

WebNovelBench: Placing LLM Novelists on the Web Novel Distribution

Leon Lin¹, Jun Zheng², Haidong Wang²

¹Nanyang Technological University, ²Sun Yat-Sen University
liangtao002@e.ntu.edu.sg, {zhengj98, wanghd7}@mail2.sysu.edu.cn

 <https://github.com/OedonLestrange42/webnovelbench>

 <https://huggingface.co/datasets/Oedon42/webnovelbench>

Abstract

Robustly evaluating the long-form storytelling capabilities of Large Language Models (LLMs) remains a significant challenge, as existing benchmarks often lack the necessary scale, diversity, or objective measures. To address this, we introduce WebNovelBench, a novel benchmark specifically designed for evaluating long-form novel generation. WebNovelBench leverages a large-scale dataset of over 4,000 Chinese web novels, framing evaluation as a synopsis-to-story generation task. We propose a multifaceted framework encompassing eight narrative quality dimensions, assessed automatically via an LLM-as-Judge approach. Scores are aggregated using Principal Component Analysis and mapped to a percentile rank against human-authored works. Our experiments demonstrate that WebNovelBench effectively differentiates between human-written masterpieces, popular web novels, and LLM-generated content. We provide a comprehensive analysis of 24 state-of-the-art LLMs, ranking their storytelling abilities and offering insights for future development. This benchmark provides a scalable, replicable, and data-driven methodology for assessing and advancing LLM-driven narrative generation.

1 Introduction

Can Large Language Models (LLMs) generate stories that surpass human-written ones? Recent breakthroughs, exemplified by models like GPT-4o (Hurst et al., 2024) and Deepseek-R1 (DeepSeek-AI et al., 2025a), underscore their remarkable ability to produce coherent, imaginative, and contextually nuanced narratives. This raises intriguing questions: How proficient are today’s LLMs in story generation, and how do their outputs compare to human-authored works?

Evaluating LLM performance in this open-ended domain remains a significant challenge. While prior research has explored story generation evaluation (Guan et al., 2021; Liu et al., 2024; Paech,

2024; Ismayilzada et al., 2025), these efforts often face limitations such as small dataset sizes or insufficient story diversity, hindering widespread adoption. This contrasts with fields like code generation and mathematical reasoning, where benchmarks such as CodeForces Rating (Quan et al., 2025) and American Invitational Mathematics Examination 2024 (AIME 2024) (MAA, 2024) serve as widely accepted standards.

Inspired by such successes, we propose WebNovelBench, a comprehensive and intuitive benchmark for story generation guided by three key principles:

- **Broad Data Foundation:** Utilizes diverse, popular human-authored works.
- **Representative Tasks:** Covers diverse storytelling styles, themes, and complexities (details in Section 3.1).
- **Automated and Objective Evaluation:** Minimizes subjectivity via robust, consistent automated methods.

We leverage 4,000+ popular Chinese web novels (> 10,000 readers each) for a synopsis-to-story generation task. An LLM-as-Judge approach evaluates stories across eight narrative dimensions. Validation with Mao Dun Literature Prize novels, which scored high (Figure 1), confirms alignment with human judgment. WebNovelBench thus offers automatic assessment of LLM storytelling capabilities without extensive manual intervention, establishing a standardized framework for comparison against human-authored content. In summary, our main contributions are:

- We introduce WebNovelBench, a large-scale, data-driven evaluation framework for story generation, accurately ranking human-authored and LLM-generated stories via distribution analysis.
- We define eight evaluation dimensions for Chinese story quality, employing a validated LLM-as-Judge mechanism for robust automated evaluation.

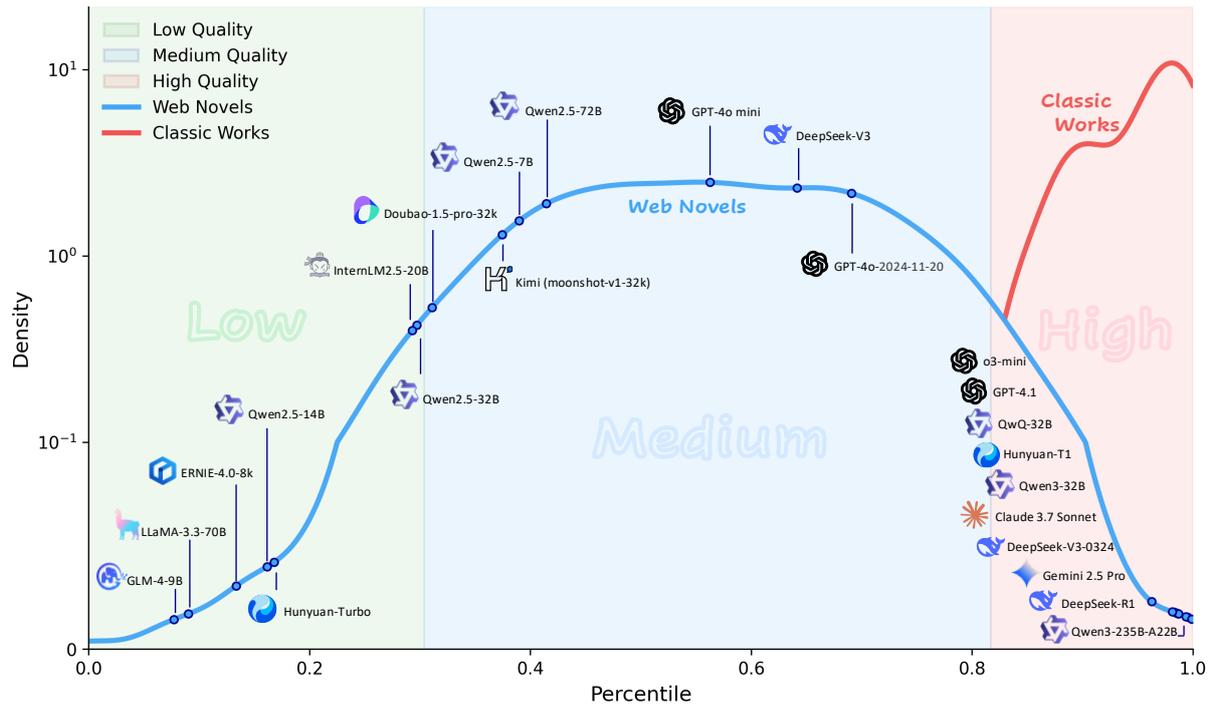


Figure 1: **Web Novel Dataset Distribution and LLM Placement.** Our web novel dataset’s quality distribution, with Low, Medium, and High zones (95% CIs). The blue curve represents the benchmark distribution constructed from 4,000 curated Chinese web novels. The red curve (Classic Works) was generated using the same pipeline on 25 Mao Dun Literature Prize novels. Positions of 24 LLMs indicate their performance relative to this corpus.

- We conduct a comprehensive evaluation of 24 advanced LLMs, ranking their storytelling abilities relative to human-authored works and offering insights for future development.

2 Related Work

2.1 LLM General Benchmark

General LLM benchmarks like MMLU (Hendrycks et al., 2021) and its variants (Wang et al., 2024; Yue et al., 2024), or dynamic benchmarks like MixEval (Ni et al., 2024), are invaluable for assessing broad capabilities such as reasoning and knowledge recall. However, they often lack the specific, nuanced criteria required to effectively evaluate creative, open-ended tasks like story generation, particularly regarding narrative quality, coherence, and creativity. This highlights the need for specialized benchmarks in creative domains, akin to CodeForces Rating (Quan et al., 2025) and SWE-Bench Verified (Jimenez et al., 2024) or AIME 2024 (MAA, 2024) for coding and mathematics.

2.2 Story Generation Benchmark

Previous work on evaluating LLM-generated stories has explored creativity (Ismayilzada et al., 2025; Paech, 2024) and the correlation of automatic

metrics with human judgment (Guan et al., 2021; Chhun et al., 2022; Xie et al., 2023; Tian et al., 2024). For instance, EQ-Bench (Paech, 2024) offers a creative writing score based on a limited set of thirty-two prompts. OpenMEVA (Guan et al., 2021) provided a framework for evaluating automatic metrics but did not focus on broad dataset diversity. Xie et al. (2023) and Tian et al. (2024) found that while LLMs can approach human-level quality, their stories often lack originality, suspense, and narrative diversity compared to human writing. These existing efforts are limited by small datasets, narrow diversity, or restricted evaluation criteria, leaving open questions about how closely LLMs match human creativity and narrative quality. To address this, we present a large, diverse web novel benchmark and a robust evaluation framework for systematic comparison between LLM and human storytelling.

