1]https://book.douban.com/

Findings of the Association for Computational Linguistics: ACL 2026, pages 41116–41134

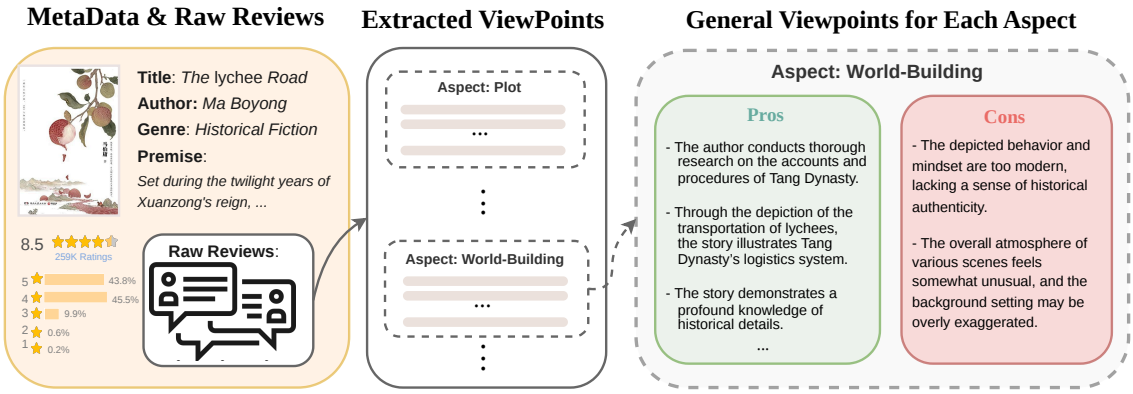

Figure 1: Our data construction process (Detailed in Section 3). English translations are provided for easy reading.

age, (2) carefully selected excerpts that better capture writing quality, and (3) genre-aware weighting that emphasizes genre-specific aspects (e.g., world-building in historical fiction). The framework generates aspect-specific viewpoints, an overall review, and a final score. Our enhanced method is both efficient and effective: its final scores achieve comparable or superior correlation with human-assigned scores than, aggregation-based methods that access the entire book. Finally, leveraging this framework and our distilled viewpoints, we fine-tune an 8B model, **CLEM**, which surpasses existing open-source baselines. Our main contributions are threefold:

- **ChangJuan: The first benchmark for lengthy Chinese story evaluation.** We construct a large-scale benchmark of 300 books, with metadata, summaries, average ratings, and user reviews.
- **Objective and general evaluation viewpoints across multiple aspects.** We introduce a novel pipeline to distill subjective user reviews into general pros and cons, covering diverse evaluation aspects (see Figure 3), enabling more objective and reproducible evaluations.
- **An enhanced summary-based evaluation framework, an expert model CLEM, and systematic explorations.** We propose a summary-based framework featuring length-detailed summaries, representative excerpts, and genre-aware weighting. Leveraging this framework and our distilled viewpoints, we fine-tune an expert evaluation model CLEM, which achieves better evaluation performance than open-source alternatives. Additionally, we provide comprehensive analyses comparing aggregation- and summary-based evaluation, as well as model performance across aspects and genres.

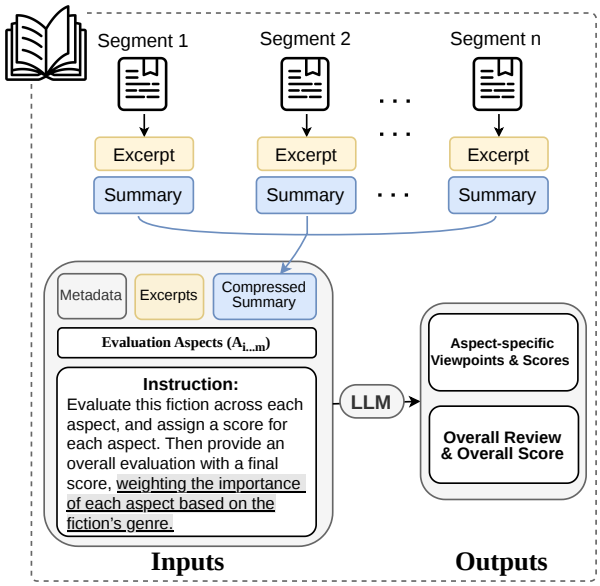

Figure 2: Overview of our evaluation framework (Detailed in Section 4.1.

2 Related Works

Story Evaluation. Story generation is a creative task without definitive answers. Unlike traditional metrics (e.g., BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020)) that compare results with annotated ground truth, evaluating based on human-preferred aspects (Chhun et al., 2022; Guan et al., 2021) is more reasonable. Recent approaches enhance evaluation performance through multi-aspect assessments or by fine-tuning on story datasets to improve comprehension (Guan and Huang, 2020). Despite this progress, most research focuses on short stories, while book-length narratives remain underexplored. Yang and Jin (2025) conducted a preliminary study in this domain. Our work aims to extend these explorations,

improve the evaluation framework, address inconsistencies due to subjective training reviews, and investigate evaluations of stories in Chinese culture.

LLM-based Evaluation. The advancement of LLMs has significantly enhanced LLM-based evaluations (Li et al., 2024; Gao et al., 2024). Leveraging their strong high-level semantic comprehension capabilities, LLMs can now serve as "annotators" for story evaluation. Evaluating across more comprehensive aspects enables LLMs to achieve better performance (Chhun et al., 2024). However, despite these improvements, current LLMs still cannot process whole-book evaluations efficiently. Two potential approaches exist: aggregation-based or summary-based methods. Since the latter offers much better efficiency, we focus on exploring more effective summary-based evaluation techniques that can achieve comparable performance to aggregation-based methods.

3 Dataset: ChangJuan

3.1 Data Collection

We collect 300 long-form Chinese fiction works [2] published in the past 5 years, gathering their metadata and evaluations from *Douban Books*, the largest and most influential Chinese book review platform. Each book includes comprehensive metadata (title, author, premise, genre, publication date) as well as its average rating, rating distribution, and multiple user reviews.

3.2 Review Processing

To convert inherently subjective and noisy reviews into more objective evaluation signals, we propose a systematic viewpoint distillation pipeline that converts multiple user reviews into aspect-aligned pros and cons. The process is detailed below.

Raw Viewpoints Extraction. For each review, we employ a powerful LLM to identify user-mentioned aspects and extract the corresponding viewpoints for each aspect. Owing to its strong Chinese comprehension capabilities, we adopt Deepseek-Chat (Liu et al., 2024) for this step, with temperature set to 0. The prompt used is shown in Figure 8. This stage converts each review into structured, aspect-guided raw viewpoints.

[2]Due to copyright restrictions, we release only summaries, not the full book content.

Figure 3: Aspect Taxonomy and Distribution of correlated viewpoints.

Aspect Taxonomy Construction and Viewpoints Organization. After collecting raw aspect labels from thousands of reviews, we observe numerous naming variants (e.g., "Plot Development" and "Plot Pacing"). To standardize them, we construct a unified aspect taxonomy by merging synonymous or related dimensions and organizing them hierarchically (see Figure 3). All extracted viewpoints are then reclassified under this taxonomy, resulting in organized, book-level viewpoint collections across eight major aspects.

General Viewpoints Extraction. Next, we derive general viewpoints that represent common reader consensus. Since review votes on Douban reflect agreement strength, we incorporate them as a weighting signal. Our analysis shows that review votes follow a strong long-tail distribution—few reviews receive many votes, while most receive few or none. Simply processing all reviews together risks overlooking minority but valuable viewpoints. To balance mainstream and diverse opinions, we process high-vote and long-tail reviews separately: For the high-vote group, all viewpoints are retained, with similar ones merged and conflicts resolved in favor of higher-vote opinions (see Figure 9). For long-tail groups, frequently mentioned viewpoints are preserved, again prioritizing the most common opinions in cases of conflict (see Figure 10). Finally, we merge the results from both groups and classify each viewpoint as a pro or con (see Figure 11). This yields objective, aspect-level general viewpoints.

