MMLU-CF: A Contamination-free Multi-task Language Understanding Benchmark

Qihao Zhao Yangyu Huang* Tengchao Lv Lei Cui Furu Wei
Qinzheng Sun Ying Xin Shaoguang Mao Xin Zhang Qiufeng Yin Scarlett Li
Microsoft Research

Abstract

Multiple-choice question (MCQ) datasets like Massive Multitask Language Understanding (MMLU) are widely used to evaluate the commonsense, understanding, and problem-solving abilities of large language models (LLMs). However, the open-source nature of these benchmarks and the broad sources of training data for LLMs have inevitably led to benchmark contamination, which is studied in our contamination evaluation experiment, resulting in unreliable evaluation. To alleviate this issue, we propose the contamination-free MCQ benchmark called MMLU-CF, which reassesses LLMs' understanding of world knowledge by averting both unintentional and deliberate data contamination. To mitigate unintentional data contamination, we source questions from a broader domain of over 200 billion webpages and apply three specifically designed decontamination rules. To prevent deliberate data contamination, we divide the benchmark into validation and test sets with similar difficulty and subject distributions. The test set remains closed-source to ensure reliable results, while the validation set is publicly available to promote transparency and facilitate independent evaluation. We evaluated over 40 mainstream LLMs on the MMLU-CF. Compared to the original MMLU, not only LLMs' performances significantly dropped but also the performance rankings of them changed considerably. This indicates the effectiveness of our approach in establishing a contamination-free and fairer evaluation standard. The GitHub repository is available at https://github.com/microsoft/MMLU-CF and the dataset refers to https://huggingface.co/datasets/microsoft/MMLU-CF.

1 Introduction

Given the emergence of powerful capabilities in large language models (LLMs) such as GPT-4 (Achiam et al., 2023), Llama (Meta, 2024), Gemini (Reid et al., 2024), and Claude-3 (Anthropic, 2023), evaluation of these models has become particularly important for understanding their strengths and limitations. Consequently, a number of benchmarks covering reasoning (Hendrycks et al.; Wang et al., 2024), reading comprehension, mathematics (Cobbe et al., 2021), science (Rein et al., 2023), and coding (Yu et al., 2023) have been explored and released. Among them, Massive Multitask Lan-

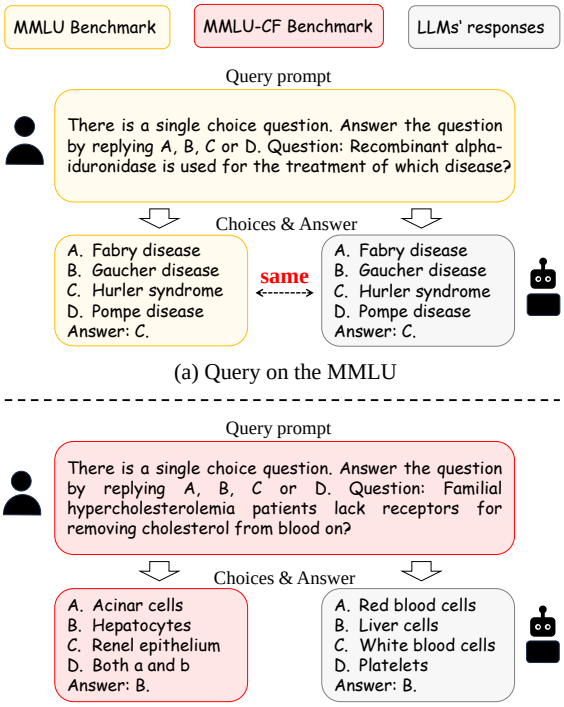


Figure 1: (a) An instance of contamination in MMLU. When questions are used as prompts from the MMLU, certain LLMs, due to their memorization capabilities, directly provide **choices identical to the original ones**. (b) When questions are used as prompts from the MMLU-CF, LLMs only provide guessed choices. This indicates that the MMLU test set suffers from data contamination and memorization by some LLMs, while the proposed MMLU-CF avoids such leakage. Further details are analyzed in Section 4.5 of Appendix.

*Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13371–13391

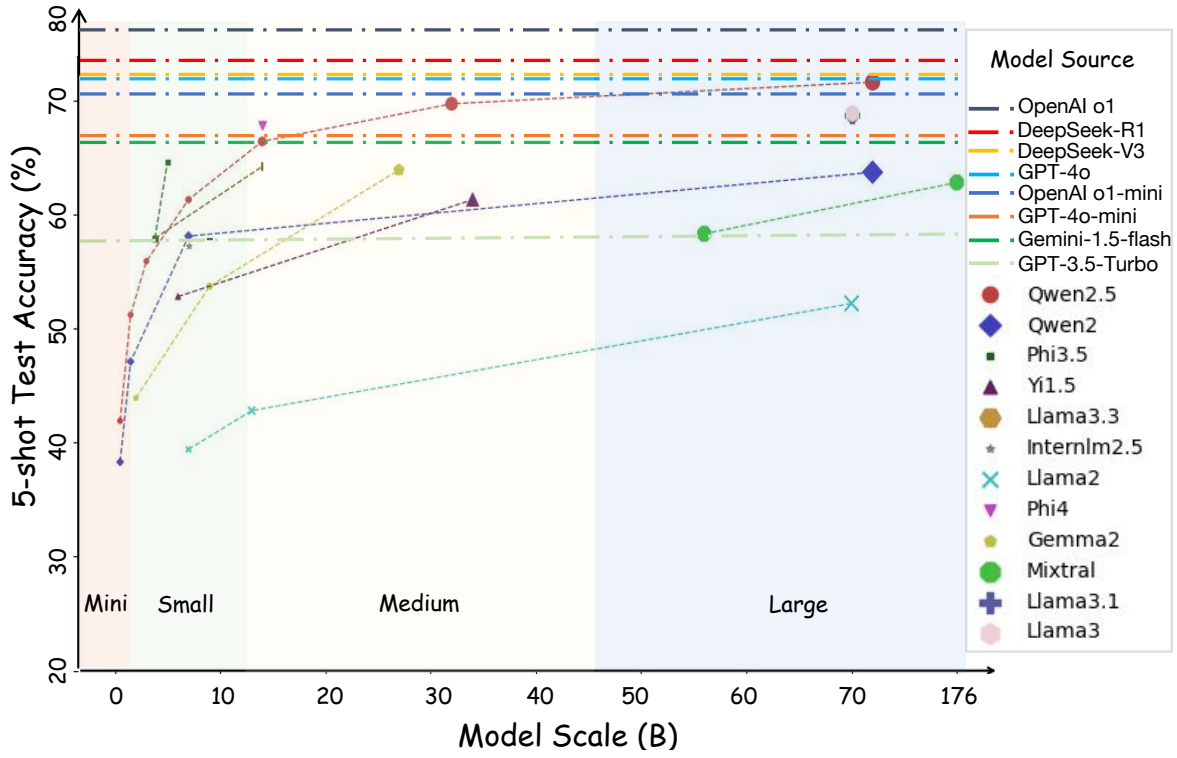


Figure 2: The 5-shot results on the MMLU-CF test set encompass mainstream open-source models ranging from 0.5 billion (B) to 176 billion (B) parameters, including closed-source API models.

guage Understanding (MMLU) (Hendrycks et al.) is a widely used multiple-choice question (MCQ) gold standard benchmark because it covers various disciplines and difficulty levels, allowing for a comprehensive evaluation of LLMs' performance across diverse domains.

