What Matters in Evaluating Book-Length Stories? A Systematic Study of Long Story Evaluation

Dingyi Yang, Qin Jin*
Renmin University of China
{yangdingyi,qjin}@ruc.edu.cn

Abstract

In this work, we conduct systematic research in a challenging area: the automatic evaluation of book-length stories (>100K tokens). Our study focuses on two key questions: (1) understanding which evaluation aspects matter most to readers, and (2) exploring effective methods for evaluating lengthy stories. We introduce the first large-scale benchmark, **LongStoryEval**, comprising 600 newly published books with an average length of 121K tokens (maximum 397K). Each book includes its average rating and multiple reader reviews, presented as critiques organized by evaluation aspects. By analyzing all user-mentioned aspects, we propose an *evaluation criteria structure* and conduct experiments to identify the most significant aspects among the 8 top-level criteria. For evaluation methods, we compare the effectiveness of three types: *aggregation-based, incremental-updated*, and *summary-based* evaluations. Our findings reveal that aggregation- and summary-based evaluations perform better, with the former excelling in detail assessment and the latter offering greater efficiency. Building on these insights, we further propose **NovelCritique**, an 8B model that leverages the efficient summary-based framework to review and score stories across specified aspects. NovelCritique outperforms commercial models like GPT-4o in aligning with human evaluations. Our datasets and codes are available at https://github.com/DingyiYang/LongStoryEval.

1 Introduction

Automatic Story Evaluation involves providing critiques and ratings to assess the quality of human-written or machine-generated stories. This process is crucial for recommendation systems or offering constructive feedback for improvement. Unlike simpler evaluation tasks that focus on fluency and accuracy (e.g., machine translation), story evaluation demands a comprehensive assessment, based on diverse human-centered criteria (Chhun et al., 2022). While recent advances have improved the evaluation of short stories (Guan and Huang, 2020; Guan et al., 2021; Chhun et al., 2022), particularly with the aid of large language models (LLMs) (Jiang et al., 2024b; Xie et al., 2023b), the evaluation of book-length stories (exceeding 100K tokens) remains significantly underexplored.

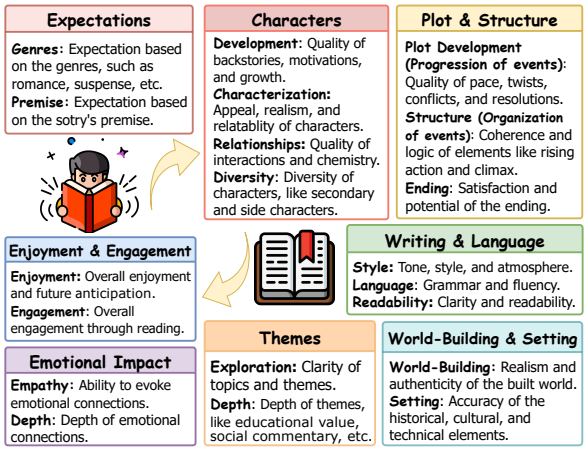

Figure 1: Our proposed *evaluation criteria structure* and the reading process: A reader approaches a book with initial **expectations** based on its genres and premise. The story unfolds through **character**-driven **plots**, revealing its **themes** and **world-building** through the author's **writing**. Through reading, the reader experiences **enjoyment**, **engagement**, and **emotional impact**, and determines whether this book meets the expectations.

Evaluating book-length stories poses three major challenges: (1) **Data Annotation Constraints**: Human evaluation, while the gold standard, is time-intensive and cognitively demanding. As shown in Table 1, existing story evaluation benchmarks focus on shorter texts (100-2,000 tokens). Scaling human annotations for stories exceeding 100K tokens is impractical. (2) **Inconsistent Evaluation Criteria**: Most prior works rely on predefined criteria for evaluation, but there is no universal standard. Evaluation criteria vary across studies and often fail to

*Corresponding Author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16375–16398
July 27 - August 1, 2025

reflect actual reader preferences. Our work aims to explore what **real readers** value in lengthy stories. (3) **Long Story Processing**: Book-length stories often exceed the 128K-token context limit of most LLMs, posing challenges for effective evaluation. Even within this limit, processing such long contexts remains challenging for models. Identifying efficient evaluation strategies for lengthy stories is therefore critical.

To address these challenges, we collect ratings and reviews for 600 newly published lengthy novels from online readers. To completely avoid data contamination issues (Chang et al., 2024) that might affect our experimental analysis, none of these books were included in the training data of evaluated LLMs. The raw review data is sourced from GoodReads[1], the largest book review platform. Using LLMs, we extract over 1000 reader-mentioned evaluation aspects, analyze the most frequent ones, and organize them into a *hierarchical criteria structure* (Figure 1). We further compare three types of processing methods for lengthy story evaluation: *aggregation-based*, *incremental-updated*, and *summary-based*. Additionally, we introduce *NovelCritique*, a specialized model for reviewing and scoring lengthy stories across specified aspects, which demonstrates superior alignment with human evaluations compared to commercial models.