Following this pipeline, we convert raw reviews into general viewpoints, each categorized as a pro

Dataset	Language	# Stories	# Length	Review Type	# Reviews	# General Viewpoints
OpenMEVA (Guan et al., 2021)	EN	2,000	143 tokens	-	-	-
HANNA (Chhun et al., 2022)	EN	1,056	375 tokens	-	-	-
StoryER-Rate (Chen et al., 2022)	EN	12,669	493 tokens	Overall	45.9K	-
Xie (Xie et al., 2023)	EN	200	79 tokens	-	-	-
Per-DOC (Wang et al., 2024)	EN	596	2.5K tokens	Overall	8.9K	-
LongStoryEval (Yang and Jin, 2025)	EN	600	121K tokens	Aspect-Guided	340K	-
ChangJuan	ZH	300	292K tokens	Aspect-Guided	97.5K	22.1K

Table 1: Comparison between ChangJuan and existing story evaluation datasets.

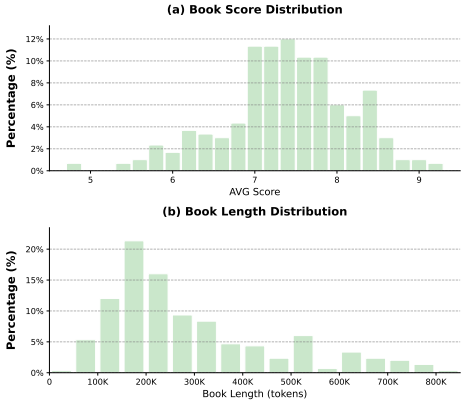

Figure 4: Score and book length distribution.

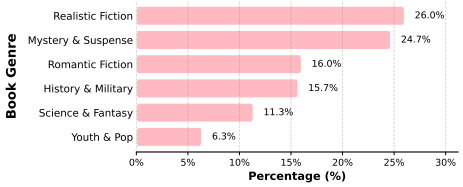

Figure 5: Book genre distribution.

or con under one of the major evaluation aspects. These distilled signals form the foundation for reproducible, interpretable, and aspect-guided story evaluation in ChangJuan.

Quality Control We assess the quality of the extracted viewpoints using Recall and Precision metrics (see Appendix A.2 for details). Our extracted viewpoints achieve a Recall of 0.932 and a Precision of 0.946. This automated precision score aligns closely with our human verification score of 0.91, demonstrating the reliability of our extraction method.

3.3 Statistics and Comparison

In summary, our ChangJuan benchmark dataset contains: (1) 300 fictions with metadata, average ratings, and rating distributions; (2) 97.5K raw reviews; and (3) 22.1K distilled, aspect-aligned general viewpoints. Compared to existing story evaluation benchmarks (Table 1), ChangJuan stands as the first large-scale benchmark for **book-length Chinese story evaluation**, with an average of 292K tokens per story. The length distribution is shown in Figure 4. The dataset encompasses six genres as illustrated in Figure 5, ensuring diversity and enabling our genre-related analysis.

4 Method

Our task requires models to evaluate book-length stories by generating reviews and scores. We adopt the efficient summary-based evaluation approach (Yang and Jin, 2025), which condenses books into plot summaries and key excerpts. We further propose an enhanced summary-based evaluation framework (Section 4.1), and develop an expert model, **CLEM**, to improve evaluation performance of open-source models (Section 4.2).

4.1 Summary-Based Evaluation Framework

Overview. As displayed in Figure 2, the framework inputs a book's metadata (title, genre, and premise), well-constructed summaries, and representative excerpts into the tested models. The models are instructed to generate aspect-specific reviews and scores, followed by an overall assessment and overall score (see Figure 15). Below, we describe key design enhancements.

Detail-Length Balanced Summaries. While humans prefer incremental-updated summaries [3] (Chang et al., 2024) in terms of level of detail, we find that such iterative summarization confuses models—causing over-compression and inconsistent information retention. This leads to unstable summary lengths and detail loss, ultimately degrading evaluation quality.

To address this, we design a two-stage summarization strategy. First, we segment the narrative into coherent units. For short chapters, we attempt

[3]This means incrementally updating an overall summary across chapters.

to retain chapter completeness – if adjacent shorter chapters will not exceed 8192 tokens after combining, they are merged; otherwise, each chapter serves as a unit. For very long chapters exceeding 8192 tokens, we retain paragraph completeness and segment at the paragraph level. We then generate a detailed summary for each unit (See Figure 12 and 13). This approach preserves essential plot developments, character arcs, and contextual details while avoiding omissions of crucial events. Second, we iteratively compress segment summaries into a unified long-form summary of about 4,096 tokens (Figure 14), removing redundancy but retaining pivotal information. This balanced strategy ensures sufficient detail coverage and manageable length, giving models coherent yet information-rich inputs for evaluation.

Representative Excerpt Selection. Summaries alone cannot capture stylistic elements such as tone, narrative rhythm, or character voice. To complement them, we select **representative excerpts** during summarization (Figure 13). Rather than random sampling, excerpts are chosen to reflect stylistic diversity—dialogue, description, and pacing—and to showcase the author's writing quality. We select three excerpts from the beginning, middle, and end of each work. By pairing these with detailed summaries, the model receives both structural and stylistic cues, enabling evaluations that better approximate human literary judgment.

Genre-Aware Weighting. Different genres emphasize distinct narrative elements: e.g., worldbuilding in science fiction versus plotting in mystery. To account for this, we introduce a genre-aware weighting scheme when aggregating aspect-level scores. We do not ask models to output explicit weights for each aspect, but rely on their inner implicit weights for different genres. This could better leverage the LLM's latent knowledge to prioritize aspects—rather than reducing to a rigid linear weighting (See ablations in Sec C).

Our genre-aware weighting ensures that the evaluation aligns with audience expectations and literary conventions, making the framework more faithful to real-world reading practices.

4.2 CLEM: Chinese Long-form Evaluation Model

Objective Review Construction. As described in Section 3, we transform subjective book reviews into objective viewpoints (pros and cons). Using these viewpoints, we build objective review samples for training. For each main aspect A with n_A^+ pros and n_A^- cons, we sample T viewpoints (set as 3 in our experiments). The sampled pros and cons maintain the same ratio as the overall distribution within this aspect. That is, the quota for pros and cons is assigned by:

$$t_A^+ = \lfloor T \cdot \frac{n_A^+}{n_A} \rfloor, \quad t_A^- = \lfloor T \cdot \frac{n_A^-}{n_A} \rfloor. \qquad (1)$$

If $t_A^+ + t_A^- < T$, we will randomly select the remaining quota from the pros or cons, with the higher counts having a higher probability.

Each book thus produces multiple balanced, multi-aspect reviews through iterative sampling. Books with more viewpoints are sampled proportionally more times[4]. This process yields diverse and representative training samples reflecting empirical distribution.

Instruction Tuning Applying our proposed summary-based framework and the constructed objective reviews, we finetune the Qwen3-8B model (Yang et al., 2025) via instruction tuning (Ouyang et al., 2022) with cross-entropy loss. The instruction details are provided in Figure 15.

5 Experiments

We conduct experiments using the ChangJuan dataset. Specifically, we sample 90 books published in the past 2 years as our test set, with the remaining books serving as the training set. To prevent contamination issues in our test set, we conduct experiments to investigate this potential problem. The results are presented in Appendix A.1. Following the previous works, we evaluate the Pearson, Spearman, and Kendall Tau correlations (Pearson, 1895; Spearman, 1961; Kendall, 1938) between human-assigned ratings (the average rating) and model-generated scores for the test books.