However, data leakage or contamination, where LLMs inadvertently encounter benchmark data during training, can compromise the effectiveness, reliability, and fairness of these evaluations (Deng et al., 2024; Roberts et al., 2023; Srivastava et al., 2024), termed **unintentional contamination**. Additionally, due to the public availability of benchmarks and the ability of LLMs to memorize data (Carlini et al., 2023), instances of **deliberate contamination** may occur. As illustrated in Figure 1, we observe that when only given the questions, some LLMs directly provide the choices and answers, where the choices are exactly the same as those in the MMLU test set. This indicates that the benchmark may have been deliberately added to the training set and that LLMs have memory for these questions.

To fairly investigate the world knowledge of LLMs, we propose *MMLU-CF*, a contamination-free multiple-choice question benchmark for LLMs. To minimize the risk of benchmark exposure and contamination, we perform five key processing steps for the data: (1) MCQ Collection, (2) MCQ Cleaning, (3) Difficulty Sampling, (4) LLMs Checking, and (5) Contamination-Free Processing. In the contamination-free processing step, we employ three rules to rewrite the questions, which could be referred to as Section 3.2. For humans, rewriting questions without changing their meaning does not affect their ability to answer. However, for LLMs, if they do not understand the question in depth, they may memorize the original question and struggle to answer the rewritten question. **Therefore, the rewriting helps mitigate the unintentional contamination**. Based on the rewritten questions, we construct the MMLU-CF consisting of 10,000 questions for the test set and another 10,000 questions for the validation set. **To prevent deliberate contamination, the test set remains closed-source**, while the validation set is open-source for transparency and convenient evaluation. Noteworthily, we propose a strategy to submit an evaluation request for the test set, as outlined in Section 4.7. Currently, the performance gap between the validation and test sets for each evaluated LLM is quite small. As LLM development progresses, this performance gap indicates the contamination degree of LLMs in the future.

We benchmark leading open-source and closed-source LLMs on the MMLU-CF test and validation sets, including OpenAI o1(Jaech et al., 2024), GPT-4o (OpenAI, 2024), GPT-4o-mini (OpenAI, 2024),

DeepSeek (Guo et al., 2025; Liu et al., 2024), Qwen (Bai et al., 2023; Team, 2024), Llama (Meta, 2024), Phi (Abdin et al., 2024), and many more. The 5-shot test results are briefly summarized in Figure 2. More results and analysis could refer to Section 4.3 and Section 4.5.

2 Related Work

2.1 General LLMs Benchmark

In the field of natural language processing (NLP), benchmarks play a crucial role in evaluating and comparing the performance of different large language models (Wang et al., 2018; Cobbe et al., 2021; Hendrycks et al.; Wang et al., 2024; Zhou et al., 2023; Zheng et al., 2023; Rein et al., 2023; Zhang et al., 2024; Hendrycks et al.; Phan et al., 2025). They serve as a common ground for fair comparison, fostering transparency and reproducibility in research. For instance, GLUE (Wang et al., 2018; Sarlin et al., 2020) is a collection of nine different tasks designed to evaluate the natural language understanding capabilities of models. GSM8K (Cobbe et al., 2021) is a benchmark dataset of 8,000 high-quality, linguistically diverse grade school math word problems. It is designed to evaluate the problem-solving abilities of language models, requiring a combination of language understanding and mathematical reasoning. MMLU (Hendrycks et al.) is a benchmark designed to evaluate a model's multitask learning capabilities across a diverse set of 57 tasks, including high school mathematics, college-level biology, law, and more, focusing on testing the model's generalization ability across different domains. Building upon this, MMLU-Pro (Wang et al., 2024) enhances the benchmark by introducing more challenging, reasoning-focused questions and expanding the choice set from four to ten choices, shifting the emphasis from knowledge retrieval to reasoning. Further, MMLU-Pro+ (Asgari et al.) extends MMLU-Pro by assessing shortcut learning and higher-order reasoning in large language models, offering a comprehensive evaluation of both reasoning depth and model robustness. These benchmarks have become standard tools in the evaluation of large language models due to their widespread adoption and comprehensive coverage of various domains.

However, these benchmarks, including GSM8K, MMLU, MMLU-Pro, and MMLU-Pro+, do not account for contamination prevention.

2.2 Contamination-free Benchmark

Several benchmark datasets have been introduced for contamination-free evaluation. KIEval (Yu et al., 2024) is an interactive framework with an LLM-powered "interactor" for multi-round dialogues to assess deep comprehension beyond mere recall. However, its multi-turn dialogue mechanism is time-consuming and poses reproducibility challenges. LatestEval (Li et al., 2024) creates dynamic reading comprehension evaluations from recent texts using a three-step process: collecting texts, extracting key information, and constructing questions with template-filling or LLMs. The questions are generated rather than from the real world. LiveCodeBench (Jain et al., 2024) continuously collects new coding problems from LeetCode, AtCoder, and CodeForces for a contamination-free benchmark, revealing performance drops in some models, such as DeepSeek (Guo et al., 2024), which specifically for code domain. GSM1K (Zhang et al., 2024) assesses the true reasoning ability of large language models by creating a new benchmark with a similar style and complexity to GSM8k, revealing significant accuracy drops of memorization in many LLMs. However, it is the only closed-source benchmark that focuses solely on math with a small scale of questions. LiveBench (White et al., 2024) introduces (1) frequently updated questions from recent sources, (2) automatic scoring based on ground-truth values, and (3) a variety of challenging tasks, including math, coding, reasoning, language, instruction following, and data analysis. However, it requires re-evaluating LLMs on regularly updated questions. This leads to high costs for maintenance and uncertainty for reevaluation.

Consequently, there is no contamination-free benchmark that assesses LLMs' understanding of world knowledge on a large scale, across diverse domains, and with a one-time, low-cost evaluation. Some of the existing benchmarks are small-scale for specific domains, while others rely on complex strategies, such as dynamic dialog, that incur high evaluation costs and pose reproducibility challenges. Additionally, some require regular updates to their question sets, leading to unstable difficulty distribution and high re-evaluation costs.

Unlike the methods mentioned above, we categorize data contamination into unintentional and deliberate types. To mitigate unintentional data contamination, we apply three decontamination rules while collecting data from a broader domain.

Meanwhile, our MMLU-CF benchmark keeps the test set closed-source to prevent deliberate data contamination. To the best of our knowledge, MMLU-CF is the first contamination-free benchmark for world knowledge understanding under large scale and diverse domains, which provides reproducible results and a convenient evaluation process.