2.3 LLM-as-a-Judge

Automated evaluation for open-ended text generation increasingly employs the LLM-as-a-Judge paradigm (Zheng et al., 2023), where LLMs assess outputs without reference texts (Li et al., 2024; Kasner and Dušek, 2024). While promising for align-

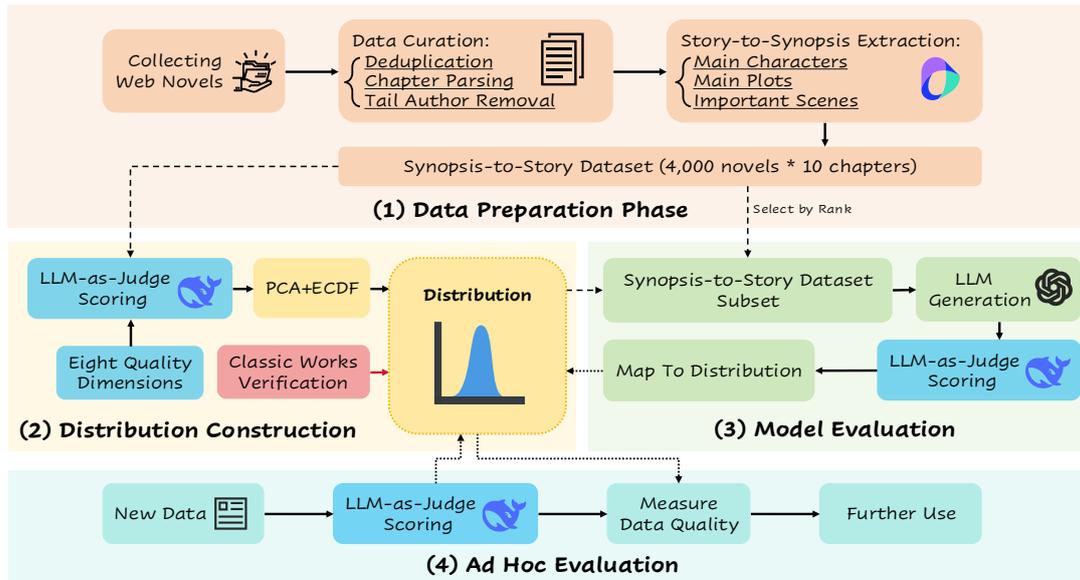


Figure 2: **Framework of Our Method.** Our benchmark framework consists of four major components: (1) *Data Preparation Phase*: We collect and curate a large web novel dataset, and use Doubao for story-to-synopsis extraction to build a 4,000 novels synopsis-to-story dataset. (2) *Distribution Construction*: Each story is scored across eight quality dimensions using LLM-as-judge, followed by PCA+ECDF to form a quality distribution benchmark. Classic literary works are used to validate the high end of the distribution. (3) *Model Evaluation*: LLMs generate stories from selected subsets of the dataset. Their outputs are scored and mapped onto the distribution to assess model performance. (4) *Ad Hoc Evaluation*: New data can be scored and aligned with the benchmark for measuring data quality and supporting further applications.

ing with human preferences, LLM-as-Judge reliability is a concern, with research focusing on fairness and potential biases (Shi et al., 2025; Zhang et al., 2023; Ye et al., 2024). For instance, Ye et al. (2024) introduced CALM to identify and quantify biases in LLM judges. Recent works (Liu et al., 2023; Kim et al., 2023; Pereira et al., 2024; Chang et al., 2023) also leverage LLM-as-a-Judge for evaluating text generation, but do not specifically address the unique challenges of story evaluation, where aspects such as narrative coherence, structural coherence, and characterisation are essential. This body of work underscores the importance of careful design and validation of the LLM-as-Judge component, which we address in our work to ensure reliable and fair story evaluation based on multi-dimensional criteria.

3 Benchmark Construction

3.1 Dataset

Based on Jun (2023)’s work, we curated a dataset from over 10,000 Chinese web novels (published 2013-2020) via several preprocessing steps:

- **Deduplication.** Some text files may contain highly similar content under different titles, in-

dicating that they represent the same novel. Directly comparing exact character matches is insufficient for identifying duplicates. To address this, we employed `difflib`¹ to compute pairwise similarity scores and removed novels with a similarity score greater than 0.9.

- **Chapter Parsing.** Web novels are typically long-form narratives presented in a serial-chapter format. However, the raw texts lacked explicit chapter delimiters. To extract chapters, we designed regular expression patterns tailored to common chapter formatting styles in Chinese web novels. Novels containing fewer than ten chapters were excluded, as they did not meet our definition of “long-form” or suggested incomplete parsing.
- **Tail Author Removal.** It is commonly observed that successful authors tend to publish multiple novels over time. Based on this observation, we compiled the list of all authors and excluded novels written by those with the fewest works. After this filtering step, we retained a final set of 4,000 web novels.

The curated 4,000 novels ensure ‘Representative Tasks’ by covering diverse genres (e.g., Eastern

¹A Python Standard Library

Fantasy: 1281, General Realism: 1255, Western Fantasy: 670, Historical: 234, plus Sci-Fi, Suspense, Romance with varied subthemes like 'Original World Setting'). This breadth reflects popular web fiction complexities. (Details in Appendix E)

For the synopsis-to-story task, we utilize Doubao-pro-32k to generate a 'synopsis' (main characters, key plot points, important scenes) for ten random consecutive chapters from each novel. This yielded ten < chapter content, synopsis > pairs per novel. (Details in Appendix A)

3.2 Metric

We evaluate narrative quality through eight carefully designed key dimensions (Table 1). These dimensions were systematically established through a multi-stage process. We began by reviewing established narrative theories (e.g., character arcs, plot structuring) and analyzing the official writing guides from major web novel platforms. This was followed by a pilot study where initial dimension candidates were tested for clarity, assessability, and orthogonality. The final eight dimensions were selected to provide a comprehensive yet non-redundant coverage of essential stylistic and structural components of storytelling, enabling nuanced analysis beyond surface-level fluency and effectively capturing the complexity and depth inherent in narrative quality.

3.3 Scoring Method

To rank LLM works against our 4,000 human-authored web novels, we combine Principal Component Analysis (PCA) (Hotelling, 1933) for multi-dimensional score aggregation and Empirical Cumulative Distribution Function (ECDF) (Conover, 1999) for percentile ranking.

Score Aggregation via PCA. Given a dataset of N samples, where each sample is evaluated along d dimensions, we first normalize all dimension scores using z-score standardization:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} (i \in [1, N], j \in [1, d]), \quad (1)$$

where x_{ij} is the raw score of the i -th sample in the j -th dimension, and μ_j and σ_j are the mean and standard deviation of the j -th dimension across all samples.

We then perform PCA on the standardized data to extract the first principal component, whose normalized loading vector $\mathbf{w} = (w_1, w_2, \dots, w_d)$ represents the relative importance of each dimension.

The aggregated composite score s_i for the i -th sample is then computed as a weighted sum:

$$s_i = \sum_{j=1}^d w_j z_{ij} \quad (2)$$

Percentile Ranking via ECDF. To translate raw scores into interpretable relative rankings, we apply the ECDF over all aggregated scores $\{s_1, s_2, \dots, s_N\}$:

$$\text{ECDF}(x) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(s_i \leq x), \quad (3)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The ECDF provides a percentile score in the range $[0, 1]$, representing the proportion of samples with scores less than or equal to x .

Given a new LLM-generated sample with aggregated score s_{new} , its percentile rank is $P_{\text{new}} = \text{ECDF}(s_{\text{new}})$.

This percentile reflects the model's performance relative to the full empirical distribution of the reference dataset, indicating the level at which stories generated by a LLM are comparable to those written by humans. By evaluating the model across batches of test samples, we can estimate the LLM's overall writing ability.

Let $\mathcal{B} = \{s_{\text{LLM}}^{(1)}, s_{\text{LLM}}^{(2)}, \dots, s_{\text{LLM}}^{(M)}\}$ be a set of aggregated scores for a batch of M LLM-generated samples. The estimated writing level is then defined as the average percentile:

$$\hat{P}_{\text{LLM}} = \frac{1}{M} \sum_{m=1}^M \text{ECDF}(s_{\text{LLM}}^{(m)}) \quad (4)$$

The value $\hat{P}_{\text{LLM}} \in [0, 1]$ represents the expected percentile rank of the LLM's output relative to the distribution of human-written texts in the reference dataset.

4 Experiments and Results

4.1 Experimental Setup

Due to resource limits, our evaluation dataset uses 100 web novels (one per percentile from the 4,000-novel distribution), with 10 synopsis-to-story pairs per novel (1,000 test samples in total). The overall rank of an LLM is its average percentile rank across these 100 books.