5.1 Training Setup

Our training set consists of 210 books and 12.0K constructed objective reviews. The summary and excerpts are achieved through the process described in Section 4.1. We finetune the Qwen3-8B for 5 epochs with a learning rate of $1e^{-5}$ and a batch size of 32. The LoRA parameters (Hu et al., 2022) are configured as $r = 8$ and $alpha = 16$. The

[4]For each book, we compute the maximum number of viewpoints k across aspects, and set the number of sampling times to $3k$.

		Pearson		Spearman		Kendall-Tau	
		r	p-value	r	p-value	r	p-value
Closed-Source	GPT-5	**66.5**	8.1e-13	**61.4**	1.2e-10	**45.6**	8.0e-10
	GPT-5-mini	41.8	4.1e-5	41.8	4.2e-5	29.8	5.5e-5
	GPT-4o	53.9	4.4e-8	55.3	1.6e-8	40.1	5.9e-8
	DeepSeek-Chat	58.6	1.3e-9	56.6	6.3e-9	41.7	1.5e-8
Open-Source	LLaMA3.3-70B	25.0	1.7e-2	35.0	7.3e-4	24.7	1.2e-3
	LLaMA3.1-70B	30.2	3.8e-3	25.4	0.02	18.3	0.01
	LLaMA3.1-8B	7.3	0.5	14.1	0.2	10.4	0.07
	Qwen3-32B	32.4	1.90e-3	32.7	1.6e-3	23.3	1.6e-3
	Qwen3-8B	35.7	5.6e-4	35.6	5.5e-4	24.8	8.2e-4
Ours	**CLEM-8B**	**46.9**	3.1e-6	**48.8**	1.1e-6	**34.1**	3.7e-6

Table 2: Correlation between human and model-assigned ratings under Pearson, Spearman, and Kendall-Tau metrics.

training was conducted on 4 A800 GPUs, taking approximately 20 hours.

5.2 Baselines

Evaluated Models We test several models across two categories: closed-source models (GPT-5, GPT-5-mini, GPT-4o, DeepSeek-Chat) and open-source models (LLaMA3.3-70B (Dubey et al., 2024a), LLaMA3.1-70B (Dubey et al., 2024b), LLaMA3.1-8B, Qwen3-32B (Yang et al., 2025), and Qwen3-8B).

Evaluation Frameworks We compare our summary-based evaluations with the aggregation-based approach (Yang and Jin, 2025). The latter evaluates each segment individually, with previous summaries included during each segment's evaluation to enhance comprehension. The final score is the average of these individual segment scores. Details of our summary-based framework are presented in Section 4.1, with summaries and excerpts generated by the DeepSeek-Chat model.

Generation Parameters For closed-source models, we apply the default generation configurations. For open-source models, we use temperature=0.7, topP=0.8, and topK=30. For all models, we calculate the average of 5 generations.

5.3 Main Results

Model Capacities. Table 2 reveals a significant performance gap between closed-source and open-source models. We suggest that scaling models and incorporating more Chinese cultural context could improve results. Furthermore, the training data focusing on story evaluation is crucial. The Qwen3-8B model shows remarkable performance gains when fine-tuned on our specially constructed review samples.

Aggregation-based vs. Summary-Based. As shown in Table 3, our proposed summary-based method can achieve comparable, even better performance than aggregation-based methods. While aggregation-based methods require processing the entire book five times, summary-based evaluation needs only one complete processing, significantly reducing computing time and API token usage. This demonstrates both efficiency and effectiveness of our summary-based framework.

Ablation Studies. We conduct ablation studies to explore the effectiveness of our summary-based framework:

a. w/o Sum Detail Enrich & b. w/o Sum Length Compress: The prior uses summaries created through incremental summarization (Chang et al., 2024), which are shorter and inconsistent in length, resulting in loss of important details; the latter omits the compression of segment-level summaries. As shown in Table 3, the results confirm the effectiveness of our length-detail balanced summaries. Specifically for GPT-5, compressing the length does not show clear improvement. This might be attributed to GPT-5's long context comprehension ability.

c. w/o Representative Excerpts: This condition randomly samples paragraphs, which may not effectively showcase the story's writing quality. As shown in Table 3, we observe a decline in performance. This demonstrate the effectiveness of our representative excerpt selection algorithm.

d. w/o Genre-weighting: This condition removes the instruction for LLMs to emphasize genre-

Model	Setting	Pearson	Spearman	Kendall-Tau
GPT-5	Aggregation-based	64.8	59.0	43.7
	Our Summary-based	**65.5**	61.4	45.6
	a. w/o Sum Detail Enrich	60.8	59.4	43.2
	b. w/o Sum Length Compress	65.0	**63.6**	**46.7**
	c. w/o Representative Excerpts	63.3	61.2	44.1
	d. w/o Genre-weighting	58.7	55.3	40.3
DeepSeek-Chat	Aggregation-based	**61.2**	**58.1**	**42.8**
	Our Summary-based	59.9	57.9	42.5
	a. w/o Sum Detail Enrich	56.6	56.3	42.4
	b. w/o Sum Length Compress	48.7	47.6	35.2
	c. w/o Representative Excerpts	53.9	53.5	39.1
	d. w/o Genre-weighting	45.4	43.3	32.8
Qwen3-8B	Aggregation-based	35.8	**37.5**	**26.2**
	Our Summary-based	35.7	35.6	24.8
	a. w/o Sum Detail Enrich	27.2	32.8	22.5
	b. w/o Sum Length Compress	31.0	29.8	21.4
	c. w/o Representative Excerpts	32.9	31.2	22.3
	d. w/o Genre-weighting	**36.5**	35.0	25.5

Table 3: Ablation studies of our summary-based evaluation framework.

	PLOT	CHAR	THE	LAN	EMO	EXP	WOR	CREA	Overall
Human (Approximation)	29.3	36.6	22.5	28.1	31.2	14.5	33.1	12.6	100
GPT-5	15.9	30.6	36.3	39.5	29.8	12.4	28.5	19.0	45.6
GPT-5-mini	-6.8	19.3	30.0	27.6	20.6	-6.1	17.7	9.8	29.8
GPT-4o	16.7	32.1	35.3	33.7	34.3	20.2	24.8	20.9	39.7
DeepSeek-Chat	15.0	34.8	40.0	39.1	27.3	10.4	33.4	27.3	42.5
LLaMA3.3-70B	-6.2	15.3	26.2	14.1	20.4	3.4	21.2	18.3	24.6
LLaMA3.1-70B	3.5	14.7	30.2	14.3	18.4	8.6	16.0	15.8	18.3
LLaMA3.1-8B	3.2	11.7	19.7	15.8	14.3	2.6	9.7	8.2	10.4
Qwen3-32B	0.1	14.4	31.3	29	15.9	2.7	22.2	19.7	23.3
Qwen3-8B	-4.8	22.8	29.2	17.2	22.8	1.8	21.9	16.0	24.8

Table 4: The system-level Kendall correlations between the human-assigned scores and model-generated evaluations. We report the correlation between aspect-specific scores and the overall score. *PLOT: Plot, CHAR: Character, THE: Theme, LAN: Language, EMO: Emotional impact, EXP: Reading experience, WOR: World-building, CREA: Creativity*.

related aspects. Results demonstrate that prioritizing genre-related aspects is important. Qwen-8B is an exception; however, we found that in its reasoning process, it states, "I need to focus more on some aspects for this genre," indicating its inherent ability for genre-weighting.

