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,
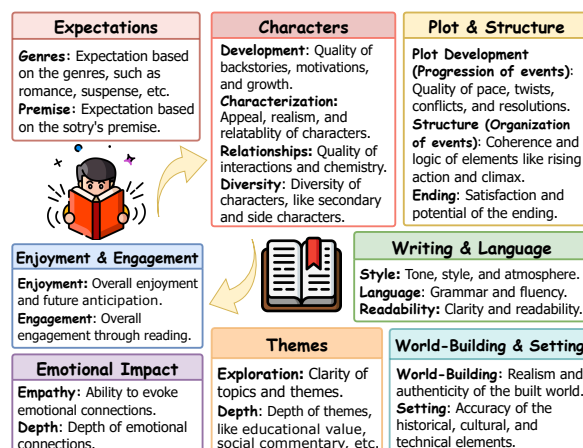

Figure 1: Our proposed *evaluation criteria structure* and the reading process: A reader approaches a book with initial **expectations** based on its genres and premise. The story unfolds through **character**-driven **plots**, revealing its **themes** and **world-building** through the author's **writing**.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

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.

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.

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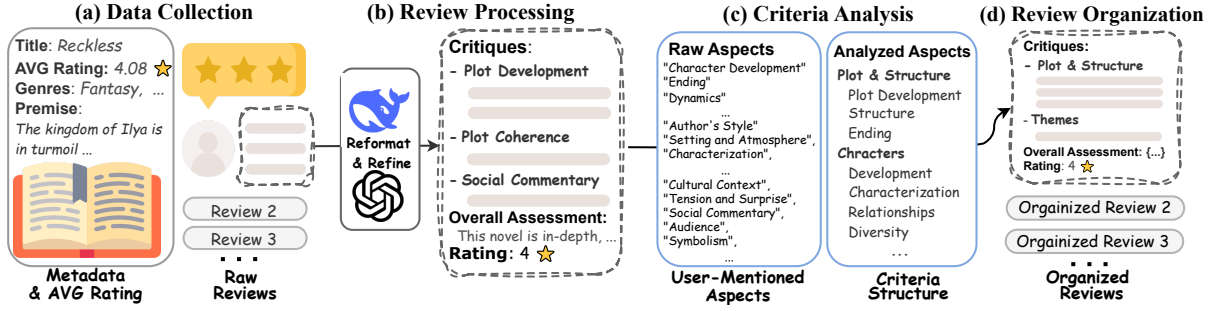
Figure 2: Our data construction process (§3).

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*

---

[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.

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.

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.
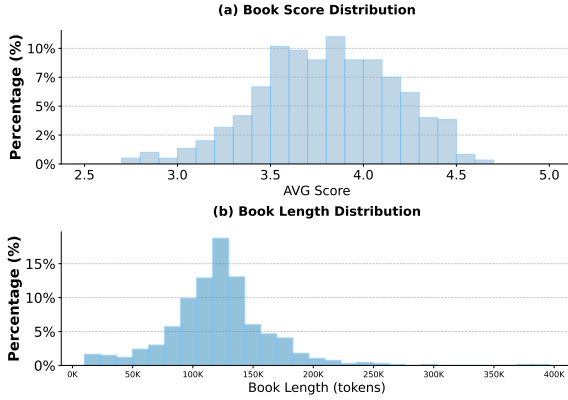

Figure 3: Average score distribution and book length distribution in LongStoryEval.

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 con-
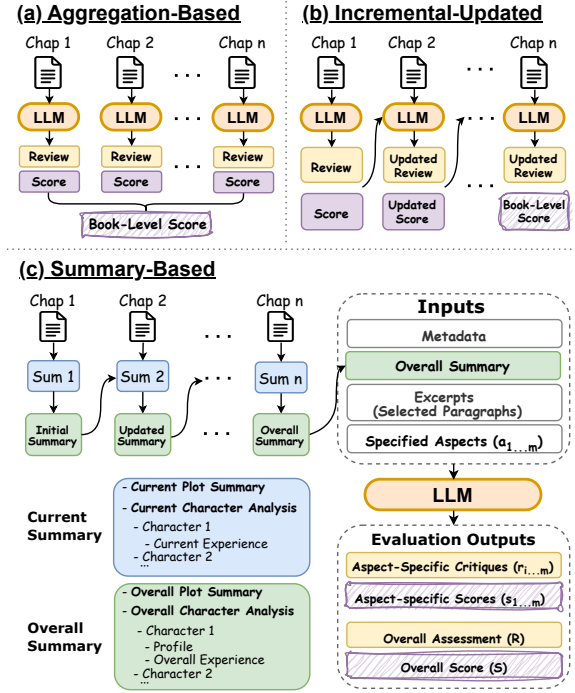

Figure 5: Overview of three evaluation methods (§4.1). Here we illustrate the complete inputs and outputs in the summary-based structure. The other two methods similarly incorporate metadata and specified aspects as inputs to generate aspect-specific and overall evaluations (reviews and scores). For 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.

sistency 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.

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 pre-dating 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

---

		PLOT	CHA	WRI	WOR	THE	EMO	ENJ	EXP	Overall
One-Pass	GPT-4o	3.3	4.1	7.9	0.8	3.3	-1.2	-3.2	8.4	5.5
(Subset)	DeepSeek-v2.5	4.4	3.5	4.8	-0.9	3.3	-1.1	-1.3	9.4	4.8
	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
Aggregation	GPT-4o-mini	14.2	17.2	7.2	4.4	9.5	8.9	8.1	15.1	12.3
-Based	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
	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
Incremental	GPT-4o-mini	7.9	10.8	6.7	7.4	8.5	11.6	8.5	10.7	9.3
-Updated	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
	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
Summary	GPT-4o-mini	8.7	7.5	5.4	4.8	11.1	11.6	8.3	7.9	9.7
-Based	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.

16381

GPT-4o	Llama 3.1-8B	*NovelCritique-8B*

📝 Plot and Structure:

GPT-4o: 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)

Llama 3.1-8B: 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)

NovelCritique-8B: 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:

GPT-4o: 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)

Llama 3.1-8B: 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)

NovelCritique-8B: 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:

GPT-4o: 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)

Llama 3.1-8B: 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)

NovelCritique-8B: 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.
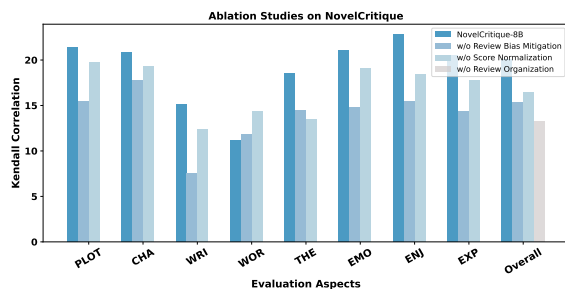

Figure 7: Ablation studies on NovelCritique.

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 *Novel-Critique*, 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

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

Wayne C Booth. 1983. The rhetoric of fiction. *University of Chicago Press*.

Peter Brooks. 1992. *Reading for the plot: Design and intention in narrative*. Harvard University Press.

Janet Burroway, Elizabeth Stuckey-French, and Ned Stuckey-French. 2019. *Writing fiction: A guide to narrative craft*. University of Chicago Press.

Joseph Campbell. 2008. *The hero with a thousand faces*, volume 17. New World Library.

John G Cawelt. 2014. *Adventure, mystery, and romance*. University of Chicago Press.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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.

Seymour Benjamin Chatman and Seymour Chatman. 1980. *Story and discourse: Narrative structure in fiction and film*. Cornell University Press.

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.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 - Findings, Nusa Dua, Bali, November 1-4, 2023*, pages 361–374. 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.

Paul Coulton, Joseph Lindley, Miriam Sturdee, and Michael Stead. 2017. Design fiction as world building. In *Proceedings of the 3nd Biennial Research Through Design Conference*, pages 1–16.