LLMs (open-source and proprietary) were accessed via APIs with a standardized system prompt

Metric	评估维度	Explanation	Weight
D1: Use of Literary Devices	修辞手法	Based on the quantity and quality of rhetorical devices like metaphor and symbolism	<u>0.1304</u>
D2: Richness of Sensory Detail	感官描述丰富度	Frequency of visual, auditory, and other sensory descriptions	0.1160
D3: Balance of Character Presence	角色平衡度	Frequency, dialogue proportion, and psychological depth of each character	0.1152
D4: Distinctiveness of Character Dialogue	角色对白独特性	Whether dialogue reflects distinct personalities	0.1171
D5: Consistency of Characterisation	角色一致性	Whether language and actions align with character identity	0.1377
D6: Atmospheric and Thematic Alignment	意境匹配度	Whether scenes support the overall atmosphere and themes	0.1290
D7: Contextual Appropriateness	语境适配度	Whether settings match time/place/cultural background	0.1281
D8: Scene-to-Scene Coherence	跨场景衔接度	Whether scene transitions are smooth and natural	0.1263

Table 1: **Narrative Evaluation Metrics and PCA-Derived Weights.** This table lists the eight dimensions used to evaluate narrative quality. Each weight reflects the metric’s relative importance, derived through PCA on web novels scores.

and constant generation settings (4096 max tokens, temperature 0.6, eight evaluation criteria in prompt; see Appendix A). Outputs were evaluated by Deepseek-V3 using a consistent evaluation prompt. For details on the robustness analysis, see Section 5.3 below.

4.2 Main Experiments

We evaluate a total of 13 open source models and 11 closed source models on our benchmark. Figure 3 illustrates the performance of these frontier models across eight distinct narrative evaluation dimensions and overall effect.

Top models (Qwen3-235B-A22B (Yang et al., 2025), Deep-Seek R1 (DeepSeek-AI et al., 2025a), Gemini-2.5-Pro) score high (3.5-4.6) across dimensions; Qwen3-235B-A22B achieves a 5.21 norm score, indicating strong alignment with high-quality human writing. Mid-tier models (e.g., GPT-4o, DeepSeek-V3 (DeepSeek-AI et al., 2025b)) show varied performance (scores 2.5-3.8), highlighting areas like sensory detail and literary device use for improvement. Lower-ranked models (e.g., GLM-4-9B-chat (GLM et al., 2024), LLaMA-3-8B (Grattafiori et al., 2024)) perform consistently poorly (less than 2.0 norm score), especially in literary devices and character dialogue, indicating

significant room for improvement, particularly for open-source or smaller LLMs.

An insightful pattern emerging from the heatmap lies in the varying levels of difficulty across evaluation dimensions. Most models perform relatively well in dimensions such as D6 and D8, suggesting these aspects are easier for current LLMs to capture. In contrast, D3 consistently shows lower and more uniform scores across models, indicating a shared weakness in maintaining psychological depth and balanced development of multiple characters. Similarly, dimensions like D1 and D2 also pose challenges, particularly for smaller models. These findings point to concrete directions for future improvement that incorporate richer narrative techniques and improve character-level modeling may help LLMs achieve more human-like storytelling abilities.

In general, this benchmark effectively captures distinct strengths and weaknesses among current LLMs. While the most advanced models achieve near-perfect scores within our distribution, demonstrating their strong performance on the story-to-synopsis dataset, this does not diminish the value of the benchmark. In contrast, it validates our original intuition. Our primary goal is to evaluate the story generation capabilities of contemporary

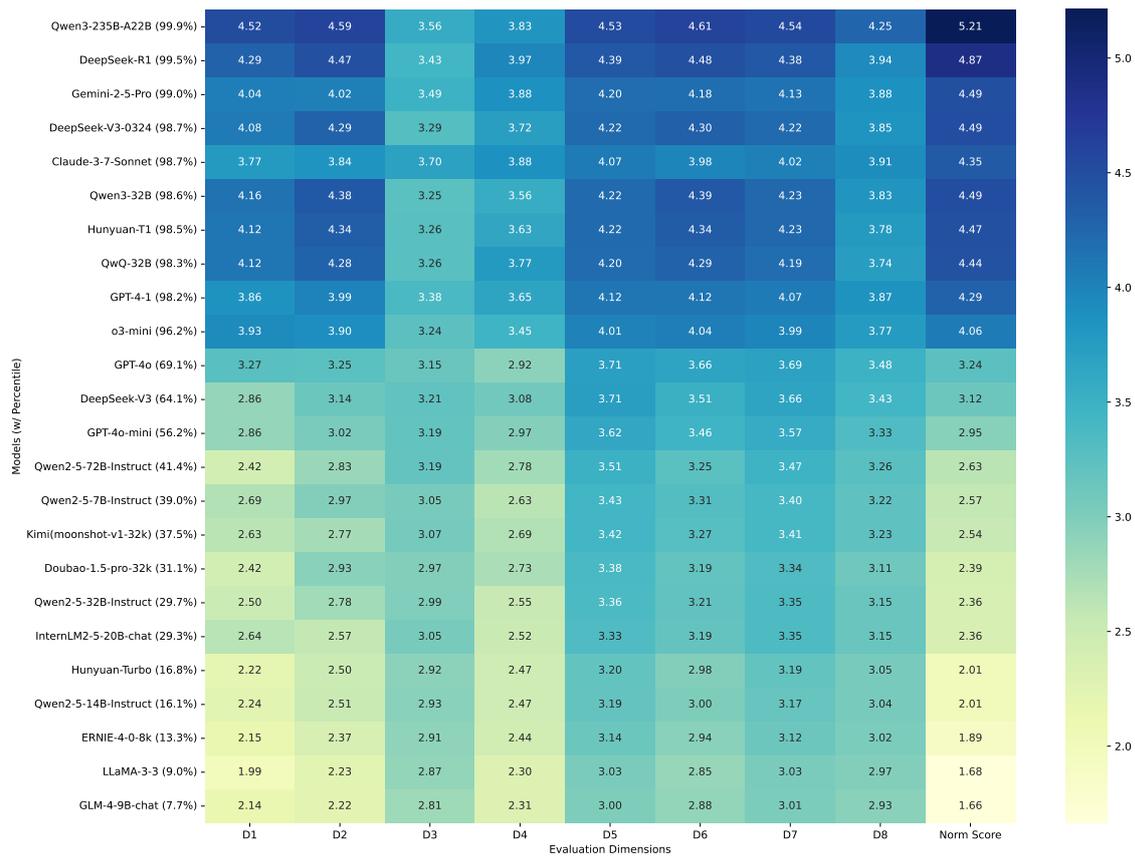


Figure 3: **LLM Performance Heatmap Across Narrative Dimensions.** Shows average scores (on a 1-5 scale) for 24 LLMs on eight dimensions, sorted by percentile rank. Final column is PCA-derived weighted norm score. Higher scores indicate better alignment with quality human writing.

LLMs, and the results suggest that leading models have reached a level comparable to top-tier works in web novels. This study primarily presents a methodological framework; future work may extend it through the inclusion of higher-quality reference texts or more fine-grained and diverse evaluation criteria.

4.3 Comparison with Other Benchmarks

To contextualize our contribution, we compare WebNovelBench with existing benchmarks for story generation evaluation, as summarized in Table 2. OpenMEVA (Guan et al., 2021) provides a framework for evaluating automatic metrics using an existing dataset of 400 stories. It aligns with human preference and utilizes 8 evaluation dimensions. However, it does not employ weighted dimensions for score aggregation, potentially treating all aspects of narrative quality as equally important, which might not reflect nuanced human judgment. AlignBench Writing Ability (Liu et al., 2024) uses a self-constructed dataset of 75 stories. While it considers human preference and evaluates across

5 dimensions, its dataset size is relatively small, which might limit the diversity of narrative styles and scenarios covered. Similar to OpenMEVA, it does not incorporate dimension weights in its scoring. EQ-Bench Longform Creative Writing (Paech, 2024) focuses on long-form creative writing with a self-constructed dataset of 12 stories, each comprising 8 chapters (totaling 96 evaluation instances). It employs a comprehensive set of 14 dimensions. However, according to the available information, it does not explicitly align its dataset construction with broad human preference (e.g., via popularity metrics of source texts) and also lacks a weighted approach to aggregating its numerous dimensional scores. The work by Ismayilzada et al. (2025) evaluates creative short story generation using an existing dataset. While it aligns with human preference and uses 4 evaluation dimensions, its very small test set of only 4 stories per LLM significantly limits the robustness and generalizability of its findings. It also does not use weighted dimensions. In contrast, WebNovelBench offers several key advantages:

Benchmark	Dataset Information		Evaluation Method		
	Data Source	Testing Samples	Human Preference Alignment	Dimension Weights	Evaluation Dimension
OpenMEVA (Guan et al., 2021)	Existing dataset	400 stories	✓	✗	8
AlignBench Writing Ability (Liu et al., 2024)	Self-constructed	75 stories	✓	✗	5
EQ-Bench Longform Creative Writing (Paech, 2024)	Self-constructed	12 stories, each with 8 chapters	✗	✗	14
Ismayilzada et al. (2025)	Existing dataset	4 stories	✓	✗	4
WebNovelBench (ours)	Web novels	100 stories, each with 10 chapters	✓	✓	8

Table 2: Comparison of Other Benchmarks.