5.4 Further Analysis

Aspect Importance and Aspect-level Evaluation Performance. To explore which aspects readers consider important in Chinese long stories, we calculate the correlation between aspect-specific LLM-evaluated scores and overall ratings. We also use user viewpoints to approximate human-evaluated scores for each aspect. As shown in Table 4, despite reading experience and creativity being frequently mentioned in user reviews (Figure 3), they appear less significant for overall evaluation. Other aspects show no clear differences in importance. A notable distinction between Chinese and English literature emerges in the aspects of writing and world-building (Yang and Jin, 2025), which matters less in English evaluations. This difference may stem from Chinese readers' tendency to focus on details, not just main plots. Regarding model performance on different aspects, we find that for plot and character aspects, LLMs tend to *merely summarize content rather than provide effective evaluation,* especially for less prominent LLMs, resulting in poor evaluation performance in these two categories. Some examples can be found in Figure 6, where Qwen3-8B spends excessive context summarizing the plot and characters in the generated review.

Plot	Character		Score
GTP-5	**GTP-5**		**GTP-5**
多线并进，几条线索（鹰/"小子"、虎头鞋、旧帽子）相互回环，**反转铺垫扎实，城市漫游式推进自然。缺点是巧合感偏重、支线繁密致节奏忽紧忽缓，个别情节承载量过大，主线张力偶被分散。** Multiple narrative threads unfold simultaneously. Several clues — the hawk/"kid," the tiger-head shoes, and the old hat — loop and intersect, while twists are solidly foreshadowed. The story progresses naturally in a city-roaming style. The downside is that coincidences feel a bit overused, and the dense subplots make the pacing fluctuate between tension and slack. Some scenes carry too much narrative weight, occasionally dispersing the main thread's dramatic tension.	人物谱系**丰厚而各具棱角**：刘建国的负罪—悯恤—自惩线**完整有力**；黄娥复杂、悖德与慈爱的拉扯**极具戏剧张力**；于大卫温润内敛，承载跨文化记忆；翁子安"子/锤"双重身份使其既是谜面也是答案。**少数配角略工具化，但总体群像立得住，关系脉络清晰。** The cast of characters is rich and sharply defined. Liu Jianguo's arc of guilt–compassion–self-punishment is coherent and powerful; Huang E's inner conflict between transgression and maternal tenderness brims with dramatic tension; Yu Dawei, gentle and restrained, embodies cross-cultural memory; and Weng Zian's dual identity as "Zi" and "Hammer" makes him both the riddle and its answer. A few supporting roles feel somewhat functional, yet overall the ensemble stands solidly, with relationships clearly traced.	...	**8.0/10**
Qwen3-8B	**Qwen3-8B**		**Qwen3-8B**
小说采用多线叙事，将刘建国寻子、黄娥的过往、大卫的家族秘密、榆樱院的矛盾等线索交织，**情节复杂但逻辑严密。**高潮如铜锤身份揭示、黄娥的死亡、刘建国的救赎等。部分支线（如刘骄华的婚姻危机）**稍显冗长，但整体结构完整。** The novel adopts a multi-threaded narrative, intertwining Liu Jianguo's search for his son, Huang E's past, David's family secret, and the conflicts surrounding Yuying Courtyard. The plot is intricate yet logically coherent. Climaxes unfold one after another — the revelation of Copper Hammer's identity, Huang E's death, and Liu Jianguo's redemption. Some subplots (such as Liu Jiaohua's marital crisis) feel slightly protracted, but the overall structure remains complete and well-balanced.	**人物塑造富有层次，**刘建国的执念与救赎、黄娥的愧疚与成长、于大卫的宽恕与矛盾、杂拌儿的纯真与叛逆、次要角色如刘骄华、老郭头、翁子安等**均刻画深刻。**但部分角色（如谢楚薇）**动机略显单薄。** The characterization is richly layered. Liu Jianguo's obsession and redemption, Huang E's guilt and growth, Yu Dawei's forgiveness and inner conflict, and Zaban'er's innocence and rebelliousness are all portrayed with depth and subtlety. Secondary characters such as Liu Jiaohua, Old Guo, and Weng Zian are also vividly drawn. However, some characters—like Xie Chuwei—have relatively shallow motivations.	...	**8.5/10**
CLEM-8B (Ours)	**CLEM-8B (Ours)**		**CLEM-8B (Ours)**
小说中的**故事转折令人惊讶，结局出乎意料，增加了阅读的惊喜。** 存在**大量重复内容，影响阅读体验**，例如关于"铜锤"的线索过多，**导致故事显得啰嗦。** **部分情节发展较为突兀，缺乏必要的铺垫**，如刘建国与铜锤的身份揭晓显得仓促，**未能充分建立情感基础。** The story's twists are surprising, and the unexpected ending adds a strong sense of intrigue and delight to the reading experience. However, there is a considerable amount of repetition that weakens the narrative flow—for example, the clues surrounding "Copper Hammer" are overly reiterated, making the story feel verbose. Certain plot developments feel abrupt and lack sufficient buildup; for instance, the revelation of Liu Jianguo's connection with Copper Hammer appears rushed and emotionally underdeveloped.	黄娥**这个角色塑造深刻**，展现出坚韧、勇敢、智慧的特点，对孩子的爱超越了亲情，体现了母爱的伟大，**令读者动容。** **杂拌儿这个角色的刻画非常成功**，展现了孩童纯真的特质，**他的形象鲜明，具有很强的感染力。** **多数角色如同扁平化剪影，**如刘建国的性格被简单定义为懦弱、窝囊，而刘骄华则被描绘为唠叨、固执，**这种刻板印象削弱了角色的真实感。** The portrayal of Huang E is remarkably deep, highlighting her resilience, courage, and wisdom. Her love for her child transcends blood ties, embodying the greatness of maternal affection, making readers compressed. The depiction of Zaban'er is successful—his innocence and childlike purity are vividly drawn, giving his character great emotional appeal. Many characters resemble flat silhouettes. Liu Jianguo's personality is narrowly defined by timidity and weakness, while Liu Jiaohua is portrayed as merely nagging and stubborn. Such stereotypical portrayals diminish the sense of authenticity and depth in the character ensemble.	...	**7.2/10.0**

Figure 6: Reviews and score assigned for "When Fire Rolls Across the Sky" (human rating 7.7). We highlight the effective review points.

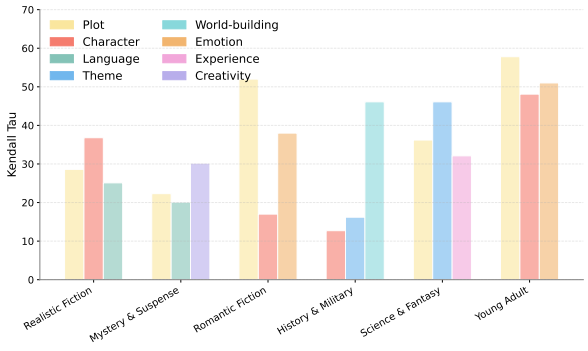

Figure 7: The aspect important of different genres, showing the 3 most important aspects across different genres.

Genre-related exploration Figure 7 displays the genre-related aspect correlations, revealing clear differences across genres. For instance, historical and military fiction emphasize world-building, while mystery and suspense fiction prioritize creativity and plot.

Qualitative Analysis As displayed in Figure 6, less powerful models tend to provide more concise reviews that lack effective analysis of books' pros and cons. Our proposed model CLEM, which has been trained on our well-constructed reviews, shows much better performance.

6 Conclusion

In this work, we present ChangJuan, the first comprehensive benchmark for evaluating book-length Chinese stories. We collect 300 fictions with metadata, summaries, average ratings, and multiple user reviews. To better utilize these subjective user reviews, we convert them into more general viewpoints (pros and cons), which can be applied for further analysis and training. We conduct systematic explorations of current models' performance, examining overall effectiveness across evaluation aspects and genres. For methodology, we propose

an effective and efficient summary-based framework that improves the performance of summary-based methods. Based on this foundation and our constructed general viewpoints, we develop CLEM, an expert 8B evaluation model that significantly improves correlation with human judgments. We hope ChangJuan will serve as a stepping stone toward more objective, interpretable, and culturally adaptive story evaluation, and inspire further research on long-form narrative understanding in both Chinese and multilingual contexts.

