

# ESSAYBENCH: Evaluating Large Language Models in Multi-Genre Chinese Essay Writing

Anonymous ACL submission

## Abstract

Chinese essay writing and its evaluation are critical in educational contexts, yet the capabilities of Large Language Models (LLMs) in this domain remain largely underexplored. Existing benchmarks often rely on coarse-grained text quality metrics, largely overlooking the structural and rhetorical complexities of Chinese essays, particularly across diverse genres. To address this gap, we propose ESSAYBENCH, a multi-genre benchmark specifically designed for Chinese essay writing across four major genres: *Argumentative*, *Narrative*, *Descriptive*, and *Expository*. We curate and refine a total of 728 real-world prompts to ensure authenticity and meticulously categorize them into the *Open-Ended* and *Constrained* sets to capture diverse writing scenarios. To reliably evaluate generated essays, we develop a fine-grained, genre-specific scoring framework that hierarchically aggregates scores. We further validate our evaluation protocol through a comprehensive human agreement study. Finally, we benchmark 15 large-sized LLMs, analyzing their strengths and limitations across genres and instruction types. With ESSAYBENCH<sup>1</sup>, we aim to advance LLM-based Chinese essay evaluation and inspire future research on improving essay generation in educational settings.

## 1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Team., 2023; Team, 2023) have achieved impressive results in text generation, with growing applications in education, including automated writing support and feedback (Gao et al., 2024). Among these tasks, essay writing plays a central role in language learning and assessment (Venkatraman et al., 2025; Miura et al., 2025; Wen et al., 2025). However, the lack of robust evaluation frameworks for generated essays

<sup>1</sup><https://anonymous.4open.science/r/EssayBench-2B14>

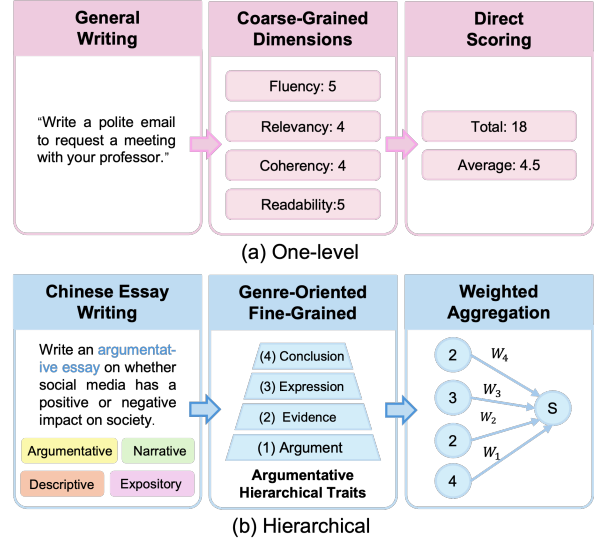


Figure 1: Comparison between coarse-grained evaluation methods (a) and our fine-grained and genre-oriented framework for ESSAYBENCH (b).

limits the development and deployment of LLMs in real-world educational settings (Kim et al., 2025).

As shown in Figure 1, current predominant LLM-as-a-judge strategies (Zheng et al., 2023; Li et al., 2025) for assessing texts mainly fall into two paradigms. One relies on meta-evaluation to judge response quality in terms of fluency, relevancy, coherency, readability, and hallucination (Liu et al., 2023; Chen et al., 2023; Hashemi et al., 2024; Fu et al., 2024), while the other employs downstream tasks (e.g., question-answering) as proxies for measuring informational richness and accuracy (Tan et al., 2024; Que et al., 2024; Lee et al., 2025). Although these methods yield valuable insights, they exhibit two fundamental weaknesses. First, the evaluation criteria remain overly coarse-grained, i.e., current LLMs consistently achieve high scores in fluency, relevancy, and coherency (Gu et al., 2025), making it difficult to reveal failure modes or specific weaknesses. Second, existing evaluation methods fail to capture the unique characteristics

| Benchmark                          | Num.       | Dataset Composition  |           |             | Evaluation Method |             |                 |
|------------------------------------|------------|----------------------|-----------|-------------|-------------------|-------------|-----------------|
|                                    |            | Domain               | Language  | Constraints | LLM               | F.G. Traits | Scoring         |
| C-Eval (Huang et al., 2023)        | 13,948     | General Tasks        | ZH        | ✗           | ✗                 | ✗           | -               |
| AlignBench (Liu et al., 2024b)     | 683        | General Tasks        | ZH        | ✗           | ✓                 | ✗           | Direct          |
| LongBench-Write (Bai et al., 2024) | 120        | General Writing      | ZH&EN     | ✗           | ✓                 | ✗           | Direct          |
| HelloBench (Que et al., 2024)      | 647        | General Tasks        | EN        | ✗           | ✓                 | ✗           | Weighted        |
| WritingBench (Wu et al., 2025)     | 1239       | General Writing      | ZH&EN     | ✓           | ✓                 | ✗           | Direct          |
| <b>ESSAYBENCH (Ours)</b>           | <b>728</b> | <b>Essay Writing</b> | <b>ZH</b> | <b>✓</b>    | <b>✓</b>          | <b>✓</b>    | <b>Weighted</b> |

Table 1: Comparison of ESSAYBENCH with other benchmarks in terms of size, composition, and evaluations.

of essays like logographic characters, complex constructions, and rhetorical traditions, although several benchmarks like *AlignBench* (Liu et al., 2024b) and *WritingBench* (Wu et al., 2025) have turned attention to evaluating general Chinese writing.

Moreover, Chinese literary and expository practices differ markedly across genres: argumentative essays demand logical structure and persuasive rhetoric (Wachsmuth et al., 2017); narratives require compelling plot development and character voice (Somasundaran et al., 2018); descriptive writings emphasize vivid imagery and sensory detail (McCarthy, 1998); and expository passages call for clarity, organization and factual precision (Balepur et al., 2023). However, existing evaluation frameworks largely overlook genre-specific criteria, limiting their ability to reflect the nuanced demands of Chinese essay writing. This motivates our central research question as follows:

*How can we reliably assess the quality of LLM-generated Chinese essays in ways that truly reflect genre-specific conventions?*

In this paper, we introduce ESSAYBENCH, a **fine-grained** and **multi-genre** benchmark tailored for Chinese essay writing. ESSAYBENCH covers four widely recognized genres in Chinese education: *Argumentative*, *Narrative*, *Descriptive*, and *Expository* writing. To ensure alignment with real-world educational scenarios, we collect and manually refine a total of 728 essay prompts. These prompts are further categorized into two types based on their instruction style: *Open-Ended* and *Constrained*, allowing us to examine LLMs’ behavior under different writing conditions, as introduced in Section 2. Additionally, to overcome the limitations of existing evaluation methods for Chinese essay writing, we propose a fine-grained and genre-oriented evaluation framework, as shown in Figure 1. We define

multiple evaluation traits with hierarchical dependencies based on their complexity, ranging from basic to advanced requirements for each essay genre. For each trait, we design targeted sub-questions that reflect genre-specific writing expectations at different levels. To account for the hierarchical nature of these traits, we further introduce a dependency-weighted score aggregation strategy to better capture the writing quality, as introduced in Section 3.

We conduct two key experiments to validate the proposed framework. First, to assess its effectiveness and robustness, we perform a comprehensive human agreement study and a quality sensitivity analysis. The results demonstrate that our evaluation protocol aligns closely with human judgments, especially when applied to more advanced LLMs. More importantly, it significantly improves the ability to distinguish essay quality across high-, medium-, and low-level responses (See Section 4). Second, we benchmark 15 large-scale LLMs on the Chinese essay writing using our framework, offering detailed comparisons of their capabilities in writing Chinese essays (See Section 5).

In Table 1, we highlight the key differences between our work and existing approaches. In summary, our main contributions are as follows:

- We present ESSAYBENCH, a multi-genre benchmark tailored for Chinese essay writing, covering *Argumentative*, *Narrative*, *Descriptive*, and *Expository* genres. The benchmark is curated from real-world scenarios and is suitable for practical use in educational applications.
- We propose an effective and robust evaluation protocol for Chinese essays that aligns closely with human judgments and greatly improves the ability to distinguish essays of varying quality.
- We benchmark 15 widely used large-scale LLMs to evaluate their strengths and weaknesses in Chinese essay writing, and highlight areas for future improvement.

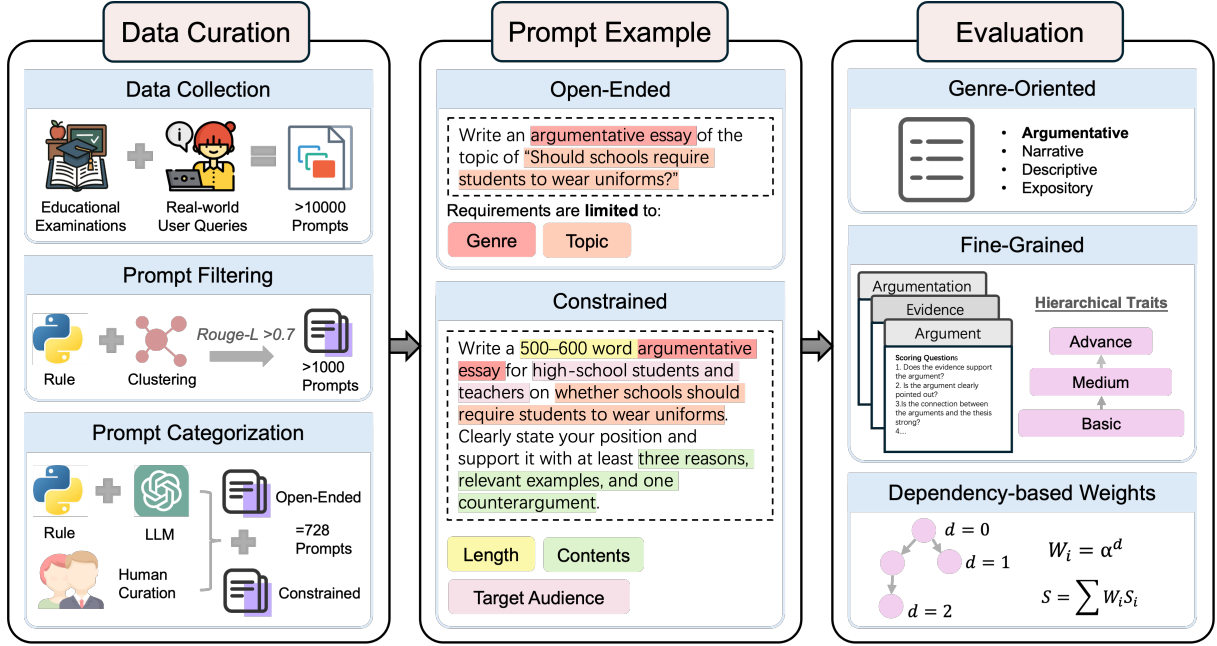


Figure 2: Overview of the ESSAYBENCH dataset curation, representative prompts, and the evaluation framework.

## 2 ESSAYBENCH Dataset

ESSAYBENCH originally contributes to developing the datasets specifically tailored for Chinese Essay Writing. While prior benchmarks (Wu et al., 2025) have largely provided queries on creative writing tasks in general domains, they do not adequately capture the structure, purpose, and constraints of formal essays, particularly within educational and academic contexts. To effectively benchmark the essay generation abilities, ESSAYBENCH introduces a comprehensive set of essay prompts that span four major and widely recognized genres in Chinese writing instruction (Chadbourne, 1983): *Argumentative*, *Narrative*, *Descriptive*, and *Expository* essays, which cover the majority of Chinese prose compositions in educational settings. Furthermore, to support comprehensive evaluation, we categorize prompts into two distinct sets based on their multiple constraints. In this section, we describe the essay prompt construction process in detail, including data collection and quality control, and the two-phase query categorization procedures.

### 2.1 Prompt Collection

As shown in Figure 2, to reflect real-world usage and align with educational settings, we choose to collect prompts from practical and authentic resources. Specifically, we collect data from two primary resources, namely 1) real-world user queries obtained through online chatbot interactions, re-

flecting informal and user-generated prompts in tutoring or self-study contexts. 2) educational examination materials, including official Chinese essay prompts, represent standardized and curriculum-aligned writing tasks used in formal assessments.

