

DyBenchEval: A Comprehensive Evaluation Suite for Measuring LLM-as-Dynamic-Benchmark

Pei Ke^{1,2}, Jingye Wang¹, Maolin Sun¹, Lan Wang¹, Haitao Peng¹, Bosi Wen³,
Enping Hu¹, Ruyi Luo¹, Minlie Huang³, Ke Qin^{1,2,*}

¹Laboratory of Intelligent Collaborative Computing,

University of Electronic Science and Technology of China, Chengdu, China

²Ubiquitous Intelligence and Trusted Services Key Laboratory of Sichuan Province

³The Conversational Artificial Intelligence (CoAI) Group, Tsinghua University

kepei@uestc.edu.cn, 202522900230@std.uestc.edu.cn, qinke@uestc.edu.cn

Abstract

Dynamic benchmarking is an essential research direction in the evaluation of large language models (LLMs) to avoid data contamination. Although existing works have proposed LLM-based methods for dynamic benchmarking, a unified paradigm for performance comparison among different LLMs remains lacking, which hinders researchers from model selection for specific domains and tasks. To address this challenge, we propose a comprehensive evaluation suite called DyBenchEval, including a novel seed dataset across 4 domains and 7 tasks as well as multi-aspect evaluation metrics. A cross data filling method is utilized for high-quality data augmentation with domain/task-level similarity, aiming to overcome the data scarcity problem. For automatic metrics, we consider coherence, correctness, and diversity of LLM-generated evaluation samples, simultaneously measuring the respective quality of generated questions / answers and the relationship between them. Extensive experimental results on DyBenchEval reveal deficiencies in 11 mainstream closed-source and open-source LLMs when dealing with dynamic benchmarking across different domains and tasks.

1 Introduction

With rapid development of large language models (LLMs), they have achieved remarkable performance across various natural language processing (NLP) tasks, including open-domain question answering (Yang et al., 2025), mathematical reasoning (Shao et al., 2025), and code generation (Zhu et al., 2024b). Thereby, how to effectively benchmark LLMs’ capacities in different tasks becomes a key research focus and attracts much attention (Chang et al., 2024). High-quality benchmarking datasets are expected to reflect LLMs’ performance accurately, guide the selection of LLMs for specific

tasks, and reveal LLMs’ deficiencies for future improvement (Ahuja et al., 2023; Wang et al., 2025).

Although there are already numerous benchmark datasets for evaluating LLMs’ performance, these static datasets are increasingly inadequate for a thorough and precise assessment of LLMs (Zhu et al., 2024a) due to the potential data contamination issues (Dong et al., 2024; Chen et al., 2025a). Since there is a pressing demand for data use in the persistent improvement of LLMs, public benchmark datasets may be improperly included during the training process, leading to unreliable evaluation results (Wang et al., 2025). To address this limitation, a line of work on *LLM-as-Dynamic-Benchmark* has emerged recently, which dynamically generates evaluation samples instead of using a fixed set of data and thus provides contamination-free benchmarking (Chen et al., 2025b). These works employ LLM-based methods to update existing benchmark datasets and acquire diversity- and complexity-controlled new samples (Bai et al., 2023; Cheng et al., 2024; Chen et al., 2025b).

However, a unified evaluation paradigm for comparing the performance of different LLMs in terms of dynamic benchmarking remains lacking. This hinders researchers from effectively selecting the most suitable one when they tend to dynamically update benchmark datasets for specific domains and tasks. Most of the existing works focus on restricted domains / tasks such as reasoning (Zhu et al., 2024a; Wang et al., 2025) and code generation (Chen et al., 2024, 2025b; Jain et al., 2025), and adopt task-specific seed datasets and metrics for automatic judgments. This evaluation paradigm limits the application scenarios of these methods and ignores the comparison with potentially similar baselines, leading to missing consensus on the performance ranking of LLMs for dynamic benchmarking across various domains and tasks.

To address this issue, we propose a comprehensive evaluation suite called DyBenchEval, which

*Corresponding author

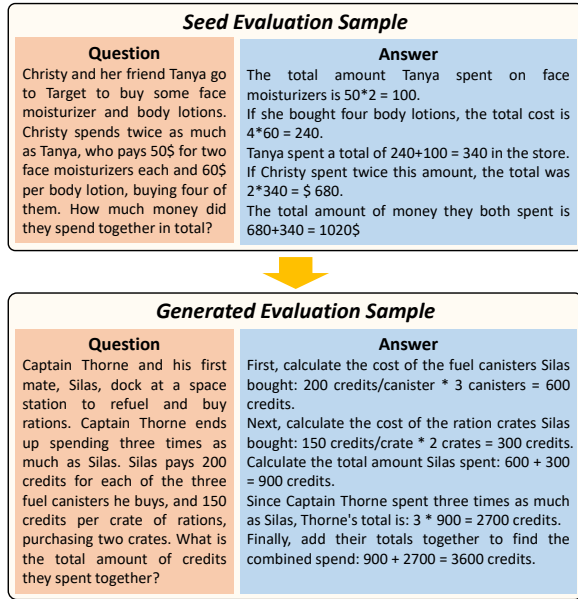


Figure 1: Illustration of dynamic benchmarking, aiming to generate new evaluation samples from seed ones.

includes a novel seed dataset to comprehensively assess the performance of LLMs on dynamic benchmarking across different domains and tasks as well as multi-aspect evaluation metrics. Specifically, our dataset is constructed based on a two-dimensional domain-task taxonomy, containing 4 domains and 7 task types, which provides a broad perspective to measure LLMs in evaluation sample generation. To overcome the data scarcity in uncommon domains / tasks, we propose a simple yet effective method called cross data filling, which augments diverse and high-quality data samples based on the demonstrations from similar domains / tasks. Then, we introduce evaluation metrics that correspond to three essential aspects for the quality of generated evaluation samples, including the coherence of sample questions (Cheng et al., 2024), the correctness of sample answers with respect to questions (Chen et al., 2025a), and the diversity of generated samples (Chen et al., 2025b). These multi-aspect metrics are expected to achieve high agreement with human judgments, thereby providing reliable automatic evaluation results of LLM-as-Dynamic-Benchmark on our proposed dataset.

Our main contributions are as follows¹:

- We make the first attempt to propose an evaluation suite to comprehensively measure the performance of LLM-as-Dynamic-Benchmark

¹Our dataset and code are available at <https://github.com/wangjingye342/dybencheval>.

across different domains and tasks, revealing their deficiencies for future improvements.

- We construct a multi-domain multi-task seed benchmark dataset with a cross filling method to augment data in uncommon domains / tasks, as well as multi-aspect evaluation metrics.
- We conduct extensive experiments on 11 mainstream LLMs, providing insights into the selection of LLMs for dynamic benchmarking in terms of specific scenarios.

2 Related Work

2.1 Static Benchmarking

Static benchmarking, as a mainstream method to measure the performance of NLP models, attracts much more attention recently due to the emergence of LLMs. Traditional static benchmark datasets are designed based on the characteristics of each NLP task (Ahuja et al., 2024a), such as text classification (Socher et al., 2013), reading comprehension (Rajpurkar et al., 2016), machine translation (Ma et al., 2019), text summarization (Hermann et al., 2015), and dialogue generation (Dinan et al., 2019). Since LLMs have achieved remarkable performance on different NLP tasks, researchers resort to construct benchmark datasets to assess the capability of LLMs instead of targeting at specific NLP tasks, such as instruction following (Zheng et al., 2023; Jiang et al., 2024; Qin et al., 2024), reasoning (Cobbe et al., 2021; Hendrycks et al., 2021c), and code generation (Chen et al., 2021b; Jimenez et al., 2024). As LLMs develop rapidly, recently proposed benchmarks also focus on the assessment of LLMs' advanced capabilities like performing as agents (Liu et al., 2024b; Zeng et al., 2025).

Compared with the above benchmarks, our work instead focuses on assessing an important but under-explored ability of LLMs to generate evaluation samples, which can be applied to dynamically updating existing benchmark datasets and has the potential to alleviate data contamination issues.