Limitations

In this work, we focus on summary-based methods. Our proposed designs achieve comparable or even better performance than aggregation-based methods. For book-length story evaluation, the ultimate goal is to evaluate in one pass. However, due to current challenges and limitations in model capacity for extremely long context comprehension, this remains difficult to address. In future work, we will try to explore the key limitations of one-pass evaluation.

Ethical Problems

We acknowledge and strictly adhere to the Code of Ethics and Professional Conduct throughout this research. The potential ethical concerns are addressed as follows:

Data Source and Copyrights. All of our collected data comes from publicly available websites. All user names and IDs are anonymized to protect personal information. Regarding copyright issues, we will only release summaries of the book contents and processed versions of user reviews. We will release all metadata to ensure it can be accessed.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 62576347).

References

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hong Chen, Duc Minh Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. Storyer: Automatic story evaluation via ranking, rating and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1739–1753. Association for Computational Linguistics.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5794–5836. International Committee on Computational Linguistics.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Trans. Assoc. Comput. Linguistics*, 12:1122–1142.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024a. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Jian Guan and Minlie Huang. 2020. UNION: An unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9157–9166. Association for Computational Linguistics.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6394–6407. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for nlg evaluation: Advances and challenges. *Preprint*, arXiv:2401.07103.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, arXiv:2405.04434.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.

Charles Spearman. 1961. The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2024. Learning personalized story evaluation. *Preprint*, arXiv:2310.03304.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. *Preprint*, arXiv:2301.09790.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dingyi Yang and Qin Jin. 2025. What matters in evaluating book-length stories? a systematic study of long story evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16375–16398, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Model	Input Setting	Pearson	Spearman	Kendall
GPT-5	Whole Content	66.6	61.4	45.6
	Title+Premise	36.9	36.2	20.9
DeepSeek-Chat	Whole Content	58.6	56.6	41.7
	Title+Premise	26.6	29.5	24.6

Table 5: Experiments about contamination issues.

A Dataset

A.1 Contamination Exploration

To explore whether the books in the test set have possible contamination issues, we conduct an experiment by providing the LLMs with only the book title and premise, then analyzing the evaluation results. The results in Table 5 show a significant performance drop, demonstrating minor issues.

A.2 Quality Control

To ensure the quality of the automatic review-to-viewpoints conversion process, we carefully design the prompt and few-shot examples. By sampling 5 books and having our two major authors check the processed results, we demonstrated that the extracted viewpoints have a high accuracy of around 91%—sufficient for model training.

We also evaluate the performance using recall and precision:

- **Recall:** We first remove raw reviews with only 0–1 votes that are mentioned only once. As noted in Figure 10 in our paper, such reviews are too subjective to represent the general opinion. For the remaining reviews, we utilize GPT-5 to check whether each review is covered by the final extracted viewpoints.

Model	Explicit Weights	Implicit Weights (Ours)
GPT-5	44.1	**45.6**
GPT-4o	33.9	**40.1**
DeepSeek-Chat	37.4	**41.7**
Qwen3-32B	19.1	**23.2**

Table 6: Comparison of model performance (Kendall-Tau Correlation with human evaluations) using Explicit vs. Implicit weighting strategies.

If mentioned, it is considered recalled. Sampling 50 books with 6,598 reviews, we achieve a recall score of 0.932.

- **Precision:** We use GPT-5 to verify whether each extracted viewpoint has a source in the original reviews. If it does, it is precise; if it contains hallucinations, it is not. For the sampled 50 books with 1,647 extracted reviews, the precision score is 0.946, which is quite close to our human verification score of 0.91.

B Quality of Book Summaries

We evaluate the quality of book summaries based on coherence and faithfulness:

- **Coherence:** Following BOOKSCORE (Chang et al., 2024), we measure whether each sentence exhibits coherence issues. The coherence score is calculated as the percentage of sentences without such problems. By sampling 50 books, we achieve a coherence score of 0.922.

- **Faithfulness:** Since ground truth summaries are unavailable, we selected 10 popular books and engaged 3 annotators (college students who had read the books within the past 3 months) to verify whether each sentence contained unfaithful content. The final faithfulness score—the percentage of faithful sentences—is 0.938.

C Implicit Genre-Aware Weighting

In current evaluations, we do not ask models to output explicit weights for each aspect, but based on their inner implicit weights for different genres. We consider that implicit weight could be a better choice than explicit. Here we make ablations for these two settings, the results are shown in Table 6. We can find that explicit weights show a clear drop, demonstrating that implicit weighting will be better.

D Prompts

We provide the prompts used in our data processing and story evaluation from Figure 8 to Figure 15.

作为故事创作领域的专家，你的任务是将读者对长篇小说的评论重新编排为结构化JSON格式，包括评价维度和总结。

处理规范
1. 严格识别评论中涉及的评价维度，提取维度对应的核心观点
2. 完整保留重要观点，仅进行结构化重组，禁止添加任何新信息
3. 注意保持原人称视角（如第一人称），就好像你是原始读者，而非转述者
4. 必要时可对语言进行润色（部分评论语言风格较为随意，请修改为正式且规范的表达）

输入示例：
"我追更了三个月，作者对情感线的把控绝了！特别是女主从懦弱到觉醒的蜕变，好几次让我看哭。不过男主后期的黑化感觉铺垫不够，第五章突然就崩人设了。群像戏里工匠老张的支线特别动人，但其他配角就像NPC。环境描写很有电影感，不过打斗场景老是那套'剑气如虹'的模板。最惊喜的是伏笔回收，我猜了二十章的谜底居然是个双反转！"

输出示例：
```
{
    "评价维度": {
        "情感刻画": "女主成长经历完整且富有感染力，情感递进处理细腻",
        "人物塑造": "男主后期转变缺乏充分铺垫，第五章存在人设断裂感",
        "配角塑造": "工匠老张支线刻画深刻，但多数配角缺乏独立人格",
        "场景描写": "环境描写极具镜头美感，但战斗场景重复使用固定模板",
        "伏笔设计": "多层反转的谜底设计超出预期，伏笔回收令人惊喜",
        "节奏把控": "主线推进张弛有度，关键转折具有足够情绪蓄力"
    },
    "总结": "作品在女性角色塑造和伏笔设计方面表现突出，情感传递极具穿透力，但在男性角色逻辑自洽性和战斗场景创新性上仍需完善，配角塑造存在明显不均衡现象"
}
```

As an expert in story writing, your task is to reorganize readers' comments on long novels into a structured JSON format, including evaluation dimensions and a summary.

 # Guidelines
1. Strictly identify the evaluation dimensions mentioned in the comments and extract the core opinions corresponding to each dimension.
2. Retain all important viewpoints completely; only restructure them in a structured format. Do not add any new information.
3. Maintain the original narrative perspective (e.g., first-person), as if you are the original reader, not paraphrasing.
4. If necessary, polish the language (some comments may be casual or informal; revise them into formal and standardized expressions).

Input example:
"I followed the updates for three months, and the author's handling of the emotional line was amazing! Especially the heroine's transformation from timid to awakened, which made me cry several times. However, the male protagonist's dark turn in the later chapters felt underdeveloped; in Chapter 5, his characterization suddenly collapsed. Among the ensemble cast, the subplot of the craftsman Lao Zhang was particularly moving, but other supporting characters felt like NPCs. The setting descriptions were very cinematic, though the fight scenes repeatedly used the same 'sword aura like a rainbow' template. The biggest surprise was the payoff of foreshadowing; I guessed the mystery for twenty chapters, and it turned out to be a double twist!"