Our contributions are summarized as follows:

- **LongStoryEval: A benchmark for lengthy story evaluation.** We introduce a large-scale benchmark comprising 600 books (published between 2024 and January 2025), with average rating scores and 340K reader reviews. Raw reviews are converted into structured critiques, overall assessments, and ratings, as shown in Figure 2 (d). Metadata, including book details (e.g., title, genres, premise) and reviewer profiles, is provided to facilitate future research.
- **A hierarchical structure of evaluation criteria, and analysis of significant aspects.** By analyzing all aspects raised by real readers, we develop a hierarchical evaluation criteria structure with 8 main aspects and 20 sub-aspects (Figure 1; Table 9). Our experiments reveal that *plot* and *characters* are the most influential objective aspects, while subjective aspects — *emotional impact*, *overall enjoyment & engagement*, and *expectation fulfillment* — are also critical to overall ratings.
- **Explorations on effective methods for lengthy story evaluation.** Among the three types of lengthy story processing methods, *aggregation-* and *summary-based* evaluations perform best. Our findings suggest that the most cost-efficient method involves generating a concise summary and averaging multiple summary-based evaluation results. Further experimental analysis is provided in §5.4.
- ***NovelCritique*: A specialized model for lengthy story evaluation.** We propose NovelCritique, an 8B model capable of reviewing and scoring lengthy stories across specified aspects. It outperforms commercial LLMs such as GPT-4o in aligning with human ratings.

2 Related Works

Story Evaluation. Story generation is a creative and open-ended task, making it more appropriate to explore metrics based on specific human standards. Traditional lexical-based metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) correlate poorly with human judgments. More recent metrics based on pre-trained neural networks, like BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), achieve better semantic comprehension. However, they still struggle to align well with human standards in story evaluation. To address this, several works (Guan and Huang, 2020; Ghazarian et al., 2021; Chen et al., 2022; Maimon and Tsarfaty, 2023) have conducted further training on story evaluation datasets or explored methods based on detailed analysis (Jiang et al., 2024b; Xie et al., 2023b) to improve performance. However, these explorations remain limited to *short stories* generated from ROC and WP datasets. The *criteria* used might also be restricted to predefined ones (Chhun et al., 2022; Xie et al., 2023a; Wang et al., 2024). These evaluation standards are inconsistent (Yang and Jin, 2024), and how well they align with actual readers' preferences remains unclear.

LLM-Based Evaluation. The development of large language models also boosts LLM-based evaluations (Li et al., 2024; Gao et al., 2024). Through carefully designed prompts (Chen et al., 2023; Kim et al., 2023) and helpful strategies (Chan et al., 2024; Saha et al., 2024; Lee et al., 2024), existing methods can achieve good correlation with humans. However, methods based on closed-source models

[1] https://www.goodreads.com

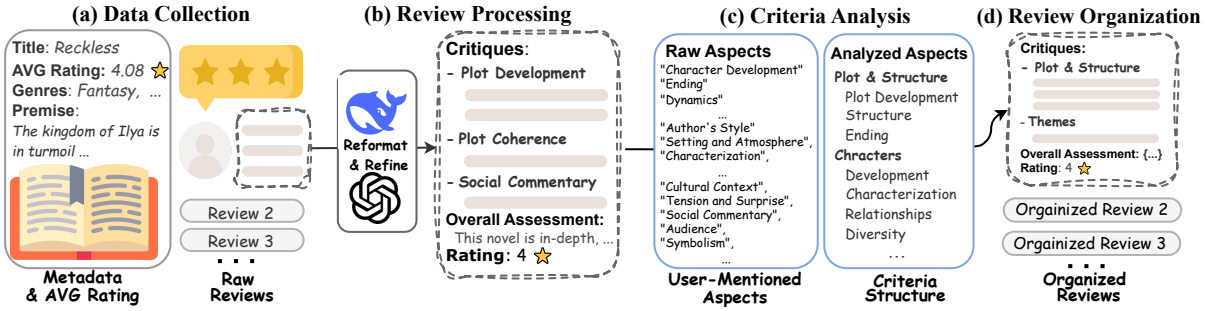

Figure 2: Our data construction process (§3).

can face problems of bias and inconsistency (Stureborg et al., 2024). Open-source LLM evaluators (Li et al., 2023; Kim et al., 2024), on the other hand, include only a small portion of creative story evaluation data in their pretraining. Considering the important role of lengthy stories in people's daily lives, we attempt to explore how current LLMs handle lengthy story evaluation, compare different evaluation strategies, and propose a specialized evaluation model.

3 LongStoryEval Dataset

Data Collection. Considering the high cost and time constraints of human annotations, we leverage large-scale online reviews from real readers. Our dataset comprises 600 newly published novels. Due to copyright restrictions, we release only plot and character summaries rather than the full book content. To ensure fairness in our experimental analysis, these books are verified to be absent from the pretraining dataset of our evaluated LLMs, avoiding data contamination issues[2] (Chang et al., 2024). For each book, we collect its *average rating score* and *multiple reviews* from Goodreads, the largest book review platform. Each raw review consists of the reader's written critique along with a rating on a 1-5 scale.