### 2.2 Prompt Filtering

Building on the collected prompts from these two sources, we construct a broad candidate pool containing several thousand raw entries. To ensure the quality and representativeness of the datasets, we implement a multi-step filtering pipeline. First, we apply heuristic-based rules to remove irrelevant and low-quality prompts. We then employ clustering methods (e.g., *K*-means (Hastie et al., 2009) with elbow method) to detect and eliminate duplicate or near-duplicate entries. To further enhance prompt diversity, we compute pairwise ROUGE-L scores between prompts and retain only those pairs with a similarity score below 0.7 (Jiang et al., 2024). In this stage, we get over 1000 relative prompts covering essay writing.

### 2.3 Prompt Categorization

To better evaluate how LLMs perform at different levels of writing difficulty, we divide the prompts into two subsets: *Open-Ended* and *Constrained*. To support this categorization, we first analyze the collected prompts and define five key factors that influence writing complexity and reader expectations: (1) Genre Specification. Each prompt clearly

defines the target genres, including argumentative, narrative, descriptive, or expository, which guide the structural and rhetorical style of the expected response. (2) Topic Specification. Prompts indicate a central topic that the essay should focus on. For example, an argumentative prompt may require elaborating on a specific viewpoint, while an expository prompt asks for the introduction of a particular object or concept. (3) Content Constraints. These constraints specify required elements or themes within the essay. For instance, an argumentative prompt may instruct to include a historical example. (4) Length Requirements. Some prompts include explicit word or paragraph limits, adding structural constraints that impact the planning and execution of essay writing. (5) Target Audience. Prompts may specify the intended readership, such as middle school students or readers of a children’s literary magazine, influencing the tone, vocabulary, and complexity of the writing. In particular, each prompt explicitly specifies both the writing genre and the topic, ensuring clarity in the contents.

Building on the above-mentioned factors, we categorize each prompt into the either set based on the presence of constraints beyond the genre and topic, i.e., prompts in the *Open-Ended* set include only basic instructions (genre and topic), while those in the *Constrained* set contain additional requirements, such as length, content focus, or stylistic constraints. To perform this classification, we adopt a hybrid approach that combines rule-based parsing with LLM-based analysis. Specifically, rule-based methods are applied to identify explicit length constraints, while LLMs are used to detect more nuanced elements, such as topic- and content-related restrictions. All prompts are then manually reviewed by the authors to correct any misclassifications and ensure the overall consistency and quality of the dataset. After manual curation, we totally get 728 prompts that capture a wide range of topics, genres, and instructional objectives in real-world Chinese writing tasks. The statistics of the dataset are shown in Figure 3.

### 3 ESSAYBENCH Evaluation Protocol

In this section, we present the design of our evaluation framework for assessing Chinese essays. Due to the open-ended and reference-free nature of essay writing, we adopt the LLM-as-a-judge paradigm (Chen et al., 2024; Gu et al., 2025) as our evaluation approach. Despite its growing popular-

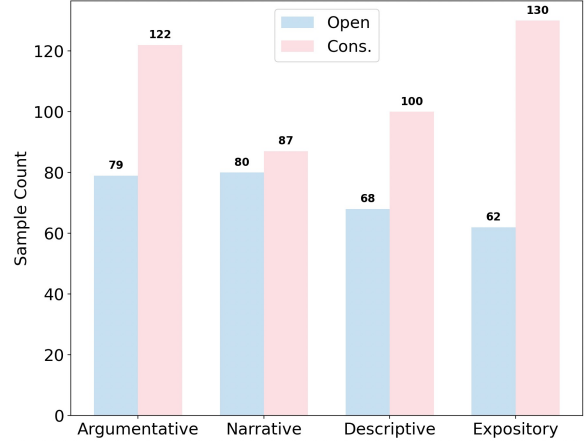


Figure 3: Dataset Statistics. Note that **Open** denotes Open-Ended sets, **Cons.** refers to Constrained sets.

ity, existing protocols for evaluating essay generation remain insufficient, particularly in the context of Chinese writing, which involves distinct linguistic features and culturally rooted rhetorical conventions (Liu et al., 2024a). To meet these evaluation needs, we propose a genre-oriented, fine-grained, and dependency-aware evaluation framework for ESSAYBENCH, capturing structural, linguistic, and hierarchical aspects of Chinese essays.

**Genre-Oriented Evaluation.** In practical essay evaluation, the criteria for assessing quality often vary across genres, as different genres emphasize distinct aspects of writing based on their inherent characteristics. As a result, our framework is adapted to different genres accordingly. Following the principal rubrics outlined in (Blanchard et al., 2013; Hamner et al., 2012), we refine and construct genre-specific evaluation traits that align with Chinese writing conventions. Specifically, we define six genre-specific evaluation dimensions, each designed with expectations that range from basic to advanced requirements, as detailed in Appendix A. This setup allows our framework to effectively capture the distinctive features of different essay types and evaluate essays across varying quality levels.

**Fine-Grained Evaluation.** Existing methods to evaluate individual dimensions typically rely on direct scoring or binary questions (Que et al., 2024), but these approaches are often limited by their coarse granularity (Kim et al., 2025). Inspired by the multi-trait evaluation design (Lee et al., 2024), we introduce a set of sub-questions ( $q_i$ ) under each evaluation dimension to enable more nuanced assessments (See Appendix B). We adopt the Chain-of-Thought (CoT) (Wei et al., 2023) prompting



| Methods            | Overall      |              | Argumentative |              | Narrative    |              | Descriptive  |              | Expository   |              |
|--------------------|--------------|--------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                    | $\rho$       | $\tau$       | $\rho$        | $\tau$       | $\rho$       | $\tau$       | $\tau$       | $\rho$       | $\rho$       | $\tau$       |
| <b>DeepSeek-V3</b> |              |              |               |              |              |              |              |              |              |              |
| Align-Score        | <b>0.674</b> | <b>0.599</b> | <b>0.744</b>  | <b>0.674</b> | 0.635        | <b>0.559</b> | 0.656        | 0.580        | <b>0.656</b> | <b>0.578</b> |
| Ours w/o WT.       | 0.646        | 0.529        | 0.701         | 0.576        | 0.596        | 0.464        | 0.778        | 0.672        | 0.509        | 0.405        |
| Ours               | 0.667        | 0.549        | 0.670         | 0.546        | <b>0.648</b> | 0.518        | <b>0.796</b> | <b>0.676</b> | 0.554        | 0.458        |
| <b>GPT-4o</b>      |              |              |               |              |              |              |              |              |              |              |
| Align-Score        | 0.628        | 0.546        | 0.587         | 0.516        | 0.582        | 0.514        | 0.642        | 0.563        | 0.700        | 0.594        |
| Ours w/o WT.       | 0.706        | 0.596        | 0.747         | 0.643        | 0.747        | 0.645        | 0.688        | 0.576        | 0.643        | 0.520        |
| Ours               | <b>0.733</b> | <b>0.627</b> | <b>0.754</b>  | <b>0.662</b> | <b>0.773</b> | <b>0.658</b> | <b>0.700</b> | <b>0.594</b> | <b>0.707</b> | <b>0.601</b> |
| <b>DeepSeek-R1</b> |              |              |               |              |              |              |              |              |              |              |
| Align-Score        | 0.749        | 0.667        | 0.745         | 0.667        | 0.764        | 0.695        | 0.709        | 0.617        | 0.778        | 0.686        |
| Ours w/o WT.       | 0.803        | 0.685        | 0.789         | 0.648        | 0.830        | 0.719        | 0.817        | 0.702        | 0.785        | 0.669        |
| Ours               | <b>0.816</b> | <b>0.704</b> | <b>0.795</b>  | <b>0.673</b> | <b>0.838</b> | <b>0.724</b> | <b>0.839</b> | <b>0.731</b> | <b>0.791</b> | <b>0.690</b> |

Table 2: Comparison of human agreement evaluation across different scoring methods on sampled data.  $\rho$  refers to Spearman’s  $\rho$ ,  $\tau$  denotes the Kendall’s  $\tau$ , while WT. represents the dependency-based weights.

technique to guide LLMs in analyzing responses and identifying linguistic evidence in support of the assigned scores. The final score for  $t$ -th dimension  $S_t$  is computed by aggregating the scores of the corresponding sub-questions as  $S_t = \sum_i q_i$ .

**Dependency-Aware Evaluation.** Many existing works determine the overall response quality by simply summing or averaging the scores of individual dimensions. However, based on our observations and preliminary experiments, we find that hierarchical traits contribute unequally, and treating them independently often fails to capture nuanced features in high-quality essays. To address this limitation, we propose a dependency-aware scoring approach inspired by (Saaty, 1980; Žižović and Pamucar, 2019), which assigns weights to each trait based on its position in the evaluation hierarchy. For example, traits at the base level are assigned a depth ( $d$ ) of 0, while mid-level traits have a depth of 1. The weights ( $W_t$ ) are computed using Equation 1, with the hyperparameter  $\alpha$  controlling the importance of basic and advanced levels. The final score is a weighted sum of all trait scores.

$$W_t = \alpha^d. \quad (1)$$

## 4 Human Agreement Evaluation

To validate the effectiveness of our evaluation protocol, we conduct a comprehensive human agreement study in Chinese essays. Specifically, the study focuses on two aspects: **1) Ranking Agreement**, which measures how closely the rankings produced by our evaluation framework align with human judgments; and **2) Sensitivity Evaluation**,

which assesses the robustness of the framework in distinguishing essays of varying quality.

### 4.1 Experiment Setup

**Datasets.** We randomly sample 80 prompts across different categories, selecting ten prompts per genre per difficulty level. For each prompt, we evaluate essays generated by seven language models, including both open- and closed-source models: LLaMA-3.1-70B-Instruct (Meta, 2024), Qwen-2.5-72B-Instruct (Qwen, 2025), GPT-3.5-turbo (Brown et al., 2020), Claude-3.5-Sonnet (Ouyang et al., 2022), Deepseek-v3 (DeepSeek-AI, 2025b), Grok-3 (xAI, 2025), and GPT-4o (OpenAI, 2024). We then recruit 14 professional annotators with rich backgrounds in Chinese linguistics to assess the generated essays. To ensure reliability and consistency, we adopt a pairwise comparison annotation method (Wen et al., 2024), assigning each essay pair to three annotators. In total, the annotation process results in 5,040 labeled data. Finally, in Table 4, a Fleiss’ Kappa Agreement (Fleiss, 1971) is used to measure the agreements among three evaluators to ensure the annotation quality.

**Baselines.** As the first to propose an evaluation protocol specifically tailored for Chinese essay writing, we compare our method against two baseline approaches: (1) **Align Scoring** (Liu et al., 2024b) from AlignBench, which evaluates general Chinese writing quality, particularly, we slightly modify it to evaluate reference-free essays; and (2) **Ours w/o Weights**, which applies the same evaluation rubrics as our method but without dependency-based weighting.