Output example:
```
{
    "evaluation_dimensions": {
        "emotional_portrayal": "The heroine's growth is complete and emotionally compelling, with nuanced emotional progression",
        "characterization": "The male protagonist's later transformation lacks sufficient setup, causing a sense of characterization",
        "supporting_characters": "Lao Zhang's subplot is deeply developed, but most supporting characters lack independent personalities",
        "scene_description": "The setting is vividly cinematic, though fight scenes repeatedly rely on fixed templates",
        "foreshadowing_design": "The multi-layered twist exceeded expectations, with foreshadowing payoff being impressive",
        "pacing_control": "The main plot advances with balanced tension, and key turning points are emotionally well-built"
    },
    "summary": "The work excels in female character development and foreshadowing design, delivering strong emotional impact, but the male character's logical consistency and fight scene innovation need improvement, and supporting character development is uneven."
}
```

Figure 8: The prompt for review processing. English translation is provided for easy reading.

以下是针对“{theme}”主题的高票评论。请对这些评论进行聚合与总结，提炼出能代表大众的核心观点。

请遵循以下规则：
- 聚合相似观点：将不同评论中的相似观点合并为一个核心观点。
- 观点不重复：确保每个输出观点都代表一个独特的视角，避免观点内容重复。
- 新观点处理：只要是新颖且不与更高赞观点冲突的观点，都可以被采纳。如果存在冲突，优先采纳更高赞的观点。
- 语言润色：对原始评论进行规范化润色，注意保留其原有意义和细节（比如具体的情节、人名地名、对观点的支撑例子等），不要添加或丢失任何信息。
- 观点返回：输出总结的核心观点及其对应的原始评论。核心观点应直接呈现内容本身，不要添加不要附加“作品”、“多数读者认为”、“部分读者认为”这类前缀。
- 支撑依据：若原始评论包含观点的具体支撑依据或细节描述（例如，评论“语言幽默”并指出原因“主人公买房贷款的吐槽十分幽默”），合并相似观点时不可以丢失这些内容，请在聚合后的观点中完整保留这些细节。
- 主题相关：若观点与“{theme}”这一主题不相关，则丢弃。

示例（以“人物刻画”这一主题为例，体现了 聚合、冲突观点、支撑依据、主题相关等的处理）
高票评论：
- 主人公李长安的性格非常立体，既有坚强的一面也有脆弱的时刻 (votes=374)
- 客栈老板的动机描写很到位，每个决定都符合其性格设定，他在客房前的心里纠结刻画到位 (votes=260)
- 反派不是脸谱化的坏，有自己的苦衷和信念 (votes=189)
- 李长安的成长弧线清晰可见，从天真到成熟的转变很自然 (votes=156)
- 深刻揭露了官场权力的运作机制 (votes=51)
- 客栈老板的行为有时让人难以理解，缺乏一致性 (votes=45)

总结的核心观点：
- 主人公李长安性格立体，既有坚强面也有脆弱时刻，从天真到成熟的成长弧线清晰自然
  - 原始评论：主人公李长安的性格非常立体，既有坚强的一面也有脆弱的时刻 (votes=374)；李长安的成长弧线清晰可见，从天真到成熟的转变很自然 (votes=156)
- 客栈老板动机描写到位，行为符合性格设定，客房前的心里纠结刻画到位
  - 原始评论：客栈老板的动机描写很到位，每个决定都符合其性格设定，他在客房前的心里纠结刻画到位 (votes=260)
- 反派并非脸谱化的坏，有自己的苦衷和信念
  - 原始评论：反派不是脸谱化的坏，有自己的苦衷和信念 (votes=189)

高票评论：
{high_text}

请输出总结的核心观点：（直接以列表形式输出，不要有任何前缀）

Below are high-voted reviews for the theme “{theme}.” Aggregate and summarize these reviews to extract core viewpoints that reflect the general audience’s consensus.

Follow these guidelines:
1. Merge similar viewpoints: Combine overlapping opinions from different reviews into a single cohesive core viewpoint.
2. Avoid redundancy: Ensure each output viewpoint represents a unique perspective—no duplicate content.
3. Resolve conflicts: Adopt novel viewpoints only if they do not contradict higher-voted opinions. If conflicts arise, prioritize higher-voted reviews.
4. Refine language: Standardize phrasing while preserving the original meaning and details (e.g., specific plot points, character names, supporting examples). Do not add or omit information.
5. Format requirements: Output core viewpoints alongside their corresponding original reviews. Present viewpoints directly (no prefixes like “The work” or “Most readers agree”).
6. Preserve evidence: If a review includes supporting details (e.g., “The language is humorous—specifically, the protagonist’s complaints about taking out a mortgage are funny”), retain these details in the aggregated viewpoint.
7. Stay on theme: Discard any viewpoint unrelated to “{theme}.”

Example (Theme: “Characterization” — demonstrates aggregation, conflict resolution, and evidence preservation)
High-voted reviews:
- Li Chang’an (the protagonist) has a well-rounded personality—he shows both strength and vulnerability (votes=374)
- The innkeeper’s motivations are clearly established; every decision aligns with his character, and his inner struggle is vividly portrayed (votes=260)
- The antagonist is not a one-dimensional villain—he has clear苦衷 (motivations) and beliefs (votes=189)
- Li Chang’an’s character arc is distinct: his growth from naivety to maturity feels natural (votes=156)
- The book exposes how power operates in officialdom (votes=51)
- The innkeeper’s actions sometimes feel inconsistent and hard to follow (votes=45)

Summarized core viewpoints:
- Li Chang’an (protagonist) has a well-rounded personality (showing strength and vulnerability) and a clear character arc (growing from naivety to maturity)
  - Original reviews: Li Chang’an (the protagonist) has a well-rounded personality—he shows both strength and vulnerability (votes=374); Li Chang’an’s character arc is distinct: his growth from naivety to maturity feels natural (votes=156)
- The innkeeper’s motivations are clear, with actions that align with his character (his inner struggle is vividly portrayed)
  - Original reviews: The innkeeper’s motivations are clearly established; every decision aligns with his character, and his inner struggle is vividly portrayed (votes=260)
- The antagonist is not one-dimensional—he has clear motivations and beliefs
  - Original reviews: The antagonist is not a one-dimensional villain—he has clear motivations and beliefs (votes=189)

High-voted reviews:
{high_text}

Please output summarized core viewpoints: (List format only; no prefixes)

Figure 9: The prompts for viewpoints extraction in high-vote comments. English translation is provided for easy reading.

请从以下“{theme}”这一主题的长尾评论中，提炼出能代表较多读者看法的核心观点。

请严格遵循以下规则：
1. 优先提取多次出现的观点：优先提取在多条评论中以不同形式表达的相似观点，并将其总结为一个核心观点。如果观点间存在冲突，应采纳出现频率较高的观点。
2. 忽略无意义或孤立的评论：严格忽略那些只代表个人情感（如“我感觉不好看”）或只出现过一次且不具普遍性的评论。
3. 语言润色：对原始评论进行规范化润色，注意保留其原有意义和细节（比如具体的情节、人名地名、对观点的支撑例子等），不要添加或丢失任何信息。
4. 观点返回：输出总结的核心观点及其对应的原始评论，仅当原始评论数量不少于两条时保留该观点。核心观点应直接呈现内容本身，不要添加不要附加“作品”、“多数读者认为”、“部分读者认为”这类前缀。
5. 支撑依据：若原始评论包含观点的具体支撑依据或细节描述（例如，评论“语言幽默”并指出原因“主人公买房贷款的吐槽十分幽默”），合并相似观点时不可以丢失这些内容，请在聚合后的观点中完整保留这些细节。
6. 主题相关：若观点与“{theme}”这一主题不相关，则丢弃。

