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

Fan Gao[1, 2], Dongyuan Li[2], Ding Xia[2], Fei Mi[1], Yasheng Wang[1], Lifeng Shang[1], Baojun Wang[1]
[1]Huawei Technologies Ltd.
[2]The University of Tokyo
fangao0802@gmail.com

Abstract

Prompt-based essay writing is an effective and common way to assess students' critical thinking skills. Recent work has evaluated the impressive capabilities of Large Language Models (LLMs) on this task. However, most studies focus primarily on English. Those examining LLMs' performance in Chinese often rely on coarse-grained text quality metrics, overlooking the structural and rhetorical complexities of Chinese essays, particularly across diverse genres. We therefore propose ESSAYBENCH, a multi-genre benchmark specifically designed for Chinese essay writing, along with a fine-grained, genre-specific scoring framework that hierarchically aggregates scores to better align with human preferences. The dataset comprises 728 real-world prompts across four major genres (Argumentative, Narrative, Descriptive, and Expository), and includes both Open-Ended and Constrained types. Our evaluation protocol is validated through a comprehensive human agreement study. The results show that our protocol aligns well with human judgments, achieving a highest Spearman's correlation of 0.816 and outperforming coarse-grained evaluation methods by an average of 8.6%. Finally, we benchmark 15 large LLMs, analyzing their strengths and limitations across genres and instruction types. We believe ESSAYBENCH offers a more reliable framework for evaluating Chinese essay generation and provides valuable insights for improving LLMs in this domain.

Introduction

Prompt-based writing is widely used to assess students' critical thinking and reasoning skills (Dunham 1997), representing a natural evaluation setting for assessing the text-generation capabilities of Large Language Models (LLMs) (Brown et al. 2020; Touvron et al. 2023). While recent studies have introduced datasets and evaluated LLMs for this task (Almegren et al. 2024; Sari et al. 2025), most focus on English. In contrast, research on Chinese remains limited and lacks robust evaluation frameworks aligned with human preferences. This gap limits practical deployment, particularly in educational contexts, where high-quality model-generated essays could exercise students' reading and writing abilities (Xiao et al. 2023; Woo et al. 2023).

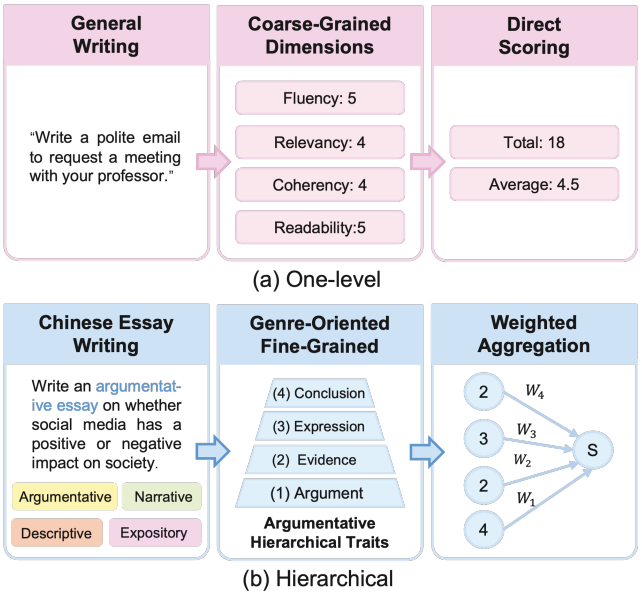

Figure 1: Comparison between coarse-grained evaluation methods (a) and our fine-grained and genre-oriented framework for ESSAYBENCH (b).

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 (Chen et al. 2023; Hashemi et al. 2024; Fu et al. 2024), while the other employs pair-wise comparisons (Liu et al. 2024b; Chiang et al. 2024). 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 specific weaknesses. Second, pairwise comparisons typically depend on human-annotated references, limiting their adaptability and scalability in evaluating essay writing.