**Judges.** To verify how well the proposed eval-

| Method              | DeepSeek-V3                     |                             | GPT-4o                          |                             | DeepSeek-R1                     |                             |
|---------------------|---------------------------------|-----------------------------|---------------------------------|-----------------------------|---------------------------------|-----------------------------|
|                     | $U_p \uparrow$                  | $MD_{std} \uparrow$         | $U_p \uparrow$                  | $MD_{std} \uparrow$         | $U_p \uparrow$                  | $MD_{std} \uparrow$         |
| Align-Score<br>Ours | <i>high&amp;medium</i>          |                             |                                 |                             |                                 |                             |
|                     | 0.56 <sub>&lt;0.05</sub>        | 0.17 <sub>0.62</sub>        | 0.56 <sub>=0.14</sub>           | 0.25 <sub>0.84</sub>        | 0.64 <sub>=1.43</sub>           | 0.42 <sub>0.77</sub>        |
|                     | <b>0.57</b> <sub>&lt;0.10</sub> | <b>0.24</b> <sub>0.74</sub> | <b>0.66</b> <sub>&lt;0.05</sub> | <b>0.45</b> <sub>1.05</sub> | <b>0.79</b> <sub>&lt;0.05</sub> | <b>0.70</b> <sub>0.79</sub> |
| Align-Score<br>Ours | <i>medium&amp;low</i>           |                             |                                 |                             |                                 |                             |
|                     | <b>0.90</b> <sub>&lt;0.05</sub> | 1.42 <sub>1.26</sub>        | 0.87 <sub>&lt;0.05</sub>        | 2.16 <sub>1.48</sub>        | 0.93 <sub>&lt;0.05</sub>        | 1.98 <sub>1.05</sub>        |
|                     | 0.78 <sub>&lt;0.05</sub>        | <b>1.96</b> <sub>1.41</sub> | <b>0.93</b> <sub>&lt;0.05</sub> | <b>2.46</b> <sub>1.42</sub> | <b>0.97</b> <sub>&lt;0.05</sub> | <b>2.79</b> <sub>1.32</sub> |
| Align-Score<br>Ours | <i>high&amp;low</i>             |                             |                                 |                             |                                 |                             |
|                     | <b>0.92</b> <sub>&lt;0.05</sub> | 1.66 <sub>1.41</sub>        | 0.93 <sub>&lt;0.05</sub>        | 2.41 <sub>1.35</sub>        | 0.97 <sub>&lt;0.05</sub>        | 2.41 <sub>1.06</sub>        |
|                     | 0.82 <sub>&lt;0.05</sub>        | <b>2.13</b> <sub>1.42</sub> | <b>0.98</b> <sub>&lt;0.05</sub> | <b>2.90</b> <sub>1.41</sub> | <b>0.99</b> <sub>&lt;0.05</sub> | <b>3.49</b> <sub>1.37</sub> |

Table 3: Comparison of sensitivity analysis results between baselines and our proposed evaluation method, with the best-performing scores highlighted in bold.  $p$  denotes statistical significance, and  $std$  indicates standard deviation.

| <i>Pair-Wise Kappa Score</i> |               |           |             |            |
|------------------------------|---------------|-----------|-------------|------------|
| Overall                      | Argumentative | Narrative | Descriptive | Expository |
| 0.469                        | 0.477         | 0.457     | 0.464       | 0.475      |

Table 4: Fleiss’ Kappa Agreement on pairwise annotations. A score between 0.41 to 0.60 indicates moderate inter-annotator agreement (Qin et al., 2024).

uation method works, we employ three LLMs as judges, including DeepSeek-V3 (DeepSeek-AI, 2025b), DeepSeek-R1 (DeepSeek-AI, 2025a) and GPT-4o (OpenAI, 2024) to assign scores 1~10 to each sub-question within every evaluation trait. Each model analyzes all sub-questions in a single turn. Specifically, we convert annotated pairwise comparisons into model rankings using a voting-based scoring approach to facilitate more effective comparisons. In all experiments, the temperature is set to 0.2, and the parameter  $\alpha$  is fixed to 3.

## 4.2 Ranking Agreement

To assess the ranking agreement, we use **Spearman’s Rank Correlation** (Spearman, 1904) and **Kendall’s  $\tau$**  (Kendall, 1938), which capture monotonic relationships between rankings. As shown in Table 2, our fine-grained and genre-oriented evaluation framework shows strong alignment with human judgments (Shen et al., 2023), achieving high correlations in both Spearman’s  $\rho$  and Kendall’s  $\tau$ . From these results, we draw three key conclusions: **(1) Our protocol performs better with stronger LLMs.** Our method crafts dimension-specific sub-questions and uses the CoT strategy to analyze the text and then assign all scores in a single turn. More powerful models exhibit a superior understanding of this complex and fine-grained process. Notably, DeepSeek-R1 achieves an almost

perfect alignment with human annotations, with  $\rho = 0.816$  and  $\tau = 0.704$ . **(2) Dependency-based score aggregation improves performance by approximately 2%.** Incorporating trait-level weights consistently improves alignment across different judges and essay genres, indicating that when assessing essays, the higher-level dimensions contribute more significantly to accurate evaluation. **(3) Our framework achieves higher alignment in Narrative and Descriptive genres.** Unlike argumentative and expository essays that emphasize logical structure and coherence and are effectively handled by general text evaluation method, narrative and descriptive writing focus on vivid imagery, rhetorical richness, and lexical complexity, which benefit more from our evaluation approach.

## 4.3 Sensitivity Analysis

Accurately determining an LLM’s proficiency in specific capabilities is essential for identifying its limitations and guiding improvements (Kim et al., 2025). Therefore, it is crucial that the evaluation protocol reliably reflects both high- and low-quality output. To this end, we conduct a sensitivity analysis to examine how effectively our evaluation protocol distinguishes essays of varying quality.

Accordingly, we categorize the essays into three quality tiers: high-, medium-, and low-quality based on the top-ranked, median-ranked, and bottom-ranked essays from the manually annotated data. Then we apply **Mann-Whitney  $U$  test** (Mann and Whitney, 1947) and compute the **Mean Difference** ( $MD$ ) to assess the robustness of the methods, as shown in Table 3. Take the *high&medium* set as an example. The  $U$  score indicates the proportion of cases in which high-quality data receive a higher score than medium-quality data. The mean difference reflects the average score difference between

| Models                              | Overall | Argumentative |       | Narrative |       | Descriptive |       | Expository |       |
|-------------------------------------|---------|---------------|-------|-----------|-------|-------------|-------|------------|-------|
|                                     |         | Open          | Cons. | Open      | Cons. | Open        | Cons. | Open       | Cons. |
| English Language Models             |         |               |       |           |       |             |       |            |       |
| Claude-3.7-sonnet (Anthropic, 2025) | 76.6    | 77.7          | 78.8  | 75.7      | 75.3  | 74.6        | 73.6  | 77.5       | 79.0  |
| Claud-3.5-sonnet (Anthropic, 2024)  | 75.4    | 73.4          | 73.8  | 75.3      | 73.6  | 74.8        | 73.4  | 77.1       | 80.4  |
| Grok-2 (xAI, 2024)                  | 75.3    | 75.6          | 78.5  | 71.5      | 73.6  | 70.2        | 73.5  | 75.1       | 79.3  |
| Grok-3 (xAI, 2025)                  | 74.6    | 74.9          | 78.1  | 73.6      | 72.8  | 73.1        | 72.0  | 73.3       | 76.4  |
| GPT-4o (OpenAI, 2024)               | 74.2    | 74.8          | 76.9  | 72.8      | 72.4  | 70.5        | 71.7  | 75.8       | 76.7  |
| GPT-4o-mini (OpenAI, 2024)          | 71.7    | 72.0          | 74.1  | 71.6      | 68.4  | 69.9        | 65.9  | 72.8       | 76.7  |
| GPT-3.5-turbo (Brown et al., 2020)  | 51.5    | 49.4          | 51.4  | 56.5      | 53.1  | 51.1        | 46.8  | 50.0       | 52.9  |
| Gemini-2.0-flash (Gemini., 2025)    | 72.9    | 74.5          | 76.3  | 71.5      | 71.1  | 68.4        | 67.6  | 76.7       | 75.4  |
| LLaMa-3.3-70B (Meta, 2024)          | 61.4    | 61.2          | 64.1  | 62.3      | 60.3  | 56.2        | 53.8  | 63.2       | 67.1  |
| LLaMa-3.1-70B (Meta, 2024)          | 40.5    | 37.6          | 46.6  | 35.1      | 28.6  | 45.0        | 42.2  | 39.6       | 44.8  |
| Chinese Language Models             |         |               |       |           |       |             |       |            |       |
| Qwen-Max (Qwen, 2025)               | 75.6    | 74.5          | 78.7  | 73.5      | 74.7  | 74.1        | 72.6  | 77.1       | 77.6  |
| Qwen2.5-72B-Instruct (Qwen, 2025)   | 72.7    | 73.1          | 75.2  | 71.7      | 71.4  | 68.8        | 68.8  | 74.5       | 75.5  |
| DeepSeek-V3 (DeepSeek-AI, 2025b)    | 75.1    | 77.2          | 77.9  | 71.2      | 71.8  | 72.7        | 67.8  | 80.4       | 79.4  |
| Doubao-1.5 (Doubao Team, 2025)      | 73.3    | 75.1          | 76.2  | 72.4      | 70.8  | 70.8        | 69.5  | 75.4       | 75.1  |
| ChatGLM-turbo (GLM, 2024)           | 71.2    | 70.0          | 70.8  | 70.0      | 69.6  | 69.2        | 68.7  | 74.2       | 75.8  |

Table 5: Benchmarking Results on Chinese Essay Writing. In each column, the highest and the second highest performance is highlighted in **bold** and is underlined. **Open** denotes Open-Ended and **Cons.** denotes Constrained.

the high- and medium-quality data.

The sensitivity analysis in Table 3 shows that **our evaluation method is effective at distinguishing essays of varying quality compared to the baseline**. Notably, our method shows significantly better performance in the high- and medium-quality essay classification, with an improvement ranging from approximately 2% to 10%. Furthermore, it yields a larger mean difference, suggesting that the score distributions between quality levels are more distinguishable. These trends hold consistently across all judge models, highlighting the robustness and sensitivity of our framework when evaluating outputs from strong LLMs. Overall, R1 emerges as the top-performing model, achieving the highest  $U$  score and exhibiting a pronounced distinction across all quality levels.

## 5 Benchmarking

### 5.1 Experiment Setup

**Baselines.** To explore how current state-of-the-art LLMs perform in Chinese essay writing, we meticulously select 15 popular large-scale LLMs for evaluation, including English language models and Chinese language models. We access proprietary LLMs via their official APIs and open-source LLMs through their public repositories. During writing, we set the temperature to 0.8 to encourage creativity in generation.

**Metrics.** Considering the inference time cost and overall performance, we adopt GPT-4o as the evaluation judge model. The temperature is set to 0.2 to ensure deterministic output, while all other parameters remain in their default settings. To facilitate fair comparison across models, we normalize the aggregated scores to a 100-point scale.

**Main Results.** The benchmark results are presented in Table 5. Notably, state-of-the-art proprietary models achieve strong performance on the Chinese essay writing task, with Claude-3.7-sonnet attaining the highest overall score. Moreover, most newer versions outperform their predecessors, with the exception of Grok, as Grok-3 places greater emphasis on reasoning. It is worth highlighting that Chinese LLM families also perform competitively: Qwen-max ranks as the second-best model, DeepSeek surpasses Grok-3 and GPT-4o on this task, and Qwen-2.5-72B-Instruct outperforms both the GPT-4o-mini and its similarly sized counterpart, LLaMA-3.1-70B-Instruct.

**Genre-based Performance.** LLMs demonstrate stronger capabilities in writing argumentative and expository essays, while they fall short in narrative and descriptive genres. This disparity likely stems from the inherent characteristics of these genres: argumentative and expository essays emphasize structural coherence, logical reasoning, and clear topic development, where LLMs typically excel. In contrast, narrative and descriptive

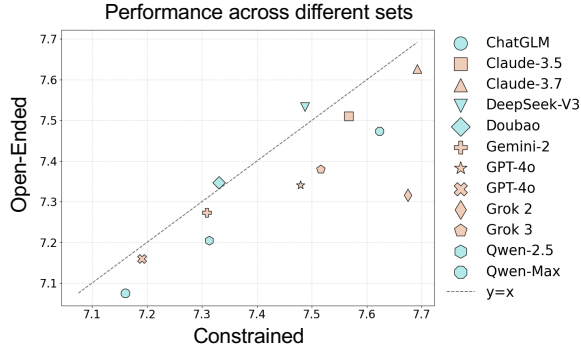


Figure 4: Comparison of Performance by Sets.

essays require creativity, emotional nuance, and context-aware storytelling. These challenges are further amplified in Chinese writing, where expressive richness, metaphorical language, and cultural context play a more significant role, especially in narrative and descriptive forms. Such features are difficult to model with LLMs, leading to degraded performance in these genres.

**Open-Ended versus Constrained.** Interestingly, LLMs perform better in constrained sets than open-ended sets, as shown in Figure 4. This is likely because constrained prompts provide more explicit requirements and clearer guidance, which help the models organize content, maintain relevance, and follow a well-defined structure. In contrast, open-ended prompts offer greater flexibility but less direction, placing higher demands on the model’s ability to plan, generate diverse content, and maintain coherence without external constraints.