Review Processing. As illustrated in Figure 8, raw reviews are often unstructured and lack clarity. Prior works (Gong and Mao, 2023; Lee et al., 2024) have shown that aspect-guided critiques enhance both readability and evaluation accuracy compared to direct scoring or relying on an unstructured overall review. Building on this insight, we reformat raw reviews by identifying user-mentioned aspects, extracting *viewpoints* for each aspect, and summarizing these viewpoints into a concise *overall assessment*. This process also involves refining the original language for clarity and brevity. The detailed processing prompt is provided in Table 10, with the temperature set to zero to prevent the introduction of new information.

We first apply DeepSeek-v2.5 (Liu et al., 2024) to process the raw reviews. If a raw review is too ambiguous and the reformatted version has less than 40% word overlap with the original text, we apply GPT-4o to process this raw review. If the overlap remains below the threshold after this second pass, the sample is filtered out.

Criteria Analysis. Through our review process, we extract over 1000 user-mentioned aspects and analyze the most frequently referred ones. We organize these aspects into a hierarchical criteria structure, referring to existing evaluation works (Guan et al., 2021; Chhun et al., 2022) and literary studies (Halliwell, 1998; Herman, 2011). Specifically, we begin by analyzing the eight top-level aspects and use LLMs to identify potential sub-aspects. After further analysis and refinement, we establish our criteria structure (Table 9). Figure 4 shows the distributions of these aspects. Additional discussion can be found in §A. Among the top-level aspects, some focus on objective qualities of the novel (i.e., plot & structure, characters, writing & language, world-building & setting, and themes), while others capture more subjective reader experiences (i.e., emotional impact, overall enjoyment & engagement, and expectation fulfillment).

Review Organization. After criteria analysis, we organize the extracted viewpoints by grouping them under the same criteria as their corresponding *critiques*. For example, as shown in Figure 2, the separate viewpoints for "plot development" and "plot coherence" are listed under "Plot & Structure", forming the corresponding critique.

[2]We also propose an anonymized test set for evaluating future LLMs, which can significantly mitigate data contamination concerns (Wang et al., 2024). Details are in §B.3.

Dataset	# Stories	# Samples	# AVG Length	Review	Criteria
OpenMEVA (Guan et al., 2021)	2,000	2.0K	143 tokens	-	PLOT(COH), **CHA**, WRI(FLU)
HANNA (Chhun et al., 2022)	1,056	19.0K	375 tokens	-	PLOT(COH,SUR), WOR(COM), **EMO(EMP)**, ENJ(ENG), **EXP(REL)**
StoryER-Rate (Chen et al., 2022)	12,669	45.9K	493 tokens	Overall	PLOT(STR), CHA(CHAR), WRI(STY), EXP(GENRE)
Xie (Xie et al., 2023a)	200	1K	79 tokens	-	PLOT(COH), WRI(FLU), WOR(COMM), ENJ(INT), **EXP(REL)**
Per-DOC (Wang et al., 2024)	596	8.9K	2.5K tokens	Overall	PLOT(ADAP,SUR,END), **CHA**, ENJ(INT)
LongStoryEval	600	340K	121K tokens	Aspect-Guided	**PLOT, CHA, WRI, THE, WOR, EMO, ENJ, EXP**

Table 1: LongStoryEval and existing story evaluation datasets. "Criteria" denotes the considered aspects – *PLOT: plot & structure, CHA: characters, WRI: writing & language, THE: themes, WOR: world-building & setting, EMO: emotional impact, ENJ: enjoyment & engagement, EXP: expectation fulfillment*. Abbreviations of existing datasets' aspects are detailed in Table 3. Our criteria structure encompasses the previous inconsistent criteria, with overlapping top-level aspects shown in **bold** and covered sub-aspects underlined.

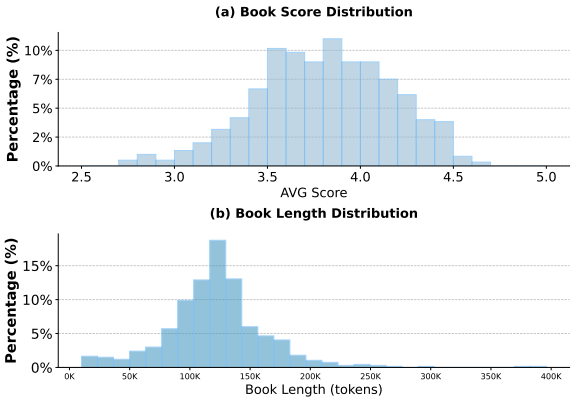

Figure 3: Average score distribution and book length distribution in LongStoryEval.