Extended version: https://arxiv.org/abs/2506.02596

Benchmark	Num.	Dataset Composition			Evaluation Method		
		Domain	Language	Constraints	LLM	Fine Grained 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
ESSAYBENCH (Ours)	**728**	**Essay Writing**	**ZH**	✓	✓	✓	**Weighted**

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

Moreover, as introduced in Table 1, although recent benchmarks like *AlignBench* (Liu et al. 2024b) and *WritingBench* (Wu et al. 2025) have turned attention to assess LLM's capability in Chinese writing, their evaluation methods fail to capture the unique characteristics of essays like logographic characters, complex constructions, and rhetorical traditions. 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, Huang, and Chang 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: *Argumentative*, *Narrative*, *Descriptive*, and *Expository* writing. To ensure alignment with real-world 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. Furthermore, to overcome the limitations of existing evaluation methods for Chinese essay writing, we propose a fine-grained and genre-oriented evaluation framework (see 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. To account for the hierarchical nature of these traits, we further introduce a dependency-weighted score aggregation strategy to better capture the writing quality.

We conduct two key experiments. First, to assess our framework's 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. 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.

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, ensuring practical relevance and applicability.
- 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.

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 underexplored 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

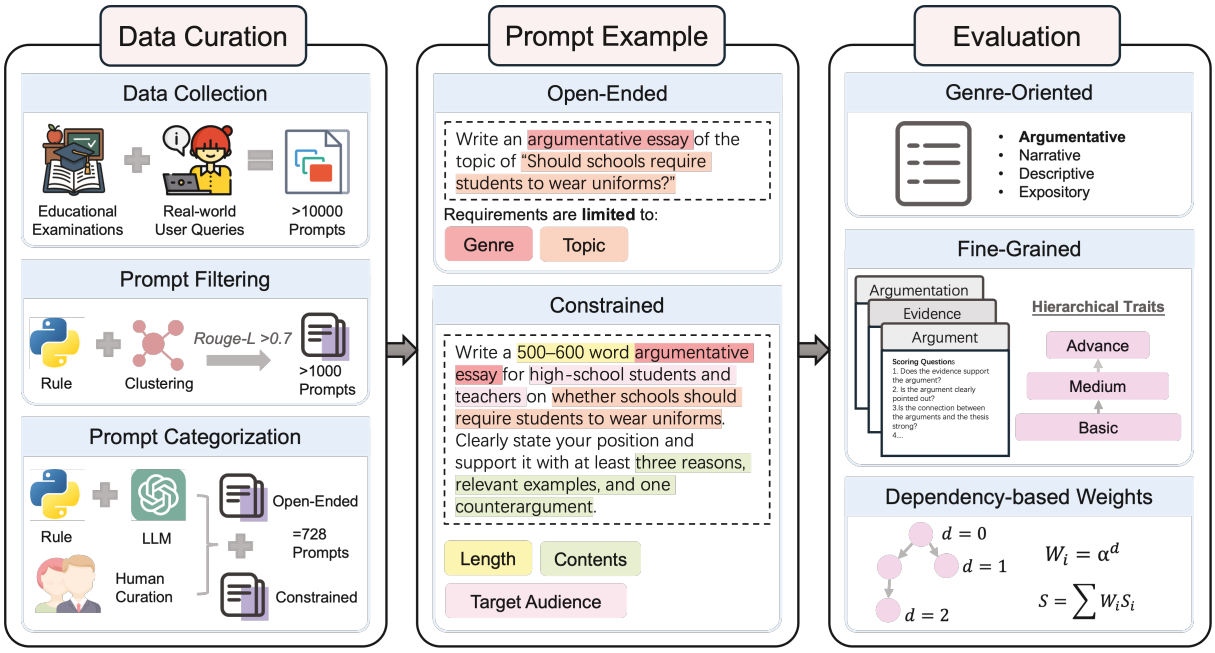

Figure 2: Overview of the ESSAYBENCH dataset curation, representative prompts, and the evaluation framework.

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, Briscoe, and Medlock 2011; Persing and Ng 2013) to neural, trait-specific models (Taghipour and Ng 2016; Uto, Xie, and Ueno 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.

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.

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

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.

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.