Mihaly Csikszentmihalyi. 1997. Flow and the psychology of discovery and invention. *HarperPerennial, New York*, 39:1–16.

Jerome De Groot. 2009. *The historical novel*. Routledge.

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

Peter Elbow. 1998. *Writing without teachers*. Oxford University Press, USA.

Syd Field. 2005. *Screenplay: The foundations of screenwriting*. Delta.

Edward Morgan Forster. 1927. *Aspects of the novel*. Harcourt, Brace.

Gustav Freytag. 1895. *Technique of the drama: An exposition of dramatic composition and art*. S. Griggs.

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

Gérard Genette. 1980. Narrative discourse: An essay in method. *Cornell University Press*.

Sarik Ghazarian, Zixi Liu, Akash SM, Ralph M. Weischedel, Aram Galstyan, and Nanyun Peng. 2021. Plot-guided adversarial example construction for evaluating open-domain story generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4334–4344. Association for Computational Linguistics.

Peiyuan Gong and Jiaxin Mao. 2023. Coascore: Chain-of-aspects prompting for nlg evaluation.

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.

Stephen Halliwell. 1998. *Aristotle's poetics.* University of Chicago Press.

David Herman. 2011. *Basic elements of narrative.* John Wiley & Sons.

Patrick Colm Hogan. 2003. *The mind and its stories: Narrative universals and human emotion.* Cambridge University Press.

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

Henry James. 1884. The art of fiction. *Longman's magazine, 1882-1905,* 4(23):502–521.

Fredric Jameson. 2013. *The political unconscious: Narrative as a socially symbolic act.* Routledge.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088.*

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024b. Tigerscore: Towards building explainable metric for all text generation tasks. *Trans. Mach. Learn. Res.,* 2024.

Suzanne Keen. 2007. *Empathy and the novel.* Oxford University Press.

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

Joonghoon Kim, Sangmin Lee, Seung Hun Han, Saeran Park, Jiyoon Lee, Kiyoon Jeong, and Pilsung Kang. 2023. Which is better? exploring prompting strategy for llm-based metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP 2023, Bali, Indonesia, November 1, 2023,* pages 164–183. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.* OpenReview.net.

Stephen King. 2000. *On writing: A memoir of the craft.* Pocket Books.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2024. Checkeval: Robust evaluation framework using large language model via checklist. *Preprint,* arXiv:2403.18771.

Geoffrey N Leech and Mick Short. 2007. *Style in fiction: A linguistic introduction to English fictional prose.* Pearson Education.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment.

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.

Aviya Maimon and Reut Tsarfaty. 2023. COHESEN-TIA: A novel benchmark of incremental versus holistic assessment of coherence in generated texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* pages 5328–5343. Association for Computational Linguistics.

Nicole Matthews and Nickianne Moody. 2007. *Judging a book by its cover: Fans, publishers, designers, and the marketing of fiction.* Ashgate Publishing, Ltd.

Allyssa McCabe and Carole Peterson. 1984. What makes a good story. *Journal of Psycholinguistic Research,* 13(6):457–480.

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

Victor Nell. 1988. *Lost in a book: The psychology of reading for pleasure.* Yale University Press.

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.

James Phelan. 1996. *Narrative as rhetoric: Technique, audiences, ethics, ideology.* Ohio State University Press.

Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin uK.

Shlomith Rimmon-Kenan. 2003. *Narrative fiction: Contemporary poetics*. Routledge.

Hanna-Riikka Roine. 2016. Imaginative, immersive and interactive engagements. the rhetoric of worldbuilding in contemporary speculative fiction.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation.

Luai A. Al Shalabi and Zyad Shaaban. 2006. Normalization as a preprocessing engine for data mining and the approach of preference matrix. In *2006 International Conference on Dependability of Computer Systems (DepCoS-RELCOMEX 2006), 24-28 May 2006, Szklarska Poreba, Poland*, pages 207–214. IEEE Computer Society.

John Skorupski. 1976. *Symbol and theory*. Cambridge University Press.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

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

James Wood. 2008. How fiction works. *Jonathan Cape*.

Wendy Wyatt. 2016. The ethics of trigger warnings. *Teaching Ethics*, 16(1).

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

Zhuohan Xie, Miao Li, Trevor Cohn, and Jey Lau. 2023b. Deltascore: Fine-grained story evaluation with perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5317–5331, Singapore.

Cong Xu, Gayathri Saranathan, Mahammad Parwez Alam, Arpit Shah, James Lim, Soon Yee Wong, Foltin Martin, and Suparna Bhattacharya. 2024. Data efficient evaluation of large language models and text-to-image models via adaptive sampling. *Preprint*, arXiv:2406.15527.

Dingyi Yang and Qin Jin. 2024. What makes a good story and how can we measure it? a comprehensive survey of story evaluation.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

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.

Lisa Zunshine. 2006. Why we read fiction: Theory of mind and the novel. *Ohio State University Press*.

A Evaluation Criteria

A.1 Human Criteria

Table 9 presents our detailed criteria structure with raw aspects mapped to standardized categories. Below, we detail our analyzed criteria structure, including the aspect names, definitions, and their corresponding references:

Plot and Structure:

• **Plot Development**: Assess the progression of events within a story through their pacing, twists, conflicts, and resolutions, as mentioned in Bell (2004); Forster (1927); Brooks (1992).

• **Structure**: Assess the organization of events through exposition, rising action, climax, falling action, and resolution, examining their coherence and logical flow, as mentioned in Freytag (1895); Genette (1980).

• **Ending**: How well the ending reflects the consequences of the story's main events and themes. A well-crafted ending provides closure or shows attractive potential, as mentioned in Rimmon-Kenan (2003); Phelan (1996).

Characters:

• **Character Development**: Examine how well characters develop throughout a story by examining their backstories, motivations, and character growth

arcs, as mentioned in McKee (1997); James (1884); Campbell (2008).

• **Characterization**: Assess the representation of the characters, ensuring appeal, realism, and reliability, as mentioned in Wood (2008); Burroway et al. (2019).

• **Relationships**: Assess the interactions and chemistry between the characters, examining how their relationships drive the plot forward and engage readers, as mentioned in James (1884); Chatman and Chatman (1980).

• **Diversity**: How effectively the story portrays a diverse cast of characters, including not only the protagonist but also secondary and side characters, mentioned in Hogan (2003); Adichie (2009).

Writing and Language:

• **Writing Style**: Whether the author's choice of words, sentence structure, and paragraph organization conveys meaning effectively and creates an engaging atmosphere for readers, as mentioned in Leech and Short (2007); Skorupski (1976).

• **Language**: Assess the quality of vocabulary and syntax, focusing on issues like grammar and fluency, mentioned in Pinker (2003); Elbow (1998).

• **Readability**: Assess the clarity and readability of the entire story, as mentioned in Burroway et al. (2019); King (2000).

Themes:

• **Exploration**: How well the author explores clear topics and themes through the narrative, as mentioned in Forster (1927); Booth (1983).

• **Depth**: Measure the depth of the themes, such as their educational value and social commentary, as mentioned in Booker (2004); Jameson (2013).

World-Building and Setting:

• **World-Building**: Assess the realism and authenticity of the built world by examining its consistency and descriptive detail, as mentioned in Roine (2016); James (1884); Coulton et al. (2017).

• **Setting**: Evaluate the accuracy of historical, cultural, geographic, and technical settings, as discussed in De Groot (2009); James (1884). This might be more considered in historical stories.

Emotional Impact:

• **Empathy**: Evaluate how well the stories evoke emotional connections between readers and charac-

ters, as mentioned in McCabe and Peterson (1984); Keen (2007); Hogan (2003).