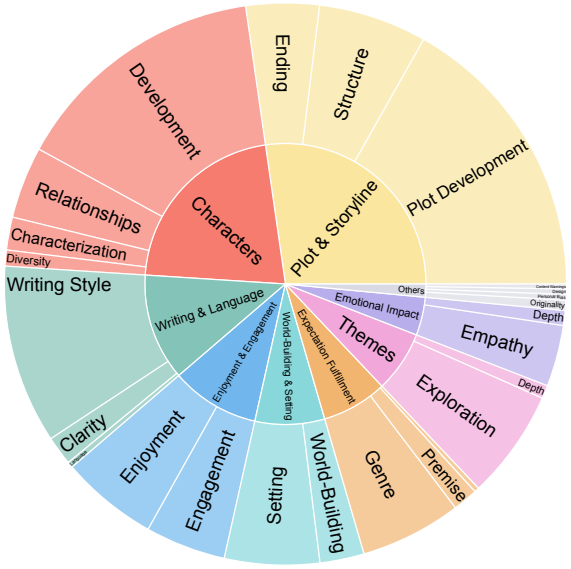

Figure 4: The distribution [4]of the evaluation aspects in readers' reviews.

Statistics and Comparison. Our benchmark dataset includes: (1) 600 newly published books with their metadata, including titles, genres, and premises; (2) An average rating score for each book, along with its rating distribution from 1-5 stars; (3) Multiple reviews for each book, organized as aspect-guided critiques, an overall assessment, and a final rating score; (4) Reviewer metadata, including rating score distribution and self-introduction (if available).

Compared to existing story evaluation benchmarks (Table 1), **LongStoryEval** is the first to focus specifically on lengthy stories, with an average length of 121K tokens and a maximum of 397K. The length distribution is shown in Figure 3 (b). Unlike previous benchmarks, which rely on annotators to evaluate stories based on limited predefined criteria[3], we collect real-world reader reviews and derive evaluation criteria through systematic analysis. This data-driven approach ensures that the criteria structure better reflects actual reader standards. As shown in Table 1, our evaluation criteria structure covers all key aspects identified in prior works. Additionally, our organized reviews demonstrate a multi-aspect-guided reasoning process for the evaluation score, enhancing interpretability and providing greater granularity for training story evaluation models.

4 Method

Given a book-length story consisting of several chapters $\{c_1, ..., c_n\}$ and an evaluation criteria list $\{a_1, ..., a_m\}$, aspect-specific critique/review r_i and aspect-specific score s_i will be generated for each a_i. All critiques will then be summarized into an overall assessment R, accompanied by an overall rating score S. In this work, we explore and compare three methods for evaluating lengthy stories

[3]While some benchmarks allow user-defined criteria, this is typically on a very small scale.

[4]To avoid genre bias, this distribution includes equal books from each of the genres: Romance, Fantasy, Thriller, Mystery, Historical Fiction, Science Fiction, and Young Adult.

(§4.1; Figure 5): aggregation-based, incremental-updated, and summary-based evaluations. We then propose a specialized model that uses the efficient summary-based strategy, as detailed in §4.2.

4.1 Lengthy Story Evaluation Methods

Aggregation-Based Evaluation. As illustrated in Figure 5 (a), each chapter is evaluated individually, and the chapter-level scores are subsequently averaged as the book-level score. These chapter-level scores can refer to either aspect-specific scores or the overall score. For each chapter's evaluation, we provide the LLMs with the book's metadata, the current chapter, and a plot summary of previous chapters to ensure contextual awareness.

Incremental-Updated Evaluation. This method assumes that a reader's opinion of a book evolves during the reading process. As illustrated in Figure 5 (b), the model updates evaluations (both reviews and scores) progressively as it processes each chapter. At each step, the model receives the summary and evaluations from the previous chapters, processes the current chapter, and updates the reviews and scores. This process continues iteratively until the final chapter is reached.

Summary-Based Evaluation. A more intuitive approach involves reading the entire book first to form an overall impression before evaluation. A comprehensive overview of a lengthy story should include key aspects such as plot, characters, and writing style, similar to Wikipedia-style novel introductions. As illustrated in Figure 4, these elements are also frequently mentioned by real readers. Therefore, we condense the story into: *plot summary, character analysis*, and *writing excerpts* (selected paragraphs to reflect the writing style). These elements can effectively capture additional aspects such as themes and overall enjoyment.

Our summary is generated through incremental summarization, which aligns better with human preferences (Chang et al., 2024). As shown in Figure 5 (c), at each summarization step, we provide the current chapter and the previous summary (plot and character), generate a summary of the current chapter, and update the overall summary. Detailed explanations and prompts are displayed in §C.

4.2 Proposed Model: *NovelCrtique*

As mentioned in §2 and confirmed by our experiments in §5.3, closed-source methods, while outperforming open-source alternatives, still lack consistency and exhibit poor alignment with human evaluations. To address these limitations, we introduce *NovelCritique*, a specialized model for evaluating long-form stories. NovelCritique follows the summary-based structure, which can achieve comparable results to aggregation-based evaluations (Table 2) while being much more efficient. The only deviation from the framework in Figure 5 (c) is that our real-reader reviews lack aspect-specific scores. For each training sample with a criteria list $\{a_1, ..., a_m\}$, the model outputs consist of: aspect-specific critiques, an overall assessment, and an overall score. During inference, when an aspect-specific score is needed, the model applies the generated critiques of this aspect, summarizes them into an overall assessment, and produces a score that becomes the aspect-specific score.