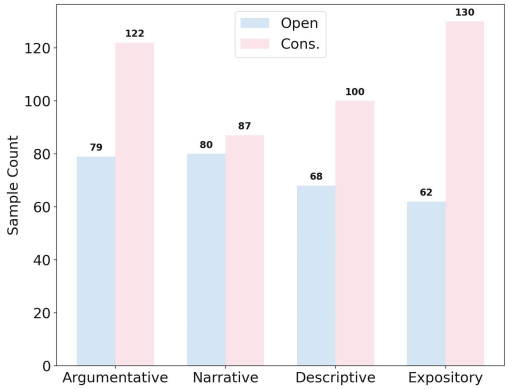

Figure 3: Dataset Statistics. Note that **Open** denotes Open-Ended sets, **Cons.** refers to Constrained sets.

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 popularity, 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) and the Chinese high school curriculum standards (MOE 2020), we refine and develop genre-specific evaluation traits that align with Chinese writing conventions. For each genre, we summarize these guidelines into six evaluation dimensions that range from basic to advanced requirements. For instance, in the case of argumentative essays, the dimensions include *Argument*, *Evidence*, *Argumentation Methods*, *Logical Development*, *Expression*, and *Conclusion*. The detailed definitions of these dimensions are provided in the Appendix. 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, as detailed in the Appendix. We adopt the Chain-of-Thought (CoT) (Wei et al. 2023) prompting 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

Methods	Overall		Argumentative		Narrative		Descriptive		Expository	
	ρ	τ	ρ	τ	ρ	τ	τ	ρ	ρ	τ
	DeepSeek-V3									
Align-Score	**0.674**	**0.599**	**0.744**	**0.674**	0.635	**0.559**	0.656	0.580	**0.656**	**0.578**
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	**0.648**	0.518	**0.796**	**0.676**	0.554	0.458
	GPT-4o									
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	**0.733**	**0.627**	**0.754**	**0.662**	**0.773**	**0.658**	**0.700**	**0.594**	**0.707**	**0.601**
	DeepSeek-R1									
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	**0.816**	**0.704**	**0.795**	**0.673**	**0.838**	**0.724**	**0.839**	**0.731**	**0.791**	**0.690**

Table 2: Comparison of human agreement evaluation across different scoring methods on sampled data. ρ refers to Spearman's ρ, τ denotes the Kendall's τ, while *WT.* represents the dependency-based weights.

(W_t) are computed using 1, with the hyperparameter α controlling the importance of basic and advanced levels. The final score is a weighted sum of all trait scores.

$$W_t = \alpha^d. \tag{1}$$

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.

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 et al. 2024).

Human Annotation. We 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 based on the same prompt to three annotators. Annotators label each essay pair by selecting whether essay A is better, essay B is better or if there is a tie. Each annotator is assigned approximately 70 sample pairs per day within an 8-hour work schedule. Over five days, the annotation process results in a total of 5,040 fully labeled sample pairs. We achieve a high level of human agreeement in our annotations, with at least two annotators agreeing on 95% of the data. Specifically, the full agreement rate (all three annotators agreeing) is 58.1%, and the partial agreement rate (two annotators agreeing) is 36.9%. Finally, to effectively leverage the annotations, we use a majority voting strategy: each essay receive one point each time it is judged superior to its counterpart by an annotator, enabling us to derive a clear ranking of models for each essay prompt.

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 evaluation method works, we employ three LLMs as judges, including DeepSeek-V3 (DeepSeek-AI 2025b), DeepSeek-R1 (DeepSeek-AI 2025a) and GPT-4o (OpenAI et al. 2024) to assign scores $1\sim10$ to each sub-question within every evaluation trait. Each model analyzes all sub-questions in a single turn. Specifically, we convert the scores into a model ranking. In all experiments, the temperature is set to 0.2, and the parameter α is fixed to 3.

Ranking Agreement

To assess the ranking agreement, we use **Spearman's Rank Correlation** (Spearman 1904) and **Kendall's** τ (Kendall 1938), which capture monotonic relationships between rankings. As shown in 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 ρ and Kendall's τ. 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 annota-

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

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.

tions, 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.

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 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 the high- and medium-quality data.

The sensitivity analysis in 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.

Benchmarking

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 4. 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 (see Figure 4). This disparity likely stems from the inherent characteristics of these genres: argumentative and expository essays emphasize structural

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 et al. 2024)	74.2	74.8	76.9	72.8	72.4	70.5	71.7	75.8	76.7
GPT-4o-mini (OpenAI et al. 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 4: 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.