示例：（以“叙事节奏”这一主题为例，注意有些观点只出现一次就没有留在核心观点里面）
长尾评论：
- 故事开头节奏太慢，差点放弃阅读
- 中间部分情节紧凑，一环扣一环
- 高潮部分转折太快，有点突兀
- 结尾太仓促，很多事情没交代清楚
- 节奏忽快忽慢，影响阅读体验
- 前半部分铺垫太长，后面又进展太快
- 情节推进很自然，张弛有度
- 中期节奏把握得很好，既不拖沓也不仓促
- 前半部分十分拖沓，一直在写日常琐事，没有切入主线剧情

提炼出的核心观点：
- 故事开头节奏慢，前半部分铺垫过长，写太多日常琐事而不切入主线剧情
  - 对应评论（3条）：故事开头节奏太慢，差点放弃阅读；前半部分铺垫太长，后面又进展太快；前半部分十分拖沓，一直在写日常琐事，没有切入主线剧情
- 中间部分情节紧凑，节奏把握得当
  - 对应评论（2条）：中间部分情节紧凑，一环扣一环；中期节奏把握得很好，既不拖沓也不仓促
- 整体节奏忽快忽慢，影响阅读体验
  - 对应评论（2条）：节奏忽快忽慢，影响阅读体验；前半部分铺垫太长，后面又进展太快

长尾评论：
{chunk_text}

请输出提炼出的核心观点：（直接以列表形式输出，不要有任何前缀，因数量不足或主题不相关被忽略的观点不需要附加说明）

Extract core viewpoints that reflect the consensus of many readers from the following long-tail reviews for the theme “{theme}.”

Follow these strict rules:
1. Prioritize recurring opinions: Focus on viewpoints expressed (in varied phrasing) across multiple reviews. If conflicts occur, adopt the more frequently mentioned opinion.
2. Filter irrelevant content: Ignore reviews that only express personal preferences (e.g., “I didn’t enjoy this”) or appear once (no generalizability).
3. Refine language: Standardize phrasing while preserving original meaning and details (e.g., plot points, character names). Do not add or omit information.
4. Format requirements: Output core viewpoints with their original reviews—only retain viewpoints supported by ≥2 reviews. Present viewpoints directly (no prefixes like “The work”).
5. Preserve evidence: If a review includes supporting details (e.g., “The pacing is off—too much time is spent on side stories”), retain these details in the aggregated viewpoint.
6. Stay on theme: Discard any viewpoint unrelated to “{theme}.”

Example (Theme: “Narrative Pace” — Note: Single-occurrence viewpoints are excluded)
Long-tail reviews:
- The beginning drags on—I almost stopped reading
- The middle section is tight and engaging, with seamless plot progression
- The climax feels rushed and abrupt
- The ending is too hurried; many plot threads are unresolved
- The pace is inconsistent, which harms the reading experience
- The first half has excessive setup; the second half moves too fast
- The plot flows naturally, with a good balance of tension and calm
- The middle section’s pacing is perfect—neither slow nor rushed
- The first half wastes time on trivial daily scenes instead of advancing the main plot

Extracted core viewpoints:
- The beginning drags, and the first half has excessive setup (focusing on trivial scenes instead of the main plot)
  - Corresponding reviews (3): The beginning drags on—I almost stopped reading; The first half has excessive setup; the second half moves too fast; The first half wastes time on trivial daily scenes instead of advancing the main plot
- The middle section is tightly paced and engaging
  - Corresponding reviews (2): The middle section is tight and engaging, with seamless plot progression; The middle section’s pacing is perfect—neither slow nor rushed
- The overall pace is inconsistent, harming the reading experience
  - Corresponding reviews (2): The pace is inconsistent, which harms the reading experience; The first half has excessive setup; the second half moves too fast
---
Long-tail reviews:
{chunk_text}

Please output extracted core viewpoints: (List format only; no prefixes or explanations for omitted content)

Figure 10: The prompts for viewpoints extraction in long-tail comments. English translation is provided for easy reading.

我将为你提供来自高赞评论的核心观点和来自长尾评论的独立观点。请将长尾观点中【新的且与高赞观点不冲突】的部分，合并到高赞观点列表中。

请遵循以下规则：
1. 如果长尾观点与高赞观点矛盾，则不被采纳。
2. 如果长尾观点与高赞观点或其他长尾观点内容相似，则合并为一条更全面的观点；如内容完全重复，则直接删除。
3. 注意保留观点的原有意义和细节（比如具体的情节、人名地名、对观点的支撑例子等），不要添加或丢失任何信息。
4. 直接输出合并后的观点列表，不要附加“多数读者认为”、“部分读者认为”这类前缀，无需添加任何解释。
5. 若原始评论包含观点的具体支撑依据或细节描述（例如，评论“语言幽默”并指出原因“主人公买房贷款的吐槽十分幽默”），合并相似观点时不可以丢失这些内容，请在聚合后的观点中完整保留这些细节。
6. 请将正面评价和负面评价分开总结，并标注在观点开头（正面评价为1，负面评价为0）；若一条原始观点中同时包含正负面评价，请拆分。

示例：
高赞核心观点：
- 语言风格干净利落，意境优美，对天空之城场景的描写让人如临其境
- 文笔细腻，富有想象力，善于运用生动比喻和意象
- 作者文笔老辣成熟，文字功底深厚，但是语言说教意味略重，会让人有些不适

长尾独立观点：
- 文笔老辣，对官场权贵心理的刻画入木三分
- 文笔生动形象，将草叶上的晨露比作碎钻让人印象深刻
- 语言多用比喻，引人入胜

合并后的完整观点：
- 1语言风格干净利落，意境优美，对天空之城场景的描写让人如临其境
- 1文笔细腻，富有想象力，善于运用生动比喻和意象，将草叶上的晨露比作碎钻让人印象深刻
- 1作者文笔老辣成熟，文字功底深厚，对官场权贵心理的刻画入木三分
- 0语言说教意味略重，让人有些不适

高赞核心观点：
{high_text}

长尾独立观点：
{chunk_text}

请输出合并后的完整观点：（直接以列表形式输出，不要有任何前缀）

You will receive two sets of viewpoints: core viewpoints from high-voted reviews and independent viewpoints from long-tail reviews. Merge the long-tail viewpoints that are **new and non-conflicting** with the high-voted viewpoints into the high-voted list.

Follow these rules:
1. Reject long-tail viewpoints that conflict with high-voted ones.
2. Merge similar viewpoints (high-voted ↔ long-tail or long-tail ↔ long-tail) into a more comprehensive viewpoint; delete exact duplicates.
3. Preserve original meaning and details (e.g., plot points, supporting examples). Do not add or omit information.
4. Format requirements: Output the merged list directly (no prefixes like “Readers think”) and no extra explanations.
5. Preserve evidence: If a viewpoint includes supporting details (e.g., “The prose is vivid—comparing dew on grass to broken diamonds”), retain these details in the merged version.
6. Label sentiment: Mark positive evaluations with “1” and negative ones with “0” at the start of each viewpoint. Split viewpoints that mix positive and negative sentiment into two separate entries.

Example:
High-voted core viewpoints:
- The language is concise and elegant, with vivid imagery—descriptions of the “sky city” feel immersive
- The prose is exquisite and imaginative, using rich metaphors and imagery
- The author’s writing is mature and skilled, but the didactic tone feels heavy and uncomfortable

Long-tail independent viewpoints:
- The writing is sharp, with incisive portrayals of officials’ psychology
- The prose is vivid—comparing dew on grass to broken diamonds leaves a strong impression
- The language uses frequent metaphors, making it engaging

Merged comprehensive viewpoints:
- 1The language is concise and elegant, with vivid imagery—descriptions of the “sky city” feel immersive
- 1The prose is exquisite and imaginative, using rich metaphors and imagery (e.g., comparing dew on grass to broken diamonds)
- 1The author’s writing is mature and skilled, with sharp, incisive portrayals of officials’ psychology
- 0The didactic tone feels heavy and uncomfortable

High-voted core viewpoints:
{high_text}

Long-tail independent viewpoints:
{chunk_text}

Please output merged comprehensive viewpoints: (List format only; no prefixes)

Figure 11: The prompts for viewpoints merging. English translation is provided for easy reading.