• **Depth**: Evaluate the stories' ability to create deep emotional connections, as discussed in Ahmed (2013); Keen (2007).

Enjoyment and Engagement:

• **Enjoyment**: Evaluate how well a story creates enjoyment and interests, and builds anticipation for upcoming events, as mentioned in Nell (1988); Zunshine (2006).

• **Engagement**: Evaluate how well the story engages the readers through the reading process, as mentioned in Nell (1988); Zunshine (2006).

Expectation Fulfillment (Or Relevance):

• **Genre**: How well the story fulfills genre expectations for elements like romance, suspense, and other genre-specific features. As mentioned in Cawelt (2014); Phelan (1996).

• **Premise**: How well the story fulfills expectations based on the premise, as mentioned in Forster (1927); Field (2005).

Others:

• **Originality (Or Creativity)**: Assess the story's uniqueness and imaginative ideas, as mentioned in Csikszentmihalyi (1997).

• **Content Warnings**: Whether the story contains rude, unreasonable, or disrespectful elements, as mentioned in Wyatt (2016).

• **Designment**: Assess the design of the book cover (Matthews and Moody, 2007).

• **Personal Bias**: Assess the feelings influenced by very subjective factors, such as individual preferences and personal experiences (Wang et al., 2024).

A.2 Condensed Definitions

We condense the definition of the top-level aspects as follows, which can be applied in our evaluation prompts:

1. **Plot and Structure (PLOT)**: Evaluate the plot development by examining pace, twists, conflicts, and their resolutions. Evaluate the story structure for coherence, logic, and complexity, paying attention to key elements like climax and ending.

2. **Characters (CHA)**: Evaluate how well the characters are drawn, considering their development, characterization (including realism,

appeal, and relatability), relationships, and diversity.

3. **Writing and Language (WRI)**: Evaluate the writing style's ability to engage and captivate readers. Assess language quality by examining descriptions and dialogue. Measure the story's clarity and readability.

4. **World-Building and Setting (WOR)**: Evaluate the world-building and setting by assessing how detailed and well-described they are. Consider their authenticity, accuracy, or realism.

5. **Themes (THE)**: Evaluate how well the themes are explored throughout the story and assess their depth.

6. **Emotional Impact (EMO)**: Evaluate the story's ability to evoke strong and deep emotional impact.

7. **Enjoyment and Engagement (ENJ)**: Evaluate how engaging and enjoyable the story is for readers.

8. **Expectation Fulfillment (EXP[7])**: Evaluate how effectively the story meets the readers' expectations based on the premise and genres.

A.3 Criteria of Existing Benchmarks

The considered aspects of existing benchmarks and their abbreviation are listed in Table 3.

B Dataset: LongStoryeval

B.1 Review Processing

Figure 8 displays an example of our raw review processing, while the processing prompt is detailed in Table 10. Our dataset sampling shows high accuracy (96%), which is sufficient for our model training. Our experimental improvements demonstrate its quality, though further human refinement could be considered in the future.

B.2 Statistics of Review Bias

Review bias refers to the selection bias in providing reviews, where users who give moderate ratings are less likely to leave reviews. We evaluate the average ratio of review numbers among ratings for each rating scale. The statistics show: 5 stars

---

[7]This can also be considered as the relevance to the "Inputs"(premise and genres).

Dataset	Aspects
OpenMEVA (Guan et al., 2021)	COH: coherence [consistency, causal and temporal relationship], CHA: character behavior, FLU: fluency [semantic repetition, paraphrases]
HANNA (Chhun et al., 2022)	COH: coherence, SUR: surprise, COM: complexity, EMP: empathy, ENG: engagement, REL: relevance.
StoryER-Rate (Chen et al., 2022)	STR: structure, CHAR: characterization [character shaping], STY: writing style, GENRE: genre expectations.
Xie (Xie et al., 2023a)	FLU: fluency, COH: coherence, COMM: commonsense, INT: interestingness, REL: relevance.
Per-DOC (Wang et al., 2024)	ADAP: plot adaptability, SUR: surprise, END: ending, CHA: characters, INT: interestingness.

Table 3: Existing story evaluation benchmarks and their considered aspects.

PLOT	CHA	WRI	WOR	THE	EMO	ENJ	EXP
79.9	83.2	72.0	66.4	72.3	74.0	79.1	72.4

Table 4: Approximate correlation scores (§B.4).

(22%), 4 stars (19%), 3 stars (17%), 2 stars (22%), and 1 star (31%).

B.3 Contamination Issues

Contamination problems mean the datasets are exposed to LLMs. For contaminated cases, LLMs might show unfairly high performance (Chang et al., 2024; Wang et al., 2024). However, anonymization and summarization can largely mitigate these issues (Wang et al., 2024). Since our dataset is already summarized, we conduct additional anonymization by replacing identifiable elements such as character names and locations. This anonymized test set helps ensure fair evaluation of future LLMs. Additionally, we will release our processing code to facilitate the collection of newly published books.

B.4 Approximate Exploration of Significant Aspects

Since individual aspect ratings are unavailable in the ground truth data, we attempt to approximate aspect-specific ratings by calculating the average of ratings from these reviews covering each specific aspect. As the results shown in Table 4, plot, characters, and overall enjoyment show clear significance. However, this approximation is imperfect, as the scores are influenced by multiple mentioned aspects rather than representing a single aspect. Therefore, these findings should be considered as a reference, with the results in Table 2 offering more reliable evidence.

Raw Review	Processed Review
Review: I was so so excited to read reckless and it did not disappoint!! this was a masterpiece and I absolutely devoured it 😋while powerless is all high stakes, adventure and sharp edges reckless is exploration, angst, and sweet moments and I think that was amazing! The focus on the characters just grew my connection to characters I already adored, and I feel like I could literally SEE Pae and Kai's connection changing and growing with them. and the banter!!! I was giggling out loud and kicking my feet, and then five seconds later I was on the edge of my seat because Kai, Paedyn or both of them were in some type of pickle…it was the best reading experience. And somehow there was this suspense that grew the whole book and I just had to read, the days that I read this book, those nights I went to sleep thinking about this world. and the writing!! Lauren's writing is so beautiful and the way she describes things had my heart sobbing, these characters and this world seriously came off the page for me and that's magical 😊 and the ending. I don't even want to talk about it because what was that!!!! I seriously didn't predict that and immediately started sobbing like a baby and that's how I know a book is a favorite. When I think about how much I love this book I literally get heart palpitations. It was insanely good and this series WILL live rent free in my mind until the third one 😋🫶 🎧—songs that remind me of this series•favorite crime by Olivia Rodrigo•sad **Rating Score:** 5 ⭐	**- Characters:** 　**(Character Development)** 　- The focus on the characters deepened my connection to characters I already adored. 　**(Character Connection)** 　- I could literally see Pae and Kai's connection changing and growing with them. **- Writing & Language:** 　**(Writing Style and Descriptions)** 　- Lauren's writing is so beautiful, and the way she describes things had my heart sobbing. The characters and this world felt like they came off the page, which was magical. **- Emotional Impact:** 　**(Emotional Impact)** 　- I was moved by the the ending, it had me sobbing like a baby, which is how I know a book is a favorite. **- Enjoyment & Engagement:** 　**(Engagement)** 　- The banter had me giggling out loud and kicking my feet, but five seconds later, I was on the edge of my seat because Kai, Paedyn, or both were in some type of pickle. 　- The suspense grew throughout the book, and I couldn't stop reading. On the days I read this book, I went to sleep thinking about this world. 　**(Overall Enjoyment)** 　- When I think about how much I love this book, I literally get heart palpitations. This series will live rent-free in my mind until the third book comes out. **Overall Assessment:** *Reckless* is a masterpiece that delivered everything I hoped for and more. It's a perfect blend of exploration, angst, and sweet moments. It shows incredible character development, suspense, and emotional depth. Lauren's writing is stunning, and the ending left me in tears. This book and series have left an indelible mark on me, and I can't wait for the next installment. **Rating Score:** 5 ⭐

Figure 8: Example of raw review processing. We identify user-mentioned aspects (shown in brackets), then extract viewpoints (underlined texts) for each aspect, and summarize these viewpoints into an overall assessment. These viewpoints will be organized based on the criteria structure shown in Table 9.