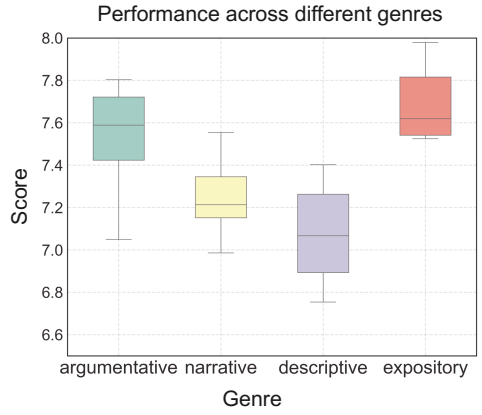

Figure 4: Comparison of LLMs' Performance across genres.

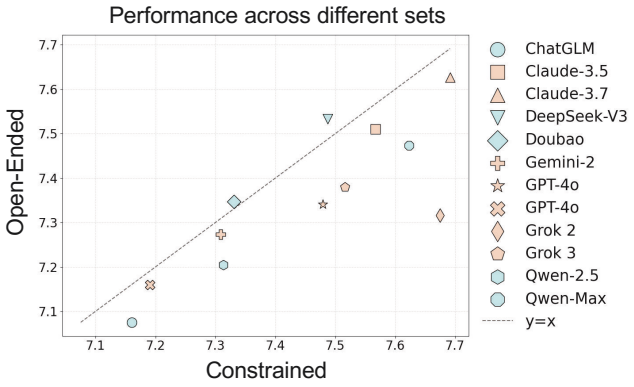

Figure 5: Comparison of Performance by Sets.

coherence, logical reasoning, and clear topic development, where LLMs typically excel. In contrast, narrative and descriptive 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 5. 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.

Conclusion

In this work, we present ESSAYBENCH, the first comprehensive benchmark for evaluating the capabilities of LLMs in the Chinese essay generation. 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 real-world contexts.

References

Almegren, A.; Mahdi, H. S.; Hazaea, A. N.; Ali, J. K.; and Almegren, R. M. 2024. Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International*, 1–16.

Anthropic. 2024. Introducing Claude 3.5 Sonnet. Accessed: 2025-05-19.

Anthropic. 2025. Claude 3.7 Sonnet. Accessed: 2025-05-19.

Attali, Y.; and Burstein, J. 2006. Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Bai, Y.; Zhang, J.; Lv, X.; Zheng, L.; Zhu, S.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2024. LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs. *arXiv preprint arXiv:2408.07055*.

Balepur, N.; Huang, J.; and Chang, K. 2023. Expository Text Generation: Imitate, Retrieve, Paraphrase. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11896–11919. Singapore: Association for Computational Linguistics.

Blanchard, D.; Tetreault, J.; Higgins, D.; Cahill, A.; and Chodorow, M. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2): i–15.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chadbourne, R. M. 1983. A puzzling literary genre: comparative views of the essay. *Comparative literature studies*, 20(2): 133–153.

Chen, G. H.; Chen, S.; Liu, Z.; Jiang, F.; and Wang, B. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–8327. Miami, Florida, USA: Association for Computational Linguistics.

Chen, Y.; Wang, R.; Jiang, H.; Shi, S.; and Xu, R. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. arXiv:2304.00723.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M.; Gonzalez, J. E.; et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Chu, S.; Kim, J. W.; Wong, B.; and Yi, M. Y. 2025. Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, 5796–5814. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7.

DeepSeek-AI. 2025a. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

DeepSeek-AI. 2025b. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Dikli, S. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Doubao Team. 2025. Doubao-1.5-pro: A High-Performance Sparse MoE Large Language Model. Accessed: 2025-05-19.

Dunham, R. A. 1997. Assessing EFL Student Progress in Critical Thinking With the Ennis-Weir Critical Thinking Essay Test! *JALT Journal*, 19(1): 43.

Fu, W.; Wei, B.; Hu, J.; Cai, Z.; and Liu, J. 2024. QGEval: Benchmarking Multi-dimensional Evaluation for Question Generation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11783–11803. Miami, Florida, USA: Association for Computational Linguistics.