3 The MMLU-CF Benchmark

3.1 Overview

The MMLU-CF benchmark contains 20,000 data points and spans 14 fields, screened from 200+ billion documents on public open websites. To produce this diverse, high-quality, safety and contamination-free benchmark, we employ a series of steps, shown in Figure 3. Ultimately, we curate a dataset comprising 10,000 questions for the test set and 10,000 questions for the validation set respectively. The test set remains closed-source to prevent deliberate exposure of the questions (Zhang et al., 2024), while the validation set is open-source to validate the authenticity and effectiveness of the questions. The following sections outline the steps to process the raw data.

3.2 Dataset Construction Pipeline

MCQ Collection. The data sources are diversified broadly to preliminary mitigate the data exposure in existing LLMs. Specifically, over 200 billion documents are collected from open-source websites, and rule-based methods are employed to extract 2.7 million multiple-choice questions with the corresponding answers as the raw data [1]. Unlike previous efforts, such as those by (Hendrycks et al.; Wang et al., 2024), which collect data from a few sources, these 2.7 million questions encompassed over 3,000 different website domains, ensuring a wide variety of knowledge, which spans 14 fields, including Health, Math, Physics, Business, Chemistry, Philosophy, Law, Engineering, etc.

MCQ Cleaning. With the 2.7 million raw questions, a series of filtering and standardization strategies are employed for initial data cleaning. Firstly, we filtered out questions with length less than 10 or larger than 512 characters and removed question numbers if they existed. Secondly, we removed questions that had fewer than four choices, contained empty content, or had inconsistent choice numbers. Meanwhile, we also eliminated questions whose choice numbers do not start with "a,b,c,d," "1,2,3,4," or "i,ii,iii,iv" in lower characters. Then, we mapped the choice numbers and corresponding answers to "A, B, C, D.". Thirdly, we discarded questions without answers or with answers that were not among the choices and further formalized the answer to ensure including both choice number and content. Finally, we only kept English questions and performed deduplication. Through these steps, the data scale was reduced to 1.66 million.

Difficulty Sampling. Firstly, GPT-4o is used to categorize the difficulty levels of the original MMLU under prompt in Table 6. The difficulty distribution of MMLU is illustrated in Figure 4, that shows nearly one-third of the questions have a difficulty level below [4]. To categorize our data using the same difficulty standard as MMLU, the questions in MMLU dataset were selected as samples for few-shot labeling under the same scoring prompt and LLM. We selected questions using a normal distribution centered around a difficulty level of [6], as indicated in Figure 4, to construct an appropriate difficulty distribution. During sampling process, we maintained a balanced distribution of question categories, maximized the diversity of domains, and ensured that questions had corresponding explanations whenever possible for the diversity, reliability, and quality. Ultimately, 50,000 questions are sampled from the 1.6 million questions.

LLMs Checking. To mitigate single LLM judgment bias, we employed multiple LLMs, including GPT-4o, Gemini, and Claude, to review the quality and harmlessness of these questions.
For the quality of questions, we assessed them based on the following criteria:

- Context and Clarity: Are the question and choices consistent and unambiguous, providing enough context for understanding?
- Logical Consistency: Are the question and choices structured without contradictions?
- Factual Accuracy: Are the question and choices factually correct and not misleading?
- Mutual Exclusivity: Are choices mutually exclusive without overlap?
- Correct Answer: Is the correct answer included in the choices?

From the perspective of harmlessness, we reviewed the content from the following four aspects:

- Non-hatred: Ensure the content does not contain hate speech.

[1] To avoid replicating the raw data collection pipeline, we will not disclose the data sources or extraction rules.

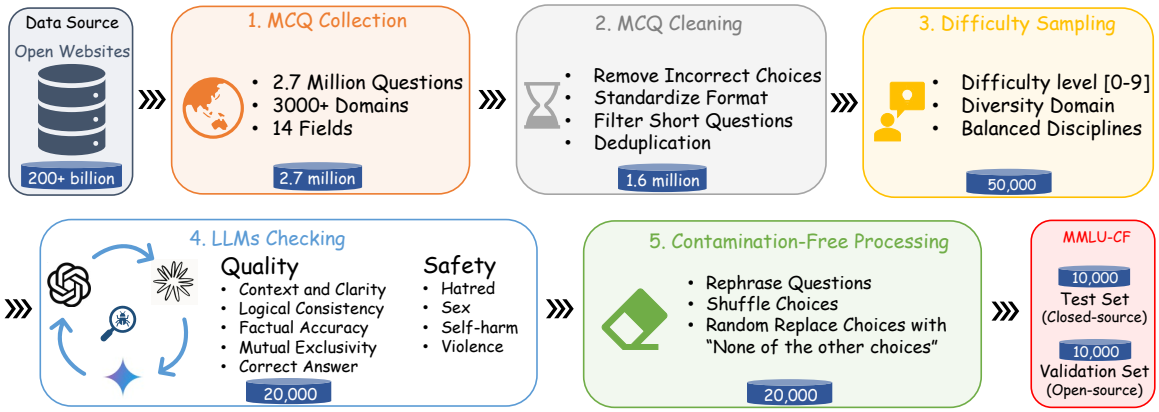


Figure 3: The construction pipeline of the MMLU-CF Benchmark. The pipeline involves (1) MCQ Collection to gather a diverse set of questions; (2) MCQ Cleaning to ensure quality; (3) Difficulty Sampling to ensure an appropriate difficulty distribution for questions; (4) LLMs checking: The LLMs, including GPT-4o, Gemini, and Claude, are reviewing the accuracy and safety of the data; and (5) Contamination-Free Processing to prevent data leakage and maintain dataset purity. Ultimately, this process results in the MMLU-CF, consisting of 10,000 questions for the closed-source test set and 10,000 for the open-source validation set.

- Non-sex: Ensure the content does not inappropriate sexual suggestions or content.
- Non-selfharm: Ensure the content neither contains self-harm nor encourages self-harm.
- Non-violence: Ensure the content does not contain violence or incite violence.

Specifically, these three LLMs are applied to rate each question on a scale from level 1 to 5, where level 5 represents the highest quality. Then, questions with an average score greater than level 4 were selected to construct validation and test sets of MMLU-CF. Additionally, inspired by Decontaminator (Yang et al., 2023), GPT-4o is utilized to perform redundancy detection (Yang et al., 2023) on semantically identical test and validation questions. To further ensure diversity, the selected questions came from over 1,000 web domains. The detailed prompts could refer to the Section A.8 in the Appendix.

Contamination-Free Processing. Although the questions are sourced from diverse domains, they are publicly available and could still be included in LLMs training data. We hope the constructed benchmark could assess the LLMs' understanding ability rather than their memorization of answers (Carlini et al., 2023). To mitigate this unintentional contamination, we implemented the following three decontamination rules based on the principles of simplicity and understandability, as shown in Figure 5:

(1) Rule 1: Rephrase Question. To avoid the model rote learning (memorizing) the question rather than understanding it, we rephrased the questions using powerful GPT-4o. The manual check in the Section A.1 of Appendix ensure that the rephrased questions retain their original meaning.