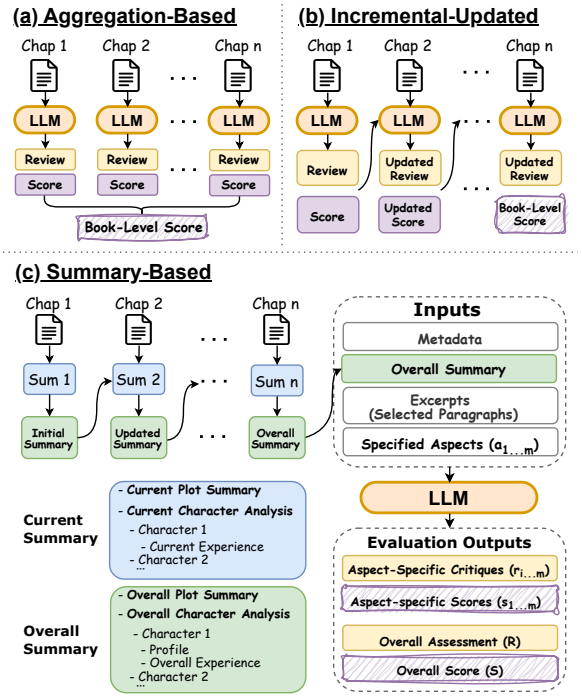

Figure 5: Overview of three evaluation methods (§4.1). Here we illustrate the complete inputs and outputs in the summary-based structure. The other two methods similarly incorporate metadata and specified aspects as inputs to generate aspect-specific and overall evaluations (reviews and scores). For these two types, each chapter's evaluation includes the current content and previous summaries as input, ensuring contextual awareness. Detailed prompt appears in Table 14.

Review Bias Mitigation. We address the selection bias in providing reviews, as users who give moderate ratings are less likely to write reviews (§B.2). Training on all collected reviews would create bias in model predictions. To counter this,

we filter training reviews across all rating levels (1-5) to match each book's rating distribution. For instance, if a book has mostly 3-star ratings, but 5-star reviews are disproportionately overrepresented due to this bias, we filter out extra 5-star reviews based on the book's rating distribution.

Rating Score Normalization[5]**.** Users' rating standards vary significantly, with some being strict and others more moderate. To normalize the ratings (Shalabi and Shaaban, 2006), we adjust each rating score S as follows:

$$S' = \frac{S - \mu_u}{\sigma_u} \times \sigma_{\text{plat}} + \mu_{\text{plat}}, \quad (1)$$

where μ_u, σ_u denote the mean and standard deviation of the current user's rating distribution, and $\mu_{\text{plat}}, \sigma_{\text{plat}}$ represent platform-wide statistics.

Training. We train NovelCritique via instruction tuning (Ouyang et al., 2022) using cross-entropy loss. The instruction details are provided in Table 13. We select Llama 3.1-8B (Dubey et al., 2024) as the base model due to its strong performance among open-source alternatives. The training loss of each sample is calculated as:

$$-\log P(r_{i\leq m}, R, S' | X_{\text{Instruct,Metadata,Sum,Exceprts}}, a_{i\leq m}). \quad (2)$$

5 Experiments

We conduct experiments using the LongStoryEval dataset. From the 600 books, we designate 150 books as the test set (Tables 15-16), ensuring diversity across genres and score distributions.

5.1 Training Setup of NovelCritique

Our training set consists of the remaining 450 books and 176K filtered reviews (after mitigating review bias). Input summaries are generated via incremental summarization (§C) using the GPT-4o model. We finetune the base Llama 3.1-8B model for three epochs with a learning rate of $1e^{-5}$ and a batch size of 32. The LoRA parameters (Hu et al., 2022) are configured as $r = 64$ and $alpha = 16$. The training was conducted on four A6000 GPUs, taking approximately 125 hours.

5.2 Baselines

LLM-Based Lengthy Story Evaluation. We conduct experiments to compare the effectiveness of three evaluation methods (detailed in §4.1): aggregation-based, incremental-updated, and summary-based evaluations. Building on research showing that LLM-based evaluation benefits from detailed criteria definitions (Chhun et al., 2024), we establish evaluation criteria for eight top-level aspects based on literature standards, as detailed in §A.2. The evaluation prompts are provided in Table 14.

Backbone LLMs. We experiment with five backbone models: GPT-4o, GPT-4o-mini, DeepSeek-v2.5 (Liu et al., 2024), Mixtral 8×7B-Instruct (Jiang et al., 2024a), Llama 3.1-70B-Instruct, and Llama 3.1-8B-Instruct (Dubey et al., 2024). As shown in Table 5, all these models were trained on data predating year *2024* to ensure fairness in evaluation. To improve stability, we apply greedy decoding for open-source models and set the temperature to zero for closed-source models. Since closed-source LLMs still produce variations (they do not use pure greedy decoding), we report the average rating score across five generations.