Gao, F.; Jiang, H.; Yang, R.; Zeng, Q.; Lu, J.; Blum, M.; She, T.; Jiang, Y.; and Li, I. 2024. Evaluating large language models on Wikipedia-style survey generation. In *Findings of the Association for Computational Linguistics ACL 2024*, 5405–5418.

Gemini. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

GLM, T. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; Wang, S.; Zhang, K.; Wang, Y.; Gao, W.; Ni, L.; and Guo, J. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594.

Hamner, B.; Morgan, J.; lynnvandev; Shermis, M.; and Ark, T. V. 2012. The Hewlett Foundation: Automated Essay Scoring. https://kaggle.com/competitions/asap-aes. Kaggle.

Hashemi, H.; Eisner, J.; Rosset, C.; Van Durme, B.; and Kedzie, C. 2024. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13806–13834. Bangkok, Thailand: Association for Computational Linguistics.

Hastie, T.; Tibshirani, R.; Friedman, J. H.; and Friedman, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In *Advances in Neural Information Processing Systems*.

Islam, P.; Kannappan, A.; Kiela, D.; Qian, R.; Scherrer, N.; and Vidgen, B. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Jiang, Y.; Wang, Y.; Zeng, X.; Zhong, W.; Li, L.; Mi, F.; Shang, L.; Jiang, X.; Liu, Q.; and Wang, W. 2024. FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4667–4688.

Kendall, M. G. 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1–2): 81–89.

Kim, S.; Suk, J.; Cho, J. Y.; Longpre, S.; Kim, C.; Yoon, D.; Son, G.; Cho, Y.; Shafayat, S.; Baek, J.; Park, S. H.; Hwang, H.; Jo, J.; Cho, H.; Shin, H.; Lee, S.; Oh, H.; Lee, N.; Ho, N.; Joo, S. J.; Ko, M.; Lee, Y.; Chae, H.; Shin, J.; Jang, J.; Ye, S.; Lin, B. Y.; Welleck, S.; Neubig, G.; Lee, M.; Lee, K.; and Seo, M. 2025. The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5877–5919. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Krishna, S.; Krishna, K.; Mohananey, A.; Schwarcz, S.; Stambler, A.; Upadhyay, S.; and Faruqui, M. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*.

Lee, S.; Cai, Y.; Meng, D.; Wang, Z.; and Wu, Y. 2024. Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring. *arXiv preprint arXiv:2404.04941*.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; Shu, K.; Cheng, L.; and Liu, H. 2025. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv:2411.16594.

Li, S.; and Ng, V. 2024. Automated Essay Scoring: A Reflection on the State of the Art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17876–17888.

Ling, Z.; Liu, K.; Yan, K.; Yang, Y.; Lin, W.; Fan, T.-H.; Shen, L.; Du, Z.; and Chen, J. 2025. LongReason: A Synthetic Long-Context Reasoning Benchmark via Context Expansion. *arXiv preprint arXiv:2501.15089*.

Liu, N.; Chen, X.; Wu, H.; Sun, C.; Lan, M.; Wu, Y.; Bai, X.; Mao, S.; and Xia, Y. 2024a. CERD: A Comprehensive Chinese Rhetoric Dataset for Rhetorical Understanding and Generation in Essays. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6744–6759. Miami, Florida, USA: Association for Computational Linguistics.

Liu, X.; Lei, X.; Wang, S.; Huang, Y.; Feng, A.; Wen, B.; Cheng, J.; Ke, P.; Xu, Y.; Tam, W. L.; Zhang, X.; Sun, L.; Gu, X.; Wang, H.; Zhang, J.; Huang, M.; Dong, Y.; and Tang, J. 2024b. AlignBench: Benchmarking Chinese Alignment of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11621–11640. Bangkok, Thailand: Association for Computational Linguistics.

Liu, Y.; Fabbri, A.; Chen, J.; Zhao, Y.; Han, S.; Joty, S.; Liu, P.; Radev, D.; Wu, C.-S.; and Cohan, A. 2024c. Benchmarking Generation and Evaluation Capabilities of Large Language Models for Instruction Controllable Summarization. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 4481–4501. Mexico City, Mexico: Association for Computational Linguistics.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.

Mann, H. B.; and Whitney, D. R. 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.