## 6 Related Work

**LLM Generation Evaluation.** The rapid progress of LLMs prompts the need for a comprehensive evaluation of their text generation (Liu et al., 2023; Kim et al., 2025). Existing frameworks are often task-specific: instruction-following is assessed via diverse prompts and constraint scenarios (Qin et al., 2024; Wen et al., 2024; Jiang et al., 2024), while reasoning is tested through multi-hop question answering (Krishna et al., 2024; Ling et al., 2025). In this work, we turn our attention to the issue of generated text quality evaluation. Previous research has addressed quality assessment in specific contexts: e.g., summarization (Liu et al., 2024c), financial content (Islam et al., 2023; Xie et al., 2024), Wikipedia-style writing (Gao et al., 2024; Zhang et al., 2025), and long-form text (Tan et al., 2024; Que et al., 2024). In contrast, we address the under-

explored challenge of evaluating Chinese writing across literary genres, offering a systematic framework for multilingual LLM assessment.

**Automatic Essay Evaluation.** Automated Essay Scoring (AES) uses computer systems to assess written text in educational settings (Dikli, 2006; Attali and Burstein, 2006). While datasets like ASAP (Hamner et al., 2012) and TOEFL11 (Blanchard et al., 2013) provide valuable English essay prompts, they are limited in scale and unsuitable for assessing LLM-generated essays, especially in non-English contexts. AES methods have progressed from hand-crafted features (Yannakoudakis et al., 2011; Persing and Ng, 2013) to neural, trait-specific models (Taghipour and Ng, 2016; Uto et al., 2020), and recently to LLM-based evaluation (Lee et al., 2024; Chu et al., 2025). These typically score coarse-grained aspects like grammar, coherence, content, and creativity (Li and Ng, 2024), but remain English-centric and overlook the rhetorical and cultural complexities of Chinese writing. In addition, although recent frameworks like *WritingBench* (Wu et al., 2025) and *BigGen Bench* (Kim et al., 2025) offer fine-grained evaluation strategies through prompt-specific assessment instances, they fall short in covering a broader range of writing prompts, limiting their applicability to various essay tasks.

## 7 Conclusion

In this work, we present ESSAYBENCH, the first comprehensive benchmark for evaluating the capabilities of LLMs in the Chinese essay writing and evaluation across four distinct literary genres. To address the challenges of analytic and accurate essay evaluation, ESSAYBENCH adopts a genre-oriented, hierarchical multi-trait evaluation approach that enables fine-grained scoring. Specifically, we introduce a dependency-based aggregation strategy to compute the final scores. Our comprehensive human agreement study and sensitivity analysis demonstrate that the framework achieves high alignment with human judgment and effectively distinguishes essays of varying quality. Furthermore, we benchmark 15 large-size LLMs on Chinese essay writing, revealing notable limitations in descriptive and narrative essays, particularly for open-ended prompts. Overall, ESSAYBENCH offers a diverse dataset and a robust evaluation framework for Chinese essay, with practical implications for educational applications and future research.



## Limitations

Despite the contributions presented in our work, several minor limitations remain:

- First, the datasets and evaluation dimensions used in this study are primarily based on Chinese essay prompts. While the widely adopted essay categorization framework can be applied to other languages such as English and Japanese, the current work focuses on Chinese essay writing. This is due to the significant differences in idioms, linguistic conventions, and cultural expressions across languages. Nonetheless, the framework has the potential to be translated and extended to multilingual settings in future work.
- Second, although this work proposes a more fine-grained evaluation method for Chinese essays, the designed evaluation traits primarily focus on overall expression and structural aspects, such as paragraph organization and comprehensive performance of the essays from multiple perspectives. However, it overlooks more granular analyses at the lexical and sentence levels. Future research could incorporate finer-grained evaluations that consider sentence-level coherence and word-level richness
- Third, ESSAYBENCH primarily focuses on evaluating the overall quality of essays, while overlooking instruction-following abilities. For example, whether the generated essays adhere strictly to the prompt requirements has not been thoroughly assessed. To enable a more comprehensive evaluation, future research could address this gap by incorporating the essay prompt following ability as an explicit evaluation dimension.

## Ethics Statement

To mitigate potential ethical concerns, all essay prompts were carefully reviewed and filtered by manual inspection. We ensured that none of the prompts contained offensive, gender-biased, harmful, or otherwise ethically inappropriate content. In addition, all participants involved in the human agreement study were professional annotators who were fairly compensated for their contributions.

## References

Anthropic. 2024. [Introducing claude 3.5 sonnet](#). Accessed: 2025-05-19.

- Anthropic. 2025. [Claude 3.7 sonnet](#). Accessed: 2025-05-19.
- Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).
- Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.
- Nishant Balepur, Jie Huang, and Kevin Chang. 2023. [Expository text generation: Imitate, retrieve, paraphrase](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11896–11919, Singapore. Association for Computational Linguistics.
- Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Richard M Chadbourne. 1983. A puzzling literary genre: comparative views of the essay. *Comparative literature studies*, 20(2):133–153.
- Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. [Humans or LLMs as the judge? a study on judgement bias](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. 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](#).
- SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025. [Rationale behind essay scores: Enhancing S-LLM’s multi-trait essay scoring with rationale generated by LLMs](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5796–5814, Albuquerque, New Mexico. Association for Computational Linguistics.
- DeepSeek-AI. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- DeepSeek-AI. 2025b. [Deepseek-v3 technical report](#).
- Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).



- Nuowei Liu, Xinhao Chen, Hongyi Wu, Changzhi Sun, Man Lan, Yuanbin Wu, Xiaopeng Bai, Shaoguang Mao, and Yan Xia. 2024a. [CERD: A comprehensive Chinese rhetoric dataset for rhetorical understanding and generation in essays](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6744–6759, Miami, Florida, USA. Association for Computational Linguistics.
- Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew 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. 2024b. [Align-Bench: Benchmarking Chinese alignment of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11621–11640, Bangkok, Thailand. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024c. [Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.
- Henry B. Mann and Donald R. Whitney. 1947. [On a test of whether one of two random variables is stochastically larger than the other](#). *Annals of Mathematical Statistics*, 18(1):50–60.
- T. McCarthy. 1998. *Descriptive Writing*. Scholastic Inc.
- Meta. 2024. [The llama 3 herd of models](#).
- Yusuke Miura, Chi-Lan Yang, Masaki Kuribayashi, Keigo Matsumoto, Hideaki Kuzuoka, and Shigeo Morishima. 2025. [Understanding and supporting formal email exchange by answering ai-generated questions](#).
- OpenAI. 2024. [Gpt-4o system card](#).
- Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).
- Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. [InFoBench: Evaluating instruction following ability in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048, Bangkok, Thailand. Association for Computational Linguistics.
- Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, et al. 2024. [HeloBench: Evaluating long text generation capabilities of large language models](#). *arXiv preprint arXiv:2409.16191*.
- Qwen. 2025. [Qwen2.5-max:exploring the intelligence of large-scale moe model](#). Accessed: 2025-05-19.
- Qwen. 2025. [Qwen2.5 technical report](#).
- Thomas L. Saaty. 1980. *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill International Book Co., New York & London.
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. [Large language models are not yet human-level evaluators for abstractive summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.
- Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. [Towards evaluating narrative quality in student writing](#). *Transactions of the Association for Computational Linguistics*, 6:91–106.
- Charles Spearman. 1904. [The proof and measurement of association between two things](#). *The American Journal of Psychology*, 15(1):72–101.
- Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.
- Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, et al. 2024. [Proxyqa: An alternative framework for evaluating long-form text generation with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6806–6827.
- InternLM Team. 2023. [Internlm: A multilingual language model with progressively enhanced capabilities](#).



|     |  |     |
|-----|--|-----|
| 881 | Qwen Team. 2023. Qwen technical report. <i>arXiv preprint arXiv:2309.16609</i> . |     |
| 882 |  |     |
| 883 | Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier                            |     |
| 884 | Martinet, Marie-Anne Lachaux, Timothée Lacroix,                                  |     |
| 885 | Baptiste Rozière, Naman Goyal, Eric Hambro,                                      |     |
| 886 | Faisal Azhar, et al. 2023. Llama: Open and effi-                                 |     |
| 887 | cient foundation language models. <i>arXiv preprint</i>                          |     |
| 888 | <i>arXiv:2302.13971</i> .  |     |
| 889 | Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020.                                    |     |
| 890 | Neural automated essay scoring incorporating hand-                               |     |
| 891 | crafted features. In <i>Proceedings of the 28th interna-</i>                     |     |
| 892 | <i>tional conference on computational linguistics</i> , pages                    |     |
| 893 | 6077–6088.   |     |
| 894 | Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon                            |     |
| 895 | Lee. 2025. <i>Collabstory: Multi-llm collaborative story</i>                     |     |
| 896 | <i>generation and authorship analysis</i> .                                      |     |
| 897 | Henning Wachsmuth, Nona Naderi, Yufang Hou,                                      |     |
| 898 | Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberd-                                |     |
| 899 | ingh Thijm, Graeme Hirst, and Benno Stein. 2017.                                 |     |
| 900 | <i>Computational argumentation quality assessment in</i>                         |     |
| 901 | <i>natural language</i> . In <i>Proceedings of the 15th Con-</i>                 |     |
| 902 | <i>ference of the European Chapter of the Association</i>                        |     |
| 903 | <i>for Computational Linguistics: Volume 1, Long Pa-</i>                         |     |
| 904 | <i>pers</i> , pages 176–187, Valencia, Spain. Association                        |     |
| 905 | for Computational Linguistics.   |     |
| 906 | Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten                                 |     |
| 907 | Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and                               |     |
| 908 | Denny Zhou. 2023. <i>Chain-of-thought prompting elic-</i>                        |     |
| 909 | <i>its reasoning in large language models</i> .                                  |     |
| 910 | Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao                                    |     |
| 911 | Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu,                                    |     |
| 912 | Wendy Gao, Jiaying Xu, et al. 2024. Benchmark-                                   |     |
| 913 | ing complex instruction-following with multiple con-                             |     |
| 914 | straints composition. <i>Advances in Neural Informa-</i>                         |     |
| 915 | <i>tion Processing Systems</i> , 37:137610–137645.                               |     |
| 916 | Zhiyuan Wen, Jiannong Cao, Zian Wang, Beichen Guo,                               |     |
| 917 | Ruosong Yang, and Shuaiqi Liu. 2025. <i>Interactivesur-</i>                      |     |
| 918 | <i>vey: An llm-based personalized and interactive sur-</i>                       |     |
| 919 | <i>vey paper generation system</i> .   |     |
| 920 | Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li,                                   |     |
| 921 | Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang,                                   |     |
| 922 | Mengyue Wu, Qin Jin, et al. 2025. Writingbench:                                  |     |
| 923 | A comprehensive benchmark for generative writing.                                |     |
| 924 | <i>arXiv preprint arXiv:2503.05244</i> .   |     |
| 925 | xAI. 2024. <i>Grok-2 beta release</i> . Accessed: 2025-05-19.                    |     |
| 926 | xAI. 2025. <i>Grok 3 beta — the age of reasoning agents</i> .                    |     |
| 927 | Accessed: 2025-05-19.  |     |
| 928 | Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu                                  |     |
| 929 | Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong                                   |     |
| 930 | Li, Yongfu Dai, Duanyu Feng, et al. 2024. Finben:                                |     |
| 931 | A holistic financial benchmark for large language                                |     |
| 932 | models. <i>Advances in Neural Information Processing</i>                         |     |
| 933 | <i>Systems</i> , 37:95716–95743.   |     |
|     | Helen Yannakoudakis, Ted Briscoe, and Ben Medlock.                               | 934 |
|     | 2011. A new dataset and method for automatically                                 | 935 |
|     | grading esol texts. In <i>Proceedings of the 49th annual</i>                     | 936 |
|     | <i>meeting of the association for computational linguis-</i>                     | 937 |
|     | <i>tics: human language technologies</i> , pages 180–189.                        | 938 |
|     | Jiebin Zhang, J Yu Eugene, Qinyu Chen, Chenhao                                   | 939 |
|     | Xiong, Dawei Zhu, Han Qian, Mingbo Song, Weimin                                  | 940 |
|     | Xiong, Xiaoguang Li, Qun Liu, et al. 2025. Wikigen-                              | 941 |
|     | bench: Exploring full-length wikipedia generation                                | 942 |
|     | under real-world scenario. In <i>Proceedings of the 31st</i>                     | 943 |
|     | <i>International Conference on Computational Linguis-</i>                        | 944 |
|     | <i>tics</i> , pages 5191–5210.   | 945 |
|     | Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan                                | 946 |
|     | Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,                                     | 947 |
|     | Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,                                 | 948 |
|     | Joseph E. Gonzalez, and Ion Stoica. 2023. <i>Judging</i>                         | 949 |
|     | <i>llm-as-a-judge with mt-bench and chatbot arena</i> .                          | 950 |
|     | Mališa Žižović and Dragan Pamucar. 2019. <i>New model</i>                        | 951 |
|     | <i>for determining criteria weights: Level based weight</i>                      | 952 |
|     | <i>assessment (lbwa) model</i> . <i>Decision Making: Appli-</i>                  | 953 |
|     | <i>cations in Management and Engineering</i> , 2(2):126–                         | 954 |
|     | 137.   | 955 |



## A Hierarchical Traits Design

We developed a comprehensive, genre-specific multi-trait evaluation framework tailored explicitly to the distinctive features, objectives, and contexts of each essay genre. This system aligns closely with educational standards and requirements, ensuring a precise and relevant assessment approach.