- **Scale and Diversity:** Built on 4,000+ web novels, testing with 100 distinct stories (10 chapters each, 1,000 instances), ensuring wide genre/style coverage.
- **Inherent Human Preference Alignment:** By using popular web novels (each with over 10,000 readers) as the source, our benchmark inherently captures broad human literary preferences.
- **Data-Driven Dimension Weighting:** PCA derives weights for 8 narrative dimensions, providing a nuanced, objective assessment reflecting their relative importance in human-authored works.
- **Comprehensive and Focused Evaluation:** Eight carefully defined dimensions provide a thorough yet focused assessment of key storytelling elements.

These features make WebNovelBench a robust, scalable, and replicable solution for assessing and advancing LLM-driven narrative generation, particularly for long-form stories in the Chinese web novel domain.

5 Rationality Analysis

5.1 Metric Analysis

To assess the rationality and effectiveness of the proposed eight evaluation metrics, we conducted a detailed statistical analysis combining PCA and distributional visualization.

Principal Component Analysis.² PCA shows the first component explains 75.6% of variance (first three greater than 90%), indicating the metrics capture a dominant quality factor. This high concentration of variance in the first principal component is a significant finding. It suggests that while our metrics capture distinct facets of narrative, they are strongly correlated and align with a unified dimension interpretable as ‘overall narrative quality.’ This supports our use of dimensional aggregation, as it reflects the most significant shared signal

²See Appendix B for further details.

across all aspects of writing. The derived weights (11.5%–13.8%) remain well-balanced, allowing each metric to contribute meaningfully to the final score. This approach preserves the strengths of averaging while significantly reducing tie cases in actual evaluation, thereby improving ranking resolution and enabling more fine-grained differentiation between models.

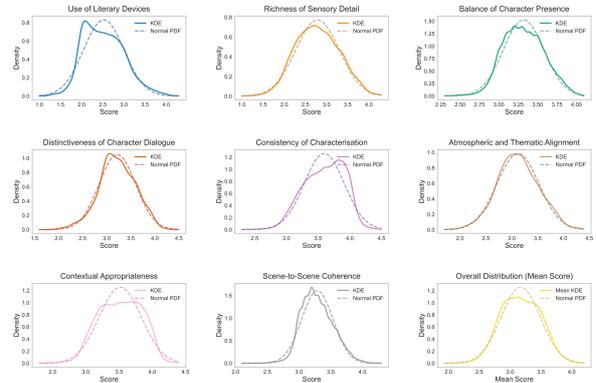


Figure 4: Distributions of Narrative Metrics and Fitted Normal Curves. Solid lines show the empirical distribution; dashed lines show the fitted normal distribution.

Distributional Characteristics. The probability density function (PDF) of each individual metric, obtained via kernel density estimation (KDE) (Parzen, 1962), is plotted against its best-fit normal distribution to evaluate shape characteristics (Figure 4). The majority of metrics (e.g., D2, D3, D4, and D6) display near-Gaussian behavior, implying smooth and well-behaved scoring distributions conducive to comparative assessment. Metrics such as D1 and D7 exhibit mild deviations from normality, with signs of skewness or slight multimodality. These deviations likely reflect the existence of content subgroups, for instance, differences in stylistic density between human and LLM-generated texts or variations in how explicitly context is embedded. Importantly, the averaged score across all dimensions yields an aggregate dis-

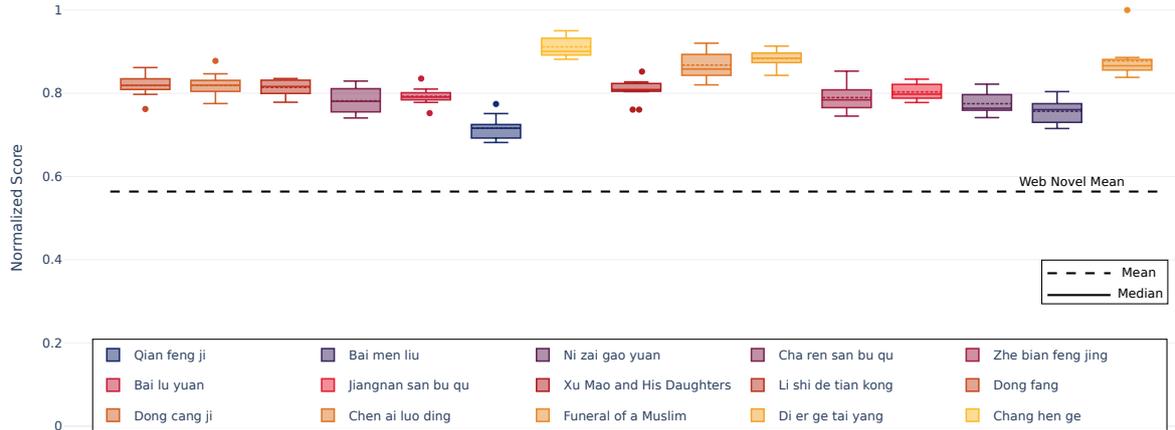


Figure 5: **Robustness Assessment of LLM-as-Judge.** Boxplot of normalized scores for selected classic Chinese novels, based on 11 times repeated evaluations using the LLM-as-judge framework. Each box shows the interquartile range (IQR) with the median and mean marked. The majority of works demonstrate consistently high scores with narrow IQRs and minimal outliers, indicating the robustness and stability of model evaluations.

tribution that closely aligns with the Gaussian, further validating the integration of the metrics into a coherent composite score.

PCA and distributional analyses confirm our metrics are well-structured, diverse, complementary, and robust, suitable for large-scale evaluation of human and LLM novels.

5.2 Classic Literature Comparison

We validated our benchmark against 25 Mao Dun Literature Prize-winning novels (first 10 chapters each). As shown in Figure 1, these classics consistently scored in the high range, confirming our framework’s ability to capture acknowledged literary merit.

This comparative analysis not only aligns with human evaluative judgments but also affirms the anticipated quality hierarchy among the three text categories under study: classic works, web novels, and LLM-generated content. The findings validate the benchmark’s sensitivity to nuanced differences in textual quality and its robustness in reflecting the relative literary value of diverse sources. Moreover, the stratification of LLM-generated outputs into three distinct quality tiers based on this distribution appears both credible and well-justified.

5.3 LLM-as-Judge

To create a scalable and fully automated evaluation pipeline while ensuring reliability, we adopted the LLM-as-Judge paradigm. Acknowledging the potential biases of a single judge model, our primary concern was to validate its consistency and robustness. Therefore, we conducted a rigorous

cross-validation experiment by having GPT-4.1, a powerful model from a different developer, evaluate and rank all 24 LLMs using the same methodology. The resulting rankings showed exceptionally strong correlation with those from our primary evaluator (Spearman ρ : 0.986, Kendall tau: 0.922), as detailed in Figure 6. This high level of agreement across different leading models substantiates the stability and reliability of our evaluation results.

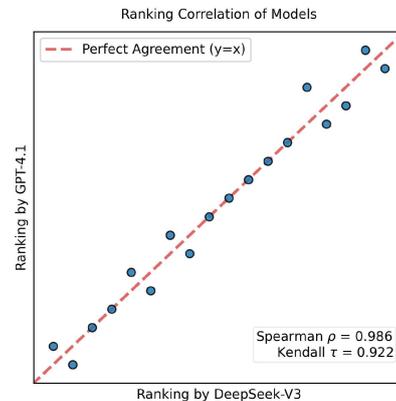


Figure 6: **Cross-Evaluator Consistency between DeepSeek-V3 and GPT-4.1 Rankings.** The scatter plot shows strong agreement in model rankings, with Spearman ρ = 0.986 and Kendall τ = 0.922, indicating high consistency between the two evaluators.

For our primary evaluation pipeline, we selected Deepseek-V3 as the judge. This decision was based on its state-of-the-art performance, particularly its nuanced understanding of the Chinese language and cultural contexts prevalent in web novels, which is crucial for high-quality assessment.

Empirical evidence also indicates that direct scoring, without explicit chain-of-thought reasoning, effectively mitigates position and context-length biases highlighted by Ye et al. (2024). This design enhances evaluation flexibility, scalability, and accuracy, especially for single-generation outputs essential to dataset curation.

To further corroborate the validity of our automated ranking against human judgment, we conducted a supplementary human study. We employed a stratified sampling approach, selecting three distinct groups of models, each containing 5 LLMs. For each group, we selected six novels (two each from low, medium, and high-quality tiers) and evaluated the synopsis-to-story output for one chapter.

Three human annotators independently assessed the outputs following the same eight-dimensional criteria as the LLM-as-Judge framework. We averaged the human scores for each model and computed rank correlations within each group. As detailed in Table 3, the results demonstrate high alignment, with an average Spearman’s ρ of 0.93 and Kendall’s τ of 0.87. Notably, Group 2 achieved perfect correlation. This confirms that the rankings produced by WebNovelBench align closely with human preference, validating the ecological validity of our automated judge.