McCarthy, T. 1998. *Descriptive Writing*. Scholastic Inc.

Meta. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

MOE. 2020. General High School Chinese Language Curriculum Standards (2017 Edition, Revised in 2020). Ministry of Education of the People's Republic of China.

OpenAI; :; Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; Madry, A.; Baker-Whitcomb, A.; Beutel, A.; Borzunov, A.; Carney, A.; Chow, A.; Kirillov, A.; Nichol, A.; Paino, A.; Renzin, A.; Passos, A. T.; Kirillov, A.; Christakis, A.; Conneau, A.; Kamali, A.; Jabri, A.; Moyer, A.; Tam, A.; Crookes, A.; Tootoochian, A.; Tootoonchian, A.; Kumar, A.; Vallone, A.; Karpathy, A.; Braunstein, A.; Cann, A.; Codispoti, A.; Galu, A.; Kondrich, A.; Tulloch, A.; Mishchenko, A.; Baek, A.; Jiang, A.; Pelisse, A.; Woodford, A.; Gosalia, A.; Dhar, A.; Pantuliano, A.; Nayak, A.; Oliver, A.; Zoph, B.; Ghorbani, B.; Leimberger, B.; Rossen, B.; Sokolowsky, B.; Wang, B.; Zweig, B.; Hoover, B.; Samic, B.; McGrew, B.; Spero, B.; Giertler, B.; Cheng, B.; Lightcap, B.; Walkin, B.; Quinn, B.; Guarraci, B.; Hsu, B.; Kellogg, B.; Eastman, B.; Lugaresi, C.; Wainwright, C.; Bassin, C.; Hudson, C.; Chu, C.; Nelson, C.; Li, C.; Shern, C. J.; Conger, C.; Barette, C.; Voss, C.; Ding, C.; Lu, C.; Zhang, C.; Beaumont, C.; Hallacy, C.; Koch, C.; Gibson, C.; Kim, C.; Choi, C.; McLeavey, C.; Hesse, C.; Fischer, C.; Winter, C.; Czarnecki, C.; Jarvis, C.; Wei, C.; Koumouzelis, C.; Sherburn, D.; Kappler, D.; Levin, D.; Levy, D.; Carr, D.; Farhi, D.; Mely, D.; Robinson, D.; Sasaki, D.; Jin, D.; Valladares, D.; Tsipras, D.; Li, D.; Nguyen, D. P.; Findlay, D.; Oiwoh, E.; Wong, E.; Asdar, E.; Proehl, E.; Yang, E.; Antonow, E.; Kramer, E.; Peterson, E.; Sigler, E.; Wallace, E.; Brevdo, E.; Mays, E.; Khorasani, F.; Such, F. P.; Raso, F.; Zhang, F.; von Lohmann, F.; Sulit, F.; Goh, G.; Oden, G.; Salmon, G.; Starace, G.; Brockman, G.; Salman, H.; Bao, H.; Hu, H.; Wong, H.; Wang, H.; Schmidt, H.; Whitney, H.; Jun, H.; Kirchner, H.; de Oliveira Pinto, H. P.; Ren, H.; Chang, H.; Chung, H. W.;