C Summarization Method

As mentioned in Section 4.1, in our summary-based methods, we apply the summaries generated through an incremental-update process. Specifically, we update both plot and character summarizations. The detailed prompts are shown in Table 11 and 12. For characters, we summarize their profiles and overall experiences. For plots, considering that non-linear narrative structures can affect readers' reading experience, we instruct the model to maintain the original plot structure.

D Experimental Details

D.1 Knowledge Cutoff of Baseline Models.

As shown in Table 5, all of the baseline models are trained with data before 2024, not containing our dataset.

Models	Knowledge Cutoff
GPT-4o	October, 2023
GPT-4o-mini	October, 2023
DeepSeek-v2.5 (Liu et al., 2024)	November, 2023
Mixtral 8×7B (Jiang et al., 2024a)	December, 2023
Llama 3.1-8B (Dubey et al., 2024)	December, 2023

Table 5: Knowledge cutoff of baseline models.

D.2 Additional Discussions

We find that during the incremental evaluation, if the LLMs generate positive opinions for the earlier paragraphs, then they are likely to maintain this tendentiousness. Even if there are problems in the latter chapters, they will not significantly change their assessment. The same pattern occurs in reverse.

E Ablation Studies

Criteria Definitions.   To investigate the impact of detailed criteria definitions, we conducted experiments both with and without them. As shown in Table 6, providing definitions did not significantly improve performance for powerful models like GPT-4o and DeepSeek-v2.5. However, less powerful models such as Llama 3.1-8B and Mixtral-8×7B show marked improvement with detailed definitions. We attribute this difference to powerful models having already acquired a comprehensive understanding of these aspects through their training, while others have not.

Summary Quality.   To explore how input summaries affect the summary-based evaluations, we conducted experiments with three types: detailed chapter-level summaries, an overall summary, and a lower-quality overall summary (generated by

		PLOT	CHA	WRI	WOR	THE	EMO	ENJ	EXP	Overall
GPT-4o	w/ definitions	15.3	17.8	4.5	5.0	7.2	12.6	11.8	14.0	13.4
	w/o definitions	11.1	15.0	2.2	4.5	6.0	10.7	8.5	9.4	11.5
DeepSeek-v2.5	w/ definitions	13.4	12.2	1.8	-3.8	7.1	8.9	13.2	15.1	14.4
	w/o definitions	16.8	13.7	0.8	5.0	8.1	13.8	10.9	11.9	10.4
Llama 3.1-8B	w/ definitions	10.4	14.1	4.9	9.1	9.6	15.3	14.5	12.3	12.4
	w/o definitions	6.4	6.1	6.9	4.1	5.5	4.0	4.2	5.9	7.1
Mixtral 8×7B	w/ definitions	7.8	7.4	7.1	-0.5	-4.0	5.6	9.4	6.7	8.3
	w/o definitions	0.0	-5.1	-2.1	-7.1	3.3	5.7	-3.2	1.0	1.0

Table 6: The system-level Kendall correlations between human-assigned scores and summary-based evaluation results, comparing scenarios with and without provided criteria definitions.

		PLOT	CHA	WRI	WOR	THE	EMO	ENJ	EXP	Overall
GPT-4o	Chapter Summaries (4o)	14.8	14.5	8.3	10.7	9.0	15.6	8.2	16.6	14.9
	Overall Summary (4o)	15.3	17.8	4.5	5.0	7.2	12.6	11.8	14.0	13.4
	Overall Summary (4o-mini)	9.7	9.2	-1.6	6.0	8.7	8.5	7.2	14.3	10.1
DeepSeek-v2.5	Chapter Summaries (4o)	14.4	15.0	2.9	1.3	11.5	10.8	10.2	12.2	15.1
	Overall Summary (4o)	13.4	12.2	1.8	-3.8	7.1	8.9	13.2	15.1	14.4
	Overall Summary (4o-mini)	12.3	9.7	-1.5	1.4	6.9	10.6	6.1	12.9	10.8

Table 7: The system-level Kendall correlations between human-assigned scores and summary-based evaluation results, comparing scenarios that offer different types of summaries.

GPT-4o-mini). As shown in Table 7, more detailed and higher-quality summaries can be helpful, though not significantly so.

F Efficiency Comparison of Different Evaluation Strategies

The cost comparison of different strategies (§4.1) follows this order: Incremental-Updated > Aggregation-Based > Summary-Based. As mentioned in §5.2, we mitigate inconsistencies in LLM-generated evaluations by averaging 5 results. Thus, the summary-based method is more efficient because it requires only one high-cost summary of the lengthy book, with the 5 evaluations processing only short summaries. In contrast, for aggregation-based evaluation, each of the 5 evaluations must process both the lengthy book and previous summaries. For the incremental-updated method, each evaluation additionally processes the previous evaluations. Taking GPT-4o-based evaluations as an example, the runtime and money cost of 5 evaluations on the test set are shown in Table 8. In conclusion, while aggregation-based methods perform better than summary-based methods (Table 2), their high cost makes summary-based methods a more cost-effective choice.

Methods	Input Tokens	Runtime	Cost
Summary-based	3,940K	770min	$94
Incremental-updated	12,720K	4,268min	$499
Aggregation-based	11,480K	3,056min	$416

Table 8: The runtime and money cost of different evaluation methods (taking GPT-4o-based evaluations as an example).

G Additional Cases

In Figure 6, we present an example of a poorly written story, while Figure 9 shows a well-written one. It can be found that NovelCritique offers more concrete and less formulaic comments.

H Prompts

H.1 Review Processing

The review processing prompt is shown in Table 10.

H.2 Instructions of NovelCritique

The instruction of NovelCritique is shown in Table 13.

H.3 LLM-based Evaluation

The prompt for evaluation is displayed in Table 14.