小说开头片段：

{}

指令：
为上述长篇小说开头创作一份全面摘要，涵盖情节摘要和1段原文精选。

指南：
1. 情节：总结关键事件、冲突、背景、场景和主要人物。若有非线性叙事、闪回或视角切换，请说明其时间和场景。**保持原书叙述人称与视角。**
2. 语言要求：减少无意义辞藻和过多描述，确保文字信息密度高。保持情节发展连贯，让未读过小说的人也能完整了解主要人物及情节走向。
3. 原文精选：选出1段（或连续几段）最能体现小说写作风格的片段（如心理、场面、对话、动作描写等），直接返回原文，无需解释。
4. 由于格式原因，在小说正文开始前可能会包含版权、目录、前言等信息，**忽略**即可，不要作为摘要总结进来。

输出格式：
情节摘要：
原文精选：

 # Start of the Novel:

{}

 # Instruction:
Create a comprehensive summary for the opening of the above full-length novel, including a plot summary and one selected excerpt from the original text.

 # Guidelines:
1. Plot: Summarize key events, conflicts, backgrounds, settings, and main characters. If there is non-linear narration, flashbacks, or perspective shifts, specify their corresponding timeframes and settings. **Retain the original narrative perspective and point of view of the novel**.
2. Language Requirements: Reduce meaningless phrases and excessive descriptions to ensure high information density. Maintain the coherence of plot development so that readers who have not read the novel can fully understand the main characters and plot direction.
3. Original Excerpt Selection: Choose one excerpt (or several consecutive excerpts) that best reflects the novel's writing style (e.g., psychological descriptions, scene depictions, dialogue, action descriptions, etc.). Return the original text directly without additional explanations.
4. Due to formatting reasons, the content before the novel's main text may include copyright information, table of contents, preface, etc. **Ignore such content** and do not include it in the summary.

 # Output Format:
Plot Summary:
Representative Excerpt:

Figure 12: The prompts for story beginning summarization. English translation is provided for easy reading.

 # 当前小说片段：

{}

 # 前文摘要：
情节摘要：{}

 # 指令：
分析当前小说片段，更新整体摘要（包括情节和原文精选）。
注意，有些较长章节可能进行了拆分，前后拆分块间可能保留了部分的重合段落。

 # 指南：
1. 当前片段摘要： 总结当前片段。
2. 情节更新： 在前文摘要基础上，纳入当前片段中关键事件、冲突、背景、场景和人物等重要信息，更新后的整体摘要原则上字数要比更新前字数多。若有非线性叙事、闪回或视角切换，请说明其时间和场景。**保持原书叙述人称与视角。** 注意**合理分段**以提高清晰度。
3. 语言要求： 减少无意义辞藻和过多描述，确保文字信息密度高。保持情节发展连贯，让未读过小说的人也能完整了解主要人物及情节走向。
4. 原文精选： 选出1段（或连续几段）最能体现小说写作风格的片段（如心理、场面、对话、动作描写等），直接返回原文，无需解释。

 # 输出格式：
当前片段摘要： （不要超过200字）
更新后的整体情节摘要：
原文精选：

 # Current Segment:

{}

 # Previous Summary:
Plot Summary: {}

 # Instruction:
Analyze the current novel excerpt and update the overall summary (including plot and selected original excerpt). Note that some longer chapters may have been split, and there may be partial overlapping paragraphs between consecutive split sections.

 # Guidelines:
1. Current Excerpt Summary: Summarize the current excerpt.
2. Plot Update: Build on the previous summary by integrating key information from the current excerpt, such as key events, conflicts, backgrounds, settings, and characters. In principle, the word count of the updated overall summary should be greater than that before the update. If there is non-linear narration, flashbacks, or perspective shifts, specify their corresponding timeframes and settings. **Retain the original narrative perspective and point of view of the novel**. Pay attention to **reasonable paragraph division** to improve clarity.
3. Language Requirements: Reduce meaningless phrases and excessive descriptions to ensure high information density. Maintain the coherence of plot development so that readers who have not read the novel can fully understand the main characters and plot direction.
4. Original Excerpt Selection: Choose one excerpt (or several consecutive excerpts) that best reflects the novel's writing style (e.g., psychological descriptions, scene depictions, dialogue, action descriptions, etc.). Return the original text directly without additional explanations.

 # Output Format:
Current Excerpt Summary: (Do not exceed 200 words)
Updated Overall Plot Summary:
Representative Excerpt:

Figure 13: The prompts for latter segments summarization. English translation is provided for easy reading.

 # 长篇小说情节：

{concatenated_summaries}

你是一名资深的文学编辑，擅长撰写精炼的内容摘要。你的核心任务是将提供的长篇小说情节，提炼成约4000字的连贯摘要。

 # 指令：
1. 保留重点：尽可能保留核心剧情、主要人物关系、关键背景与重要转折点。可以压缩或适当删除冗余的叙述和次要的情节。
2. 禁止新增：不能引入任何原文没有的情节、角色或设定。
3. 保持风格：沿用原有的语言风格和叙述节奏，确保自然流畅。
4. 长度控制：目标长度为 4000字。可接受500字的浮动范围。

 # 输出格式：
请直接开始你的摘要正文，无需任何开场白、标题或结束语。

 # 压缩后的摘要：

 # Long Fiction Plot:

{concatenated_summaries}

You are a senior literary editor skilled in writing concise content summaries. Your core task is to condense the provided long novel plot into a coherent summary of about 4000 **Chinese characters**.

 # Instructions:
Preserve Key Points: Retain the core storyline, main character relationships, crucial background details, and major turning points. You may compress or appropriately remove redundant narration and minor subplots.
No Additions: Do **not** introduce any new events, characters, or settings not present in the original text.
Maintain Style: Follow the original linguistic tone and narrative rhythm to ensure a natural and smooth flow.
Length Control: The target length is *4000* characters, with an acceptable range of ±500 characters.

 # Output Format:
Please begin directly with the summary text — no introduction, title, or conclusion.

 # Compressed Summary:

Figure 14: The prompts for story summary compression. English translation is provided for easy reading.

下面是一本长篇故事的书名、类别和故事简介。

书名: {Title}

类别: {Genre}

故事简介: {Premise}

下面是这本长篇故事的总体情节概要和原文精选片段。

情节概要:
{Plot_Summary}

原文精选片段（主要体现语言风格和文笔技巧）：
{Excerpt}

作为小说读者，请对这本小说的各个维度进行评价和评分，综合考虑优缺点，全面客观地评价，并在1-10分范围内给出评分。
最后，请提供整体评价和评分。在评定整体分数时，请根据小说的类型决定各维度对整体分数的影响。
请按以下格式进行输出：

1. 情节：
- 评价：
- 评分：x.x/10

2. 人物：
- 评价：
- 评分：x.x/10
...

整体评价：
- 评价：
- 整体评分：x.x/10

Below is the title, genre, and premise of a long-form story.

Title: {Title}

Genre: {Genre}

Story Premise: {Premise}

The following section provides the overall plot summary and selected excerpts from the original text.

Plot Summary: {Plot_Summary}

Selected Excerpts (highlighting the author's writing style and literary techniques): {Excerpt}

As a novel reader, please evaluate this work across multiple dimensions. Consider both strengths and weaknesses to provide a comprehensive and objective assessment.
Assign a **score from 1 to 10** for each dimension.
Finally, provide an **overall reviewand score**, taking into account the novel's genre when determining the relative weight of each dimension.

Please follow the format below:

1. Plot:
Review:
Score: x.x / 10

2. Characters:
Review :
Score: x.x / 10
…

Overall Review:
Review :
Overall Score: x.x / 10

Figure 15: The prompts for story evaluation. English translation is provided for easy reading.