Kivlichan, I.; O'Connell, I.; O'Connell, I.; Osband, I.; Silber, I.; Sohl, I.; Okuyucu, I.; Lan, I.; Kostrikov, I.; Sutskever, I.; Kanitscheider, I.; Gulrajani, I.; Coxon, J.; Menick, J.; Pachocki, J.; Aung, J.; Betker, J.; Crooks, J.; Lennon, J.; Kiros, J.; Leike, J.; Park, J.; Kwon, J.; Phang, J.; Teplitz, J.; Wei, J.; Wolfe, J.; Chen, J.; Harris, J.; Varavva, J.; Lee, J. G.; Shieh, J.; Lin, J.; Yu, J.; Weng, J.; Tang, J.; Yu, J.; Jang, J.; Candela, J. Q.; Beutler, J.; Landers, J.; Parish, J.; Heidecke, J.; Schulman, J.; Lachman, J.; McKay, J.; Uesato, J.; Ward, J.; Kim, J. W.; Huizinga, J.; Sitkin, J.; Kraaijeveld, J.; Gross, J.; Kaplan, J.; Snyder, J.; Achiam, J.; Jiao, J.; Lee, J.; Zhuang, J.; Harriman, J.; Fricke, K.; Hayashi, K.; Singhal, K.; Shi, K.; Karthik, K.; Wood, K.; Rimbach, K.; Hsu, K.; Nguyen, K.; Gu-Lemberg, K.; Button, K.; Liu, K.; Howe, K.; Muthukumar, K.; Luther, K.; Ahmad, L.; Kai, L.; Itow, L.; Workman, L.; Pathak, L.; Chen, L.; Jing, L.; Guy, L.; Fedus, L.; Zhou, L.; Mamitsuka, L.; Weng, L.; McCallum, L.; Held, L.; Ouyang, L.; Feuvrier, L.; Zhang, L.; Kondraciuk, L.; Kaiser, L.; Hewitt, L.; Metz, L.; Doshi, L.; Aflak, M.; Simens, M.; Boyd, M.; Thompson, M.; Dukhan, M.; Chen, M.; Gray, M.; Hudnall, M.; Zhang, M.; Aljubeh, M.; Litwin, M.; Zeng, M.; Johnson, M.; Shetty, M.; Gupta, M.; Shah, M.; Yatbaz, M.; Yang, M. J.; Zhong, M.; Glaese, M.; Chen, M.; Janner, M.; Lampe, M.; Petrov, M.; Wu, M.; Wang, M.; Fradin, M.; Pokrass, M.; Castro, M.; de Castro, M. O. T.; Pavlov, M.; Brundage, M.; Wang, M.; Khan, M.; Murati, M.; Bavarian, M.; Lin, M.; Yesildal, M.; Soto, N.; Gimelshein, N.; Cone, N.; Staudacher, N.; Summers, N.; LaFontaine, N.; Chowdhury, N.; Ryder, N.; Stathas, N.; Turley, N.; Tezak, N.; Felix, N.; Kudige, N.; Keskar, N.; Deutsch, N.; Bundick, N.; Puckett, N.; Nachum, O.; Okelola, O.; Boiko, O.; Murk, O.; Jaffe, O.; Watkins, O.; Godement, O.; Campbell-Moore, O.; Chao, P.; McMillan, P.; Belov, P.; Su, P.; Bak, P.; Bakkum, P.; Deng, P.; Dolan, P.; Hoeschele, P.; Welinder, P.; Tillet, P.; Pronin, P.; Tillet, P.; Dhariwal, P.; Yuan, Q.; Dias, R.; Lim, R.; Arora, R.; Troll, R.; Lin, R.; Lopes, R. G.; Puri, R.; Miyara, R.; Leike, R.; Gaubert, R.; Zamani, R.; Wang, R.; Donnelly, R.; Honsby, R.; Smith, R.; Sahai, R.; Ramchandani, R.; Huet, R.; Carmichael, R.; Zellers, R.; Chen, R.; Chen, R.; Nigmatullin, R.; Cheu, R.; Jain, S.; Altman, S.; Schoenholz, S.; Toizer, S.; Miserendino, S.; Agarwal, S.; Culver, S.; Ethersmith, S.; Gray, S.; Grove, S.; Metzger, S.; Hermani, S.; Jain, S.; Zhao, S.; Wu, S.; Jomoto, S.; Wu, S.; Shuaiqi; Xia; Phene, S.; Papay, S.; Narayanan, S.; Coffey, S.; Lee, S.; Hall, S.; Balaji, S.; Broda, T.; Stramer, T.; Xu, T.; Gogineni, T.; Christianson, T.; Sanders, T.; Patwardhan, T.; Cunninghman, T.; Degry, T.; Dimson, T.; Raoux, T.; Shadwell, T.; Zheng, T.; Underwood, T.; Markov, T.; Sherbakov, T.; Rubin, T.; Stasi, T.; Kaftan, T.; Heywood, T.; Peterson, T.; Walters, T.; Eloundou, T.; Qi, V.; Moeller, V.; Monaco, V.; Kuo, V.; Fomenko, V.; Chang, W.; Zheng, W.; Zhou, W.; Manassra, W.; Sheu, W.; Zaremba, W.; Patil, Y.; Qian, Y.; Kim, Y.; Cheng, Y.; Zhang, Y.; He, Y.; Zhang, Y.; Jin, Y.; Dai, Y.; and Malkov, Y. 2024. GPT-4o System Card. arXiv:2410.21276.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Persing, I.; and Ng, V. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 260–269.