2.2 Dynamic Benchmarking

Dynamic benchmarking methods aim to dynamically transform existing benchmark datasets to new ones, avoiding possible data contamination (Ahuja et al., 2024b; Chen et al., 2025a). As an emerging research direction recently, researchers have made preliminary attempts to design rule-based and LLM-based methods. Rule-based methods adopt

Research Work	Seed Dataset				Evaluation Aspect		
	Size	#Domain	#Task	Source	Coherence	Correctness	Diversity
StructEval (Cao et al., 2024)	600	1	3	Existing	✓	✓	-
Auto-Dataset (Ying et al., 2024)	3,317	2	3	Existing	✓	✓	-
DyCodeEval (Chen et al., 2025b)	2,500	3	1	Existing	-	✓	✓
Self-Evolving (Wang et al., 2025)	400	2	2	Existing	-	✓	-
DyBenchEval (Ours)	1,665	4	7	Existing+Augmented	✓	✓	✓

Table 1: Comparisons between our work and recent research works on the seed dataset and evaluation aspects used in the experiments. Current research works mostly utilize existing datasets that involve restricted domains or tasks, while our work additionally augments examples to cover multiple domains and tasks.

task-specific pre-defined rules to devise templates and update existing benchmark datasets accordingly (Kurtic et al., 2024; Mirzadeh et al., 2025; Zhao et al., 2025). Although these methods can acquire high-quality evaluation samples, the diversity is largely limited. Other public datasets which adopt similar rules for data synthesis may cause in-domain contamination (Chen et al., 2025a). Another line of work utilizes LLMs to rewrite evaluation samples from existing benchmarks. Most of the existing works depend on prompting LLMs directly to style- or difficulty-controlled evaluation samples (Ying et al., 2024; Zhu et al., 2024c; Qian et al., 2024), while others employ LLM-based multi-agent frameworks to yield diverse and high-quality benchmarks (Wang et al., 2025).

Although existing works have explored various LLMs in the dynamic benchmarking of specific tasks, an evaluation paradigm for performance comparison among LLMs across different domains and tasks remains lacking. For comparison, our work makes the first attempt to build a comprehensive evaluation suite for LLM-as-Dynamic-Benchmark, including a multi-domain multi-task benchmark as well as multi-aspect evaluation metrics. This suite is expected to provide a unified standard for guiding the selection of LLMs for dynamic benchmarking in specific domains and tasks.

3 DyBenchEval Framework

3.1 Overview

To comprehensively measure the ability of LLMs to generate evaluation examples from seed benchmark datasets, we propose a two-dimensional domain-task taxonomy for constructing a high-coverage seed dataset. Our proposed dataset spans all the combinations of domains and tasks, which provides a comprehensive perspective of LLMs’ performance in both context understanding of multiple

domains and text generation of different tasks. We collect the information of existing works and compare our work with them in terms of this taxonomy, as shown in Table 1.

3.2 Domain

We adopt the same domain taxonomy as existing works like MMLU (Hendrycks et al., 2021b) and C-Eval (Huang et al., 2023), which contains 4 main categories, i.e., STEM, Humanity, Social Science, and Others. The definition and scope of each domain are borrowed from Hendrycks et al. (2021b). **STEM.** We consider four detailed subdomains within STEM including science, technology, engineering, and mathematics.

Humanity. The humanity consists of the disciplines which utilize analytic methods and qualitative analysis instead of scientific empirical methods. We select four common branches as our subdomains, i.e., art, history, culture, and logic.

Social Science. Social science aims to examine human behavior and society. Considering the broad scope of social science, we choose six representative subdomains, i.e., finance, business administration, news, law, commercial, and education.

Others. There are other domains which may not fit into any of the above categories, such as daily life, which are put into *Others*.

3.3 Task

We follow recent works on LLM evaluation including AlignBench (Liu et al., 2024a) and CritiqueLLM (Ke et al., 2024), and merge their task taxonomies into 7 task types. Specifically, we acquire the union of the task taxonomy in these two works, and remove language-specific task types since our work is not targeted at specific languages. Then, we merge two reasoning-related tasks (i.e., Logical Reasoning and Mathematics) into a Reasoning Ability category. The definition and scope

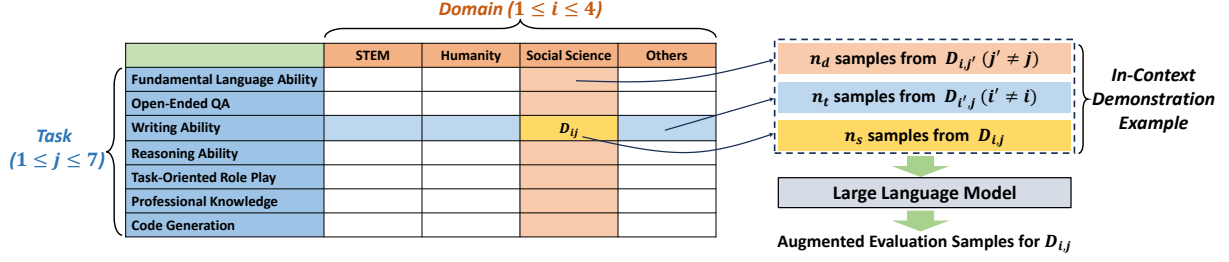


Figure 2: Overview of cross data filling, which augments the dataset D_{ij} with the samples from the same domain / task / cell using in-context learning.

of each task both refer to the original works (Liu et al., 2024a; Ke et al., 2024).

Fundamental Language Ability. This category contains the basic NLP tasks such as text summarization (Hermann et al., 2015), information extraction (Chen et al., 2022), and data-to-text generation (Dusek et al., 2020), which support the common needs of users.

Open-Ended Question Answering. The category indicates subjective question answering in terms of open-domain topics (Zheng et al., 2023; Liu et al., 2024a), which is a common scenario of LLMs especially in the daily usage.

Writing Ability. Writing plays an important role in the open-domain LLMs’ applications, including narratives, fictions, and poems, which should meet the high requirement of consistency and coherence (Liu et al., 2024a).

Reasoning Ability. As a core ability to deal with complex problems using a step-by-step solving process, reasoning has become a research focus for LLMs, including mathematical (Cobbe et al., 2021; Hendrycks et al., 2021c), commonsense (Sap et al., 2019), and logical (Clark et al., 2020) reasoning.

Task-Oriented Role Play. As an emerging capability of LLMs, role play requires to make LLMs play as a specific identity for corresponding tasks, aiming to follow user instructions better (Zhou et al., 2024; Tu et al., 2024).

Professional Knowledge. This category focuses on measuring the ability of LLMs to answer questions with professional knowledge in various domains (Liu et al., 2024a).

Code Generation. As an important application direction of LLMs, code generation benchmarks assess the ability to automatically generate codes based on various inputs (Chen et al., 2021b).

4 DyBenchEval Construction

4.1 Data Collection with Cross Data Filling

We construct the seed benchmark dataset of DyBenchEval based on the taxonomy described in §3. The construction process involves two steps, i.e., existing data collection and cross data filling.

Existing Data Collection. We first span existing LLM benchmark datasets in different domains and tasks, and select high-quality and diverse ones as representatives, which are then categorized into our taxonomy. The detailed dataset list for each pair of domains and tasks is shown in Appendix A. We merge the datasets that belong to the i -th domain ($1 \leq i \leq 4$) and j -th task ($1 \leq j \leq 7$) into a combined one $D_{ij} = \{q_{ij}^k, a_{ij}^k\}_{k=1}^{|D_{ij}|}$, where (q_{ij}^k, a_{ij}^k) indicates the question and answer of the k -th data pair in D_{ij} .

Cross Data Filling. To alleviate the data scarcity problem in uncommon domains / tasks, we devise a simple yet effective method called cross data filling. As shown in Figure 2, our main idea is to utilize the data samples at the same domain (row) or task (column) to conduct data augmentation for the cell with scarce data.

Specifically, if we need to augment the data in D_{ij} , we compute the domain/task-level similarity within the *cross* (as shown in Figure 2), and acquire demonstration samples based on the similarity for in-context learning of LLMs, which can finally obtain the augmented data samples. We follow Sannigrahi et al. (2023) and Liu et al. (2016) to acquire the dataset embedding $\text{Emb}(D_{i,j})$ based on the average of each data sample:

$$\text{Emb}(D_{i,j}) = \frac{1}{|D_{i,j}|} \sum_{k=1}^{|D_{i,j}|} \text{Emb}(q_{i,j}^k \oplus a_{i,j}^k) \quad (1)$$

where $\text{Emb}(q_{i,j}^k \oplus a_{i,j}^k)$ denotes the embedding of the concatenation of the question and answer. We adopt Qwen3-Embedding-8B (Zhang et al., 2025)

to compute embeddings. This model is one of the best-performing ones on commonly-recognized multi-domain and multi-task text embedding benchmarks like MMTEB (Enevoldsen et al., 2025). It supports the context length of 32K tokens and has the strong long-context capability due to the effectiveness of its base model Qwen3 (Yang et al., 2025), which is sufficient for the datasets considered in our paper.