Metric	G1	G2	G3	Mean
Spearman ρ	0.90	1.00	0.90	0.93
Kendall τ	0.80	1.00	0.80	0.87

Table 3: **Correlation between WebNovelBench automated rankings and human evaluation across three sampled model groups.** The high correlation coefficients indicate strong alignment with human preference.

Additionally, to validate robustness, we conducted a stability analysis on repeated evaluations over classic literary texts, observing high consistency with an interquartile range below 0.05 and variance around 0.001 (Figure 5).

While our current framework demonstrates reliable performance, we note that it is constrained by available resources. Future work could further enhance robustness by implementing an ensemble-judge system that aggregates scores from multiple diverse LLMs to better average out model-specific biases.

5.4 Length Analysis

To avoid introducing bias, we conducted an analysis of the length of model-generated outputs.³ The results indicate that most models closely adhered to the requested length or context window, producing outputs averaging between 800 and 1200 words. Notable exceptions include Claude 3.7 Sonnet and Gemini 2.5 Pro, which consistently generated significantly longer texts. Overall, output length remained relatively stable across models and did not emerge as a major differentiating factor under the 4096-token constraint. In future work, a scoring regularization term based on output length may be introduced to enhance robustness.

6 Conclusion

WebNovelBench addresses challenges in evaluating LLM long-form storytelling. It uses 4,000+ Chinese web novels for a synopsis-to-story task. An automated pipeline with eight LLM-judged narrative dimensions and PCA+ECDF scoring provides percentile rankings against human content. Experiments show WebNovelBench effectively distinguishes classic literature, web fiction, and LLM outputs, providing stable rankings for 24 SOTA LLMs. WebNovelBench is a valuable tool for benchmarking progress and guiding LLM development in creative narrative generation. While focused on Chinese web novels, the methodological framework of WebNovelBench, which involves leveraging a large corpus of popular works to construct a distributional benchmark and applying PCA with ECDF for ranking, is highly extendable. These principles offer a valuable template for developing similar benchmarks in other languages and literary domains, although the specific narrative dimensions would likely require careful cultural and genre-specific adaptation. Future work includes diverse judge models, genre expansion, and fostering more engaging LLM storytellers, catalyzing innovation in machine-generated narratives.

Limitations

Here we outline several limitations of our work. First, our benchmark relies exclusively on Chinese web novels as the evaluation dataset. While this provides a rich, diverse, and representative corpus for our purposes, future work should extend the benchmark to other languages and literary forms

³See Appendix C for further details.

to improve its generalizability. Second, due to resource and time constraints, our experimental scale is limited: we evaluated performance only on the subset using a single LLM-as-Judge model. Although our results demonstrate robustness, evaluating additional subsets with multiple judge models in future studies would further strengthen and validate our conclusions. Lastly, while we propose that our benchmark can directly assess data quality, we have not yet explored its broader applications. Future research will investigate how these benchmark datasets can be leveraged to enhance model performance and other downstream tasks.

References

- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Boookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.
- 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*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- William Jay Conover. 1999. *Practical nonparametric statistics*. john wiley & sons.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadao Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. [Openmeva: A benchmark for evaluating open-ended story generation metrics](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Harold Hotelling. 1933. [Analysis of a complex of statistical variables into principal components](#). *Journal of Educational Psychology*, 24:498–520.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2025. [Evaluating creative short story generation in humans and large language models](#). *Preprint*, arXiv:2411.02316.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. [SWE-bench: Can language models resolve real-world github issues?](#) In *The Twelfth International Conference on Learning Representations*.
- Jun. 2023. [webnovel_cn \(revision 745338c\)](#).
- Zdeněk Kasner and Ondřej Dušek. 2024. [Beyond traditional benchmarks: Analyzing behaviors of open llms on data-to-text generation](#). *Preprint*, arXiv:2401.10186.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.
- Ruosun Li, Teerth Patel, and Xinya Du. 2024. [Prd: Peer rank and discussion improve large language model based evaluations](#). *Preprint*, arXiv:2307.02762.
- Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie

- Huang, Yuxiao Dong, and Jie Tang. 2024. [Align-bench: Benchmarking chinese alignment of large language models](#). *Preprint*, arXiv:2311.18743.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- MAA. 2024. [American invitational mathematics examination - aime](#). In *American Invitational Mathematics Examination - AIME 2024*.
- Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. [Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures](#). *arXiv preprint arXiv:2406.06565*.
- Samuel J. Paech. 2024. [Eq-bench: An emotional intelligence benchmark for large language models](#). *Preprint*, arXiv:2312.06281.
- Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.
- Jayr Pereira, Andre Assumpcao, and Roberto Lotufo. 2024. [Check-eval: A checklist-based approach for evaluating text quality](#). *arXiv preprint arXiv:2407.14467*.
- Shanghaoran Quan, Jiayi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, and 1 others. 2025. [Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings](#). *arXiv preprint arXiv:2501.01257*.
- Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2025. [Optimization-based prompt injection attack to llm-as-a-judge](#). *Preprint*, arXiv:2403.17710.
- Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. [Are large language models capable of generating human-level narratives?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). *arXiv preprint arXiv:2406.01574*.
- Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. [The next chapter: A study of large language models in storytelling](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. [Justice or prejudice? quantifying biases in llm-as-a-judge](#). *Preprint*, arXiv:2410.02736.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. 2024. [Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark](#). *arXiv preprint arXiv:2409.02813*.
- Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. [Wider and deeper llm networks are fairer llm evaluators](#). *Preprint*, arXiv:2308.01862.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

A System Prompt

To ensure consistency and clarity in our evaluation pipeline, we standardised the prompts used across all stages of the benchmark. Figure 11 shows the system prompt used to guide story generation, while Figure 12, 13 presents the evaluation prompt employed by the LLM-as-judge to assess narrative quality. Additionally, Figure 10 illustrates the system prompt used for Story-to-Synopsis extraction, which serves as a foundation for constructing our generation dataset. These prompts were carefully designed to minimise ambiguity and ensure the reproducibility of our benchmark results.

B Principal Component Analysis Details

To determine the relative importance of each evaluation dimension, we applied Principal Component Analysis (PCA) to the score distributions across our web novel dataset. The scree plot in Figure 7 shows that the first principal component accounts for over 75% of the total variance. This indicates that while each of our eight evaluation dimensions captures a distinct and meaningful aspect of narrative quality, they also collectively reflect a strong underlying evaluative signal. The high explained variance supports the internal coherence of our metric design and justifies the use of PCA-derived weights for aggregating narrative quality scores. This balance suggests that the dimensions are complementary rather than redundant, each contributing uniquely to the overall narrative assessment.

The radar chart embedded in the same figure visualises the PCA-derived weights assigned to each of the eight narrative dimensions (D1–D8). These weights, used throughout our benchmark scoring, reflect each dimension’s contribution to the primary variance component and therefore represent their relative importance in the overall evaluation framework.

C Length Analysis Details

To support the main findings in Section 5.4, we provide a visual summary of the mean output lengths across all evaluated models in Figure 8. The green and red dashed lines represent the expected length bounds (800–1200 words). As shown, the vast majority of models generated outputs that fall within or near this range, indicating consistent adherence to the specified context length.

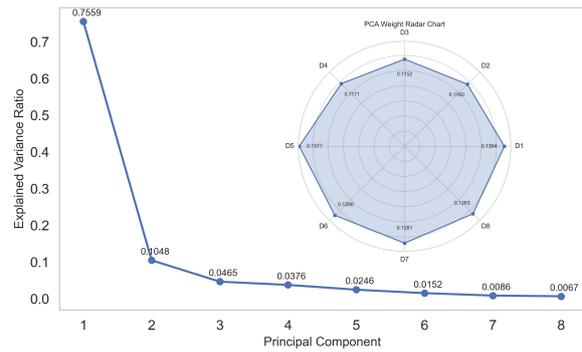


Figure 7: **PCA of Evaluation Metrics.** The bar chart shows the explained variance ratio for each principal component. The radar chart visualises the relative weights of the eight narrative dimensions used in our benchmark.

Given this overall consistency, we do not delve into detailed length-based comparisons in the main text. Notable outliers such as Claude 3.7 Sonnet (around 2,700) and Gemini 2.5 Pro produced (around 2,000) significantly longer outputs, while models like LLaMA 3.3 and GLM-4-9B-chat tended to under-generate. These deviations are exceptions rather than the norm and had limited impact on the overall evaluation results.

While length was not found to be a major differentiating factor in narrative quality, future iterations of the benchmark may consider applying soft constraints or regularisation mechanisms to penalise excessively long or short outputs.