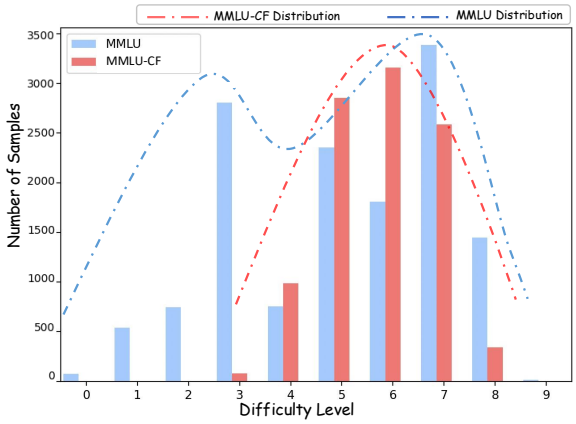


Figure 4: The difficulty levels produced by GPT-4o for MMLU and MMLU-CF are analyzed. In our data, we randomly sampled 10,000 questions for visualization.

(2) Rule 2: Shuffle Choices. To prevent the model from answering correctly based on memorizing the order of choice sequence, we shuffled the choices (Gupta et al., 2024). If the last choice was 'None of the above' or 'All of the above,' only the first three choices are randomly shuffled.

(3) Rule 3: Random Replace Choices. To further make the questions different from the original ones, we randomly replaced one of their choices with 'None of the other choices' with a 50% probability. If the last choice was 'None of the above' or 'All of the above', we skipped this question. When the

replaced the choice is the correct answer, the correctness of this choice remains. Similarly, when an incorrect choice is replaced, it acted as a distractor. Both of them require the model to employ more understanding and reasoning to answer correctly.

The first two rules do not change the difficulty of the questions, while the last rule does alter it. We place the Contamination-Free Processing at the final step due to its high computational cost. Although these rules in it are simple, they effectively mitigate the unintentional contamination and make the dataset more challenging. For a detailed analysis, refer to Section 4.4. The difficulty distribution can be found in the Section A.9.

After that, we divided the final data into 10,000 questions as validation set and 10,000 questions as test set, which maintain the similar difficulty and discipline distribution. The test set is kept closed-source to prevent deliberate contamination.

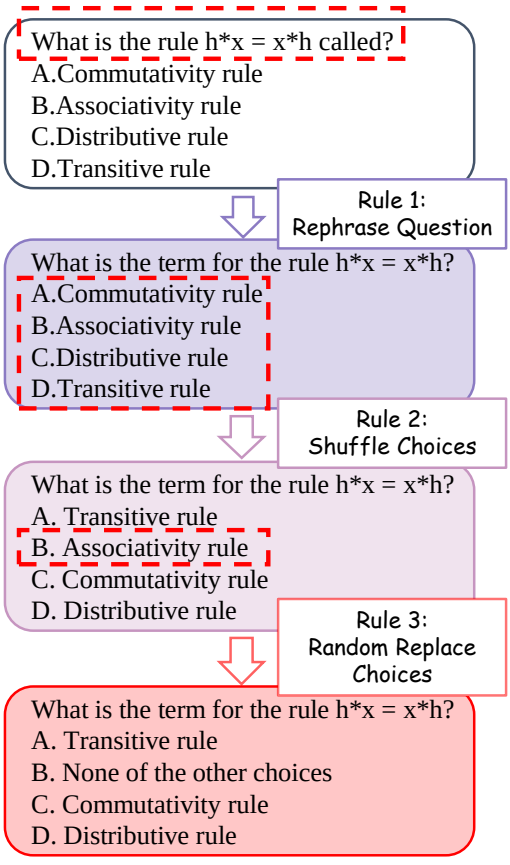


Figure 5: A MCQ instance by Contamination-free Processing. The top box is the input MCQ, and the bottom box is the decontaminated MCQ.

4 Experiments

4.1 Evaluation Models

We evaluate 40+ models across various sizes by the evaluation platform OpenCompass (Contributors, 2023), including open-source models ranging from 0.5B to 72B and closed-source APIs. The experiments include models with different classes, such as GPTs (Achiam et al., 2023) (GPT-4o (v2024-10-1), GPT-o1-mini (v2024-09-12), GPT-4o-mini (v2024-10-1), GPT-4-Turbo (v2024-2-15), Gemini (Reid et al., 2024) (Gemini-1.5-Flash), and public models like Llama-3-{8, 70}B-chat (Meta, 2024), Llama-3.1-{8, 70}B-chat (Meta, 2024), Mixtral-{7, 8x7, 8x22}B-instruct, Phi-4 (Abdin et al., 2024), Phi-3.5-{mini, small} (Abdin et al., 2024), Gemma-2-{2, 9, 27}B (Team et al., 2024), Qwen2.5-{0.5, 1.5, 7, 14, 70}B (Team, 2024).

4.2 Evaluation Metrics

We employ both 5-shot and 0-shot approaches to measure the performance of LLMs on the MMLU-CF test and validation set. Additionally, the open-source models are categorized based on their parameter size into four sections: Large (>50B), Medium (13B-50B), Small (6B-12B), and Mini (0.5-5B). The results on the validation set, 0-shot evaluation results, evaluation prompt, and corresponding settings are detailed in Section A.6 of the Appendix.

4.3 Evaluation Results and Analysis

As shown in Table 4 and Figure 2, the results on a range of mainstream LLMs are presented, which reveal several noteworthy findings:

- The performance ranks on MMLU of evaluated LMs are disrupted by the new benchmark, **MMLU-CF**. The top 3 LMs — **OpenAI o1**, **Deepseek-R1**, and **Deepseek-V3** — maintain their leading positions without any changes in their rankings.

- Interestingly, among the notable rank changes (>= 3 positions), the decreases in rank tend to be more significant than the increases. On average, the LMs that decreased in rank fell by **5.14** positions, while those that increased in rank rose by **3.78** positions. This asymmetry suggests that performance is *easier to significantly drop than to rise*, potentially due to data contamination in pretraining corpora.