Then, we obtain the similarity between the dataset requiring augmentation and other datasets within the same domain ($D_{i,j'}$) or task ($D_{i',j}$):

$$\begin{aligned} \text{Sim}(D_{i,j}, D_{i,j'}) \\ = \frac{\text{Emb}(D_{i,j}) \cdot \text{Emb}(D_{i,j'})}{\|\text{Emb}(D_{i,j})\|_2 \|\text{Emb}(D_{i,j'})\|_2} \end{aligned} \quad (2)$$

$$\begin{aligned} j' = 1, \dots, j-1, j+1, \dots, 7 \\ \text{Sim}(D_{i,j}, D_{i',j'}) \\ = \frac{\text{Emb}(D_{i,j}) \cdot \text{Emb}(D_{i',j'})}{\|\text{Emb}(D_{i,j})\|_2 \|\text{Emb}(D_{i',j'})\|_2} \end{aligned} \quad (3)$$

$$i' = 1, \dots, i-1, i+1, \dots, 4$$

Based on the above similarity, we adopt $\text{softmax}(\cdot)$ to acquire the normalized probability distributions over domains (p^d) and tasks (p^t), respectively:

$$p_{i'}^d = \text{softmax}(\text{Sim}(D_{i,j}, D_{i',j})) \quad (4)$$

$$i' = 1, \dots, i-1, i+1, \dots, 4$$

$$p_{j'}^t = \text{softmax}(\text{Sim}(D_{i,j}, D_{i,j'})) \quad (5)$$

$$j' = 1, \dots, j-1, j+1, \dots, 7$$

where $p_{i'}^d / p_{j'}^t$ denotes the probability of choosing the samples as demonstration examples from the dataset $D_{i',j} / D_{i,j'}$, respectively. To further illustrate the effectiveness of dataset embeddings in similarity computation, we provide the clustering metrics and 2D heatmap in Appendix B.

Once we acquire the n_d / n_t samples within the same domain / task based on the above probability distributions, we devise an in-context learning method which makes LLMs generate new evaluation data (including the questions and corresponding answers) based on the above $n_d + n_t$ samples as well as n_s samples from the same cell.

To ensure the quality of cross data filling, we manually check all the augmented evaluation samples in terms of four evaluation aspects, i.e., the coherence of questions, the correctness of answers, the faithfulness to the domain / task label, and the

Category	Size	Len-Q	Len-A
Overall	1,665	636.18	281.86
<i>Domain</i>			
STEM	418	1,367.47	633.27
Humanity	410	588.02	180.33
Social Science	419	419.95	64.44
Others	418	168.88	247.97
<i>Task</i>			
Fundamental Language Ability	239	1,921.28	175.08
Open-Domain QA	240	217.42	33.67
Writing Ability	238	1,197.59	1,374.76
Reasoning Ability	240	169.40	17.52
Task-Oriented Role Play	239	346.12	143.70
Professional Knowledge	238	259.97	59.13
Code Generation	231	335.94	171.19

Table 2: Statistics of DyBenchEval including the dataset size (**Size**) and the average tokens of questions (**Len-Q**) / answers (**Len-A**).

ethics of the data sample. If there are only minor errors², we revise generated data to obtain high-quality ones. Since some of the data may contain numerous errors or unethical content that are not easy to repair, we directly abandon these cases³. Throughout all the 240 augmented data samples, 5% of them are revised, while 6.25% are abandoned. Finally, 93.75% (except the abandoned ones) are added to the benchmark dataset, demonstrating the overall high quality of augmented data based on our method.

After quality control, the final seed benchmark dataset of DyBenchEval contains 1,665 data pairs of questions and answers. The data distributions of different tasks and domains, as well as the length of questions and answers, are shown in Table 2.

4.2 Multi-Aspect Evaluation Paradigm

To further assess the quality of LLMs' generated evaluation samples based on the seed dataset, we devise evaluation metrics of three aspects, i.e., the coherence of input questions (Cheng et al., 2024), the correctness of output answers with respect to corresponding questions (Chen et al., 2025a), and the diversity of generated evaluation samples (Chen

²Our identified typical errors that correspond to the first three aspects contain logical contradictions of questions, numerical / knowledgeable errors of answers, and unfaithfulness to the domain / task label. We also find some unethical content, including violent, sexual, and hateful contents.

³The abandoned cases commonly encompass multiple types of errors / unethical content throughout the questions and answers (such as logical contradictions between multiple pairs of conditions in the question and numerical errors in multiple parts of the answer), which cannot be improved via simple manual revision.

Aspect	Coherence	Correctness
Method	Pairwise Agreement	Pointwise Agreement
Direct Prompting	0.66	0.71
CoT Prompting	0.82	0.84
Ours	0.90	0.89

Table 3: Agreement between automatic and manual evaluation results on coherence and correctness.

et al., 2025b). These metrics comprehensively judge the respective quality of questions and answers, as well as the relationship between them.

Coherence. Coherence aims to measure whether the sentences in the generated question are semantically coherent without logical contradictions.

Correctness. Correctness is targeted at assessing whether the generated answer is correct with respect to the corresponding question.

Diversity. Diversity aims to evaluate whether the generated questions and answers cover diverse linguistic patterns and semantic information. Inspired by Chen et al. (2025a), we categorize the diversity metric into two parts, including external and internal diversity. External diversity measures the difference between the generated dataset and the seed one, while internal diversity quantifies the variation among the generated dataset.

For coherence and correctness, a strong LLM is utilized to obtain the evaluation results (Zheng et al., 2023). We design the evaluation prompts to make the judge LLM decompose the aspect into sub-aspects based on each sample and recombine the results of each sub-aspect to acquire the final automatic judgments. The evaluation score of coherence ranges from 1 to 10, while the evaluation result of correctness is a binary classification label, where 1 / 0 denotes correct / wrong samples, respectively. The prompt templates for coherence and correctness are shown in Appendix E.

As for diversity, we follow the existing works (Chen et al., 2025a) to adopt distinct n-grams (Dist-n) (Li et al., 2016) and Vendi Score (Friedman and Dieng, 2023) of generated datasets for measuring internal diversity, while utilizing BLEU scores (Papineni et al., 2002) and semantic similarity (Sem-Sim) (Chen et al., 2025b) to compute the diversity of generated samples with respect to seed samples. These metrics comprehensively capture both the linguistic variation and semantic diversity.

5 Experiment

5.1 Agreement Evaluation

To assess the agreement between our LLM-as-a-Judge methods and manual evaluation, we sample 200 data pairs from the DyBenchEval dataset randomly for building a meta-evaluation seed dataset, where 100 samples for coherence and the others for correctness. We collect the generation results from 11 mainstream LLMs with different families and scales, including GPT, Gemini, Claude, DeepSeek (Liu et al., 2025), GLM (Zeng et al., 2024), Qwen (Yang et al., 2025) and Llama (AI@Meta, 2024), which vary from 8B to 685B. We employ Gemini-3-Pro, which is one of the most effective LLMs currently, as our judge model to acquire the automatic evaluation results of coherence and correctness.

For coherence, we follow Wen et al. (2024) to adopt pairwise agreement. Specifically, given the seed question-answering (QA) pair, two generated QA pairs (denoted as A and B), the annotators are required to select from 3 options based on the coherence of generated questions, i.e., A is better than B, B is better than A, and tie. Then, the automatic evaluation scores for these two questions are also converted into pairwise comparison results, which can be used to compute the agreement rate with manual preference. As for correctness, given a generated QA pair, annotators are instructed to judge whether the answer is correct with respect to the question. Then, we can compute the agreement between automatic metrics and human judgments.

As shown in Table 3, we can observe that our LLM-as-a-Judge method described in §4.2 achieves high agreements with manual evaluation results, demonstrating the reliability of our automatic metrics. The improvement of our method over vanilla direct prompting and chain-of-thought (CoT) prompting (Zheng et al., 2023) indicates the contribution of our design on aspect decomposition, which provides insights into the understanding of aspects in complex evaluation scenarios.

5.2 Experimental Setting

We utilize our devised prompt combined with Gemini-3-Pro as the judge model⁴ to measure the performance of 11 mainstream LLMs, covering closed-source and open-source ones:

⁴We additionally adopt GPT-5.2 as the judge model and show the results in Appendix F, indicating almost the same performance ranking.

Aspect	Coherence	Correctness	Diversity			
			Internal		External	
Metric	LLM Judge ($\uparrow$)	LLM Judge ($\uparrow$)	Dist-2 ($\uparrow$)	Vendi ($\uparrow$)	BLEU-4 ($\downarrow$)	SemSim ($\downarrow$)
<i>Closed-Source Language Models</i>						
GPT-5.2-thinking	5.3081	0.7760	0.4940	36.09	0.1977	0.7369
Gemini-3-pro-preview	7.8420	0.9225	0.5440	38.82	0.1231	0.6224
Claude-opus-4-5-20251101-thinking	7.0384	0.9405	0.5259	38.89	0.2781	0.6953
Qwen3-Max	5.5387	0.8697	0.5806	39.83	0.2006	0.8028
<i>Open-Source Language Models</i>						
DeepSeek-V3.2	6.4379	<u>0.9099</u>	0.5487	39.97	0.2419	0.7094
GLM-4.6	<u>7.2366</u>	0.8835	0.5171	40.42	0.2176	<u>0.6375</u>
Qwen3-235b-a22b-instruct-2507	6.1742	0.8462	0.6186	40.75	<u>0.1655</u>	0.7464
Qwen3-30b-a3b-instruct-2507	5.2006	0.7946	0.4982	41.61	0.2582	0.7920
Qwen3-8b	4.5273	0.7706	0.4772	40.92	0.3626	0.8202
Llama-3.1-70b-instruct	5.2979	0.7423	0.4390	40.38	0.3903	0.7264
Llama-3.1-8b-instruct	3.8583	0.5610	0.3516	16.78	0.2334	0.6947

Table 4: Coherence, correctness, and diversity of different LLMs on DyBenchEval. The coherence / correctness score is computed via LLMs, where coherence ranges from 1 to 10 and correctness indicates a binary label including correct (1) and incorrect (0) answers for each generated evaluation sample. The highest performance overall is **bold**, while the highest performance among open-source models is underlined.