### A.1 Argumentative Essays

For Argumentative essays, we focus on:

- **Argument** ( $d = 0$ ): Clarity, precision, and relevance of the central viewpoint.
- **Evidence** ( $d = 1$ ): Strength, appropriateness, and reliability of supporting details and examples.
- **Argumentation Methods** ( $d = 2$ ): Effective use of logical strategies such as comparison, causality, and deduction.
- **Logical Development** ( $d = 2$ ): Coherent and logically sequenced progression of ideas.
- **Expression** ( $d = 3$ ): Clarity, precision, and stylistic appropriateness of language.
- **Endings** ( $d = 3$ ): Ability to summarize arguments effectively and deliver a compelling conclusion.

### A.2 Narrative Essays

For Narrative essays, we emphasize:

- **Language and Style** ( $d = 0$ ): Effectiveness of vocabulary, tone, and stylistic choices in storytelling.
- **Structural Layer and Narrative Techniques** ( $d = 0$ ): Sophisticated use of narrative structures and literary techniques.
- **Theme and Emotional Expression** ( $d = 1$ ): Depth of thematic content and authenticity of emotional portrayal.
- **Overall Structure and Plot Development** ( $d = 2$ ): Clear progression, effective pacing, and cohesive plot structure.
- **Characterization and Detail** ( $d = 2$ ): Rich, vivid portrayal of characters and setting details.
- **Choice of Material** ( $d = 3$ ): Originality, relevance, and effectiveness in selecting narrative content.

### A.3 Descriptive Essays

For Descriptive essays, we highlight:

- **Clarity of Subject and Central Theme** ( $d = 0$ ): Distinct and clearly communicated central image or idea.

- **Rhythm and Overall Fluency** ( $d = 1$ ): Smooth flow and harmonious pacing throughout the essay.
- **Content and Unique Perspective** ( $d = 1$ ): Original insights and distinctive angles in the descriptions.
- **Structure and Organization** ( $d = 2$ ): Effective and logical arrangement enhancing readability.
- **Emotional Expression and Atmosphere Description** ( $d = 2$ ): Authentic depiction of the atmosphere and emotional tone.
- **Sensory Details** ( $d = 3$ ): Use of vivid and engaging sensory imagery.

### A.4 Expository Essays

Finally, for Expository essays, we prioritize:

- **Clarity of Topic and Purpose** ( $d = 0$ ): Clearly defined subject matter and objectives.
- **Practicality and Relevance** ( $d = 0$ ): Real-world applicability and pertinence of the provided information.
- **Scientific Accuracy and Credibility of Content** ( $d = 0$ ): Validity and trustworthiness of the facts and data presented.
- **Logical Structure and Coherence** ( $d = 1$ ): Systematic and logically sound organization of ideas.
- **Clarity and Appropriateness of Language** ( $d = 2$ ): Use of a clear, accessible, and appropriate academic language.
- **Diversity and Appropriateness of Explanatory Methods** ( $d = 3$ ): Variety and suitability of explanatory techniques, enhancing comprehension and reader engagement.

## B Evaluation Prompt

### B.1 CoT Prompting

In our evaluation method, we implement the Chain-of-Thought (CoT) prompting strategy, which first guides the large language models (LLMs) to systematically analyze essays before assigning scores. This structured analytical step provides LLMs with robust reasoning and clear justifications, facilitating accurate scoring decisions. In addition, the detailed CoT reasoning process serves as a valuable reference, allowing evaluators and users to better understand and verify the rationale behind each assigned score. The specific prompts used for the CoT strategy are illustrated in Figure 5.

## **B.2 Trait-based sub-questions**

For each hierarchically designed trait, we carefully develop a series of detailed, targeted evaluation questions, addressing multiple dimensions and perspectives relevant to each trait. These questions are crafted to comprehensively assess the specific characteristics and nuances inherent in each genre of essays. The specific questions tailored for Argumentative, Narrative, Descriptive, and Expository essays are illustrated in Figures 6, 7, 8, and 9, respectively.

## **C Human Annotation**

For our human agreement study, we recruited 14 professional annotators with strong backgrounds in Chinese linguistics and language education. Each annotator was assigned approximately 70 data samples per day, working within an 8-hour schedule. The complete annotation of 5,040 data items was completed over five days. This rigorous process ensured consistency, reliability, and high-quality annotations across the dataset. The comprehensive annotation guidelines provided to annotators are illustrated in Figure 10.

## **D Case Study**

In this section, we conduct a qualitative case study of essays across different quality levels, using outputs from various LLMs. Specifically, we examine high-quality essays generated by DeepSeek Chat (score: 8.7), median-quality essays by GPT-3.5-turbo (score: 7.0), and low-quality essays by Llama-3.1-70B-Instruct (score: 6.3), as shown in Figures 11, 12, and 13. These scores are derived from the evaluations conducted by DeepSeek-R1. We observe that the evaluator provides detailed and consistent analyses across essays of varying quality, highlighting both strengths and weaknesses. This type of evaluative feedback demonstrates strong interpretability and reliability, making it valuable for future educational applications such as formative writing assessment and personalized feedback generation.

#### Chain-of-thought Evaluation Prompt (Chinese)

你是一个擅长评价文本质量的专家，负责客观公正地评估AI助手对用户关于议论作文时的回答质量。请根据以下评价标准，对助手的回答进行评估。

每个问题请根据以下评分标准给出分数：

- 1~2分：非常差（完全没有达到预期标准）
- 3~4分：较差（部分满足预期标准，但存在明显问题）
- 5~6分：中等（满足基本标准但仍有改进空间）
- 7~8分：较好（大部分满足标准，仅有少量小问题）
- 9~10分：优秀（完全符合或超过预期标准）

评估原则：

- 回答必须需要遵循用户的指令，尤其是作文回答的读者人群和适用场景要与指令相符合；
- 采用“证据-推论”模式，每个评分必须有文本依据；
- 你的打分需要尽可能严格，必须有理有据；
- 每个问题都需要提供评分，并附上简短的解释说明；
- 输出每个问题的说明和评分，每个维度的最终评分是该维度下所有问题分数的平均值，计算得到每个维度的分数值

# 以下是作文题目：

</题目开始>

{Essay Prompt}

</题目结束>

# 以下是作文题目：

</题目开始>

{Essay Prompt}

</题目结束>

# 以下是评分标准：

{议论文问题}

#### Chain-of-thought Evaluation Prompt (English Translation Version)

You are an expert skilled in evaluating text quality, responsible for objectively and impartially assessing the quality of an AI assistant's responses to users regarding argumentative essays.

Please evaluate the assistant's responses based on the following criteria. For each question, provide a score according to the following grading scale:

- 1~2 points: Very poor (completely fails to meet expected standards)
- 3~4 points: Poor (partially meets expected standards but has significant issues)
- 5~6 points: Average (meets basic standards but still has room for improvement)
- 7~8 points: Good (mostly meets standards with only minor issues)
- 9~10 points: Excellent (fully meets or exceeds expected standards)

Evaluation principles:

- Responses must adhere to the user's instructions, especially ensuring the target audience and applicable scenarios for the essay response align with the instructions.
- Adopt an "evidence-inference" model—each score must be supported by textual evidence.
- Your scoring must be as strict as possible and well-justified.
- Each question requires a score accompanied by a brief explanation.
- Provide an explanation and score for each question. The final score for each dimension is the average of all question scores under that dimension. Calculate the score for each dimension accordingly.

# The following is the essay prompt:

</Prompt Start>

{Essay Prompt}

</Prompt End>

# The following is the argumentative essay:

</Essay Start>

{Essay}

</Essay End>

The following is the evaluation rubrics:

{Sub-questions of Argumentative Essays}

Figure 5: CoT Prompt Strategy for Evaluation.

|  |   |
|--|---|
| Argument   | 中心论点  |
| <ol style="list-style-type: none"> <li>1. Is the central thesis clearly articulated?</li> <li>2. Is the thesis easy to extract and summarize?</li> <li>3. Does the essay state its thesis explicitly at the very beginning?</li> <li>4. Does the argument engage with deeper social, cultural, or philosophical issues?</li> <li>5. Does the thesis demonstrate original insight and depth of understanding?</li> </ol>  | <ol style="list-style-type: none"> <li>1. 文章中心观点是否清晰文章</li> <li>2. 中心观点是否易于提炼和概括?</li> <li>3. 文章是否在开篇直接明确地提出了论点?</li> <li>4. 文章观点是否深入探讨了深层次的社会、文化或哲学思考?</li> <li>5. 文章观点是否体现出思想见解和对问题的洞察</li> </ol>                     |
| Evidence   | 论据  |
| <ol style="list-style-type: none"> <li>1. Are the pieces of evidence closely tied to the thesis?</li> <li>2. Is the amount of evidence sufficient?</li> <li>3. Is the evidence presented in enough detail?</li> </ol> <p>Is the evidence representative of the points being made?<br/>Are cited data, sources, and examples credible and reliable?</p>   | <ol style="list-style-type: none"> <li>1. 论据与论点的关联是否紧密?</li> <li>2. 所使用的论据是否足够</li> <li>3. 所使用的论据是否详实?</li> <li>4. 所使用的论据是否具有代表性?</li> <li>5. 引用的数据、文献、实例是否真实可靠?</li> </ol>   |
| Argumentation Methods  | 论证方法  |
| <ol style="list-style-type: none"> <li>1. Does the essay employ a variety of reasoning techniques (e.g. exemplification, comparison, citation, cause-effect, induction/deduction) effectively?</li> <li>2. When using statistics or others' opinions, are they accompanied by clear explanation and analysis?</li> <li>3. Are opposing viewpoints analyzed and critiqued objectively?</li> <li>4. Does the essay consider multiple perspectives to strengthen its persuasiveness?</li> </ol> | <ol style="list-style-type: none"> <li>1. 是否灵活运用了举例论证、对比论证、引用论证、因果论证、归纳演绎等多种论证手法</li> <li>2. 如果运用了统计数据或者他人观点, 是否配合解释和分析</li> <li>3. 对于对立观点是否客观分析和批判?</li> <li>4. 在文章能否兼顾不同观点, 从而增加说服力?</li> </ol>                       |
| Logical Development  | 论证过程  |
| <ol style="list-style-type: none"> <li>1. Are the sub-arguments in each paragraph coherent with the central thesis?</li> <li>2. Do the sub-arguments form a clear, progressive logical chain?</li> <li>3. Does each paragraph's sub-argument have appropriate supporting evidence?</li> <li>4. Are there smooth transitions and logical connections between paragraphs?</li> </ol>   | <ol style="list-style-type: none"> <li>1. 文章的小论点是否与中心论点保持连贯?</li> <li>2. 小论点之间是否构成合理的层层递进的逻辑链条?</li> <li>3. 每个段落小论点是否有对应的论据支撑?</li> <li>4. 段落之间有无合适的过渡与承接?</li> </ol>   |
| Expression   | 语言表达  |
| <ol style="list-style-type: none"> <li>1. Is the word choice appropriate and precise?</li> <li>2. Is the language logically coherent?</li> <li>3. Are sentence structures and tones varied and engaging?</li> </ol>  | <ol style="list-style-type: none"> <li>1. 文章用词是否贴切、精确?</li> <li>2. 文章语言是否具有逻辑性?</li> <li>3. 文章句式、语气是否丰富多样?</li> </ol>   |
| Endings  | 文章结尾  |
| <ol style="list-style-type: none"> <li>1. Does the conclusion echo the issues or claims raised in the introduction to form a cohesive whole?</li> <li>2. Does it summarize the essay's main theme rather than merely restating the thesis?</li> <li>3. Does it synthesize the arguments from multiple angles?</li> <li>4. Does it offer meaningful insights or a call to action that guides the reader's further reflection or behaviour?</li> </ol>   | <ol style="list-style-type: none"> <li>1. 结尾是否呼应文章开头提出的问题或论断, 使文章形成整体?</li> <li>2. 结尾是否总结了文章主题, 而非做了简单的重复论点阐述?</li> <li>3. 结尾是否对文中多个角度的论证内容加以概括提炼?</li> <li>4. 结尾是否提出了有价值的启示和呼吁, 以引导读者思考或行动? 把上面内容翻译成合适的英语版本</li> </ol> |