Model	MMLU Rank	MMLU (%)	MMLU-CF Rank	MMLU-CF (%)	Diff. (%)	Rank Change
OpenAI o1 (Jaech et al., 2024)	1	92.3	1	80.3	-12.0	—
Deepseek-R1 (Guo et al., 2025)	2	90.8	2	76.3	-13.5	—
Deepseek-V3 (Liu et al., 2024)	3	88.5	3	73.9	-14.6	—
GPT-4o (OpenAI, 2024)	4	88.0	4	73.4	-14.6	—
OpenAI o3-mini (OpenAI, 2025)	5	86.9	5	73.4	-13.5	—
GPT-4-Turbo (Achiam et al., 2023)	6	86.5	8	70.4	-16.1	↓2
Llama-3.3-70B-instruct (Meta, 2024)	7	86.3	11	68.8	-17.5	↓4
Llama-3.1-70B-instruct (Meta, 2024)	8	86.0‡	12	68.7	-17.3	↓4
Qwen2.5-72B-instruct (Team, 2024)	9	85.3	6	71.6	-13.7	↑3
OpenAI o1-mini (Jaech et al., 2024)	10	85.2	7	71.2	-14.0	↑3
Phi-4-14B (Abdin et al., 2024)	11	84.8	13	67.8	-17.0	↓2
Qwen2.5-32B-instruct (Team, 2024)	12	83.9†	9	69.7	-14.2	↑3
Qwen2-72B-instruct(Bai et al., 2023)	13	82.3	20	63.7	-18.6	↓7
Llama-3-70B-instruct	14	82.0	10	68.9	-13.1	↑4
GPT-4o-mini (OpenAI, 2024)	15	81.8	15	65.5	-16.3	—
Qwen2.5-14B-instruct (Team, 2024)	16	79.9†	14	66.4	-13.5	↑2
Phi-3.5-MoE-instruct (Abdin et al., 2024)	17	78.9	17	64.6	-14.3	—
Gemini-1.5-Flash (Reid et al., 2024)	18	78.7	16	64.8	-13.9	↑2
Phi-3-medium-instruct (Abdin et al., 2024)	19	77.9	18	64.2	-13.7	↑1
Qwen2.5-7B-instruct (Team, 2024)	20	76.8†	22	61.3	-15.5	↓2
Yi-1.5-34B-chat (Young et al., 2024)	20	76.8	22	61.3	-15.5	↓2
Internlm-3-8B-chat (Cai et al., 2024)	22	76.6†	24	60.3	-16.3	↓2
Mixtral-8x22B-instruct(Jiang et al., 2024)	23	76.2	21	62.8	-13.4	↑2
Qwen1.5-72B-chat(Bai et al., 2023)	24	75.6	23	59.8	-15.8	↓1
Gemma2-27B (Team et al., 2024)	25	75.2	19	63.9	-11.3	↑6
Internlm-2.5-7B-chat (Cai et al., 2024)	26	72.8	32	57.3	-15.5	↓6
Glm-4-9B-chat (GLM et al., 2024)	27	72.4	31	57.8	-14.6	↓4
GPT-3.5-Turbo(OpenAI, 2023)	28	71.4	27	58.2	-13.2	↑1
Gemma-2-9B (Team et al., 2024)	29	71.3	36	53.7	-17.6	↓7
Phi-3-mini-instruct (3.8B) (Abdin et al., 2024)	30	70.9	29	57.9	-13.0	↑1
Qwen2-7B-instruct (Bai et al., 2023)	31	70.5	28	58.1	-12.4	↑3
Mixtral-8x7B-instruct-v0.1 (Jiang et al., 2024)	31	70.5	26	58.3	-12.2	↑5
Phi-3.5-mini-instruct (3.8B)	33	69.1	29	57.9	-11.2	↑4
Llama-2-70B-chat (Meta, 2024)	34	68.9	38	52.2	-16.7	↓4
Llama-3-8B-instruct (Meta, 2024)	35	68.4	32	57.3	-11.1	↑3
Llama-3.1-8B-instruct (Meta, 2024)	36	68.1	34	57.1	-11.0	↑2
Qwen2.5-3B-instruct (Team, 2024)	37	64.4†	35	55.9	-8.5	↑2
Yi-1.5-6B-chat (Young et al., 2024)	38	62.8	37	52.8	-10.0	↑1
Mistral-7B-instruct-v0.3 (Jiang et al., 2024)	39	60.3	40	50.7	-9.6	↓1
Qwen2.5-1.5B (Team, 2024)	40	58.5†	39	51.2	-7.3	↑1
Baichuan-2-13B-chat (Yang et al., 2023)	41	57.3	42	48.3	-9.0	↓1
Deepseek-v2-lite-chat (DeepSeek-AI, 2024)	42	55.7	41	49.3	-6.4	↑1
Llama-2-13B-chat (Meta, 2024)	43	54.8	46	42.8	-12.0	↓3
Baichuan-2-7B-chat (Yang et al., 2023)	44	52.9	44	44.5	-8.4	—
Qwen2-1.5B-instruct (Bai et al., 2023)	45	52.4	43	47.1	-5.3	↑2
Gemma-2-2B (Team et al., 2024)	46	51.3	45	43.9	-7.4	↑1
Internlm-2-chat-1.8b (Cai et al., 2024)	47	47.1	48	40.5	-6.6	↓1
Llama-2-7B-chat (Meta, 2024)	48	45.3	49	39.4	-5.9	↓1
Qwen2.5-0.5B (Team, 2024)	49	45.1†	47	41.9	-3.2	↑2
Qwen2-0.5B-instruct (Bai et al., 2023)	50	37.9	50	33.3	-4.6	—

Table 1: Performance of various models on MMLU and MMLU-CF (ours) in 5-shot test set evaluations without using CoT (Kojima et al., 2022), except for additional explanations. **Diff.** means the score difference of models between MMLU and MMLU-CF. ‡ Denotes 0-shot with COT. † Indicates employing MMLU-redux (Gema et al., 2024), the results are from the Qwen2.5 homepage (Team, 2024).

- Additionally, smaller LMs appear to be more disruptive in the new **MMLU-CF** benchmark compared to their larger counterparts.
- Notable rank **decreases** (>= 3 positions) were observed for the following LMs: **Qwen2-72B-instruct** (↓7), **Gemma-2-9B** (↓7), **Internlm-2.5-7B-chat** (↓6), **Glm-4-9B-chat** (↓4), **Llama-2-70B-chat** (↓4), **Llama-3.3-70B-instruct** (↓4), and **Llama-3.1-70B-instruct** (↓4).
- Notable rank **increases** (>= 3 positions) were observed for the following LMs: **Gemma2-27B** (↑6), **Mixtral-8x7B-instruct-v0.1** (↑5), **Phi-3.5-mini-instruct (3.8B)** (↑4), **Llama-3-70B-instruct** (↑4), **Qwen2.5-72B-instruct** (↑3), **GPT-o1-mini** (↑3), **Qwen2.5-32B-instruct** (↑3), **Qwen2-7B-instruct** (↑3), and **Llama-3-8B-instruct** (↑3).

4.4 Effective of Decontamination Rules

Comprehensive experiments are conducted on MMLU and MMLU-CF to evaluate different LLMs' performance under three decontamination rules: Rephrase Question (Rule 1), Shuffle Choices (Rule 2), and Random Replace Choices (Rule 3).

The detailed results summarized in Table 2 demonstrate the impact of each rule to the different benchmarks on different LLMs.

Firstly, all rules caused a performance drop across both the MMLU and MMLU-CF datasets and for all three LLMs.