Closed-Source LLMs: Gemini-3-Pro (Comanici et al., 2025), GPT-5.2, Claude-4.5, and Qwen3-Max (Yang et al., 2025).

Open-Source LLMs: DeepSeek-V3.2 (Liu et al., 2025), GLM-4.6 (Zeng et al., 2024), Qwen3-(8B, 30B, 235B) (Yang et al., 2025), Llama-3.1-(8B, 70B) (AI@Meta, 2024).

Note that these selected LLMs are the same as the agreement evaluation in §5.1. We use greedy search for reproducibility, and the maximum length for generated evaluation samples is 8,192. For the hyper-parameters of cross data filling, we set $n_d = 2$, $n_t = 2$, $n_s = 1$ and explore other numbers of in-context examples in §5.6.

5.3 Main Result

We present the main results in Table 4. **Firstly**, closed-source LLMs still have advantages over open-source ones in most of the metrics, showing their effectiveness on dynamic benchmarking. However, the absolute performance still has a lot of room for improvement. For example, the best-performing LLM in coherence (i.e. Gemini-3-Pro) only achieves the score of 7.8420 / 10. The quality of generated questions needs further improving.

Secondly, LLMs perform differently in all these metrics. Comparing coherence and correctness, Gemini-3-Pro / Claude-4.5 achieves the best performance, respectively. We also find that DeepSeek-V3.2 whose coherence score is relatively unsatisfactory obtains 0.9099 correctness score, beating

other open-source LLMs and even approaching the best performance of closed-source LLMs. This demonstrates that the abilities to generate coherent questions and correct answers may not be highly correlated. As for diversity, Qwen3-235B performs notably well especially in the internal diversity.

Thirdly, LLMs with larger scales show better performance (e.g., Qwen3 ranging from 8B to 235B), preliminarily demonstrating the scaling property of LLMs on dynamic benchmarking.

5.4 Analysis on Domain/Task-Level Score Distribution

To analyze the difficulty and characteristics of different domains and tasks, we compute the average scores across all the LLMs for each domain or task. The results are shown in Table 5. For domains, we observe that LLMs generally obtain high-quality evaluation samples in humanity and social science, while struggling in the domain of STEM. Since the evaluation samples in STEM commonly have professional backgrounds and may require complex derivation, generating evaluation samples for STEM seems more difficult than the other domains.

As for tasks, LLMs are generally good at open-domain tasks, such as task-oriented role play, open-domain QA, and writing ability. These tasks may hardly involve uncommon knowledge and be not limited to specific topics, providing sufficient space for new evaluation data generation. For comparison, code generation and professional knowledge

Aspect	Coherence	Correctness	Diversity			
			Internal		External	
Metric	LLM Judge ($\uparrow$)	LLM Judge ($\uparrow$)	Dist-2 ($\uparrow$)	Vendi ($\uparrow$)	BLEU-4 ($\downarrow$)	SemSim ($\downarrow$)
<i>Domain</i>						
STEM	5.4752	0.7621	0.5603	20.36	0.2847	0.7565
Humanity	5.9148	0.8317	0.6157	49.43	0.2125	0.7068
Social Science	6.2256	0.8495	0.5844	31.79	0.2399	0.7067
Others	5.8256	0.8358	<u>0.5208</u>	29.20	0.2285	0.7332
<i>Task</i>						
Fundamental Language Ability	6.0474	0.7581	0.6156	41.84	0.2188	0.6866
Open-Domain QA	6.0529	0.8859	0.5379	52.45	0.1831	0.6838
Writing Ability	6.0973	0.8740	0.6942	34.79	0.1717	0.7190
Reasoning Ability	5.7871	0.8023	0.6953	38.76	0.2215	0.6872
Task-Oriented Role Play	6.1278	0.8494	0.5050	16.53	0.2836	0.7759
Professional Knowledge	5.7146	<u>0.7449</u>	<u>0.4737</u>	28.35	0.2788	0.7463
Code Generation	<u>5.1989</u>	0.8355	<u>0.6140</u>	<u>8.58</u>	<u>0.3284</u>	<u>0.7794</u>

Table 5: Average performance over all the models in terms of each domain and task. The highest scores of each metric across the domains / tasks are bold, while the lowest scores are underlined.

are the most difficult tasks, which require complex code structure understanding and domain-specific knowledge utilization, respectively.

5.5 Analysis on Performance Ranking for Specific Domains/Tasks

To provide guidance for LLM selection in specific domains and tasks, we comprehensively show the performance of top-3 LLMs for each pair of domains and tasks in Table 9 of Appendix D due to the paper limit. The results demonstrate that Gemini-3-Pro beats other LLMs in coherence scores of most domains and tasks, indicating its superior performance on high-quality question generation. However, for correctness especially in the hard tasks such as code generation, professional knowledge, and reasoning ability, Claude-4.5 and DeepSeek-V3.2 nearly dominate all the domains, suggesting their strong abilities in problem solving.

5.6 Analysis on Cross Data Filling

To further analyze the effectiveness of our proposed cross data filling method, we conduct an ablation study on two important factors, i.e., the number and source of in-context examples. For the number of examples, we select the alternative option in $\{3, 7\}$ besides 5 in our main experiment. As for the example source, we explore the settings where in-context examples are from only the same task / domain / cell. We randomly sample 100 data from the augmented dataset and adopt similar evaluation metrics including coherence, correctness, and internal diversity. Since it is not necessary to consider

Aspect	Coherence	Correctness	Diversity (Internal)
Metric	LLM Judge	LLM Judge	Dist-2
Ours	6.87	0.99	0.5985
<i>Number of In-Context Examples</i>			
w/ 3 ICL Examples	6.78	0.98	0.5929
w/ 7 ICL Examples	6.89	0.99	0.6048
<i>Source of In-Context Examples</i>			
w/ Same Domain	6.72	0.95*	0.5321
w/ Same Task	6.71	0.92**	0.5625
w/ Same Cell	6.29**	0.93*	0.5474*

Table 6: Coherence, correctness, and internal diversity of augmented data via cross data filling. * indicates that the corresponding performance is significantly lower than our method with $p < 0.05$ based on paired bootstrap resampling (Konietzschke and Pauly, 2014), while ** denotes $p < 0.01$.

the in-context examples in the assessment of augmented samples, we do not use external diversity.

The results in Table 6 show that the performance improves as the number of in-context examples increases. Since 5 examples have achieved acceptable performance compared with 7 examples, we choose this setting in our main experiment considering the API cost. As for the example source, if the examples are all from the same domain / task / cell, the quality of generated evaluation samples degrades in all the metrics. Thus, our method incorporates in-context examples of different sources to generate overall high-quality evaluation samples.

6 Conclusion

In this work, we propose DyBenchEval, a comprehensive evaluation suite for LLM-as-Dynamic-Benchmarking, involving a benchmark dataset with 4 domains and 7 tasks and multi-aspect evaluation metrics on the coherence, correctness, and diversity of generated evaluation samples. Extensive experiments on 11 mainstream LLMs reveal LLMs’ deficiencies across various domains and tasks for future improvement, and provide guidance for the selection of LLMs in different scenarios.

Limitations

We discuss the limitations of our work as follows:

- (1) Similar to other LLM-based evaluation methods, our proposed metrics on coherence and correctness are also faced with the self-enhancement bias (Zheng et al., 2023), which indicates the preference on the generated texts from the same base model. This bias commonly exists in various LLMs. We argue that multiple LLM-based evaluators with different base models may be utilized to alleviate self-enhancement bias. Since there does not exist a satisfactory solution currently, we leave this part as important future work.
- (2) Although DyBenchEval has already covered multiple domains, the domain category is still coarse-grained, which includes several subdomains. Thus, the number of data samples at each subdomain may be not sufficient enough. We will further enlarge our benchmark dataset to reflect the LLMs’ performance more stably and precisely.

Acknowledgments

This work was supported by Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0501806), Sichuan Science and Technology Program (2025ZNS-FSC1488), Fundamental Research Funds for the Central Universities (ZYGX2025XJ041), and CIPS-SMP-Zhipu Large Model Fund (CIPS-SMP20250314).

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024a. MEGEVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637.

Sanchit Ahuja, Varun Gumma, and Sunayana Sitaram. 2024b. Contamination report for multilingual benchmarks. *arXiv preprint arXiv:2410.16186*.