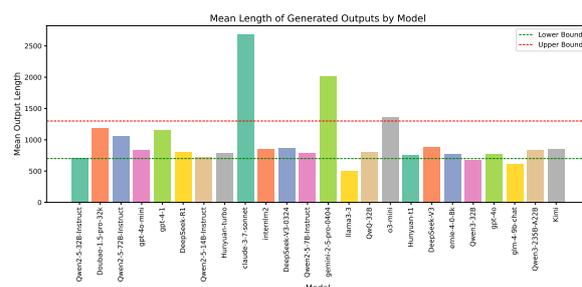


Figure 8: **The Average Length of The Outputs Generated by The Models.** The green and red dashed lines indicate the expected lower and upper bounds (800–1200 words). Most models generate outputs within or near this range, while a few, such as Claude 3.7 Sonnet and Gemini 2.5 Pro, produce significantly longer outputs. Models like LLaMA 3.3 and GLM-4-9B-chat under-generate relative to expectations. Overall, most outputs demonstrate consistent adherence to the desired context length.

D Ethical Considerations

Our WebNovelBench dataset is constructed by referencing and curating from an existing, published collection of Chinese web novels as described in Jun (2023), which is distributed under the MIT License. This ensures that our work is based on a resource that is openly accessible to the research community and permissive for academic use.

For each novel, only a limited excerpt (10 chapters) is included in our benchmark to minimise data usage while supporting comprehensive evaluation. All data is strictly used for research, evaluation, and benchmarking purposes, and is neither employed for model training nor for any commercial applications.

No personally identifiable or private information is contained in the dataset. We conducted careful reviews to exclude any sensitive or inappropriate material. Furthermore, we comply with all applicable terms of use and licensing agreements for the language models involved in our evaluation. Model outputs generated in this study are solely used for scientific analysis and reporting.

E Dataset and Classic Literature Details

E.1 Chapter Length Statistics

Our dataset comprises 40,000 chapters (10 chapters per 4,000 novels). The average chapter length is 3,139.03 characters, with a standard deviation of 1,133.80 characters, enabling robust quantitative analysis of narrative quality.

E.2 Genre Distribution

To ensure broad representativeness, the dataset spans a diverse range of genres and subgenres. The full genre distribution is shown in Figure 9 and Table 4.

E.3 The Mao Dun Literature Prize

The Mao Dun Literature Prize is one of China’s most prestigious literary awards for novels, established in 1981 and awarded by the China Writers Association. Named after the renowned novelist Mao Dun, it is typically awarded every four years, recognising outstanding achievements in contemporary Chinese fiction. Laureates of the prize are widely regarded as exemplars of literary excellence in modern Chinese literature.

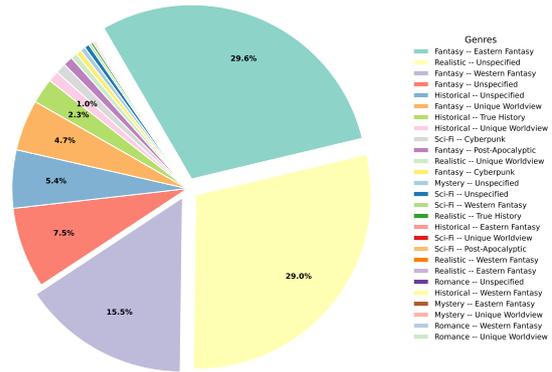


Figure 9: Genre Distribution of 4,000 Novels.

Genre (Primary – Secondary)	Count
Fantasy – Eastern Fantasy	1281
Realistic – Unspecified	1255
Fantasy – Western Fantasy	670
Fantasy – Unspecified	326
Historical – Unspecified	234
Fantasy – Unique Worldview	205
Historical – True History	101
Historical – Unique Worldview	44
Sci-Fi – Cyberpunk	42
Fantasy – Post-Apocalyptic	38
Realistic – Unique Worldview	25
Fantasy – Cyberpunk	22
Mystery – Unspecified	20
Sci-Fi – Unspecified	19
Sci-Fi – Western Fantasy	9
Realistic – True History	9
Historical – Eastern Fantasy	6
Sci-Fi – Unique Worldview	5
Sci-Fi – Post-Apocalyptic	4
Realistic – Western Fantasy	4
Realistic – Eastern Fantasy	4
Romance – Unspecified	3
Historical – Western Fantasy	2
Mystery – Eastern Fantasy	2
Mystery – Unique Worldview	1
Fantasy – Eastern Fantasy	1
Romance – Western Fantasy	1
Romance – Unique Worldview	1

Table 4: Genre distribution of the 4,000+ web novels.

E.4 Selected Mao Dun Literature Prize Novels

As high-quality human-written references, we evaluated the following 25 Mao Dun Literature Prize-winning novels (first 10 chapters each) using our benchmark pipeline:

- 战争和人 (*War and People*)
- 长恨歌 (*The Song of Everlasting Sorrow*)
- 第二个太阳 (*The Second Sun*)
- 尘埃落定 (*Red Poppies*)
- 江南三部曲 (*Jiangnan Trilogy*)
- 茶人三部曲 (*Tea Person Trilogy*)
- 穆斯林的葬礼 (*The Funeral of a Muslim*)
- 东方 (*The East*)
- 白鹿原 (*White Deer Plain*)
- 这边风景 (*The Scenery Here*)
- 你在高原 (*You Are on the Plateau*)
- 许茂和他的女儿们 (*Xu Mao and His Daughters*)
- 生命册 (*The Book of Life*)
- 李自成 (*Li Zicheng*)
- 历史的天空 (*Sky of History*)
- 白门柳 (*Baimen Willow*)
- 金瓯缺 (*The Broken Golden Basin*)
- 东藏记 (*Dongzang Chronicle*)
- 蛙 (*Frog*)
- 张居正 (*Zhang Juzheng*)
- 无字 (*No Words*)
- 牵风记 (*The Wind Walker*)
- 黄河东流去 (*The Yellow River Flows East*)
- 平凡的世界 (*Ordinary World*)
- 雪山大地 (*Snowy Land*)

These classics are widely acknowledged for their literary quality and serve as the upper-bound reference for LLM-generated narrative evaluation.

F Data Samples

Story-to-Synopsis Extraction Sample, generated by Doubao-pro-32k:

Story-to-Synopsis Extraction Input

第1498章不知好歹“萧炎哥哥，这药万归乃是四星初期的斗圣强者，实力极强，对付起来可并不容易。”……“废物一般的东西，也敢在我老师面前叫嚣，这一巴掌，就当是个小小教训！”

Story-to-Synopsis Extraction Input (English Version)

Chapter 1498: No Sense of Propriety "Brother Xiao Yan, this Yao Wangui is a four-star beginner Dou Sage. He's extremely strong and won't be easy to deal with." ... "A useless thing like you dares to boast in front of my teacher? This slap is just a small lesson for you!"

Story-to-Synopsis Extraction Output

<主要人物> 萧炎, 薰儿, 药万归, 药老, 药天</主要人物>
 <主要情节> (1) 薰儿提醒萧炎药万归是四星初期斗圣强者, 对付不易, 萧炎表示自有分寸, 只要得到并炼化净莲妖火, 灭杀药万归不在话下。…… (5) 药天愤怒让药万归杀萧炎, 萧炎灵魂分身出现, 打了药天一巴掌并教训他。</主要情节>
 <重要场景> 天空 (众人聚集在此发生冲突的地点) </重要场景>

Story-to-Synopsis Extraction Output (English Version)

<Main Characters> Xiao Yan, Xun'er, Yao Wangui, Yao Lao, Yao Tian </Main Characters>
 <Main Plots> (1) Xun'er reminds Xiao Yan that Yao Wangui is a four-star beginner Dou Sage and not easy to handle. Xiao Yan assures her he has it under control, stating that once he acquires and refines the Purifying Lotus Demon Flame, defeating Yao Wangui will be no issue. ... (5) Yao Tian, furious, orders Yao Wangui to kill Xiao Yan. Xiao Yan's soul avatar appears, slaps Yao Tian, and gives him a lesson. </Main Plots>
 <Important Scenes> Sky (the place where everyone gathers and the confrontation occurs) </Important Scenes>

LLMs synopsis-to-story generation samples, using the above extraction result as input example:

GLM-4-9B Generation

在广袤天空之下，人声鼎沸，众人的目光都被那三道身影牢牢锁定。薰儿紧紧地站在萧炎身旁，她的眼中闪烁着坚定与担忧，仿佛能感受到萧炎内心的决心。“药万归，你不是四星初期斗圣吗？对付你，我萧炎自有分寸。”萧炎的声音在空中回荡，带着一丝不容置疑的威严。……话音刚落，药天便释放出强大的灵魂力量，向萧炎发动攻击。然而，萧炎的反应却出乎所有人的意料，他轻易地捏碎了药天的灵魂力量。“你！”药天愤怒地看向药万归，“杀了他！”萧炎的灵魂分身出现在众人面前，他毫不犹豫地给了药天一巴掌：“教训你一顿。”

GLM-4-9B Generation (English Version)

Beneath the vast sky, the crowd buzzed with excitement, and all eyes were firmly fixed on three figures. Xun'er stood close beside Xiao Yan, her eyes shining with both determination and worry, as if she could sense the resolve within him.

"Yao Wangui, aren't you a four-star beginner Dou Sage? Dealing with you is well within my grasp," Xiao Yan's voice echoed through the air, laced with an undeniable authority.