Figure 6: Multi-traits and sub-questions of Argumentative Essay.



|   |   |
|---|---|
| Theme & Emotional Expression  | 主题立意与情感表达   |
| <ol style="list-style-type: none"> <li>1.Does the essay present a clear central idea or core theme?</li> <li>2.Does the narrative effectively convey the author's intended emotion or attitude (e.g., love, longing, nostalgia, awe)?</li> <li>3.Is the theme thought-provoking or inspiring reflection?</li> <li>4.Are the emotions expressed naturally, avoiding excessive sentimentality or contrivance?</li> <li>5.Does the piece resonate emotionally, allowing readers to feel its warmth and human insight?</li> </ol>   | <ol style="list-style-type: none"> <li>1.文章是否有鲜明的中心思想，核心立意？</li> <li>2.文章是否在叙事中有效地传达出作者想要表达的情感或态度，比如热爱、思念、怀旧或敬畏等？</li> <li>3.立意是否具有启发性，或给人带来思考或感悟？</li> <li>4.文章情感是否自然流露，避免过渡煽情或矫揉造作？</li> <li>5.是否能引起读者的情感共鸣，让读者感受到文章的温度与人性光辉？</li> </ol>                                    |
| Overall Structure & Plot Development  | 整体结构与情节安排   |
| <ol style="list-style-type: none"> <li>1.Does the story unfold organically with a clear beginning, middle, and end?</li> <li>2.Does the introduction capture readers' attention, and does the conclusion effectively wrap up and echo the theme?</li> <li>3.Are there key plot twists or climaxes that enhance readability?</li> <li>4.Does the use of vivid detail and emotional pacing lend unity to the events themselves?</li> </ol>  | <ol style="list-style-type: none"> <li>1.文章是否有起承转合，能够自然而有序地展开故事？</li> <li>2.文章开头能否抓住读者，结尾能否有效收束并呼应主题？</li> <li>3.叙事中是否设置了关键的情节转折或高潮，增强文章可读性？</li> <li>4.文章是否有细节描写和情感推动，让事件本身更具有内在统一性？</li> </ol>  |
| Characterization & Detail   | 人物形象与细节描写   |
| <ol style="list-style-type: none"> <li>1.Are characters portrayed vividly—through appearance, dialogue, actions, inner thoughts, and expressions?</li> <li>2.Do characters' words and behaviors align logically with their personalities?</li> <li>3.Are settings, events, and moods described with appropriate richness?</li> <li>4.Does the environment reinforce characters' emotions, advance the plot, or highlight the theme?</li> <li>5.Are distinctive details selected to lend authenticity and emotional impact?</li> <li>6.Are details handled deftly, without distracting from the main narrative?</li> </ol> | <ol style="list-style-type: none"> <li>1.是否通过外貌、语言、动作、心理活动、神态等多种方式描写人物，让形象鲜明立体？</li> <li>2.人物言行和性格特征是否一致、有逻辑？</li> <li>3.是否对故事发生的事件、地点、氛围做了恰到好处的描写？</li> <li>4.环境描写能否烘托人物情绪、推动情节发展或突出文章主题？</li> <li>5.是否抓住有典型意义的细节，让文章更具有真实感和感染力？</li> <li>6.细节刻画是否恰当，没有偏离主线或喧宾夺主？</li> </ol> |
| Language & Style  | 语言表达  |
| <ol style="list-style-type: none"> <li>1.Are literary devices (simile, personification, parallelism, etc.) used appropriately to enhance readability?</li> <li>2.Is the language both accessible and elegant, reflecting literary flair?</li> <li>3.Do sentences and paragraphs flow smoothly, avoiding abrupt jumps or unnecessary padding?</li> <li>4.Are transitional or linking words used to maintain clarity and coherence?</li> </ol>  | <ol style="list-style-type: none"> <li>1.是否通过恰当的修辞手法（比喻、拟人、排比等）增加可读性？</li> <li>2.语言是否通俗易懂而又不失文采，体现文学性？句子之间、段落之间是否衔接自然流畅，避免生硬的跳跃或内容堆砌？</li> <li>3.是否能够通过过渡词或关联词，让行文舒展、章法清晰？</li> </ol>   |
| Choice of Material  | 选材的典型性与创新性  |
| <ol style="list-style-type: none"> <li>1.Are the selected events or scenes typical and representative?</li> <li>2.Can a single small incident illuminate deeper life, social, or emotional themes?</li> <li>3.Is the angle or material fresh and inventive, avoiding clichés?</li> <li>5.Can familiar subjects be presented from a unique perspective or with a new twist?</li> <li>6.Is the material tightly connected to the theme, avoiding interesting but irrelevant details?</li> <li>7.Is real, meaningful life experience used to support the essay's purpose and emotion?</li> </ol>                             | <ol style="list-style-type: none"> <li>1.所选事件或场景是否具有代表性、典型性？</li> <li>2.能否通过一件小事折射出深层次的人生、社会或情感问题？</li> <li>3.写作材料或角度是否独具匠心，避免千篇一律？</li> <li>4.常见题材能否有独特的视角或写法，呈现新的亮点？</li> <li>5.是否用生活中真实且富有意义的素材，支撑文章的立意与情感？</li> </ol>   |
| Structural Layers & Narrative Techniques  | 结构层次与叙述方式   |
| <ol style="list-style-type: none"> <li>1.Are techniques like chronological order, flashback, or interwoven narratives used effectively?</li> <li>2.Does clever sequencing help readers immerse themselves or build suspense?</li> <li>3.Is the chosen point of view (first-person, third-person, multiple perspectives) consistent and effective for the theme?</li> <li>4.Are paragraphs organized to help readers focus on key events and emotions?</li> </ol> <p>Is the balance of long and short paragraphs appropriate, avoiding reader fatigue from overly dense text?</p>  | <ol style="list-style-type: none"> <li>1.时间顺序、倒叙、插叙、补叙等多种手法是否合理运用？</li> <li>2.是否通过巧妙的事件线调整让读者更好地进入故事场景或增强悬念？</li> <li>3.采用第一人称、第三人称或多视角叙事是否自洽且利于凸显主题？</li> <li>4.文章段落编排是否便于读者聚焦主要情节与情感？</li> <li>5.长短段落或详略搭配是否得当，避免通篇大段文字导致读者疲劳？</li> </ol>                                   |

Figure 7: Multi-traits and sub-questions of Narrative Essay.

| Clarity of Subject and Central Theme   | 描写对象的鲜明性与中心主题  |
|--|--|
| 1. Is the central theme or key image prominently highlighted?<br>2. Is the essay's main purpose or emotional tone clearly conveyed?  | 1. 中心主题或中心意象是否突出?<br>2. 文章写作的主旨或情感基调是否清晰?  |
| Sensory Detail   | 感官细节   |
| 1. Does the writing employ multiple senses—visual, auditory, tactile, gustatory, etc.—to portray the subject's details?<br>2. Does it create an immersive, “you-are-there” experience rather than remaining at a superficial or generic level?<br>3. Are word choices precise and evocative, capturing the subject's characteristics with vivid language?  | 1. 是否通过视觉、听觉、触觉、味觉等多种感官来呈现对象的细节?<br>2. 是否做到让人如临其境, 而非只停留在表面、笼统的描述?<br>3. 用词是否准确精炼, 能将对象的特征用传神的词汇表达出来?  |
| Structure and Organization   | 结构布局与层次  |
| 1. Is the overall structure logical, using spatial or chronological order to guide the reader's experience?<br>2. Is there an appropriate balance between elaboration and brevity, with major elements described in depth and minor ones summarized?<br>3. Are transitions between paragraphs smooth and natural?<br>4. Do the focal descriptions stay tightly aligned with the theme and the author's intended emotion?   | 1. 文章结构是否合理, 能否通过空间顺序或者时间顺序的组织方式让读者顺着作者的笔触去观赏或体会对象?<br>2. 是否详略得当, 主要景物或特征详写, 次要部分略写, 并在详略之间保持合适的比例?<br>3. 文章段落间的衔接、过渡是否流畅自然? ②描述重点是否紧扣主题, 并与作者的情感意图相呼应?                              |
| Emotional Expression and Atmosphere  | 情感表达与意境营造  |
| 1. Does the essay seamlessly integrate the author's emotions (e.g., affection, longing, nostalgia, wonder) into the descriptive details?<br>2. Does it evoke a distinct mood or ambiance—such as tranquility, liveliness, mystery, or solemnity?<br>3. Do the environmental descriptions resonate with the author's feelings, offering readers a unified internal-external aesthetic experience?   | 1. 文章是否将作者的情感(喜爱, 向往, 怀念, 惊叹等)自然地融于细节之中?<br>2. 文章是否营造特定的氛围和意境, 比如静谧、热闹、神秘、庄严等?<br>3. 环境描写是否与作者情感相匹配, 能够让读者感受到内外合一的审美体验?  |
| Content and Unique Perspective   | 选材典型性与独特视角   |
| 1. Does the writer uncover new insights in familiar subjects, using a fresh angle or dimension of thought?<br>2. Does it avoid clichéd phrasing or formulaic techniques?<br>3. Does it draw on personal experience or reflection to add warmth and emotional impact?<br>4. Does the chosen subject carry real-world relevance or cultural resonance that fosters reader empathy?   | 1. 所选事件或场景是否具有代表性、典型性;<br>2. 能否通过一件小事折射出深层次的人生、社会或情感问题?<br>3. 写作材料或角度是否独具匠心, 避免千篇一律?<br>4. 常见题材能否有独特的视角或写法, 呈现新的亮点? ②选材与主题能否紧密结合, 避免选材有意思但与主题脱节?<br>5. 是否用生活中真实且富有意义的素材, 支撑文章的立意与情感? |
| Rhythm and Overall Fluency   | 文字节奏与整体流畅度   |
| 1. Does the author vary sentence length to reflect the subject's characteristics or emotional pacing?<br>2. Are short, simple sentences and longer parallel constructions used judiciously to create rhythm and flow?<br>3. Are rhetorical devices—such as metaphor, personification, parallelism, and synesthesia—employed effectively to bring language to life?<br>4. Are these devices used fittingly and naturally, without feeling forced or over-embellished? | 1. 是否根据描写对象的特点或者情感节奏, 灵活运用长短句进行变化?<br>2. 是否适度使用散句、骈句, 让文章富有节奏感、韵律感?<br>3. 是否灵活运用比喻、拟人、排比、通感等修辞手法, 使语言生动形象 ②修辞手法是否用得贴切、自然、避免矫揉造作或过度堆砌?  |

Figure 8: Multi-traits and sub-questions of Descriptive Essay.