AI@Meta. 2024. [Llama 3 model card](#).

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner. In *Advances in Neural Information Processing Systems*.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. [Piqa: Reasoning about physical commonsense in natural language](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. [MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. [Taskmaster-1: Toward a realistic and diverse dialog dataset](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525.

Boxi Cao, Mengjie Ren, Hongyu Lin, Xianpei Han, Feng Zhang, Junfeng Zhan, and Le Sun. 2024. StructEval: Deepen and broaden large language model assessment via structured evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5300–5318.

- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. [LexGLUE: A benchmark dataset for legal language understanding in English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.
- Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. [CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185.
- Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021a. [Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017, Online. Association for Computational Linguistics.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021b. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Muhao Chen, Lifu Huang, Manling Li, Ben Zhou, Heng Ji, and Dan Roth. 2022. New frontiers of information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 14–25.
- Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. 2025a. Benchmarking large language models under data contamination: A survey from static to dynamic evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10091–10109.
- Simin Chen, Xiaoning Feng, Xiaohong Han, Cong Liu, and Wei Yang. 2024. PPM: automated generation of diverse programming problems for benchmarking code generation models. *Proc. ACM Softw. Eng.*, 1(FSE):1194–1215.
- Simin Chen, Pranav Pusarla, and Baishakhi Ray. 2025b. Dycodeeval: Dynamic benchmarking of reasoning capabilities in code large language models under data contamination. In *Forty-second International Conference on Machine Learning*.
- Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. [TheoremQA: A theorem-driven question answering dataset](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901.
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021c. [FinQA: A dataset of numerical reasoning over financial data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiale Cheng, Yida Lu, Xiaotao Gu, Pei Ke, Xiao Liu, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. Autodetect: Towards a unified framework for automated weakness detection in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6786–6803.
- Peter Clark, Øyvind Taffjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3882–3890.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. [A dataset of information-seeking questions and answers anchored in research papers](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.
- Franck Dernoncourt and Ji Young Lee. 2017. [PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard

- of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations*.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 12039–12050.
- Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Comput. Speech Lang.*, 59:123–156.
- Vladimir Eidelman. 2019. *Billsum: A corpus for automatic summarization of us legislation*. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, page 48–56. Association for Computational Linguistics.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatin, Ömer Çağatan, and 63 others. 2025. Mmteb: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*, pages 101715–101771.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. *ELI5: Long form question answering*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.
- Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. *Ddxplus: A new dataset for automatic medical diagnosis*. In *Advances in Neural Information Processing Systems*, volume 35, pages 31306–31318. Curran Associates, Inc.
- E. B. Fowlkes and Colin L. Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78:553–569.
- Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Trans. Mach. Learn. Res.*
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. *Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies*. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. *Samsun corpus: A human-annotated dialogue dataset for abstractive summarization*. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. *Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models*. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021a. *Measuring coding challenge competence with APPS*. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.
- Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. *Cosmos QA: Machine reading comprehension with contextual commonsense reasoning*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.
- Xi Yu Huang, Krishnapriya Vishnubhotla, and Frank Rudzicz. 2024. *The gpt-writingprompts dataset: A comparative analysis of character portrayal in short stories*. Preprint, arXiv:2406.16767.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite

- for foundation models. In *Advances in Neural Information Processing Systems*.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Live-codebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. Follow-bench: 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)*, pages 4667–4688.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. [A dataset of peer reviews \(PeerRead\): Collection, insights and NLP applications](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.
- Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054.
- Bryan Klimt and Yiming Yang. 2004. [The enron corpus: A new dataset for email classification research](#). In *Machine Learning: ECML 2004, 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004, Proceedings*, Lecture Notes in Computer Science, pages 217–226. Springer.
- Frank Konietschke and Markus Pauly. 2014. Bootstrapping and permuting paired t-test type statistics. *Stat. Comput.*, 24(3):283–296.
- Mahnaz Koupaee and William Yang Wang. 2018. [Wikihow: A large scale text summarization dataset](#). Preprint, arXiv:1810.09305.
- Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. [BOOKSUM: A collection of datasets for long-form narrative summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558.
- Eldar Kurtic, Amir Moeini, and Dan Alistarh. 2024. Mathador-lm: A dynamic benchmark for mathematical reasoning on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17020–17027.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAding comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.
- Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-Tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2023. [DS-1000: A natural and reliable benchmark for data science code generation](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18319–18345. PMLR.
- Jinyang Li, Binyuan Hui, Ge Qu, Jiayi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Ma Chenhao, Guoliang Li, Kevin Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. [Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 42330–42357. Curran Associates, Inc.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016.

- How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. [Logiqa: A challenge dataset for machine reading comprehension with logical reasoning](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.
- 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. 2024a. Alignbench: 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.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024b. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*.
- Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. [Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.
- Qingsong Ma, Johnny Wei, Ondrej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation*, pages 62–90.
- MAA. 2024. American invitational mathematics examination-aim 2024.
- Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. [MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. [Www’18 open challenge: Financial opinion mining and question answering](#). In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1941–1942. ACM.
- Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. [ExpertQA: Expert-curated questions and attributed answers](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.
- Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Walenius, and Pyry Takala. 2014. [Good debt or bad debt: Detecting semantic orientations in economic texts](#). *Journal of the American Society for Information Science and Technology*, 65(4):782–796.
- Song Mao, Lejun Cheng, Pinlong Cai, Guohang Yan, Ding Wang, and Botian Shi. 2025. [Deepwriter: A fact-grounded multimodal writing assistant based on offline knowledge base](#). Preprint, arXiv:2507.14189.
- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.
- Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. [Scigen: a dataset for reasoning-aware text generation from scientific tables](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.
- Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. [The E2E dataset: New challenges for end-to-end generation](#). In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

- 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.
- Kun Qian, Shunji Wan, Claudia Tang, Youzhi Wang, Xuanming Zhang, Maximillian Chen, and Zhou Yu. 2024. Varbench: Robust language model benchmarking through dynamic variable perturbation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16131–16161.
- 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.
- Chris Quirk, Raymond Mooney, and Michel Galley. 2015. Language to code: Learning semantic parsers for if-this-then-that recipes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 878–888, Beijing, China. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.
- Sonal Sannigrahi, Josef van Genabith, and Cristina España-Bonet. 2023. Are the best multilingual document embeddings simply based on sentence embeddings? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2261–2271.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4462–4472.
- Zhihong Shao, Yuxiang Luo, Chengda Lu, ZZ Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. 2025. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *arXiv preprint arXiv:2511.22570*.
- Eva Sharma, Chen Li, and Lu Wang. 2019. BIG-PATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.
- Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele Radicioni, and Tommaso Caselli. 2023. WikiBio: a semantic resource for the intersectional analysis of biographical events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384.
- Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617.
- Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhen-nan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19053–19061.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.
- Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordani, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850.

- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhao Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. [RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.
- Siyuan Wang, Zhuohan Long, Zhihao Fan, Xuanjing Huang, and Zhongyu Wei. 2025. Benchmark self-evolving: A multi-agent framework for dynamic LLM evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3310–3328.
- Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. [MAUD: An expert-annotated legal NLP dataset for merger agreement understanding](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16369–16382, Singapore. Association for Computational Linguistics.
- Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024b. [Scibench: Evaluating college-level scientific problem-solving abilities of large language models](#). Preprint, arXiv:2307.10635.
- Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuan Li, Binxin Hu, Wendy Gao, Jiaxing Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. 2024. Benchmarking complex instruction-following with multiple constraints composition. In *Advances in Neural Information Processing Systems*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Pengcheng Yin, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig. 2018. [Learning to mine aligned code and natural language pairs from stack overflow](#). In *International Conference on Mining Software Repositories, MSR*, pages 476–486. ACM.
- Jiahao Ying, Yixin Cao, Yushi Bai, Qianru Sun, Bo Wang, Wei Tang, Zhaojun Ding, Yizhe Yang, Xuanjing Huang, and Shuicheng Yan. 2024. Automating dataset updates towards reliable and timely evaluation of large language models. In *Advances in Neural Information Processing Systems*.
- Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020a. [Dialogue-based relation extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.
- Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. [Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921.
- Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020b. [Reclor: A reading comprehension dataset requiring logical reasoning](#). In *International Conference on Learning Representations*.
- Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, and 1 others. 2025. Futurex: An advanced live benchmark for llm agents in future prediction. *arXiv preprint arXiv:2508.11987*.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. [Character-level convolutional networks for text classification](#). In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzhen Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. 2025. MMLU-CF: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13371–13391.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.
- Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. Characterglm: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1457–1476.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2024a. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*.

Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, and 1 others. 2024b. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931*.

Qin Zhu, Qinyuan Cheng, Runyu Peng, Xiaonan Li, Ru Peng, Tengxiao Liu, Xipeng Qiu, and Xuanjing Huang. 2024c. Inference-time decontamination: Reusing leaked benchmarks for large language model evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9113–9129.