...

No sooner had he finished speaking than Yao Tian unleashed a powerful surge of soul force, launching an attack on Xiao Yan. Yet, to everyone's astonishment, Xiao Yan easily shattered Yao Tian's spiritual assault.

"You!" Yao Tian glared furiously at Yao Wangui. "Kill him!"

Xiao Yan's soul avatar appeared before the crowd. Without hesitation, he slapped Yao Tian across the face. "Take that as a lesson."

Qwen3-235B-A22B Generation (English Version)

The sea of clouds around the floating island churned like boiling lava. Xiao Yan stood with his hands behind his back by the bluestone railing, his dark robe swirling to reveal the pitch-black Fallen Heart Flame at his waist.

Suddenly, Xun'er grabbed his wrist, her fingertips trembling. "The seven-coloured pill mark on that elder's forehead..." The girl's whispered warning was torn apart by the fierce wind, yet it crystallised in Xiao Yan's ear: "Yao Wangui, Third Elder of the Yao Clan. The oppressive force of a four-star Dou Sage can bend the rings of ancient trees."

...

Yao Tian's robe was shredded by the wild gale, exposing the hideous 'Sin Pill' scar on his chest. Xiao Yan's avatar descended, stepping through the air and leaving eight afterimages. Each resounding slap sent the clouds collapsing into lotus petals. "Do you know how many crimes Yao Lao took upon himself when your grandfather secretly refined the 'Bone Melting Pill' all those years ago?" As the final blow landed, a long howl of demonic fire echoed from deep within the floating island, and the phantom of the Purifying Lotus blossomed a crimson mark upon every forehead.

Qwen3-235B-A22B Generation

浮空岛的云海翻涌如沸腾的熔岩，萧炎负手立于青石栏前，玄衣翻卷间露出腰间漆黑的陨落心炎。薰儿忽然按住他的手腕，指尖微颤，"那老者眉心的七色丹纹..."少女的呢喃被罡风撕碎，却在萧炎耳畔凝成清晰的警告："药族三长老药万归，四星斗圣的威压能压弯古树年轮。".....药天的衣袍被狂风撕裂，露出心口狰狞的"罪丹"疤痕。分身降临的萧炎踩着虚空踏出八道残影，每记响亮的耳光都让云层塌陷成莲瓣："你可知你祖父当年偷炼'化骨丹'时，药老替他背下了多少罪行？"最后一掌落下时，浮空岛深处传来妖火的长吟，净莲虚影在所有人额头绽开血色印记。

GPT-4.1 Generation

天空如同被烈焰炙烤的铁幕，高远而幽深，浮云仿佛被无形之力撕裂成碎絮，斑驳地掩映着一场即将爆发的风暴。浩渺的天际下，人影浮现，气息如潮水般搅动空气，斗气波澜层层激荡。薰儿身着淡青长裙，立于萧炎身侧，眉宇间藏着一抹忧色，她低声道："萧炎哥哥，药万归乃四星初期斗圣强者，不可大意。"她声音纤细，却带着隐隐的紧迫，如同清泉流淌过心头石隙。萧炎目光沉静如潭水，嘴角勾起自信的弧度。"放心，只要夺得并炼化净莲妖火，区区药万归，自然不在话下。".....药天面色铁青，怒吼道："杀了他！我要亲手取他性命！"话音未落，虚空骤然扭曲，一道灵魂分身悄然凝聚于药天身侧。只听"啪"的一声脆响——灵魂分身挥手给了药天一个响亮耳光。天地间仿佛静止了一瞬，只剩下风声和众人的倒吸冷气。萧炎分身俯瞰着愣神的药天，语气冰冷："自以为是，只会辱骂前辈，这就是你所谓的尊严？若再挑衅，我让你永远闭嘴。"

GPT-4.1 Generation (English Version)

The sky resembled an iron curtain scorched by raging flames—lofty, deep, and ominous. Wisps of cloud, torn apart by invisible forces, drifted raggedly across the heavens, casting mottled shadows over a storm on the verge of eruption. Beneath the vast expanse, figures appeared, their auras churning the air in tidal waves of battle energy.

Xun'er, dressed in a pale blue gown, stood by Xiao Yan's side, a hint of worry hidden between her brows. She whispered, "Brother Xiao Yan, Yao Wangui is a four-star beginner Dou Sage—do not take him lightly." Her voice, slender yet urgent, flowed like a spring through cracks in the heart's stone.

Xiao Yan's gaze was calm as still water, the corners of his lips curled in quiet confidence. "Don't worry. As long as I obtain and refine the Purifying Lotus Demon Flame, Yao Wangui will be of no concern."

...

Yao Tian's face darkened to iron as he roared, "Kill him! I'll take his life myself!" Before his words faded, the void twisted, and a soul avatar silently formed at his side. With a crisp "smack," the soul avatar struck Yao Tian across the face. For a moment, heaven and earth fell silent—only the wind and the collective gasp of the crowd remained.

Xiao Yan's avatar gazed down at the stunned Yao Tian, his voice icy: "Arrogant fool, all you know is to insult your elders. Is this what you call dignity? Provoke me again, and I'll make sure you're silenced forever."

Deepseek-V3 Evaluation (English Version)

<Extraction Results> <Main Characters and Their Dialogues> 1. Xiao Yan – "Don't worry. As long as I obtain and refine the Purifying Lotus Demon Flame, Yao Wangui will be of no concern." ... </Main Characters and Their Dialogues>
<Main Plots> 1. Scene Setting: The sky is scorched by blazing flames, a storm is about to erupt. Xiao Yan and Xun'er converse in a tense atmosphere, with Xun'er reminding Xiao Yan of Yao Wangui's strength. ... </Main Plots>
<Important Scenes> 1. Confrontation Under the Sky: Beneath the fiery sky and swirling sea of clouds, Xiao Yan and Yao Wangui stand opposed to each other. ... </Important Scenes>
</Extraction Results>

<Scoring Results> <Use of Literary Devices Score>4</Use of Literary Devices Score>
<Richness of Sensory Detail Score>4</Richness of Sensory Detail Score>
<Balance of Character Presence Score>3</Balance of Character Presence Score>
<Distinctiveness of Character Dialogue Score>4</Distinctiveness of Character Dialogue Score>
<Consistency of Characterisation Score>4</Consistency of Characterisation Score>
<Atmospheric and Thematic Alignment Score>4</Atmospheric and Thematic Alignment Score>
<Contextual Appropriateness Score>4</Contextual Appropriateness Score>
<Scene-to-Scene Coherence Score>4</Scene-to-Scene Coherence Score>
</Scoring Results>

LLM-as-Judge evaluation sample, using Deepseek-V3 as a critic:

Deepseek-V3 Evaluation

<提取结果> <主要人物及其对白> 1. **萧炎** - "放心，只要夺得并炼化净莲妖火，区区药万归，自然不在话下。"... <主要情节> 1. **场景铺垫**：天空被烈焰炙烤，风暴即将爆发，萧炎与薰儿在紧张的氛围中对话，薰儿提醒萧炎药万归的强大。... </主要情节>
<重要场景> 1. **天空下的对峙场景**：天空被烈焰炙烤，云海翻滚，萧炎与药万归双方针锋相对。... </重要场景> </提取结果>

<评分结果> <修辞手法评分>4</修辞手法评分> <感官描述丰富度评分>4</感官描述丰富度评分> <角色平衡度评分>3</角色平衡度评分> <角色对白独特性评分>4</角色对白独特性评分> <角色一致性评分>4</角色一致性评分> <意境匹配度评分>4</意境匹配度评分> <语境适配度评分>4</语境适配度评分> <跨场景衔接度评分>4</跨场景衔接度评分> </评分结果>

Story-to-Synopsis Extraction System Prompt

你的任务是从给定的小说片段中提取主要人物、情节和场景等信息，以便生成该小说的知识图谱和百科信息。请仔细阅读以下小说文本：

<小说> {text} </小说>

在提取信息时，请遵循以下步骤：

1. 仔细通读整个小说文本。
2. 识别出主要人物，主要人物是在小说中起到关键作用、有较多情节围绕的角色。
3. 梳理主要情节，主要情节是推动故事发展的核心事件和关键转折。
4. 确定重要场景，重要场景是故事发生的关键地点和环境。
5. 检查提取的信息是否准确和完整。

请在<提取结果>标签内输出你的提取结果，格式如下：

<主要人物> [列出主要人物的名字，并且用逗号分隔] </主要人物>

<主要情节> [详细描述主要情节，按照事件发展顺序。如(1)...(2)...，情节之间使用换行符分隔] </主要情节>

<重要场景> [列出重要场景的名称，并且用逗号分隔] </重要场景>

请确保提取的信息丰富、全面且准确。

Your task is to extract key information—such as main characters, major plot points, and important settings—from the given novel excerpt. This information will be used to construct a knowledge graph and encyclopaedic entry for the novel. Please read the following text carefully:

<Novel> {text} </Novel>

When extracting the information, follow these steps:

1. Carefully read through the entire novel excerpt.
2. Identify the main characters, i.e., the characters who play a central role and around whom significant parts of the plot revolve.
3. Outline the major plot points, which refer to the core events and pivotal turns that drive the story forward.
4. Determine the important settings, i.e., the key locations and environments where significant story developments occur.
5. Check the extracted information for accuracy and completeness.