|  |  |
|--|--|
| Clarity of Topic and Purpose   | 主题与目的明确性   |
| <p>1. Is the topic of the expository essay clearly identifiable and distinct?</p> <p>2. Does the essay explicitly introduce the subject or principle it seeks to explain at the beginning or clearly throughout, helping readers grasp the primary purpose?</p>  | <p>1. 说明文主题是否鲜明且容易辨识;</p> <p>2. 是否在开头或者整体行文中清晰指出了要说明的对象或原理, 让读者明了文章的主要目的?</p>  |
| Scientific Accuracy and Credibility of Content   | 内容的科学性和准确性   |
| <p>Are the data, facts, and theories cited accurate, and are their sources reliable?</p> <p>Does the essay systematically introduce the subject, covering essential background information and primary aspects?</p> <p>If the essay uses classification or step-by-step explanations, does it include all core points without omitting critical information?</p> <p>Are professional or abstract concepts sufficiently explained or simplified to be easily understandable by readers?</p> <p>Does the essay maintain academic rigor without becoming overly obscure or difficult to comprehend?</p>   | <p>1. 引用的数据、事实、理论是否准确, 是否能保证来源的可信度?</p> <p>2. 是否对说明对象做到了系统化的介绍, 包括必要的背景和主要的方面? 如果是分类型说明或分步骤说明, 是否涵盖了核心要点, 避免一漏关键信息?</p> <p>3. 对于专业或抽象的概念, 是否有足够的解释或通俗化的阐释, 让读者易于理解?</p> <p>4. 是否在保持学术严谨的同时, 避免过晦涩难懂?</p>                     |
| Logical Structure and Coherence  | 逻辑结构与条理性   |
| <p>1. Is the structure (e.g., chronological, spatial, logical, categorical) chosen appropriately for the content?</p> <p>2. Are transitions between paragraphs and sections smooth and logical?</p> <p>3. Does the essay highlight key points clearly, adequately summarizing or briefly addressing secondary details?</p> <p>4. Are explanations presented progressively from simple to complex, or from general to specific, facilitating deeper understanding as readers acquire foundational information?</p> <p>5. For complex processes or procedures, are they presented in a logical order that enhances practicality and memorability?</p>  | <p>1. 说明文常见的结构方式包括时间顺序、空间顺序、逻辑顺序、分类分述等, 是否选择了最合适的结构? 2. 段落与段落之间、章节与章节之间过渡是否顺畅?</p> <p>3. 是否能突出主要内容, 并对次要部分进行适度简化或补充说明?</p> <p>4. 说明重点是否由浅入深、由表及里, 帮助读者在逐步掌握基础信息后在理解更深入的只是?</p> <p>5. 如果有复杂过程或步骤, 是否按照可操作性或便于记忆的顺序分段进行阐述?</p> |
| Clarity and Appropriateness of Language  | 语言表达的清晰度与规范性   |
| <p>1. Given that the core aim of expository writing is efficient information delivery, is the language concise, avoiding unnecessary embellishments or overly decorative phrasing?</p> <p>2. Are sentence structures clear and easily understood, avoiding overly complex or lengthy sentences?</p> <p>3. Is word choice precise, avoiding ambiguity, overly casual expressions, or colloquial language?</p> <p>4. Are technical terms or concepts sufficiently explained, ensuring clarity and preventing reader confusion?</p>   | <p>1. 说明文的核心在于信息的高效传递, 语言是否尽量简洁, 避免过渡修饰或华丽辞藻?</p> <p>2. 句式是否易于理解, 避免复杂或者冗长的句子?</p> <p>3. 用词是否准确, 是否避免模棱两可的表述以及口语化、随意化的表达? 专业术语或概念解释是否到位, 读者阅读时不会产生歧义?</p>  |
| Practicality and Relevance   | 说明方法的多样性与合理性   |
| <p>1. Does the essay appropriately use various explanatory methods (e.g., definition, classification, examples, numerical illustration, comparison)?</p> <p>2. Do different explanatory methods complement one another effectively?</p> <p>3. Does the essay transform abstract concepts or complex techniques into understandable forms through concrete examples or relatable analogies?</p> <p>4. Does the essay avoid using irrelevant or forced examples that might confuse or distract readers?</p> <p>5. Does the essay appropriately cite authoritative sources, data, or expert opinions, clearly indicating the source or providing simple annotations to ensure transparency?</p>                           | <p>1. 是否恰当地使用定义说明、分类说明、举例说明、数字说明、比较说明等多种说明方法? 不同说明方法能否相互补充?</p> <p>2. 是否通过具体的事例、生活化的比喻, 把抽象概念或复杂技术形象化?</p> <p>3. 是否避免使用与主题无关或牵强的例子, 反而使读者感到困惑?</p> <p>4. 是否引用权威数据、数据或专家观点, 并有标明出处或进行简单注释, 保证信息透明度?</p>                       |
| Rhythm and Overall Fluency   | 实用性及针对性  |
| <p>1. Do the content and language of the essay align with the knowledge level and interests of the intended audience?</p> <p>2. If the essay is meant as popular science, instructions, or user guidelines, can readers readily practice, perform experiments, or directly apply the instructions? For purely theoretical content, does it provide sufficient examples or application scenarios to enhance comprehension or memorability?</p> <p>3. Is the content closely connected with real-life contexts or societal needs, enabling readers to appreciate its practical significance?</p> <p>4. Is the content insightful or extendable, suggesting directions for further learning or practical application?</p> | <p>1. 说明文中的内容和语言是否与目标读者群的知识水平、兴趣点相符合?</p> <p>2. 如果是科普、操作指南或使用说明, 读者是否可根据文本进行演练、实践或实验? 如果是纯理论性的说明, 是否提供了足够的事例或场景应用, 使理论更好地被理解或记忆?</p> <p>3. 内容是否与现实生活或社会需求紧密结合, 让读者在阅读后感到有现实意义?</p> <p>4. 内容是否有启发性或可拓展性, 给出进一步学习或应用的方向?</p>   |

Figure 9: Multi-traits and sub-questions of Expository Essay.

### Annotation Guideline (Chinese)

数据标注指导：文章质量评估

#### 一、数据说明

包含以下字段的数据文件：

1. query: 任务要求描述
2. writing\_1: 第一篇待比较文章
3. writing\_2: 第二篇待比较文章
4. score: 请在此填写比较结果 (1/2/tie)

#### 二、评估任务

基于query的具体要求，系统比较writing\_1和writing\_2的文本质量，并：

1. 给出评分 (1表示writing\_1更好，2表示writing\_2更好，tie表示质量相当)
2. 简要说明评分理由

#### 三、质量评估标准

##### (一) 基础要求

##### 1. 符合性要求：

- 遵循query中规定的内容要求、文体规范和目标读者/场景
- 字数要求大概估计

##### 2. 文体分类，按照文体特点进行评价：

- 议论文 (argumentative)
- 记叙文 (narrative)
- 描写文/散文 (descriptive)
- 说明文 (expository)

注意事项 1. 评分必须基于query的具体要求 2. 理由说明需简明扼要，突出关键差异点 3. 忽略格式

### Annotation Guideline (English Translation Version)

#### I. Data Description

The data file contains the following fields:

**query:** Description of task requirements

**writing\_1:** First article to be compared

**writing\_2:** Second article to be compared

**score:** Comparison result (1/2/tie)

#### II. Evaluation Task

Based on the specific requirements in the **query**, systematically compare the text quality of **writing\_1** and **writing\_2**, and:

**Provide a rating** (1 = **writing\_1** is better, 2 = **writing\_2** is better, tie = equal quality)

**Briefly explain the rationale** for the rating

#### III. Quality Evaluation Criteria

##### (1) Basic Requirements

##### **Compliance Requirements:**

Follow the **content requirements, writing style norms, and target**

**audience/scenario** specified in the query

Approximate word count estimation

##### **Genre Classification** (Evaluate based on stylistic characteristics):

**Argumentative** (persuasive essays)

**Narrative** (storytelling)

**Descriptive** (vivid depictions/prose)

**Expository** (informative/explanatory writing)

##### Key Notes:

**Ratings must be based on the specific requirements in the query**

**Explanations should be concise, highlighting key differences**

**Ignore formatting issues** (focus on content quality only)

Figure 10: Annotation Guideline for Human Agreement Study.



## Essay from Llama-3.1-70B-Instruct

**Prompt:**以人生为主题写一篇800-1200字的**议论文**，探讨人生的意义、挑战和美好之处，**适合高中生阅读**，要求**使用优美的语言和有力的论据，结构清晰，分为引言、主体和结论三部分**。

**Writing:**人生是一场充满挑战与机遇的旅程，我们每个人都在努力寻找着人生的意义。人生的意义是什么？是财富和权力？是爱情和友谊？抑或是梦想和成就？每个人对人生的意义有着不同的理解和追求。然而，无论我们如何定义人生的意义，挑战和美好之处始终是人生旅程中不可或缺的两大元素。

人生的挑战无处不在，从我们出生开始，就面临着来自各方面的考验。生活的艰辛、学习的压力、情感的波动，种种困难都可能让我们感到沮丧和无助。然而，正是这些挑战，让我们变得更加坚强和成熟。我们可以从失败中吸取教训，变得更加谦卑和努力。我们可以从挫折中找到动力，变得更加勇敢和坚韧。人生的挑战是我们成长的阶梯，是我们实现梦想的动力。

人生的美好之处同样令人着迷。生活中的每一个瞬间，都可能是美好的。一个阳光明媚的早晨，一场春雨绵绵的夜晚，一次难忘的旅行，一段真挚的友谊，都是人生美好之处的体现。人生的美好之处不仅仅在于大事，更在于细节。我们可以从一杯热咖啡的芬芳中感受到生活的温暖，从一本好书的知識中感受到求知的快乐，从一场美丽的日落中感受到大自然的壮丽。人生的美好之处是我们生活的动力，是我们前行的方向。

人生的意义是什么？答案在于每个人内心的追求。无论我们如何定义人生的意义，挑战和美好之处始终是人生旅程中不可或缺的两大元素。我们可以从挑战中变得更加坚强和成熟，可以从美好之处中找到生活的动力和方向。人生是一场旅程，我们每个人都是自己的主角。让我们勇敢地面对挑战，珍惜每一个美好的瞬间，找到属于自己的人生意义。

## Formatted Evaluation Example from DeepSeek-R1 Total Score: 6.3

1. 文章中心观点较为明确，提出人生意义与挑战、美好相关，但未明确界定三者关系（如**挑战如何定义意义**），存在模糊性
2. 中心论点可提炼为“人生意义在于挑战与美好共存”，但需从文本中二次推导，**未直接呈现核心判断句**
3. 引言段通过设问引出论点，但最终论断使用“不可或缺的两大元素”的并列式表达，未形成明确的价值判断
4. 仅停留在现象描述层面，缺乏对存在主义、斯多葛学派等哲学视角的引入，**对“意义构建机制”的探讨深度不足**
5. 对“压力转化为动力”等**常见观点复述较多**，缺少如“苦难悖论”“异化劳动”等新颖视角的洞察
6. 论据与“挑战促进成长”“美好激励前行”的关联成立，但**“春雨夜晚”等例子与论点逻辑连接较弱**
7. 仅使用6个泛化例证（学习压力、旅行等），缺乏具体人物/历史事件等实证支撑
8. 失败中吸取教训“等论据未说明具体过程，如未对比失败前后的转变机制
9. 全部采用日常生活案例，缺少跨文化（如武士道精神）或经典文本（如《活着》）等代表性例证
10. 未出现数据或文献引用，真实性无法验证但无硬伤
11. 论据来源限于个人体验，缺少学术研究、历史案例、文学典故等多元维度
12. 使用举例论证（咖啡、日落）和因果论证（挑战→坚强），但缺乏对比论证（如顺境vs逆境人生）和引用论证
13. 春雨例证过渡突兀，**未能自然衔接“美好产生动力”的论证链条**
14. 统计数据完全缺失，他人观点仅隐含在“每个人”等泛指表达中
15. 未提及“享乐主义”等对立观点，**缺乏辩证性思考**
16. 单维度论证美好与挑战的正面性，未讨论过度挑战可能带来的负面影响
17. 举例与因果论证各自独立，未形成“挑战案例→成长机制→意义构建”的协同论证
18. 主体部分“挑战”与“美好”两个小论点均服务总论点，**但第三段“意义在于内心追求”产生主题偏移**
19. 从“挑战”到“美好”采用并列结构，未构建“挑战孕育美好”或“美好化解挑战”的递进关系
20. 每个段落均有2-3个例证支撑，但存在例证密度不均问题（如压力段仅1例）
21. 然而，正是这些挑战...“等过渡句有效，但段落间缺少“不仅如此”“更重要的是”等递进连接词
22. 用整段论证“美好在细节”，但其在意义构建中的重要性未得到相应篇幅的强度匹配
23. “阶梯”“动力”等隐喻贴切，但“春雨绵绵的夜晚”等描写稍显文艺，与议论文体略有冲突
24. “我们可以从...中感受到...”的排比句式强化逻辑，**但部分结论句缺乏推理连接词（因此/由此可见）**
25. 陈述句占比90%以上，缺少反问（难道不是...）等句式变化
26. 未使用专业术语，符合高中生认知水平
27. 排比修辞增强气势，但“主角”等比喻未能深化论证深度
28. 结尾复述“挑战与美好不可或缺”，与引言设问形成闭环
29. 未升华至“意义在于创造价值”等新认知，**停留在论点归纳层面**

Figure 11: Low-quality Essay and Evaluation Example.