A Dataset List in DyBenchEval

We collect a large number of datasets to construct DyBenchEval, ensuring the quality and diversity. The detailed dataset list is shown as follows according to the domain categories, each of which contains the datasets from all the tasks:

- **STEM:** SciERC (Luan et al., 2018), PubMed (Dernoncourt and Lee, 2017), BigPatent (Sharma et al., 2019), TruthfulQA (Lin et al., 2022), Qasper (Dasigi et al., 2021), ExpertQA (Malaviya et al., 2024), PeerRead (Kang et al., 2018), SciGen (Moosavi et al., 2021), MATH (Hendrycks et al., 2021c), GSM8K (Cobbe et al., 2021), AIME24 (MAA, 2024), MathDial (Macina et al., 2023), ABCD (Chen et al., 2021a), DDxPlus (Fansi Tchango et al., 2022), SciBench (Wang et al., 2024b), TheoremQA (Chen et al., 2023), SciEval (Sun et al., 2024), MBPP (Austin et al., 2021), HumanEval (Chen et al., 2021b), DS-1000 (Lai et al., 2023).
- **Humanity:** BookSum (Kryscinski et al., 2022), WikiBio (Stranisci et al., 2023), BillSum (Eidelman, 2019), LegalBench (Guha et al., 2023), Natural Questions (Kwiatkowski et al., 2019), ROCStories (Mostafazadeh et al., 2016), ELI5 (Fan et al., 2019), WritingPrompts (Huang et al., 2024), LogiQA (Liu et al., 2020), ReClor (Yu et al., 2020b), StrategyQA (Geva et al., 2021), RoleBench (Wang et al., 2024a), LexGLUE (Chalkidis et al., 2022), MMLU (Hendrycks et al., 2021b), Apps (Hendrycks et al., 2021a).

- **Social Science:** CNN-DailyMail (Hermann et al., 2015), FiQA (Maia et al., 2018), NewsQA (Trischler et al., 2017), Enron Email Dataset (Klimt and Yang, 2004), MAUD (Wang et al., 2023), RACE (Lai et al., 2017), ESConv (Liu et al., 2021), FinQA (Chen et al., 2021c), FPB (Malo et al., 2014), Spider (Yu et al., 2018), BIRD (Li et al., 2023).
- **Others:** E2E (Novikova et al., 2017), SAM-Sum (Gliwa et al., 2019), DialogRE (Yu et al., 2020a), Yahoo! Answers (Zhang et al., 2015), Social IQA (Sap et al., 2019), WikiHow (Koupae and Wang, 2018), DeepWriting (Mao et al., 2025), CommonsenseQA (Talmor et al., 2019), PIQA (Bisk et al., 2020), Cosmos QA (Huang et al., 2019), MultiWOZ (Budzianowski et al., 2018), SGD (Rastogi et al., 2020), Taskmaster-1 (Byrne et al., 2019), CaSiNo (Chawla et al., 2021), EmpatheticDialogues (Rashkin et al., 2019), CoNaLa (Yin et al., 2018), IFTTT Recipes (Quirk et al., 2015).

B Illustration of Dataset Embeddings

To demonstrate the effectiveness of dataset embeddings, we first use k-means clustering and compute commonly-used clustering metrics including normalized mutual information (NMI) (Strehl and Ghosh, 2002) and fowlkes-mallows index (FMI) (Fowlkes and Mallows, 1983) compared with the domain and task labels, respectively. We adopt the scikit-learn toolkit (Pedregosa et al., 2011) to implement the clustering algorithm as well as the metrics. All the hyper-parameters are set as default values. The results in Table 7 show that the dataset embeddings indeed capture some latent information of the corresponding domains and tasks.

We further provide a 2D heatmap to illustrate the embedding similarities between different datasets. Figure 3 shows that the datasets of the same domain resemble each other. It is notable that the task of code generation deviates from other tasks in the embedding space, possibly due to the essential difference between texts and codes.

C Implementation Details of Semantic Diversity Metrics

We adopt the Vendi Score (Friedman and Dieng, 2023) and semantic similarity (Chen et al., 2025b) as the semantic diversity metrics for internal and external diversity, respectively. For Vendi Score,

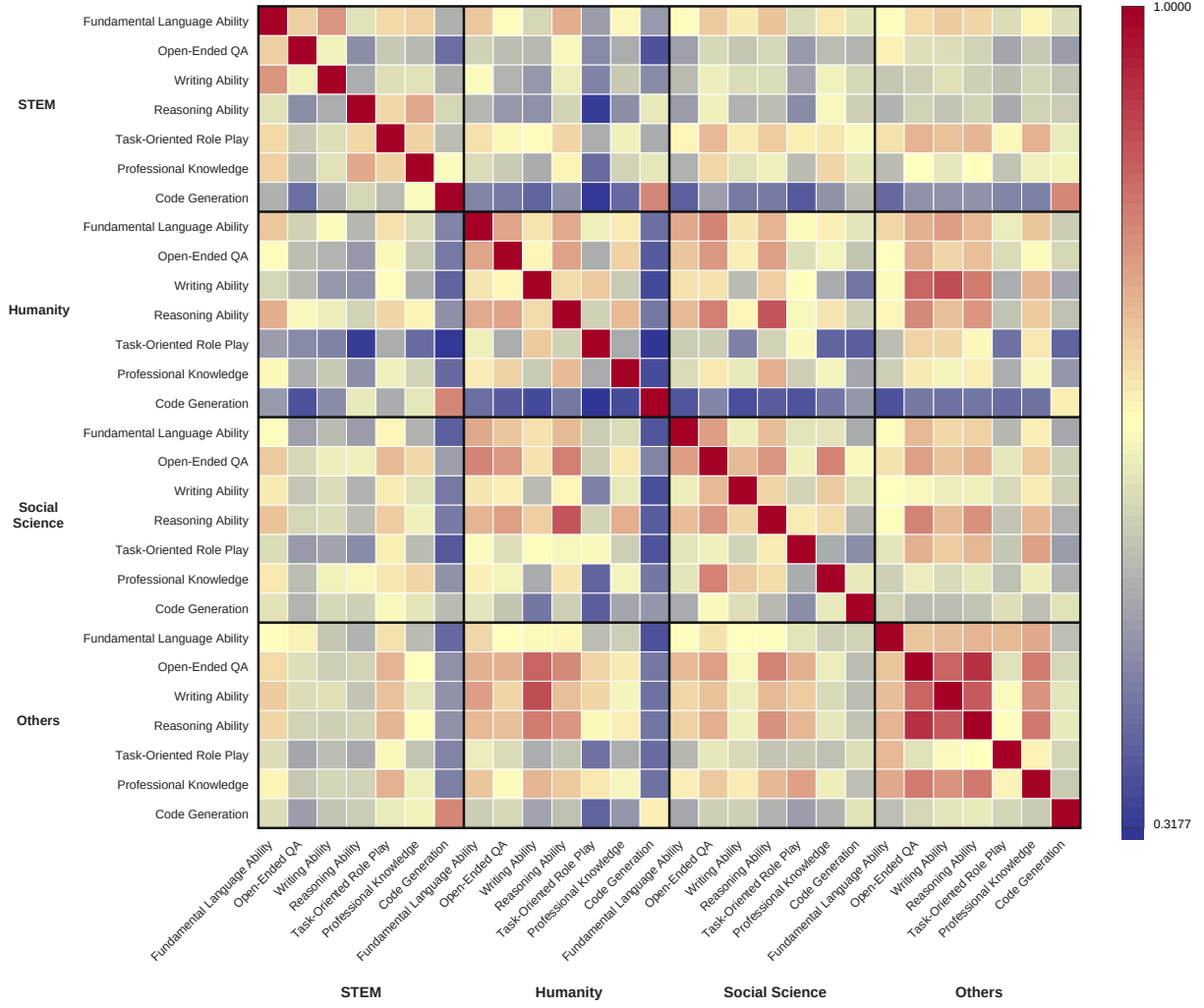


Figure 3: A 2D heatmap of embedding similarities between the datasets of different domains and tasks.

Metric	NMI	FMI
Task	0.5435	0.2996
Domain	0.3208	0.3509

Table 7: Normalized mutual information (NMI) and fowlkes-mallows index (FMI) compared with the domain and task labels, respectively.

we use the official code which adopts SimCSE (Gao et al., 2021) as the embedding model. As for semantic similarity, we adopt Qwen3-Embedding-8B (Zhang et al., 2025) as the embedding model.