Qin, Y.; Song, K.; Hu, Y.; Yao, W.; Cho, S.; Wang, X.; Wu, X.; Liu, F.; Liu, P.; and Yu, D. 2024. InFoBench: Evaluating Instruction Following Ability in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 13025–13048. Bangkok, Thailand: Association for Computational Linguistics.

Que, H.; Duan, F.; He, L.; Mou, Y.; Zhou, W.; Liu, J.; Rong, W.; Wang, Z. M.; Yang, J.; Zhang, G.; et al. 2024. Hellobench: 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. arXiv:2412.15115.

Saaty, T. L. 1980. *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. New York & London: McGraw-Hill International Book Co.

Sari, N. A. P.; et al. 2025. Using Artificial Intelligence in Writing an Academic Essay: A Literature Review. *PROJECT (Professional Journal of English Education)*, 8(2): 448–461.

Shen, C.; Cheng, L.; Nguyen, X.-P.; You, Y.; and Bing, L. 2023. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4215–4233. Singapore: Association for Computational Linguistics.

Somasundaran, S.; Flor, M.; Chodorow, M.; Molloy, H.; Gyawali, B.; and McCulla, L. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6: 91–106.

Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72–101.

Taghipour, K.; and Ng, H. T. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1882–1891.

Tan, H.; Guo, Z.; Shi, Z.; Xu, L.; Liu, Z.; Feng, Y.; Li, X.; Wang, Y.; Shang, L.; Liu, Q.; 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)*, 6806–6827.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Uto, M.; Xie, Y.; and Ueno, M. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th international conference on computational linguistics*, 6077–6088.

Wachsmuth, H.; Naderi, N.; Hou, Y.; Bilu, Y.; Prabhakaran, V.; Thijm, T. A.; Hirst, G.; and Stein, B. 2017. Computational Argumentation Quality Assessment in Natural Language. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 176–187. Valencia, Spain: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Wen, B.; Ke, P.; Gu, X.; Wu, L.; Huang, H.; Zhou, J.; Li, W.; Hu, B.; Gao, W.; Xu, J.; et al. 2024. Benchmarking complex instruction-following with multiple constraints composition. *Advances in Neural Information Processing Systems*, 37: 137610–137645.

Woo, D. J.; Susanto, H.; Yeung, C. H.; Guo, K.; and Fung, A. K. Y. 2023. Exploring AI-Generated text in student writing: How does AI help? *arXiv preprint arXiv:2304.02478*.

Wu, Y.; Mei, J.; Yan, M.; Li, C.; Lai, S.; Ren, Y.; Wang, Z.; Zhang, J.; Wu, M.; Jin, Q.; et al. 2025. WritingBench: A Comprehensive Benchmark for Generative Writing. *arXiv preprint arXiv:2503.05244*.

xAI. 2024. Grok-2 Beta Release. Accessed: 2025-05-19.

xAI. 2025. Grok 3 Beta — The Age of Reasoning Agents. Accessed: 2025-05-19.

Xiao, C.; Xu, S. X.; Zhang, K.; Wang, Y.; and Xia, L. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, 610–625.

Xie, Q.; Han, W.; Chen, Z.; Xiang, R.; Zhang, X.; He, Y.; Xiao, M.; Li, D.; Dai, Y.; Feng, D.; et al. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37: 95716–95743.

Yannakoudakis, H.; Briscoe, T.; and Medlock, B. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 180–189.

Zhang, J.; Eugene, J. Y.; Chen, Q.; Xiong, C.; Zhu, D.; Qian, H.; Song, M.; Xiong, W.; Li, X.; Liu, Q.; et al. 2025. WIKIGENBENCH: Exploring Full-length Wikipedia Generation under Real-World Scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, 5191–5210.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Žižović, M.; and Pamucar, D. 2019. New model for determining criteria weights: Level Based Weight Assessment (LBWA) model. *Decision Making: Applications in Management and Engineering*, 2(2): 126–137.