Aspects	Sub-aspects	Mapped to normalized criteria
Plot & Structure	Plot Development	Plot Development [Story Development, Plot Progression, Development], Plot Complexity [Complexity, Plot Twists, Turns, Surprises, Surprise, Plot Predictability, Plot Intensity], Plot Conflict [Conflict, Resolution, Plot Resolution, Conflict Resolution], Pacing [Pace], Plot Clarity
	Structure	Plot Structure [Story Structure, Subplots, Timelines, Flow, Narrative Perspective], Plot Elements [Beginning, Middle, Climax, Plot Completion], Plot Coherence [Coherence, Plot Continuity, Consistency, Cohesion], Logic
	Ending	Ending [Epilogue, Endings, Closure, Cliffhanger, Conclusion]
Characters	Development	Character Development [Character Progression], Backstories, Motivations [Character Dynamics], Character Growth, Character Arcs, Character Behavior
	Characterization	Characterization [Character Representation], Character Appeal [Character Appreciation, Character Identification], Character Realism, Character Relatability [Character Engagement], Character Depth [Character Complexity], Protagonist [Main Characters, Main Character, Character Focus], Character Introduction
	Relationships	Character Relationships [Character Connection, Interactions, Friendships], Character Interactions, Chemistry [Intimacy, Romance and Relationships], Relationship Development [Family Dynamics, Relationship Dynamics]
	Diversity	Character Diversity [Supporting Characters, Secondary Characters, Side Characters, Character Perspectives, Character Perspective, Character Management]
Writing & Language	Writing Style	Writing Style [Style, Author's Style, Author's Writing Style, Narrative Style, Author's Skill], Atmosphere, Tone [Author's Tone], Trope [Tropes], Writing Elements [Descriptions, Prose, Imagery, Dialogue, Dialogues]
	Language	Grammar [Language Usage, Vocabulary], Fluency [Repetition, Repetitiveness, Author's Reputation]
	Readability	Readability [Clarity, Accessibility, Length]
Themes	Exploration/Clarity	Thematic Elements [Topics, Point of View, Thematic Content, Symbolism, Motifs, Messages], Thematic Exploration [Themes Exploration]
	Depth	Thematic Depth [Social Commentary, Educational Value, Influences]
World-Building & Setting	World-Building	World-Building [Worldbuilding, World Building], Realism [Cultural Realism, Magic System, Magic, Magic Systems]
	Setting	Historical Accuracy [Historical Setting, Historical Context, Time Period], Cultural Elements [Cultural References, Cultural Context, Cultural Representation], Technological Elements...
Emotional Impact	Empathy	Emotional Response [Personal Connection, Emotional Resonance]
	Depth	Emotional Depth [Emotional Range]
Enjoyment & Engagement	Enjoyment	Enjoyment [Overall Impression, Overall Impact, Excitement, Appeal, Story Enjoyment, Interest, Humor], Anticipation for Future Works [Future Interest, Intrigue, Anticipation for Future Book, Sequel Potential]
	Engagement	Engagement [Reader Engagement, Personal Engagement, Immersion, Story Engagement, Reader Engagement, Overall Experience, Reading Experience]
Expectation & Fulfillment	Genre	Genre Expectation, [Genre Appeal, Genre Preference, Audience Appeal, Genre Suitability, Romantic Elements, Thrill, Horror Elements, Fantasy Elements, Historical, Suspense, Tension, Scare Factor, Genre Elements, Genre Classification, Mystery]
	Premise	Premise Expectations [Premise Expectation, Story Premise]
	Title	Title Relevance
Others	Originality	Originality [Creativity, Uniqueness, Originality and Creativity, Unique Elements]
	Content Warnings	Content Warnings [Warnings, Trigger Warnings]
	Designment	Design [Cover, Cover Design]
	Personal Bias	Personal Preference, Personal Experience

Table 9: The detailed criteria structure.

GPT-4o	Llama3.1-8B	NovelCritique-8B
📝Plot and Structure:	**📝Plot and Structure:**	**📝 Plot and Structure:**
The plot of *Mind Games* weaves together a suspense-filled narrative with emotionally driven character arcs. The combination of psychic abilities and a love story adds originality and depth, allowing for moments of quiet reflection amidst the tension. The central conflict between Thea and Ray Riggs builds slow-burning suspense, culminating in a gripping confrontation that offers a satisfying resolution. However, some subplots, such as Thea's game design career, feel underdeveloped and take a backseat to the central romance and suspense. The pacing is uneven at times, with stretches of inactivity that could have been trimmed to enhance momentum. Despite this, the structure follows a logical progression, balancing emotional beats with moments of true suspense. (4.0)	The plot of "Mind Games" is engaging and suspenseful, with a clear structure that effectively builds tension and conflict. The story begins with a strong premise, introducing the main character, Thea, and her unique ability to see into minds and souls. The plot unfolds at a good pace, with a clear three-act structure: setup, confrontation, and resolution. The author effectively uses the supporting characters, such as Tyler and Lucy, to add depth and complexity to the story. However, some plot points, like Thea's decision to confront Ray Riggs in prison, feel a bit convenient and lack a clear motivation. The climax, while intense, feels somewhat rushed, and the resolution could be more satisfying. Overall, the plot is well-structured, but some elements could be improved. (4.0)	The story revolves around Thea Fox, a 12-year-old girl who witnesses her parents' murder by Ray Riggs, a powerful psychic. Thea's grandmother, Lucy, shares the same psychic ability, which they've kept hidden from the world. Thea's parents' death sets her on a path of self-discovery, as she struggles to understand her gift and its implications. The narrative jumps forward in time, showing Thea's growth into a successful game designer and her ability to see into the minds of others. She eventually moves to Redbud Hollow, Kentucky, to live with her grandmother, where she meets Tyler Brennan, a single father who has moved to the town to start anew. Thea's connection to Ray Riggs, who is now in prison, continues to haunt her, and she must navigate her feelings for Tyler while confronting the evil that has shaped her life. (4.5)
👥 Characters:	**👥 Characters:**	**👥 Characters:**
Thea is a well-developed protagonist whose journey from vulnerability to empowerment is relatable and engaging. The supporting characters, including Ty, Lucy, and Rem, add depth to her world. Ty's character complements Thea as both a romantic interest and ally, while Lucy provides a grounding presence. However, the antagonist Riggs could benefit from more nuanced development to enhance his role in the psychological conflict. (4.0)	The characters in "Mind Games" are well-developed and complex, with Thea being a particularly strong protagonist. Her growth and development throughout the story are believable and relatable. These supporting characters, like Ty, Bray, and Lucy, are also well-fleshed and add depth to the story. However, some readers may find Rem and the villain, Ray, a bit one-dimensional. (4.1)	- Thea Fox: Thea was a strong and resilient protagonist who faced numerous challenges, including the trauma of her parent's murder and the ongoing threat from Ray Riggs. Her ability to see into minds and souls added depth to her character, making her both relatable and intriguing. - Lucy: Lucy was a wonderful supporting character. Her wisdom, kindness, and unwavering support for Thea made her a standout. The bond between Thea and Lucy was beautifully portrayed. - Ray Riggs: Riggs was a well-crafted antagonist. His ability to see into minds and souls, combined with his twisted thoughts and evil acts, made him a formidable and unsettling character. His ongoing threat to Thea added tension to the story. (4.5)
...	...	...
💬 Expectation Fulfillment:	**💬 Expectation Fulfillment:**	**💬 Expectation Fulfillment:**
As a work of romantic suspense, Mind Games largely fulfills its genre expectations. It delivers a mix of romance, tension, and character growth. The psychic element sets it apart from conventional narratives and provides an interesting angle that meets themes of the unknown in human relationships. (4.0)	"Mind Games" generally meets the reader's expectations, particularly in terms of the romance and suspense elements. The author effectively delivers on the premise, and the story is engaging and suspenseful. However, some readers may find the resolution somewhat convenient, and the character development could be more nuanced. (4.0)	- The novel is a romantic suspense, which I found to be a perfect blend of genres. It was engaging and kept me hooked throughout. - The story revolves around Thea Fox, a psychic who can see into minds and souls, and her grandmother Lucy, who shares the same ability. Thea's parents were murdered by a powerful psychic named Ray Riggs, and she has been living with the trauma of his actions. The novel explores themes of love, family, and the evil that exists in the world. (4.0)

Figure 9: Critiques for "Mind Games" (AVG human rating: 4.35).

Raw Review Processing

As an expert in storytelling, your task is to reformat a reader's review of a full-length novel into a more structured one.

Objective:
- Identify the evaluation aspects discussed in this review.
- Extract the relevant contents that correspond to each identified aspect.
- Organize these contents into a structured format.

Guidelines:
- Maintain the same tone and perspective (e.g., first-person) in the reformatted review, as if you are the original reader.
- Do not introduce any new information; only reformat the existing review. You may polish the language and improve clarity.
- Ensure all important viewpoints in the existing review are preserved.

Output Format:
Aspects:
- List of evaluation aspects identified in the review and the corresponding viewpoints.
Conclusion:
Brief summary of the review and a final verdict.
Rating Scores:
Any rating scores mentioned in the review.

Now, proceed with your task.

Review:
{Raw Review}

Reformatted Review:

Table 10: The prompt for review processing.

Generate Initial Summary

Below is the beginning of a lengthy story:

- - -
{}
- - -

Instructions:
Your task is to craft a comprehensive summary of the beginning of a lengthy story, including plot and characters.