D Detailed Results of Performance Ranking

We present the detailed results of performance ranking for each domain and task in Table 9, including top-3 LLMs with their coherence and correctness scores. These results support the necessity to distinguish all the domain-task pairs from another per-

Judge Model	Gemini-3-Pro	GPT-5.2
<i>Closed-Source Language Models</i>		
GPT-5.2-thinking	5.3081	5.1790
Gemini-3-pro-preview	7.8420	5.9399
Claude-opus-4-5	7.0384	5.7015
Qwen3-Max	5.5387	5.2054
<i>Open-Source Language Models</i>		
DeepSeek-V3.2	6.4379	5.5319
GLM-4.6	7.2366	5.7261
Qwen3-235b-a22b-instruct-2507	6.1742	5.4378
Qwen3-30b-a3b-instruct-2507	5.2006	5.0294
Qwen3-8b	4.5273	4.9015
Llama-3.1-70b-instruct	5.2979	5.2018
Llama-3.1-8b-instruct	3.8583	4.5796

Table 8: Coherence of different LLMs on DyBenchEval with Gemini-3-Pro / GPT-5.2 as the judge model.

spective. For example, GLM-4.6 and Qwen3-Max outperform Claude-4.5 in terms of correctness on code generation for social science, while Claude-4.5 dominates the code generation task for most of

		STEM		Humanity		Social Science		Other		ALL Domains	
		Model	Score	Model	Score	Model	Score	Model	Score	Model	Score
Fundamental Language Ability	Coherence	Gemini	8.4500	Gemini	7.9000	GLM	8.3559	Gemini	7.4000	Gemini	8.0084
		Claude	7.8000	DeepSeek	6.8167	Gemini	8.2881	GLM	6.7667	GLM	7.1757
		GLM	7.2833	Claude	6.6167	Claude	7.3559	Claude	6.6333	Claude	7.1004
	Correctness	Claude	0.7667	Gemini	0.8333	Claude	0.9661	Gemini	0.9333	Gemini	0.8619
		DeepSeek	0.7333	DeepSeek	0.7667	Gemini	0.9661	DeepSeek	0.9167	Claude	0.8410
		Gemini	0.7167	Llama-3.1-70b	0.7500	Qwen3-Max	0.9153	Claude	0.9167	DeepSeek	0.8285
Open-Domain QA	Coherence	Gemini	8.7167	Gemini	7.7541	Gemini	8.4359	Gemini	7.6000	Gemini	8.0955
		GLM	7.7833	Claude	7.3607	GLM	8.2308	GLM	7.0333	Claude	7.4091
		Claude	7.7167	GLM	7.2131	GLM	7.6667	Claude	6.6167	GLM	7.4000
	Correctness	Gemini	1.0000	Gemini	1.0000	Claude	1.0000	Claude	0.9833	Gemini	0.9955
		DeepSeek	0.9833	Claude	0.9836	Gemini	1.0000	DeepSeek	0.9833	Claude	0.9773
		Claude	0.9500	Qwen3-235b	0.9508	DeepSeek	0.9744	Gemini	0.9833	DeepSeek	0.9682
Writing Ability	Coherence	Gemini	9.2069	GLM	7.7167	Gemini	8.6167	Gemini	7.9333	Gemini	8.3571
		Claude	8.3448	Gemini	7.7000	GLM	7.6833	GLM	7.6667	GLM	7.8445
		GLM	8.3276	Claude	7.0000	Claude	7.6000	Claude	7.0500	Claude	7.4916
	Correctness	Claude	0.9310	Claude	1.0000	Gemini	1.0000	Claude	1.0000	Gemini	0.9832
		GLM	0.9310	GLM	1.0000	Qwen3-Max	0.9833	Gemini	1.0000	Claude	0.9790
		Gemini	0.9310	Gemini	1.0000	Claude	0.9833	GLM	0.9667	GLM	0.9622
Reasoning Ability	Coherence	Gemini	6.4333	Gemini	8.0167	Gemini	8.1333	Gemini	7.8667	Gemini	7.6125
		DeepSeek	6.0833	Claude	7.6000	Claude	7.4667	GLM	7.6333	GLM	6.9625
		GLM	5.7667	GLM	7.4167	GLM	7.0333	Claude	7.1833	Claude	6.9625
	Correctness	DeepSeek	0.9833	Claude	0.9833	Gemini	0.9333	Claude	0.9833	Claude	0.9542
		Claude	0.9333	Gemini	0.9667	Claude	0.9167	GLM	0.9833	DeepSeek	0.9333
		GLM	0.7833	DeepSeek	0.9500	DeepSeek	0.9000	Gemini	0.9333	GLM	0.8875
Task-Oriented Role Play	Coherence	Gemini	8.3167	Gemini	8.5085	Gemini	8.8167	Gemini	7.2833	Gemini	8.2301
		GLM	7.2833	Claude	8.3051	Claude	7.4333	GLM	6.8167	GLM	7.2343
		DeepSeek	6.7667	GLM	7.9831	GLM	6.8667	DeepSeek	6.2667	Claude	7.0544
	Correctness	Gemini	1.0000	Claude	1.0000	Claude	1.0000	Gemini	0.8333	Gemini	0.9540
		DeepSeek	0.9833	GLM	1.0000	Gemini	0.9833	Claude	0.8167	Claude	0.9456
		Claude	0.9667	Gemini	1.0000	DeepSeek	0.9667	GLM	0.7500	DeepSeek	0.9121
Professional Knowledge	Coherence	Claude	6.5000	Gemini	7.3667	Gemini	8.3625	Gemini	7.8621	Gemini	7.4651
		GLM	6.3833	GLM	7.0833	GLM	7.8500	GLM	7.5345	GLM	7.2597
		Gemini	5.9833	Claude	6.7500	Claude	7.6750	Claude	6.8103	Claude	6.9922
	Correctness	Claude	0.9667	Claude	0.9000	DeepSeek	0.8875	Gemini	1.0000	Claude	0.9186
		DeepSeek	0.9667	DeepSeek	0.8833	Claude	0.8625	Claude	0.9655	DeepSeek	0.9109
		GLM	0.7667	Qwen3-Max	0.8667	Qwen3-Max	0.8250	Qwen3-Max	0.9310	Gemini	0.8411
Code Generation	Coherence	Gemini	6.1167	Gemini	7.7843	Gemini	7.9667	Gemini	6.8500	Gemini	7.1558
		GLM	5.8167	GLM	7.1961	GLM	7.6167	GLM	6.5500	GLM	6.7792
		Claude	5.5000	Claude	7.0980	Claude	7.0833	Claude	5.5167	Claude	6.2684
	Correctness	Claude	0.9667	Claude	0.9608	GLM	0.9833	Claude	1.0000	Claude	0.9740
		DeepSeek	0.9500	Gemini	0.9608	Qwen3-Max	0.9833	Gemini	1.0000	Gemini	0.9697
		Gemini	0.9500	DeepSeek	0.9216	Claude	0.9667	GLM	0.9500	GLM	0.9264
All Tasks	Coherence	Gemini	7.5957	Gemini	7.8613	Gemini	8.3708	Gemini	7.5407	Gemini	7.8420
		GLM	6.9426	GLM	7.2749	GLM	7.5885	GLM	7.1411	GLM	7.2366
		Claude	6.8565	Claude	7.2482	Claude	7.5215	Claude	6.5311	Claude	7.0384
	Correctness	Claude	0.9258	Gemini	0.9440	Claude	0.9498	Gemini	0.9545	Claude	0.9405
		DeepSeek	0.9258	Claude	0.9343	Gemini	0.9426	Claude	0.9522	Gemini	0.9225
		Gemini	0.8493	DeepSeek	0.9051	DeepSeek	0.9187	GLM	0.8995	DeepSeek	0.9099

Table 9: Top-3 LLMs for each domain and task in terms of coherence and correctness, where each LLM binds with a color for readability.

the other domains. This demonstrates that although Claude-4.5 possesses strong general capabilities of generating codes, its understanding of social science problems may degrade the overall performance. This distinction validates the importance of distinguishing the evaluation of different domains for each task.

E Prompt Template Design

We provide the prompt templates for cross data filling, LLM generation, and the evaluation of coherence / correctness in Figure 4, 5, 6, and 7, respectively.

F Detailed Results with Different Judge Models

To show the effect of different judge models, we additionally adopt GPT-5.2 as the judge model to measure the coherence of all the LLMs. From Table 8, the spearman correlation between the evaluation results of the two judge models is 0.9909, showing almost the same performance ranking. We observe that Gemini-3-Pro still achieves the best performance with GPT-5.2 as the judge model, preliminarily demonstrating that its superior performance is not due to the self-enhancement bias (Zheng et al., 2023).