Output your extracted results within the <Extraction> tags using the following format:

<Main Characters> [List the names of the main characters, separated by commas] </Main Characters>

<Main Plots> [Describe the main plot points in detail, following the chronological order of events. Use line breaks between different events, e.g., (1)...(2)...] </Main Plots>

<Important Scenes> [List the names of the important settings, separated by commas] </Important Scenes>

Please ensure that the extracted information is rich, comprehensive, and accurate.

Figure 10: System Prompt Used for Story-to-Synopsis Extraction

Generation System Prompt

你是一个中文小说作家，你需要根据用户提供的信息进行扩写创作，创作需要满足下列条件：

1. 用户会用下面的格式给出长篇小说的主要人物、主要情节和主要场景，请仔细阅读用户提供的信息：

<主要人物>[主要人物的名字]</主要人物>

<主要情节>[主要情节，按照事件发展顺序]</主要情节>

<重要场景>[重要场景的名称]</重要场景>

2. 评论家会根据下列标准打分：

根据复杂修辞（隐喻/象征/悖论）的数量与质量提炼度，对修辞手法评分

根据文本中的视觉、听觉、嗅觉等描写数量，对感官描述丰富度评分

统计每个角色在生成内容中的出现频率、对话占比、心理描写和评估人物描述的平衡度，对角色平衡度评分

查看角色台词是否能反映本身个性，遮住名字后是否有区分度，对角色对白独特性评分

分析角色语言、动作是否匹配其身份和背景，对角色一致性评分

通过情感色谱分析，检查场景描写是否服务于整体氛围，对意境匹配度评分

通过分析环境细节是否适应时代/地域背景，对语境适配度评分

评估生成内容是否自然衔接不同场景从而避免场景割裂，对跨场景衔接度评分

3. 只需按照指定格式返回生成的小说：

<text>你生成的小说内容</text>

You are a Chinese fiction writer. Your task is to expand and create a narrative based on the information provided by the user. Your writing must adhere to the following guidelines:

1. The user will provide the key information for a long-form novel using the following format. Please read the information carefully:

<Main Characters>[Names of the main characters]</Main Characters>

<Main Plots>[Main plot points in chronological order]</Main Plots>

<Important Scenes>[Names of important scenes or locations]</Important Scenes>

2. A critic will evaluate your writing according to the following criteria:

Use of Literary Devices: Scored based on the quantity and refinement of complex rhetorical devices such as metaphor, symbolism, and paradox.

Richness of Sensory Detail: Scored based on the frequency of visual, auditory, olfactory, and other sensory descriptions.

Balance of Character Presence: Scored based on the frequency of each character's appearance, proportion of dialogue, psychological depiction, and overall balance of character portrayal.

Distinctiveness of Character Dialogue: Scored based on whether each character's dialogue reflects individual personality and remains distinguishable even if names are hidden.

Consistency of Characterisation: Scored based on whether the characters' language and actions align with their identities and backgrounds.

Atmospheric and Thematic Alignment: Scored based on whether scene descriptions support the emotional tone and thematic coherence of the narrative.

Contextual Appropriateness: Scored based on whether the setting details are appropriate for the time period and regional background.

Scene-to-Scene Coherence: Scored based on whether the narrative transitions naturally between scenes, avoiding abrupt or disjointed shifts.

3. Return only the generated novel in the following format:

<text>Your generated story content</text>

Figure 11: System Prompt Used for Generation

Evaluation System Prompt

你的任务是根据给定的指标规则对小说进行评分(1-5)。请仔细阅读以下小说文本: <小说> {chapter} </小说> 在提取信息时, 请遵循以下步骤:

1. 仔细通读整个小说文本
2. 识别出主要人物, 主要人物是在小说中起到关键作用、有较多情节围绕的角色
3. 梳理主要情节, 主要情节是推动故事发展的核心事件和关键转折
4. 确定重要场景, 重要场景是故事发生的关键地点和环境
5. 检查提取的信息是否准确和完整

请在<提取结果>标签内输出你的提取结果, 格式如下:

<主要人物及其对白>[列出主要人物的名字和对白]</主要人物及其对白>
<主要情节>[详细描述主要情节, 按照事件发展顺序]</主要情节>
<重要场景>[列出重要场景的名称]</重要场景>

请确保提取的信息丰富、全面且准确。在评分时, 请遵循以下步骤:

1. 根据复杂修辞(隐喻/象征/悖论)的数量与质量提炼度, 给出修辞手法评分
2. 根据文本中的视觉、听觉、嗅觉等描写数量, 给出感官描述丰富度评分
3. 统计每个角色在生成内容中的出现频率、对话占比、心理描写和评估人物描述的平衡度, 给出角色平衡度评分
4. 查看角色台词是否能反映本身个性, 遮住名字后是否有区分度, 给出角色对白独特性评分
5. 分析角色语言、动作是否匹配其身份和背景, 给出角色一致性评分
6. 通过情感色谱分析, 检查场景描写是否服务于整体氛围, 给出意境匹配度评分
7. 通过分析环境细节是否适应时代/地域背景, 给出语境适配度评分
8. 评估生成内容是否自然衔接不同场景从而避免场景割裂, 给出跨场景衔接度评分。

请在<评分结果>标签内输出你的评分结果, 格式如下:

<修辞手法评分>1</修辞手法评分> <感官描述丰富度评分>1</感官描述丰富度评分>
<角色平衡度评分>1</角色平衡度评分> <角色对白独特性评分>1</角色对白独特性评分>
<角色一致性评分>1</角色一致性评分> <意境匹配度评分>1</意境匹配度评分>
<语境适配度评分>1</语境适配度评分> <跨场景衔接度评分>1</跨场景衔接度评分>

请确保评分全面且准确符合要求。

Figure 12: System Prompt Used for Evaluation (Chinese Version)

Evaluation System Prompt

Your task is to rate a novel based on the given criteria (1–5). Please read the following novel text carefully: <Novel> {chapter} </Novel>. When extracting information, please follow these steps:

1. Carefully read through the entire novel text.
2. Identify the main characters, i.e., those who play key roles and around whom major events unfold.
3. Summarize the main plot, which refers to the core events and key turning points driving the story forward.
4. Determine the important scenes, i.e., key locations and settings where major story events take place.
5. Check whether the extracted information is accurate and complete.

Please output your extraction results within the <Extraction Results> tag, using the following format:
<Main Characters and Their Dialogues>[List the names of main characters and their dialogues]</Main Characters and Their Dialogues>
<Main Plots>[Describe the main plot in detail, following the sequence of events]</Main Plots>
<Important Scenes>[List the names of important scenes]</Important Scenes>

Please ensure the extracted information is rich, comprehensive, and accurate.

When scoring, please follow these steps:

1. Based on the quantity and quality of complex rhetorical devices (metaphor, symbolism, paradox), provide a score for Use of Literary Devices.
2. Based on the frequency of visual, auditory, olfactory, and other sensory descriptions, provide a score for Richness of Sensory Detail.
3. Count the appearance frequency, dialogue proportion, and psychological descriptions of each character and assess the balance of character descriptions; then provide a score for Balance of Character Presence.
4. Check whether character dialogues reflect distinct personalities and whether they remain distinguishable even without names; provide a score for Distinctiveness of Character Dialogue.
5. Analyse whether characters' language and actions match their identities and backgrounds; provide a score for Consistency of Characterisation.
6. Through emotional spectrum analysis, check if scene descriptions serve the overall atmosphere and themes; provide a score for Atmospheric and Thematic Alignment.
7. Analyse whether environmental details match the historical, locational, or cultural background; provide a score for Contextual Appropriateness.
8. Evaluate whether the generated content transitions naturally between scenes and avoids abrupt scene shifts; provide a score for Scene-to-Scene Coherence.

Please output your scoring results within the <Scoring Results> tag, using the following format:
<Use of Literary Devices Score>1</Use of Literary Devices Score> <Richness of Sensory Detail Score>1</Richness of Sensory Detail Score> <Balance of Character Presence Score>1</Balance of Character Presence Score> <Distinctiveness of Character Dialogue Score>1</Distinctiveness of Character Dialogue Score> <Consistency of Characterisation Score>1</Consistency of Characterisation Score> <Atmospheric and Thematic Alignment Score>1</Atmospheric and Thematic Alignment Score> <Contextual Appropriateness Score>1</Contextual Appropriateness Score> <Scene-to-Scene Coherence Score>1</Scene-to-Scene Coherence Score>

Please ensure your scoring is comprehensive and fully meets the requirements.

Figure 13: System Prompt Used for Evaluation (English Version)