5.3 Main Results

Following prior evaluation works, we use Kendall-Tau correlations (Kendall, 1938) to measure agreement between human evaluations and model-predicted scores. For each evaluation method, we generate aspect-specific scores and the overall scores, then compute their correlation with human-assigned ratings (i.e., the average rating of each book). As shown in Table 2, **NovelCritique** demonstrates the highest correlation with human ratings across both overall scores and most evaluation aspects, with the exception of word-building and setting. This exception is understandable, as this aspect often requires a holistic understanding of the entire book, making it particularly challenging for summary-based models. Nevertheless, NovelCritique still outperforms other summary-based methods in this aspect.

The primary issue with closed-source LLMs is their inconsistency. Even with temperature=0 and low top-p settings, the results exhibit significant variability. This inconsistency is likely due to the long context windows, which increase the likelihood for models to focus on uncertain or less relevant story elements. Notably, this inconsistency issue is less pronounced in short story evaluation tasks (Chhun et al., 2024). To mitigate this, we average scores across five evaluation runs. While this improves stability, it also significantly increases

[5]This normalization is applied only to training samples, while for evaluation, we use the average of all original ratings.

		PLOT	CHA	WRI	WOR	THE	EMO	ENJ	EXP	Overall
One-Pass (Subset)	GPT-4o	3.3	4.1	7.9	0.8	3.3	-1.2	-3.2	8.4	5.5
	DeepSeek-v2.5	4.4	3.5	4.8	-0.9	3.3	-1.1	-1.3	9.4	4.8
Aggregation -Based	GPT-4o	14.3	16.7	10.2	7.9	10.4	9.7	9.1	14.1	15.2
	DeepSeek-v2.5	17.2	15.8	7.0	7.1	11.0	14.2	11.1	16.7	15.1
	GPT-4o-mini	14.2	17.2	7.2	4.4	9.5	8.9	8.1	15.1	12.3
	Llama 3.1-70B	19.6	13.8	2.3	**13.8**	13.4	7.7	11.5	18.9	13.8
	Llama 3.1-8B	15.5	8.5	-1.4	2.8	12.3	7.5	7.0	13.7	11.6
	Mixtral 8×7B	9.5	4.0	2.5	-0.2	8.9	9.5	10.2	6.8	9.0
Incremental -Updated	GPT-4o	8.0	9.1	9.1	11.7	10.5	12.3	12.1	11.5	10.9
	DeepSeek-v2.5	8.9	12.2	9.0	8.6	12.5	12.3	6.6	12.2	11.6
	GPT-4o-mini	7.9	10.8	6.7	7.4	8.5	11.6	8.5	10.7	9.3
	Llama 3.1-70B	9.3	13.3	4.1	1.7	8.7	4.9	4.6	6.1	9.9
	Llama 3.1-8B	7.0	7.1	4.4	2.5	1.9	8.0	7.8	5.1	6.7
	Mixtral 8×7B	4.2	10.8	4.4	6.6	5.8	2.3	5.8	2.6	4.2
Summary -Based	GPT-4o	15.3	17.8	4.5	5.0	7.2	12.6	11.8	14.0	13.4
	DeepSeek-v2.5	13.4	12.2	1.8	-3.8	7.1	8.9	13.2	15.1	14.4
	GPT-4o-mini	8.7	7.5	5.4	4.8	11.1	11.6	8.3	7.9	9.7
	Llama 3.1-70B	11.2	10.8	-1.6	5.3	12.4	9.2	11.4	14.5	13.0
	Llama 3.1-8B	10.4	14.1	4.9	9.1	9.6	15.3	14.5	12.3	12.4
	Mixtral 8×7B	7.8	7.4	7.1	-0.5	-4.0	5.6	9.4	6.7	8.3
	NovelCritique-8B	**21.4**	**20.8**	**15.1**	11.2	**18.5**	**21.1**	**22.8**	**20.5**	**20.1**

Table 2: 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.

computational overhead. **The cost is particularly high for incremental-updated and aggregation-based methods** as they require processing the entire book context for each evaluation run.

5.4 Analysis

Which evaluation aspects mostly affect the final rating? As shown in Table 2, for objective aspects, *plot* and *characters* are the most influential. *Themes* have some impact but appear to be secondary considerations for most readers. *World-building* and *writing quality* are the least influential aspects, likely because most stories show similar levels in these areas (except for particularly exceptional books). For subjective aspects, *emotional impact*, *enjoyment & engagement*, and *expectation fulfillment* all play critical roles.

Which long story evaluation strategy is more effective? Prior to comparing our proposed methods, we assess whether existing models can effectively evaluate entire books in a single pass. We test a subset of books within the 128K token context. The one-pass results (lines 1-2 in Table 2) reveal a poor correlation with human ratings. Even when prompted to generate summaries first, these models often produce generic critiques that fail to capture the nuances of specific stories.