Prompt on data augmentation

```
# Role Definition
You are a professional dataset generation engine. Your responsibility is to generate NEW, ORIGINAL,
and HIGH-QUALITY dataset samples that strictly satisfy the explicitly defined Task and domain
requirements.
You MUST prioritize the formal definitions below over any example content.

# 1. Explicit Generation Requirements (HIGHEST PRIORITY)
## Target domain: {target_domain_disp}
{domain_spec}
### Sub-domain Selection Protocol (MANDATORY)
Sub-domain selection mode:
**{sub_domain_mode}**
{selected_sub_domain_block}
## Target Task: {target_task_disp}
{task_spec}

# 2. Output Contract (STRICT)
Generate exactly ONE complete dataset sample.
The sample must:
- Be self-contained
- Be answerable
- Have a clearly identifiable correct output
- Strictly match the target task and the sub-domain
DO NOT generate multiple questions.
DO NOT include meta explanations.

# 3. Reference Examples (LOW PRIORITY)
The following examples are provided only to help calibrate language style, complexity, and realism.
You MUST NOT:
- Copy entities, facts, or structures
- Rephrase or adapt example content
- Infer requirements from examples
### Example 1
[Domain: {target_domain_disp}] | [Task: {target_task_disp}]
{shot1_content}
### Example 2
[Domain: {shot2_domain_disp}] | [Task: {target_task_disp}]
{shot2_content}
### Example 3
[Domain: {shot3_domain_disp}] | [Task: {target_task_disp}]
{shot3_content}
### Example 4
[Domain: {target_domain_disp}] | [Task: {shot4_task_disp}]
{shot4_content}
### Example 5
[Domain: {target_domain_disp}] | [Task: {shot5_task_disp}]
{shot5_content}

# Final Instruction
Generate ONE new dataset sample that strictly satisfies:
- the target task
- the target domain
- the selected or derived sub-domain
Focus on reasoning depth, institutional realism, and correctness.
```

Figure 4: Prompt template design for cross data filling.

Prompt on main experiment
<pre> # Role You are an Advanced Data Augmentation Architect. Your goal is to generate a high-quality "Parallel Test Case" based on a provided SEED DATA entry. # Objective The SEED DATA is likely contaminated (memorized by models). You must create a **new** problem-answer pair that tests the *exact same* underlying skill or logic as the seed, but looks different enough to bypass memorization. # 1. Degrees of Freedom (Where you MUST be creative) You are granted freedom to modify the following elements to ensure the new data is fresh and diverse: * **Context Re-skinning:** Completely change the story or setting. (e.g., If the seed is about "calculating trajectory of a missile", you can change it to "calculating the path of a magic spell" or "robot arm movement", provided the math remains applicable). * **Variable Perturbation:** You may alter specific numbers, entity names, or attributes, as long as the change does not break the logic. (e.g., Change "3 red balls" to "5 blue crystals", adjusting the answer accordingly). * **Input Format:** You may slightly vary how the information is presented (e.g., changing a list to a paragraph description), provided it remains clear. # 2. Strict Anchors (Where you MUST NOT deviate) Despite the freedom above, you must strictly adhere to the seed's core essence: * **Core Logic Invariance:** The specific reasoning step, algorithm, or knowledge point required to solve the problem must remain identical. Do not simplify or complicate the underlying logic. * **Difficulty Anchoring:** The cognitive load required to solve the new problem must match the seed. Do not turn a 2-step reasoning problem into a 5-step one, or vice versa. * **Category Consistency:** If the seed is a "coding task", the output must be code. If it is "information extraction", the output must be extraction. # 3. Quality Assurance * The new Question must be unambiguous. * The new Answer must be rigorously derived from the new Question. * **Avoid Hallucination:** Ensure the new domain is logically sound. # Output Format Output ONLY a single valid JSON object (compact JSONL format). The structure must match the seed keys. # Input Seed Data {{SEED_DATA}} # Action Based on the guidelines above, generate the new JSONL line now. **Double-check that your generated Output is the mathematically/logically correct solution to your generated Input.** </pre>

Figure 5: Prompt template design for LLM generation in our main experiment.

```
Prompt on coherence experiment

You are a Senior Benchmark Security Expert. Your goal is to rigorously evaluate a "Rewritten Question" to ensure it serves as a valid, uncontaminated test case for LLMs.

### Input Data
**Original Seed Question:**
{seed}
**Rewritten Question:**
{rewrite}

### Evaluation Protocol
You must follow a strict 3-step evaluation process using the **Fixed Scoring Standards** provided below.

#### Step 1: Dimension Definition & Analysis
First, identify 3-4 critical dimensions for evaluating this specific type of question (e.g., "Degree of Disguise", "Logic Preservation", "Solvability").
For **EACH** dimension:
1. Analyze the rewrite's performance.
2. Assign a sub-score (1-10) for that specific dimension.

#### Step 2: Synthesis against Fixed Standards
Synthesize your dimensional analysis and map the *overall quality* to the following strict standards. Do not deviate from these definitions:
* **Score 1-2 (Invalid Question)**
  The rewrite is not a functional problem (e.g., gibberish, not a question at all).
* **Score 3-4 (High Similarity)**
  The rewrite is too similar to the original seed; it is practically identical with no meaningful variation.
* **Score 5-6 (Surface-Level Substitution)**
  The rewrite performs only simple noun replacements (e.g., changing names or objects). Crucially, the overall domain, theme, and structural framework remain unchanged.
* **Score 7-8 (Deep Transformation & Decontamination)**
  The rewrite undergoes a drastic transformation in context, variables, and setting. It is almost impossible to identify the connection to the original question. The magnitude of change is significantly larger than the 5-6 range, ensuring that a model exposed to the original question cannot solve this rewrite simply by recalling its memory.
* **Score 9-10 (Super-Human Excellence)**
  The quality is even better than the 7-8 range. The rewrite significantly exceeds the question-setting standards of human domain experts.

#### Step 3: Final Verdict
Determine the final integer score based on the standards above.

### Output Requirements
Output your response in the following XML format:
<dimensional_analysis>
[Define 3-4 dimensions. For each, provide analysis and a sub-score]
1. Dimension Name: ...
  - Analysis: ...
  - Sub-score: [x/10]
2. Dimension Name: ...
  - Analysis: ...
  - Sub-score: [x/10]
...
</dimensional_analysis>

<synthesis_and_reasoning>
[Explain why the rewrite fits into a specific score range of the Fixed Standards based on the dimensions above]
</synthesis_and_reasoning>

<final_verdict>
[Output the final integer score (1-10) in the exact format below:]
###FINAL_RESULT: [[n]]

(Replace 'n' with your integer score, e.g., 1, 6, 10).
</final_verdict>
```

Figure 6: Prompt template design for the evaluation of coherence.

Prompt on correctness experiment
You are an expert Data Quality Assurance Specialist. Your task is to verify the Correctness of the "Answer" within the modified data using a rigorous Chain-of-Thought process. Input Data: Original Data (Reference Only): \ "\ " {seed} \ "\ " Modified Data (Target for Evaluation): \ "\ " {rewrite} \ "\ " Evaluation Protocol: CRITICAL GUARDRAIL 1: INDEPENDENCE * IGNORE the Original Data's content for verification. The Modified Data may have changed numbers, logic, or facts. * Evaluate the Modified Data as a STANDALONE entity. You must recalculate or re-verify the answer strictly based on the information provided in the "Modified Data". * Example: If Original says "1+1=2" and Modified says "1+2=3", the Modified data is CORRECT (Pass). Do not compare it to the Original. CRITICAL GUARDRAIL 2: FOCUS * IGNORE Question Quality: Do not judge if the question is simple or weird. * IGNORE Style: Do not judge the writing quality. * FOCUS ONLY: Is the Answer objectively correct relative to the Question <i>in the Modified Data</i>? Step 1: Data Parsing & Extraction (From Modified Data Only) Analyze the Modified Data to lock in the target for evaluation: * Domain A (Explicit Q&A): If the text clearly separates a Question and an Answer, identify them directly. * Domain B (Unstructured/Narrative): If no explicit Q&A structure exists, you must EXTRACT a valid pair from the content: * Extracted Question: What is the core problem/topic presented in the Modified Data? * Extracted Answer: What is the specific solution provided in the Modified Data? Step 2: Autonomous Dimension Generation & Analysis Based on the content type of the Modified Data, self-define 3 specific dimensions to judge the correctness of the Answer. * Action: Define the dimension name (e.g., "Math Logic", "Code Syntax", "Historical Fact", "Internal Consistency"), then evaluate the Answer against it using only the Modified Data's context. Step 3: Synthesis Synthesize the results from your 3 self-defined dimensions. * Result 1 (Pass): The answer is objectively correct and logically sound based on the new data's premises. * Result 0 (Fail): The answer is factually wrong, contains calculation errors, or is logically invalid within the modified text. Response Template: Reasoning Process: * Phase 1: Extraction (Based on Modified Data) * Target Question: [Extracted Question] * Target Answer: [Extracted Answer] * Phase 2: Multi-Dimensional Analysis * Dimension 1 [[Self-Defined Name]]: [Detailed Check based on new context...] * Dimension 2 [[Self-Defined Name]]: [Detailed Check based on new context...] * Dimension 3 [[Self-Defined Name]]: [Detailed Check based on new context...] * Phase 3: Synthesis * Summary: [Is the answer correct according to the modified question?] FINAL_RESULT: [[x]]

Figure 7: Prompt template design for the evaluation of correctness.