Guidelines:
1. Plot:
- Summarize the vital information in the beginning part, such as key events, main conflicts, backgrounds, settings, and characters.
- The story may feature non-linear narratives, flashbacks, or switches between alternate worlds or viewpoints. In such cases, state the time and settings for these narrative shifts.

2. Characters:
- Introduce the major characters' information, including their profiles and overall experiences.

Output Format:
- - -
Plot Summary:
{}

Characters:
Name:
- **Profile**: Concisely summarize the character's profile with natural language, including the role (e.g., protagonist), attributes, personality, and relationships within 50 words.
- **Overall Experience**: Briefly describe the character's overall experience, including motivations, events, and emotional states within 100 words.
- - -

Now start your summarization:

Table 11: Prompt to generate the initial summary.

Update Summary

Below is a segment from a story:

- - -
{}
- - -

Below is a summary of the story up until this point:

- - -

Plot Summary: {}

Characters: {}

- - -

Instructions:
Your task is to sequentially analyze segments of a lengthy book, updating a comprehensive summary of the entire story, including plot and characters.

Guidelines:
1. Plot:
- First summarize the current segment. You must briefly introduce characters, places, and other major elements if they are mentioned for the first time.
- Update the overall plot summary to include vital information in this segment, related to key events, main conflicts, backgrounds, settings, and characters.
- The story may feature non-linear narratives, flashbacks, or switches between alternate worlds or viewpoints. In such cases, state the time and settings for these narrative shifts.

2. Characters:
- Update the major characters' information, including their profiles, current experiences, and overall experiences.
- Add new significant characters from this segment to the list.
- If the characters not mentioned in this segment are not crucial to the story's development, remove them from the list of major characters .

Output Format:
- - -
Summary of Current Segment:
{}
Overall Plot Summary (within 1000 words):
{}
Characters:
Name:
- **Profile:** Update the character's profile, including the role (e.g., protagonist), attributes, personality, and relationships within 50 words.
- **Current Experience:** Briefly describe the character's motivations, events, interactions, and emotional states in this segment within 50 words. If the character isn't mentioned, write "Not mentioned".
- **Overall Experience:** Update the character's overall experience and development throughout the story within 100 words, incorporating the current experience.
- - -

Now start your summarization:

Table 12: Prompt to incrementally update the summary of plot and characters.

NovelCritique

Below are the title, genres, and premise of a book-length story.

- - -

Title: {}

Genres: {}

Premise:
{}

- - -

Below are the plot summary, character analysis, and excerpts from this book-length story.

- - -

Plot Summary:
{}

Character Analysis:
{}

Excerpts: (Demonstrating Writing and Language):
{}

- - -

Write a review of this story based on the criteria below. Then conclude with an overall assessment and assign a final score from 1.0 (lowest) to 5.0 (highest).

Review:
*/*Specific Aspects*/*
Characters:
{Critiques of Characters.}

Emotional Impact:
{Critiques of Emotional Impact.}

Overall Assessment:
{Summary of the review.}

Score: X.X

Table 13: Instructions of NovelCritique. Here we display a training sample focusing on two specific aspects: Characters and Emotional Impact.

LLM-based Evaluation

Below are the title, genres, and premise of a book-length story.

- - -

Title: {}

Genres: {}

Premise:
{}

- - -

Below are the plot summary, character analysis, and excerpts from this book-length story.

- - -

Plot Summary:
{}

Character Analysis:
{}

Excerpts:
{}

- - -

Please evaluate the story based on the criteria listed below. Provide detailed reviews highlighting both strengths and weaknesses. Assign a score for each aspect on a continuous scale from 0 to 100, where 0 represents the poorest performance and 100 represents the best performance.

Evaluation Criteria:
1. Plot and Structure:
- **Review:**
- **Score:** X
- **Guidelines:** Evaluate the plot development by examining pace, twists, conflicts, and their resolutions. Evaluate the story structure for coherence, logic, and complexity, paying attention to key elements like climax and ending.

/*Definitions of other aspects*/

Conclusion:
- **Overall Assessment:**
- **Overall Score:** X
- **Guidelines:** Conclude with an overall assessment of the story, summarizing key insights and assigning an overall score on a continuous scale from 0 to 100.

Now start your evaluation:

Table 14: Prompt for LLM-based Evaluation. Here we display the prompt for summary-based evaluation, while for others, the difference is to replace the input contents and instructions. For example, in incremental evaluation, we input the current segment, previous summary, and previous evaluations, replacing "evaluate the story" as "update the evaluation".