Among the three rules, rules 1 and 2, which

do not affect the difficulty distribution of the test set, lead to performance drops, which verifies the effectiveness of data decontamination.

Rule 3, which also results in a performance drop, makes the rewritten questions more challenging.

Secondly, after applying the three rules, all evaluated LLMs exhibits a greater performance drop on MMLU than MMLU-CF, which confirms LLMs on MMLU have worse contamination. The data source of MMLU is more likely to be included in the training data of LLMs compared to MMLU-CF, which is sourced from over 200 billion documents.

Lastly, less powerful models, GPT-3.5-Turbo and Llama-3-8b, show a more pronounced performance decline, which demonstrates smaller models are more susceptible to contamination.

Rule 1	Rule 2	Rule 3	GPT-4o	GPT-3.5-Turbo	Llama-3.1-8b
			MMLU		
-	-	-	88.0	71.4	68.1
✓	-	-	86.1	68.5	66.3
✓	✓	-	85.0	67.3	65.2
✓	✓	✓	79.8 (-8.2)	62.1 (-9.3)	59.1 (-9.0)
			MMLU-CF		
-	-	-	79.8	65.3	63.8
✓	-	-	78.6	63.1	62.3
✓	✓	-	77.9	62.8	61.8
✓	✓	✓	73.4 (-6.4)	58.2 (-7.1)	57.1 (-6.7)

Table 2: 5-shot results of applying different decontamination rules to MMLU and MMLU-CF test set.

4.5 Analysis of Data Contamination

Figure 1 presents examples of data contamination in MMLU, where LLMs generate responses that exactly match the original choices of the given question under a specific prompt. To analyze this further, we tested 1,000 cases sampled from MMLU and MMLU-CF using 40 models. As shown in Figure 6, approximately 10% of the models exhibited significant contamination on MMLU, producing outputs that matched 1%–5% of the choices, while 90% matched less than 1%. When applying decontamination rules on MMLU, 97.5% of the models produced outputs that matched under 1% of the choices.In contrast, 100% of the models matched less than 0.2% of the choices in MMLU-CF. The experiment verifies the presence of contamination in MMLU and confirms the contamination-free property of our MMLU-CF.

4.6 Disciplinary Distribution of MMLU-CF

The Figure 7 demonstrates the visualization of MMLU-CF test and validation sets. We find that

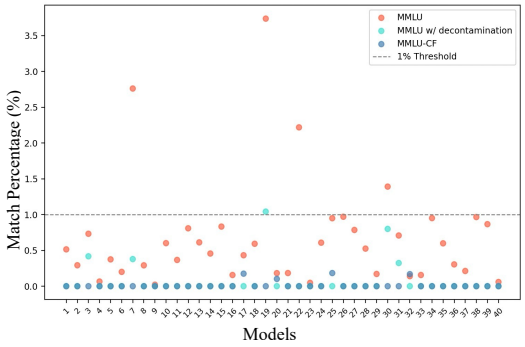


Figure 6: LLM Response Match Rates with MMLU, MMLU using decontamination rules and MMLU-CF. Evaluated LLMs have obvious data leakage on MMLU.

their disciplinary distribution proportions are quite similar. The most prevalent disciplines are Computer Science, Health, and History, with proportions of 13.1%, 12.2%, and 11.5% in the test set, respectively. This distribution may lead to slight differences in the performance of different models. Table 3, we present the performance of various large models across different disciplines. GPT-4o achieves the best performance in terms of average accuracy and different disciplines. We observe that the models perform worst in Computer Science. This is because the domain not only requires fundamental knowledge of Computer Science but also involves code understanding, which increases the difficulty. Qwen2.5-72B, -32B brings new upgrades in mathematics and coding, delivering the best results in mathematics, engineering, and computer science. Despite its small size, Phi-4 achieves competitive results compared to larger models, showcasing its efficiency in handling complex tasks.

4.7 Testset Evaluation Methods

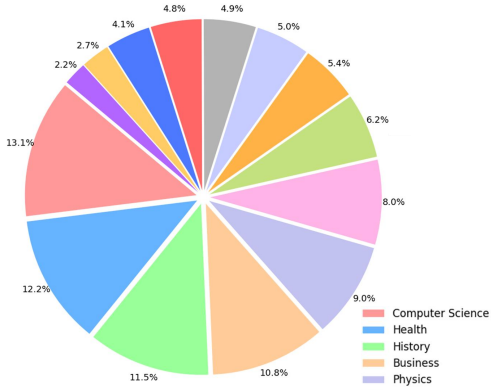
Two evaluation methods are supported for the test set of MMLU-CF. The users could submit evaluation requests by providing Hugging Face open-source model IDs or API formats through the introduction of our GitHub project homepage. Besides, we will actively evaluate the latest popular models from Hugging Face as well as mainstream APIs.

5 Conclusion

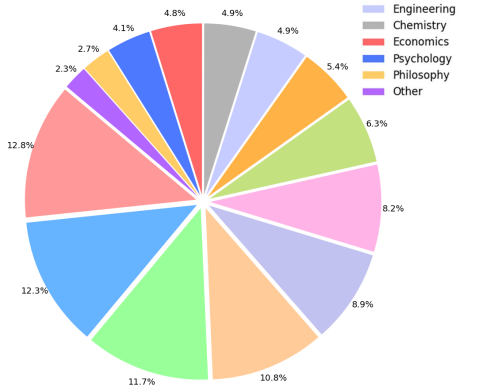
In this paper, we construct MMLU-CF, a contamination-free and challenging multiple-choice question benchmark under large scale and diverse disciplines, to reassess LLMs' understanding of world knowledge. Specifically, we categorize data contamination into unintentional and deliberate types. To mitigate unintentional con-

Subject	GPT-4o (OpenAI, 2024)	GPT-4o-mini (OpenAI, 2024)	Llama-3.3-70B (Meta, 2024)	Qwen2.5-72B (Team, 2024)	Qwen2.5-32B (Team, 2024)	Phi-4-14B (Abdin et al., 2024)
Math	56.09	45.83	56.3	**67.51**	63.10	62.18
Physics	**75.15**	64.47	69.0	74.00	71.12	69.15
Chemistry	**72.44**	66.54	68.3	69.62	68.81	67.13
Law	**81.46**	72.73	73.6	75.15	72.55	71.84
Engineering	60.15	55.41	56.7	**61.39**	57.69	54.67
Economics	**78.33**	66.31	72.5	74.95	68.90	68.88
Health	**81.09**	76.11	79.3	80.23	78.55	76.29
Psychology	**80.10**	70.28	77.5	78.95	77.94	75.45
Business	70.90	63.81	64.7	**71.00**	68.69	65.19
Biology	**82.84**	74.63	75.5	78.88	74.53	75.91
Philosophy	**81.82**	77.99	78.9	74.24	72.73	76.08
Computer Science	55.50	51.09	51.0	56.12	**68.79**	51.09
History	**77.23**	67.05	71.2	71.19	68.79	68.09
Other	**74.83**	64.74	67.9	68.15	66.88	65.55
Average	**73.42**	65.52	68.82	71.60	68.81	67.68

Table 3: Performance of different models on MMLU-CF discipline under a 5-shot test set. The best result is emphasized in bold.