Essay from GPT-3.5-turbo

**Prompt:**以人生为主题写一篇**议论文**，探讨人生的意义、挑战和美好之处，**适合高中生阅读**，要求**使用优美的语言和有力的论据**，**结构清晰**，分为**引言、主体和结论三部分**。

**Writing:**人生，是一场漫长的旅程，充满着无数的意义、挑战和美好之处。在这个世界上，每个人都在不断地探寻着自己存在的意义，面对着各种考验和挑战，同时也感受着生活中那些美好的瞬间。

引言

人生，是一场奇妙的冒险，每一个人都在这个旅程中寻找着自己的方向。我们来到这个世界上，不仅仅是为了生存，更是为了追寻那份内心深处的意义。在这个过程中，我们会遇到各种各样的挑战，但也会经历许多美好的时刻。人生的意义，就在于我们如何应对挑战，感受生活的美好。

主体

人生的意义，在于我们对世界的贡献和影响。每个人都有自己独特的才华和潜力，可以为这个世界带来改变和进步。无论是成为一名医生拯救生命，还是成为一名艺术家创造美丽的作品，每个人都可以在自己的领域里发挥作用，为社会做出贡献。人生的意义，就在于我们能否找到自己的使命，努力实现自己的梦想。

人生中的挑战，是我们成长和进步的机会。在面对困难和挫折时，我们会变得更加坚强和勇敢。挑战可以激发我们内心的潜能，让我们超越自己，变得更加优秀和成熟。正是因为挑战的存在，我们才能不断地突破自己的局限，实现自己的价值。人生的挑战，让我们变得更加坚强和有力。

人生中的美好，隐藏在生活的点滴中。每一个微笑、每一次成功、每一次相遇，都是生活赋予我们的礼物。美好不一定是璀璨的光芒，而可能是平凡中的温暖和感动。当我们用心感受生活的每一个瞬间，我们就会发现生活中有无数的美好和奇迹。人生的美好，就在于我们懂得珍惜眼前的幸福和快乐。

结论

人生，是一场充满意义、挑战和美好的旅程。在这个旅程中，我们会不断地成长和进步，感受到生活的种种风景。无论是面对困难还是享受幸福，我们都要珍惜每一个瞬间，努力追寻自己的梦想。因为只有在这样的过程中，我们才能发现人生的真正意义，挑战和美好之处。愿每个人都能在这场旅程中找到属于自己的幸福和意义。

Formatted Evaluation Example from DeepSeek-R1 **Total Score: 7.0**

- 文章中心观点较为明确，但**表述略显泛泛，未深入探讨深层思考**。例如，'人生的意义在于贡献'未结合社会结构或哲学理论展开。
- 中心论点易于提炼为'人生的意义在于贡献、挑战促进成长、美好在于点滴'，但过于直白，缺乏新意。
- 开篇首段即提出中心论点，符合'引言明确'要求。
- 仅停留在个人价值层面，未涉及如存在主义、集体主义等深层理论，**缺乏文化纵深感**。
- 对挑战与成长的关联有基本洞察，但**未触及现代青年的价值迷茫等现实痛点**。
- 论据与'贡献'挑战'美好'的关联明确，但医生/艺术家案例与高中生生活距离较远。"
- 仅使用职业案例和抽象描述，缺乏具体历史事例（如曼德拉）或文学典故支撑。医生拯救生命'等论据**未说明具体案例**，详实度不足。案例集中于职业选择，**未覆盖家庭、友谊等其他人生维度**。
- 未引用具体数据或文献，真实性无法验证但无逻辑硬伤。
- 论据来源仅限个人经历观察，缺乏跨学科引用。
- 使用举例论证（医生/艺术家）和对比论证（挑战vs美好），但缺少引用论证（名人名言）等手法
- 职业案例论证自然，但'美好在点滴'部分转为抒情，论证断层。
- 未使用统计数据或他人观点，此项不适用。
- 完全未提及对立观点（如虚无主义对人生意义的质疑）
- 单方面论证，未处理反方观点。举例与对比论证未能形成互补效应，例如未用数据佐证案例代表性。
- 三个主体段分别对应中心论点，但'**挑战促进成长'与'美好在点滴'之间逻辑断裂**。
- 采用并列结构而非递进结构，如未说明挑战如何导向美好。
- 每个段落均有案例支撑，但'美好'段以抒情代替论据。
- 段落间用'人生中的...'句式过渡，机械但有效。
- 贡献段用职业案例论证，与论点重要性匹配度不足。
- 用词规范但平淡，如重复使用'旅程'等隐喻，未出现'异化'自我实现'等精准术语**。
- 逻辑推进清晰，但'美好'段突然转为感性描写破坏理性基调。
- 以陈述句为主，缺少反问、排比等句式变化。
- 未使用专业术语，此项不适用。
- '璀璨的光芒'等比喻增强感染力，但未服务核心论证。
- 结尾用'旅程'呼应开篇'冒险'，结构完整。
- 总结三个分论点但未升华（如联系人类共同命运）。
- 呼吁'找到幸福'流于表面，未提出具体行动指引（如终身学习）。

Figure 12: Median-quality Essay and Evaluation Example.

Essay from DeepSeek-Chat

**Prompt:**以人生为主题写一篇800-1200字的议论文，探讨人生的意义、挑战 and 美好之处，**适合高中生阅读**，要求使用优美的语言和有力的论据，结构清晰，分为引言、主体和结论三部分。

Writing: \*\*人生的意义、挑战与美好\*\*\n\n引言\n\n人生，如同一场漫长的旅程，充满了未知与变数。我们每个人都在这场旅程中寻找属于自己的意义，面对各种挑战，同时也享受着其中的美好。对于高中生而言，正值人生的关键阶段，理解人生的意义、勇敢面对挑战、发现生活的美好，不仅有助于个人的成长，更能为未来的生活奠定坚实的基础。本文将从人生的意义、挑战与美好三个方面展开探讨，旨在为高中生提供一些思考和启示。

主体\n\n一、人生的意义\n\n人生的意义是什么？这是一个古老而永恒的问题。对于高中生而言，这个问题或许显得过于宏大，但正是这个问题的思考，能够帮助我们更好地理解自己，明确未来的方向。人生的意义并非一成不变，它随着我们的成长和经历而不断演变。在古希腊哲学家亚里士多德看来，人生的意义在于追求幸福，而幸福并非简单的快乐，而是通过实现自己的潜能和追求美德而获得的满足感。对于高中生而言，人生的意义或许在于通过学习、探索和成长，找到自己的兴趣和热情，为未来的生活做好准备。

然而，人生的意义并非仅仅局限于个人的追求。法国作家阿尔贝·加缪在《西西弗神话》中提出，人生虽然看似荒诞，但正是通过面对荒诞、接受挑战，我们才能找到属于自己的意义。对于高中生而言，人生的意义也在于为他人、为社会做出贡献。无论是通过志愿服务、参与公益活动，还是通过努力学习、为未来的职业做好准备，我们都能在帮助他人的过程中找到人生的意义。

二、人生的挑战\n\n人生充满了挑战，这些挑战既是成长的契机，也是我们实现自我价值的必经之路。对于高中生而言，学业压力、人际关系、自我认同等问题都是常见的挑战。面对这些挑战，我们难免会感到困惑、焦虑甚至无助。然而，正是这些挑战，塑造了我们的性格，锻炼了我们的意志。

美国心理学家卡尔·罗杰斯曾提出，人生的挑战在于不断追求自我实现，而自我实现的过程往往伴随着痛苦和挣扎。对于高中生而言，学业压力或许是最直接的挑战。面对繁重的课业和激烈的竞争，我们可能会感到疲惫和无力。然而，正是通过克服这些困难，我们才能不断提升自己的能力，为未来的生活做好准备。

此外，人际关系也是高中生面临的重要挑战。在与同学、老师、家人的相处中，我们可能会遇到误解、冲突甚至背叛。然而，正是通过这些经历，我们学会了如何与他人沟通、如何理解他人的感受，如何在复杂的人际关系中保持自我。这些技能不仅有助于我们当前的成长，也将为未来的生活提供宝贵的经验。

三、人生的美好\n\n人生尽管充满了挑战，但其中也不乏美好之处。对于高中生而言，青春本身就是一种美好。在这个阶段，我们拥有无限的潜力和可能性，能够尽情地探索世界、追求梦想。无论是通过阅读、旅行、艺术创作，还是通过参与体育活动、社团活动，我们都能在青春中找到属于自己的快乐和满足。

此外，人生的美好还在于我们与他人之间的情感联系。无论是与家人、朋友还是恋人之间的感情，都是我们生活中不可或缺的一部分。英国作家C.S.刘易斯曾说过：“友谊是人生中最好的礼物之一。”对于高中生而言，与朋友之间的相互支持、理解和陪伴，不仅能够帮助我们度过难关，更能为我们的生活增添色彩。

最后，人生的美好还在于我们能够不断成长和进步。每一次的挑战和困难，都是我们成长的机会。通过不断努力，我们能够逐渐接近自己的目标，实现自己的梦想。这种成长的过程，本身就是一种美好。

结论\n\n人生是一场充满意义、挑战与美好的旅程。对于高中生而言，理解人生的意义、勇敢面对挑战、发现生活的美好，不仅有助于个人的成长，更能为未来的生活奠定坚实的基础。人生的意义在于追求幸福、实现自我价值，同时也在于为他人、为社会做出贡献。人生的挑战虽然艰难，但正是这些挑战塑造了我们的性格，锻炼了我们的意志。人生的美好则在于青春、情感联系以及不断成长的过程。愿每一位高中生都能在这场旅程中找到属于自己的意义，勇敢面对挑战，享受生活的美好。

Formatted Evaluation Example from DeepSeek-R1 **Total Score: 8.7**

- 文章中心观点清晰，明确探讨人生的意义、挑战 and 美好，**适合高中生阅读**。
- 中心观点易于提炼，引言和结论部分明确概括。
- 开篇直接提出中心论点，结构明确。
- 虽引用哲学观点，但对社会、文化的深层次探讨较浅，例如加缪的荒诞哲学未深入展开。
- 体现出对青少年成长的洞察，但**缺乏更具批判性的思想见解**。
- 论据与论点关联紧密，如亚里士多德对应意义，罗杰斯对应挑战。
- 每个论点至少一个论据，但学业压力等例子较为常见，**创新性不足**。
- 哲学引述较简略，未展开详细分析。
- 引用加缪、刘易斯等具有代表性。
- 数据来源真实可靠。
- 涵盖哲学、心理学、文学等多领域。
- 使用引用和举例论证，但缺乏对比论证。
- 论证方法运用自然，如用罗杰斯理论解释挑战。
- 统计数据未出现，但引用观点时有简要分析。
- 未提及对立观点，如未讨论‘享乐主义’等不同人生观。
- 单一视角论证，未兼顾不同观点。
- 方法间协同效应一般，如哲学与心理学论证未形成互补。
- 小论点均围绕中心论点展开。
- 意义→挑战→美好为并列结构，递进性较弱。
- 每个段落有对应论据支撑。
- 段落间有‘此外’‘最后’等过渡词。
- 学业压力段落论证强度匹配论点重要性。
- 用词优美，如‘荒诞’‘自我实现’准确。
- 逻辑连贯，但部分语句如‘青春本身就是一种美好’稍显主观。
- 句式多样，长短句结合。
- 比喻（人生如旅程）增强感染力。
- 结尾呼应开头‘旅程’比喻，结构完整。
- 总结意义、挑战、美好而非简单重复。
- 提出‘找到属于自己的意义’的呼吁，具启发性。

Figure 13: High-quality Essay and Evaluation Example.