Title	Author	Genres	Published Date	AVG. Score	Length(Words)
Remedial Magic	Melissa Marr	Fantasy, Romance, LGBT	February 20, 2024	2.90	88407
Sanctuary of the Shadow	Aurora Ascher	Fantasy, Romance, Romantasy	January 9, 2024	2.90	96569
We Came to Welcome You	Vincent Tirado	Horror, Thriller, LGBT	September 3, 2024	3.12	160662
Womb City	Tlotlo Tsamaase	Science Fiction, Horror, Dystopia	January 23, 2024	3.20	125465
Memory Piece	Lisa Ko	Historical Fiction, Literary Fiction, Contemporary	March 19, 2024	3.20	80654
The Longest Autumn	Amy Avery	Fantasy, Romance, Adult	January 16, 2024	3.21	91973
The Girl with No Reflection	Keshe Chow	Fantasy, Young Adult, Romance	August 6, 2024	3.28	109814
Argylle	Elly Conway	Mystery, Thriller, Mystery Thriller	January 4, 2024	3.28	113821
Castle of the Cursed	Romina Garber	Fantasy, Horror, Young Adult	July 30, 2024	3.30	90165
Beach Cute	Beth Reekles	Romance, Young Adult, Contemporary	May 14, 2024	3.34	88344
All This and More	Peng Shepherd	Science Fiction, Fantasy, Adult	July 9, 2024	3.36	135548
You Know What You Did	K.T. Nguyen	Thriller, Mystery, Mystery Thriller	April 16, 2024	3.36	95189
Draw Down the Moon	P.C. Cast	Fantasy, Young Adult, Romance	April 2, 2024	3.38	80332
Big Time	Ben H. Winters	Science Fiction, Thriller, Mystery	January 9, 2024	3.41	65522
These Deadly Prophecies	Andrea Tang	Fantasy, Mystery, Young Adult	January 30, 2024	3.42	70777
Fish Out of Water	Katie Ruggle	Romance, Contemporary, Adult	February 13, 2024	3.43	75901
Adam and Evie's Matchmaking Tour	Nora Nguyen	Romance, Contemporary, Travel	September 24, 2024	3.44	129425
The Village Library Demon-Hunting Society	C.M. Waggoner	Fantasy, Mystery, Cozy Mystery	September 24, 2024	3.44	93691
This Will Be Fun	E.B. Asher	Fantasy, Romance, Adult	October 29, 2024	3.45	148053
This Ravenous Fate	Hayley Dennings	Fantasy, Lesbian, Vampires	August 6, 2024	3.45	104419
I Need You to Read This	Jessa Maxwell	Mystery, Thriller, Mystery Thriller	August 13, 2024	3.46	71906
Lady Macbeth	Ava Reid	Fantasy, Historical Fiction, Retellings	August 13, 2024	3.48	69586
Every Time I Go on Vacation, Someone Dies	Catherine Mack	Mystery, Cozy Mystery, Mystery Thriller	April 30, 2024	3.48	76346
Dark Restraint	Katee Robert	Romance, Fantasy, Mythology	August 6, 2024	3.49	78817
Infinity Alchemist	Kacen Callender	Fantasy, Young Adult, LGBT	February 6, 2024	3.49	113207
An Academy for Liars	Alexis Henderson	Fantasy, Horror, Gothic	September 17, 2024	3.50	121043
Love, Lies, and Cherry Pie	Jackie Lau	Romance, Contemporary, Adult	May 7, 2024	3.52	82144
One Big Happy Family	Jamie Day	Thriller, Mystery, Mystery Thriller	July 16, 2024	3.54	106628
Ghost Station	S.A. Barnes	Horror, Science Fiction, Thriller	April 9, 2024	3.55	105849
Granite Harbor	Peter Nichols	Mystery, Thriller, Horror	April 30, 2024	3.56	81080
A Step Past Darkness	Vera Kurian	Thriller, Mystery, Mystery Thriller	February 20, 2024	3.56	136922
Christa Comes Out of Her Shell	Abbi Waxman	Romance, Contemporary, Chick Lit	April 16, 2024	3.57	91578
The Getaway List	Emma Lord	Romance, Young Adult, Contemporary	January 23, 2024	3.57	93259
Last House	Jessica Shattuck	Historical Fiction, Historical, Family	May 14, 2024	3.59	129626
The Spare Room	Laura Starkey	Romance, Contemporary, Chick Lit	February 6, 2024	3.59	92741
The Honey Witch	Sydney J. Shields	Fantasy, Romance, LGBT	May 14, 2024	3.59	98048
The Midnight Feast	Lucy Foley	Mystery, Thriller, Mystery Thriller	June 18, 2024	3.59	90608
Our Holiday	Louise Candlish	Thriller, Mystery, Crime	July 4, 2024	3.60	114970
The Ministry of Time	Kaliane Bradley	Science Fiction, Romance, Time Travel	May 7, 2024	3.61	87170
Sweet Nightmare	Tracy Wolff	Fantasy, Romance, Young Adult	May 7, 2024	3.61	139647
Hypnotized by Love	Sariah Wilson	Romance, Contemporary, Chick Lit	March 1, 2024	3.62	79768
The Bad Ones	Melissa Albert	Horror, Fantasy, Young Adult	February 20, 2024	3.63	87730
The Cautious Traveller's Guide to the Wastelands	Sarah Brooks	Fantasy, Mystery, Historical	June 18, 2024	3.63	94373
A Friend in the Dark	Samantha M. Bailey	Thriller, Mystery, Mystery Thriller	March 1, 2024	3.64	67035
The Framed Women of Ardemore House	Brandy Schillace	Mystery, Mystery Thriller, Cozy Mystery	February 13, 2024	3.65	90841
49 Miles Alone	Natalie D. Richards	Thriller, Young Adult, Mystery	July 2, 2024	3.65	63689
The Cemetery of Untold Stories	Julia Alvarez	Fantasy, Magical Realism, Historical Fiction	April 2, 2024	3.66	72967
The Nature of Disappearing	Kimi Cunningham Grant	Thriller, Mystery, Mystery Thriller	June 18, 2024	3.66	79535
Perfect Little Monsters	Cindy R.X. He	Mystery, Young Adult, Thriller	May 7, 2024	3.66	68630
Floating Hotel	Grace Curtis	Science Fiction, Mystery, Fantasy	March 19, 2024	3.67	79893
Ocean's Godori	Elaine U. Cho	Science Fiction, Romance, Queer	April 23, 2024	3.67	82458
The Honeymoon Affair	Sheila O'Flanagan	Romance	May 9, 2024	3.67	117284
The House on Biscayne Bay	Chanel Cleeton	Historical Fiction, Mystery, Gothic	April 2, 2024	3.68	75876
I'm Afraid You've Got Dragons	Peter S. Beagle	Fantasy, Dragons, Humor	May 14, 2024	3.69	72237
Under Loch and Key	Lana Ferguson	Romance, Fantasy, Paranormal	December 3, 2024	3.69	105973
Such a Lovely Family	Aggie Blum Thompson	Thriller, Mystery, Mystery Thriller	March 12, 2024	3.69	94600
Middle of the Night	Riley Sager	Thriller, Mystery, Mystery Thriller	June 18, 2024	3.69	97895
The Summer She Went Missing	Chelsea Ichaso	Mystery, Thriller, Young Adult	March 1, 2024	3.69	83624
The Truth According to Ember	Danica Nava	Romance, Contemporary, Indigenous	August 6, 2024	3.70	91302
The Stardust Grail	Yume Kitasei	Fantasy, Space, Adult	June 11, 2024	3.70	93381
Beastly Beauty	Jennifer Donnelly	Fantasy, Young Adult, Retellings	May 7, 2024	3.70	91129
Of Jade and Dragons	Amber Chen	Fantasy, Young Adult, Romance	June 18, 2024	3.71	103917
Seven Summer Weekends	Jane L. Rosen	Romance, Contemporary, Womens Fiction	June 4, 2024	3.72	65185
The Vacancy in Room 10	Seraphina Nova Glass	Thriller, Mystery, Mystery Thriller	April 9, 2024	3.73	91666
Madwoman	Chelsea Bieker	Mystery, Thriller, Mystery Thriller	September 3, 2024	3.73	90606
Beautiful Ugly	Alice Feeney	Thriller, Mystery, Mystery Thriller	January 14, 2025	3.74	84800
Such Charming Liars	Karen M. McManus	Mystery, Young Adult, Mystery Thriller	July 30, 2024	3.74	91903
The Stars Too Fondly	Emily Hamilton	Science Fiction, Romance, LGBT	June 11, 2024	3.76	116399
Lies and Weddings	Kevin Kwan	Romance, Contemporary, Chick Lit	May 21, 2024	3.76	105919
If Something Happens to Me	Alex Finlay	Thriller, Mystery, Mystery Thriller	May 28, 2024	3.76	61908
Long Island Compromise	Taffy Brodesser-Akner	Historical Fiction, Contemporary, Literary Fiction	July 9, 2024	3.77	81230
Death at Morning House	Maureen Johnson	Mystery, Young Adult, Mystery Thriller	August 6, 2024	3.77	85578
The Wren in the Holly Library	K.A. Linde	Fantasy, Romance, Romantasy	June 4, 2024	3.78	125253
Say you'll be mine: A novel	Naina Kumar	Romance, Contemporary, Adult	January 16, 2024	3.79	86540
How to Solve Your Own Murder	Kristen Perrin	Mystery, Mystery Thriller, Thriller	March 26, 2024	3.79	99534
Love Story	Lindsey Kelk	Romance, Contemporary, Chick Lit	July 4, 2024	3.81	95533
In the Hour of Crows	Dana Elmendorf	Fantasy, Mystery, Magical Realism	June 4, 2024	3.81	78813
One Last Breath	Ginny Myers Sain	Mystery, Young Adult, Thriller	March 5, 2024	3.81	93658
A Botanist's Guide to Society and Secrets	Kate Khavari	Mystery, Historical Fiction, Cozy Mystery	June 4, 2024	3.81	96785
The School Run	Ali Lowe	Thriller, Mystery, Crime	March 5, 2024	3.82	88185
Five Broken Blades	Mai Corland	Fantasy, Romance, Adult	May 7, 2024	3.83	114493
The Broken Places	Mia Sheridan	Romance, Mystery, Romantic Suspense	December 1, 2024	3.84	100338
The Murder After the Night Before	Katy Brent	Mystery, Thriller, Mystery Thriller	February 29, 2024	3.85	79183
So Let Them Burn	Kamilah Cole	Fantasy, Young Adult, Dragons	January 16, 2024	3.85	101796
Shelterwood	Lisa Wingate	Historical Fiction, Mystery, Historical	June 4, 2024	3.86	110366
Dragonfruit	Makiia Lucier	Fantasy, Young Adult, Dragons	April 9, 2024	3.87	70344
House of Glass	Sarah Pekkanen	Thriller, Mystery, Mystery Thriller	August 6, 2024	3.88	93373
The Other Valley: A Novel	Scott Alexander Howard	Science Fiction, Fantasy, Time Travel	February 27, 2024	3.89	92041
Triple Sec	T.J. Alexander	Romance, LGBT, Queer	June 4, 2024	3.89	91915
House of Bone and Rain	Gabino Iglesias	Horror, Thriller, Mystery	August 6, 2024	3.89	104032