Regarding the three methods discussed in §4.1, aggregation-based and summary-based approaches demonstrate superior performance. While the incremental-updated method seems promising in theory, it faces two key limitations. First, it requires additional instructions, forcing models to both comprehend the current segment and consider its impact on previous evaluations. This problem is especially serious for less powerful models like Llama 3.1-8B and Mixtral 8×7B. Second, it suffers from inconsistency that accumulates over multiple updates. Given these constraints, we recommend using aggregation-based and summary-based methods until significant improvements in LLM capabilities emerge.

The main advantage of aggregation-based methods is their ability to access all details of a long story, enabling more thorough measurement and scoring. Summary-based models, on the other hand, offer two main benefits: (1) **Efficiency**—they require less time and resources per evaluation. By generating a single high-quality summary, we can reuse it across multiple evaluations, leading to more stable and robust results. (2) **Potential for early evaluation**—before completing the entire story, authors can receive reviews and scores based on their plot structure, character design, and writing samples. These early assessments can closely estimate judgments of their finished work.

GPT-4o	Llama 3.1-8B	*NovelCritique-8B*
Plot and Structure:	**Plot and Structure:**	**Plot and Structure:**
The plot of "Remedial Magic" is intricate. The setting of Crenshaw, a town reliant on magic for survival amidst environmental crises, provides a unique backdrop that is both engaging and relevant. The narrative is driven by the characters' personal quests and the overarching mystery of the Congress of Magic's deceitful tactics. The story progresses logically with well-paced storytelling, though at times it may feel overly complex due to the numerous subplots and arcs. The twists, such as the manipulation of memories and the morally ambiguous actions of characters like Prospero, add depth and intrigue. However, the resolution, particularly the forced marriages and memory alterations, may feel unsatisfying to some readers as it leaves certain moral questions unresolved and the characters' autonomy compromised. (3.5)	The plot of "Remedial Magic" is engaging and complex, with multiple storylines that intersect and converge. The narrative effectively explores the consequences of magical abilities and the moral dilemmas faced by the characters. However, the pacing can be uneven at times, with some sections feeling rushed or convenient. The plot twists are generally well-executed, but some may feel predictable to readers familiar with the fantasy genre. The resolutions are satisfying, but the ending feels somewhat abrupt, leaving some questions unanswered. Overall, the plot is well-structured and engaging, but could benefit from further refinement. (3.5)	The story revolves around Ellie, a librarian who discovers she is a witch and falls in love with a powerful witch named Prospero. The plot is filled with magical college, mysterious disappearances, and a prophecy that Ellie is the key to saving Crenshaw, a magical land facing environmental crises. However, the pacing of the novel is somewhat slow, which might deter some readers. The narrative jumps between multiple characters and their perspectives, which can be confusing at times. The story also includes a car accident that serves as a catalyst for Ellie's discovery of her magical abilities, which feels somewhat contrived. (3.5)
Characters:	**Characters:**	**Characters:**
The characters in the story are well-defined with distinctive arcs, particularly Ellie and Prospero. Ellie's transformation from a librarian to a key figure in a magical society is a compelling journey. Prospero presents a nuanced character whose motives are complex and sometimes ethically dubious. The diversity in characters, including the LGBT representation, is a notable strength. However, some secondary characters, like Dan and Maggie, while integral to the plot, could benefit from more substantial development to enhance their impact with readers. (4.0)	The characters are well-developed and complex, with distinct personalities and motivations. Ellie, Maggie, and Prospero are particularly well-written, with nuanced and relatable character arcs. Dan and Sondre are also well-developed. Craig and Hestia are somewhat underdeveloped, but their roles in the story are still impactful. The characters' relationships with each other are well-explored, and the romance between Ellie and Prospero is sweet and tender. (4.0)	Ellie's character development is somewhat lacking. She transitions from a passive librarian to a powerful witch but doesn't fully grasp her new role. Her trust issues complicate relationships, and her magic becomes crucial for survival, but she becomes a victim of Prospero's manipulations. Ellie's journey feels incomplete, and her growth is overshadowed by the actions of others. Maggie's determination to reunite with her son drives her actions. Her adaptability and magical prowess grow, but she is forced into morally complex decisions. Maggie's sacrifice for her son is poignant, but her character feels somewhat one-dimensional compared to others. (3.0)
...	...	...
Expectation Fulfillment:	**Expectation Fulfillment:**	**Expectation Fulfillment:**
The story largely fulfills expectations based on its premise and genres, offering an innovative mix of fantasy romance with LGBT representation. However, the promise of attending a magical community college is not as central as suggested, which might leave some expecting this element disappointed. Overall, it delivers on suspense and magical intrigue. (3.0)	The story meets the readers' expectations based on the premise and genres. The fantasy and romance elements are well-executed, and the story raises important questions about the nature of power and morality. However, the story could benefit from more complexity to meet reader expectations. (3.5)	I was initially drawn by its premise. The idea of a librarian discovering she's a witch and falling in love with a powerful witch from another land was intriguing. However, the execution fell short of my expectations. The story felt overly long and repetitive, with too much focus on the characters' daily lives and not enough on the magical world and the central plot. (2.5)

Figure 6: Critiques for "Remedial Magic" (AVG human rating: 2.9). The generated weaknesses are colored in red.