(a) Test Set



(b) Validation Set

Figure 7: Distribution of Disciplines in MMLU-CF.

tamination, we collect data from a broader domain and design three decontamination rules for further processing. To prevent deliberate contamination, we keep the test set closed-sourced for ensuring no leakage while making the validation set open-sourced for transparency and convenient evaluation. The evaluation results of over 40 mainstream LLMs on MMLU-CF reveal significant performance drops and performance ranking changes compared to the original MMLU benchmark. The leakage analysis of MMLU and the ablation study of three decontamination rules indicate the presence of contamination in MMLU and the effectiveness of our approach in establishing a decontaminated benchmark. We believe this contamination-free benchmark would promote fair and reliable LLMs evaluation and provide valuable insights for the design of future benchmarks.

6 Limitations

Despite being constructed with the utmost objectivity and fairness and leveraging multiple large language models to verify the correctness of the questions and answers, some errors may still remain in this dataset. To address this, we have provided a validation set that is available to the public for further scrutiny and verification. Additionally, this dataset primarily focuses on multiple-choice questions and language modalities. However, other aspects of large models' capabilities, such as math and code reasoning, multi-modal understanding (e.g., image and audio), and specific domain expertise, still require evaluation with similarly unbiased and contamination-free benchmarks.

7 Ethics Statement

MMLU-CF is constructed based on public data sources and methodologies to ensure transparency. The benchmark is designed to provide fair and reliable evaluations through decontamination rules and verification. While efforts are made to minimize errors, some may remain, and we encourage the community to review the publicly available validation set for further improvements. This benchmark focuses on language modalities, and future work is needed for unbiased evaluation in other areas. We call for the responsible use of this dataset to promote ethical and equitable AI development.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024b. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2023. Introducing the next generation of claude.

Saeid Asgari, Aliasghar Khani, and Amir Hosein Khasahmadi. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms. In *Neurips Safe Generative AI Workshop 2024*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Nathan Gan, Kai Jie, Akhil Agrawal, Jonathan Byrd, and Mark Chen. 2021. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. 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 8698–8711.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. 2024. Are we done with mmlu? *arXiv preprint arXiv:2406.04127*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. 2024. Changing answer order can decrease mmlu accuracy. *arXiv preprint arXiv:2406.19470*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. b. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al.

2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18600–18607.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Meta. 2024. Build the future of ai with meta llama 3.

OpenAI. 2023. Gpt-3.5 turbo fine-tuning and api updates.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Hello gpt-4o.

OpenAI. 2025. Introducing openai o3 and o4-mini.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In *The Twelfth International Conference on Learning Representations*.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

Saurabh Srivastava, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince, Sooraj Thomas, et al. 2024. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. 2023b. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhaojian Yu, Xin Zhang, Ning Shang, Yangyu Huang, Can Xu, Yishujie Zhao, Wenxiang Hu, and Qiufeng Yin. 2023. Wavecoder: Widespread and versatile enhanced instruction tuning with refined data generation. *arXiv preprint arXiv:2312.14187*.

Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. 2024. Kieval: A knowledge-grounded interactive evaluation framework for large language models. *arXiv preprint arXiv:2402.15043*.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*.

Jianguo Zhang, Tian Lan, Ming Zhu, Zuxin Liu, Thai Hoang, Shirley Kokane, Weiran Yao, Juntao Tan, Akshara Prabhakar, Haolin Chen, et al. 2024b. xlam: A family of large action models to empower ai agent systems. *arXiv preprint arXiv:2409.03215*.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Appendix

A.1 Human Review for LLM-Generated Outputs

In the processes of Difficulty Sampling, LLMs Checking, and Contamination-Free Processing, LLMs were employed to filter and modify instances. However, since LLMs are not flawless, false-positive cases may occur, resulting in outputs that deviate from the intended results.

To evaluate the performance of LLMs in data processing, we conducted a human review process. Specifically, three volunteers performed random sampling checks at a certain ratio. For difficulty level assessments, 100 instances were reviewed, achieving approximately 97% accuracy. For quality and safety scoring, 100 instances were checked, with an accuracy of about 99%. Additionally, we compared the 100 rephrased questions with their original versions, confirming successful rephrasing in approximately 99% of cases.

These results indicate that although LLMs have limitations, the use of a state-of-the-art GPT-4o model results in rare errors during these processes, ensuring the reliability of the generated data.

A.2 More Evaluation Results

As shown in Table 4, the 5-shot and 0-shot results on a range of mainstream LLMs are presented, which show several findings:

- The performance trend across different scales of LLMs remains consistent with the original MMLU: more powerful LLMs always have higher performance. This trend underscores the quality of constructed MMLU-CF.
- The performance of all evaluated LLMs significantly drops compared to the original MMLU benchmark, likely due to the decontamination rules and difficulty sampling process applied to the MMLU-CF dataset. Table 2 shows performance impact individually of each rule.
- The performance ranking of the evaluated LLMs varied across different model architectures and versions. This variation suggests the presence of potential contamination in the original MMLU. For instance, the early version of some models have higher score than their later version.
- The performance between the validation and test sets is narrowing gap now. The gap between them could serve as an index for assessing the contamination degree of the published validation sets in the future.

A.3 Relation between Test and Validation Sets

We partition the benchmark dataset into test and validation sets, then calculate the absolute score difference of LLMs performance as Δ. It not only demonstrates no contamination of the current LLMs on the validation set but also offers a method to assess the contamination degree of the future LLMs on the validation set. As shown in Table 4, before the validation set is publicly released, about 60% of Δ values are less than 0.5, and 96% of Δ values are below 1.0. This indicates that the evaluation results of LLMs are significantly consistent across the test and validation sets, demonstrating the effectiveness of the validation set in evaluating model performance now. Once the validation set is made public, potential data leakage can cause the models to memorize the validation set, leading to an increase in Δ values. Thus, the design Δ serves as a method to monitor whether benchmarks might be compromised. This approach helps ensure the fairness and integrity of the benchmarks, preventing models from exploiting leaked data to artificially enhance their performance.

A.4 Manual Quality Validation for Filtered Instances

To ensure the reliability of the filtered examples, we conducted a manual quality check process to verify whether the 20,000 examples were accurately rewritten or whether the choices were modified based on the original content without introducing new information, and to determine whether the score changes were caused by rule adjustments rather than increased noise. Specifically, we randomly sampled 200 LLM-checked questions and manually reviewed their quality, improving question accuracy from 86% to 99.5%. This process ensures that filtered instances are reliable and enhances the validity of score comparisons across different rule applications.