Table 15: Metadata of the test set. (Part 1/2)

Title	Author	Genres	Published Date	AVG. Score	Length(Words)
Voyage of the Damned	Frances White	Fantasy, Mystery, LGBT	January 18, 2024	3.90	116476
A Song to Drown Rivers	Ann Liang	Fantasy, Romance, Historical	October 1, 2024	3.90	93968
The Prisoner's Throne	Holly Black	Fantasy, Young Adult, Romance	March 5, 2024	3.90	92811
ASAP	Axie Oh	Romance, Young Adult, Contemporary	February 6, 2024	3.91	112194
Three Kinds of Lucky	Kim Harrison	Fantasy, Urban Fantasy, Paranormal	March 5, 2024	3.91	144316
The Queen of Sugar Hill: A Novel of Hattie McDaniel	ReShonda Tate	Historical Fiction, Historical, Race	January 30, 2024	3.92	160852
Icon and Inferno	Marie Lu	Romance, Young Adult, Mystery	June 11, 2024	3.92	82204
The Eternal Ones	Namina Forna	Fantasy, Young Adult, Young Adult Fantasy	February 13, 2024	3.92	115521
The Excitements	C.J. Wray	Historical Fiction, Mystery, Historical	January 30, 2024	3.93	88857
Hurt Mountain	Angela Crook	Mystery, Thriller, Mystery Thriller	February 27, 2024	3.93	78349
A Place for Vanishing	Ann Fraistat	Horror, Young Adult, Gothic	January 16, 2024	3.96	108387
Experienced	Kate Young	Romance, LGBT, Lesbian	June 4, 2024	3.98	93976
Still the Sun	Charlie N. Holmberg	Fantasy, Romance, Science Fiction	July 1, 2024	4.00	85867
The Bright Sword	Lev Grossman	Fantasy, Historical Fiction, Retellings	July 16, 2024	4.01	184430
Love at First Book	Jenn McKinlay	Romance, Contemporary, Books About Books	May 14, 2024	4.01	95564
Immortal	Sue Lynn Tan	Fantasy, Romance, Romantasy	January 7, 2025	4.03	189857
Where Sleeping Girls Lie	Faridah Àbíké-Íyímídé	Mystery, Young Adult, Thriller	March 14, 2024	4.03	133205
I'm Starting to Worry About This Black Box of Doom	Jason Pargin	Science Fiction, Humor, Mystery	September 24, 2024	4.05	128001
The Book of Doors	Gareth Brown	Fantasy, Magical Realism, Time Travel	February 13, 2024	4.05	175207
Deeper Than the Dead	Debra Webb	Mystery, Thriller, Mystery Thriller	August 6, 2024	4.05	112694
Summer Romance	Annabel Monaghan	Romance, Contemporary, Chick Lit	June 4, 2024	4.06	77896
Fire and Bones	Kathy Reichs	Mystery, Thriller, Crime	August 6, 2024	4.06	60889
Here One Moment	Liane Moriarty	Mystery, Contemporary, Thriller	September 10, 2024	4.06	128430
Reckless	Lauren Roberts	Fantasy, Romance, Romantasy	July 2, 2024	4.07	97628
Even If It Breaks Your Heart	Erin Hahn	Romance, Young Adult, Contemporary	February 6, 2024	4.08	79388
The Hunter	Tana French	Mystery, Ireland, Mystery Thriller	March 5, 2024	4.09	138237
A Sorceress Comes to Call	T. Kingfisher	Fantasy, Horror, Retellings	August 6, 2024	4.10	95649
The Missing Witness	Allison Brennan	Mystery, Suspense, Thriller	January 23, 2024	4.11	98108
Eleven Eleven	Micalea Smeltzer	Romance, Contemporary, Adult	January 24, 2024	4.12	109791
By Any Other Name	Jodi Picoult	Historical Fiction, Historical, Contemporary	August 20, 2024	4.13	135244
The Final Act of Juliette Willoughby	Ellery Lloyd	Historical Fiction, Mystery, Thriller	June 11, 2024	4.13	104445
I Hope This Doesn't Find You	Ann Liang	Romance, Young Adult, Contemporary	February 6, 2024	4.16	71983
The Day Shelley Woodhouse Woke Up	Laura Pearson	Romance, Contemporary, Mystery	April 6, 2024	4.18	78800
The God of the Woods	Liz Moore	Mystery, Thriller, Mystery Thriller	July 2, 2024	4.18	118863
The Unquiet Bones	Loreth Anne White	Thriller, Mystery, Mystery Thriller	March 5, 2024	4.19	93190
The Mercy of Gods	James S.A. Corey	Science Fiction, Space Opera, Fantasy	August 6, 2024	4.20	114209
The Ghost Orchid	Jonathan Kellerman	Mystery, Crime, Thriller	February 6, 2024	4.20	64841
Embers in the London Sky	Sarah Sundin	Historical Fiction, Christian Fiction, Romance	February 6, 2024	4.21	87676
Of Elves and Embers	Elle Madison	Fantasy, Romance, Romantasy	February 6, 2024	4.24	105599
Disturbing the Dead	Kelley Armstrong	Mystery, Time Travel, Historical Fiction	May 7, 2024	4.25	114262
Wisteria	Adalyn Grace	Fantasy, Romance, Young Adult	August 20, 2024	4.27	107341
Think Twice	Harlan Coben	Mystery, Thriller, Mystery Thriller	May 14, 2024	4.28	79221
Restless Stars	Caroline Peckham	Fantasy, Romance, Romantasy	April 23, 2024	4.29	308631
The Tainted Cup	Robert Jackson Bennett	Fantasy, Mystery, Mystery Thriller	February 6, 2024	4.31	110461
The Briar Club	Kate Quinn	Historical Fiction, Mystery, Historical	July 9, 2024	4.32	188275
The Act of Disappearing	Nathan Gower	Historical Fiction, Mystery, Historical	May 28, 2024	4.32	100260
Mind Games	Nora Roberts	Romance, Romantic Suspense, Suspense	May 21, 2024	4.35	127819
Exposed	Brynne Asher	Romance, Romantic Suspense, Suspense	March 26, 2024	4.36	104163
The Instruments of Darkness	John Connolly	Mystery, Thriller, Horror	May 7, 2024	4.41	127973
Toxic Prey	John Sandford	Mystery, Thriller, Crime	April 9, 2024	4.42	86433
All My Secrets	Lynn Austin	Historical Fiction, Christian Fiction, Christian	February 6, 2024	4.43	120629
The Austrian Bride	Helen Parusel	Historical Fiction, Romance, World War II	January 15, 2024	4.43	93443
Quicksilver	Callie Hart	Fantasy, Romantasy, Romance	June 4, 2024	4.43	190189
Heartbeat	Sharon Sala	Romance, Romantic Suspense, Mystery	February 6, 2024	4.45	101807
Random in Death	J.D. Robb	Mystery, Romance, Crime	January 23, 2024	4.47	91807
Heart of Frost and Scars	Pam Godwin	Reverse Harem, Dark, Romance	August 26, 2024	4.48	155855
A Calamity of Souls	David Baldacci	Mystery, Thriller, Crime	April 16, 2024	4.49	122787
Frank and Red	Matt Coyne	Contemporary, Family, Humor	February 1, 2024	4.52	100718
James	Percival Everett	Historical, Race, Retellings	March 19, 2024	4.53	58104
The Women	Kristin Hannah	Historical Fiction, Historical, Adult	February 6, 2024	4.62	124703

Table 16: Metadata of the test set. (Part 2/2)