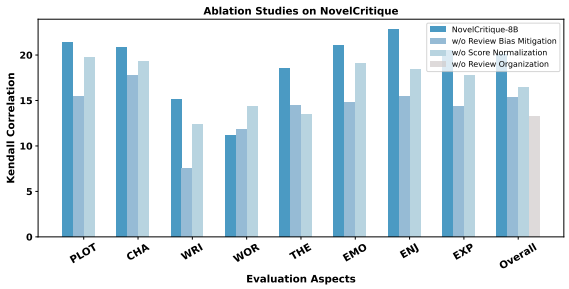

Figure 7: Ablation studies on NovelCritique.

Do detailed summaries improve summary-based evaluation? As displayed in Table 7, replacing the overall summary with detailed chapter-guided summaries leads to a slight increase. We suggest that a more detailed summary can better reflect a story's quality. However, longer summaries require more memory and involve more complex reasoning. It is important to find a balance between the level of detail and length, which could be explored in future works.

Is high-quality summaries necessary for summary-based evaluation? To assess the importance of summary quality, we replace GPT-4o-generated summaries with those produced by GPT-4o-mini (around 0.03% prices of GPT-4o). The results (Table 7) show no significant decline in performance. This suggests a cost-efficient approach: generating summaries with GPT-4o-mini and then conducting evaluations using more advanced models like DeepSeek-v2.5 or NovelCritique.

Ablation Studies on NovelCritique. We verify the effectiveness of our designs in NovelCritique, including raw review organization, review bias mitigation, and rating score normalization. The results in Figure 7 demonstrate their effectiveness.

5.5 Qualitative Results

In Figure 6, we present evaluation results from different models. We find that many models tend to focus more on the story's strengths, offering only limited commentary on its weaknesses. This tendency will also lead existing models to assign good scores for stories that humans consider poor. While we have tried to address this by asking models to provide advantages and disadvantages, this approach causes models to become excessively critical. Current models still struggle to generate nuanced critiques that closely align with human preferences, particularly in reflecting detailed evaluations. Critiques for a well-written story are displayed in §G.

6 Conclusion

This work explores the underexplored problem of evaluating book-length stories, addressing three core questions: (1) What evaluation aspects matter most to real readers? (2) What are the most effective methods for evaluating lengthy stories? (3) What challenges arise in LLM-based evaluation and how can they be addressed? To tackle these questions, we introduce *LongStoryEval*, a large-scale dataset comprising average rating scores and well-formatted reviews. Through analysis of these reviews, we propose a *criteria structure* that reflects human standards. Our experiments reveal the critical aspects influencing final ratings, and demonstrate the effectiveness of aggregation- and summary-based evaluations. While aggregation-based methods provide detailed and comprehensive evaluations, summary-based methods excel in efficiency and offer potential for early-stage evaluations. Acknowledging the limitations of existing LLMs, such as inconsistency and imperfect alignment with human preferences, we propose *NovelCritique*, an 8B model that exhibits improved correlation with human evaluations. We hope this work inspires further research into evaluating lengthy stories and fosters advancements in both lengthy story evaluation and generation.

Limitations

This work employs critiques and score generation for evaluation, which can be prone to inconsistencies. To mitigate this, we average results across multiple runs. Future work could explore alternative mitigation methods, such as employing pairwise comparison instead of direct scoring. Comparisons often yield more stable results but come with higher computational costs, underscoring the need for more efficient comparison strategies. Sampling-based approaches (Xu et al., 2024) also present a promising direction for generating more reliable scores.

Our current evaluation emphasizes general assessment over personalized preferences. However, since our dataset contains anonymized reviewer information, future studies could explore personalized evaluation approaches tailored to individual tastes and reading habits.

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. Our review data comes from publicly available content on the Goodreads website [6], which is accessible to anyone. Following previous works (Wan et al., 2019; Wan and McAuley, 2018), we anonymize user IDs and review IDs to protect personal information. To mitigate the potential dissemination of harmful content in the raw reviews, we will only release our processed versions of reviews. For the books in our dataset, we collect all metadata from public Goodreads content and purchase electronic copies of the books. Considering copyright issues, only the book-level summaries will be released (Chang et al., 2024), while full content remains accessible through publicly available titles and author information (Tables 15-16).

Copyrights. As discussed before, we will only release processed versions to avoid potential ethical and copyright issues. Our data collection from public resources is for academic use only. To prevent commercial use, we will release our dataset under highly restrictive permissions that limit its use exclusively to academic research.

Acknowledgements

We thank all reviewers for their insightful comments and suggestions. This work was partially supported by the Beijing Natural Science Foundation (No. L233008).

References

Chimamanda Ngozi Adichie. 2009. *The danger of a single story*. TED.

Sara Ahmed. 2013. *The cultural politics of emotion*. Routledge.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

James Scott Bell. 2004. *Plot & structure: Techniques and exercises for crafting a plot that grips readers from start to finish*. Writer's Digest Books.

Christopher Booker. 2004. *The seven basic plots: Why we tell stories*. A&C Black.

[6] https://www.goodreads.com