A.5 The Effect of Decontamination Rules

In the methods section, we presented three types of question modification rules applied to the MMLU-CF dataset: question rephrasing, shuffling choices, and randomly replacing a choice with "None of the other choices." To validate the effectiveness of these modifications, we first applied these three

	Model	MMLU 5-shot (%)	MMLU-CF 5-shot (%)			MMLU-CF 0-shot (%)		
		Test	Test	Validation	Δ (%)	Test	Validation	Δ (%)
API	OpenAI o1 (Jaech et al., 2024)	92.3	80.3	80.3	+0.0	80.1	79.9	+0.2
	Deepseek-R1 (Guo et al., 2025)	90.8	76.3	75.3	-1.0	75.7	76.0	-0.3
	Deepseek-V3 (Liu et al., 2024)	88.5	73.9	74.9	-1.0	72.9	71.4	+1.5
	GPT-4o (OpenAI, 2024)	88.0	73.4	73.4	+0.0	71.9	72.4	-0.5
	OpenAI o3-mini (OpenAI, 2025)	86.9	73.4	73.6	-0.2	73.7	73.6	+0.1
	OpenAI o1-mini (Jaech et al., 2024)	85.2	71.2	71.0	+0.2	71.6	71.5	+0.1
	GPT-4-Turbo (Achiam et al., 2023)	86.5	70.4	70.1	+0.3	68.9	68.7	+0.1
	GPT-4o-mini (OpenAI, 2024)	81.8	65.5	65.1	+0.4	66.0	65.3	+0.7
	Gemini-1.5-Flash (Reid et al., 2024)	78.7	64.8	64.9	-0.1	56.7	56.9	-0.2
	GPT-3.5-Turbo (OpenAI, 2023)	71.4	58.2	59.0	-0.8	57.2	58.1	-0.9
Large	Qwen2.5-72B-instruct (Team, 2024)	85.3	71.6	71.3	+0.3	70.6	70.4	+0.2
	Llama-3-70B-instruct (Meta, 2024)	82.0	68.9	68.8	+0.1	68.1	67.4	+0.7
	Llama-3.3-70B-instruct (Meta, 2024)	86.3	68.8	67.8	+1.0	67.6	67.5	+0.1
	Llama-3.1-70B-instruct (Meta, 2024)	86.0‡	68.7	68.1	+0.6	70.4	69.7	+0.7
	Phi-3.5-MoE-instruct (Abdin et al., 2024)	78.9	64.6	64.5	+0.1	63.1	62.1	+1.0
	Qwen2-72B-instruct (Bai et al., 2023)	82.3	63.7	64.3	-0.6	62.4	62.5	-0.1
	Mixtral-8x22B-instruct (Jiang et al., 2024)	76.2	62.8	62.5	+0.3	65.3	64.8	+0.5
	Qwen1.5-72B-chat (Bai et al., 2023)	75.6	59.8	60.2	-0.4	59.1	59.6	-0.5
	Llama-2-70B-chat (Meta, 2024)	68.9	52.2	51.8	+0.4	51.2	50.9	+0.3
Medium	Qwen2.5-32B-instruct (Team, 2024)	83.9†	69.7	68.8	+0.9	68.9	68.8	+0.1
	Phi-4-14B (Abdin et al., 2024)	84.8	67.8	68.5	-0.7	68.5	69.4	-0.9
	Qwen2.5-14B-instruct (Team, 2024)	79.9	66.4	66.1	+0.3	67.0	66.0	+1.0
	Phi-3-medium-instruct (Abdin et al., 2024)	77.9	64.2	64.2	+0.0	62.5	62.7	-0.2
	Gemma2-27B(Team et al., 2024)	75.2	63.9	63.5	+0.4	64.2	64.0	+0.2
	Yi-1.5-34B-chat (Young et al., 2024)	76.8	61.3	60.5	+0.8	60.6	59.5	+1.1
	Mixtral-8x7B-instruct-v0.1 (Jiang et al., 2024)	70.5	58.3	57.1	-1.2	58.9	58.5	+0.4
	Deepseek-v2-lite-chat (DeepSeek-AI, 2024)	55.7	49.3	48.7	+0.6	48.2	47.7	+0.5
	Baichuan-2-13B-chat (Yang et al., 2023)	57.3	48.3	48.6	-0.3	47.1	48.1	-1.0
	Llama-2-13B-chat (Touvron et al., 2023)	54.8	42.8	42.1	+0.7	44.8	44.6	+0.2
Small	Qwen2.5-7B-instruct (Team, 2024)	76.8†	61.3	60.4	+0.9	59.3	58.6	+0.7
	Internlm-3-8B-chat (Cai et al., 2024)	76.6†	60.3	60.5	-0.2	61.8	61.4	+0.4
	Qwen2-7B-instruct (Bai et al., 2023)	70.5	58.1	57.9	+0.2	58.3	57.4	+0.9
	Glm-4-9B-chat (GLM et al., 2024)	72.4	57.8	57.9	-0.1	58.6	58.7	-0.1
	Internlm-2.5-7B-chat (Cai et al., 2024)	72.8	57.3	56.8	+0.5	57.9	56.9	+1.0
	Llama-3-8B-instruct (Meta, 2024)	68.4	57.3	56.5	+0.8	56.4	55.4	+1.0
	Llama-3.1-8B-instruct (Meta, 2024)	68.1	57.1	57.9	-0.8	56.1	56.1	+0.0
	Gemma-2-9B (Team et al., 2024)	71.3	53.7	53.3	+0.4	32.1	31.2	+0.9
	Yi-1.5-6B-chat (Young et al., 2024)	62.8	52.8	51.4	+1.4	52.2	51.9	+0.3
	Mistral-7B-instruct-v0.3 (Jiang et al., 2023)	60.3	50.7	50.9	-0.2	51.1	50.9	+0.2
	Baichuan-2-7B-chat (Yang et al., 2023)	52.9	44.5	43.9	+0.6	43.9	44.0	-0.1
	Llama-2-7B-chat (Touvron et al., 2023)	45.3	39.4	38.5	+0.9	41.9	40.9	+1.0
Mini	Phi-3-mini-instruct (3.8B) (Abdin et al., 2024)	70.9	57.9	58.1	-0.2	58.2	57.5	+0.7
	Phi-3.5-mini-instruct (3.8B) (Abdin et al., 2024)	69.1	57.9	57.4	+0.5	58.3	57.7	+0.6
	Qwen2.5-3B-instruct (Team, 2024)	64.4†	55.9	56.4	-0.5	54.3	53.9	+0.4
	Qwen2.5-1.5B-instruct (Team, 2024)	50.7†	51.2	51.0	+0.2	50.7	50.4	+0.3
	Qwen2-1.5B-instruct (Bai et al., 2023)	52.4	47.1	47.5	-0.4	45.2	44.5	+0.7
	Gemma-2-2B (Team et al., 2024)	51.3	43.9	42.4	+1.5	30.5	29.4	+0.9
	Qwen2.5-0.5B-instruct (Team, 2024)	24.1†	41.9	41.1	+0.8	36.0	34.9	+1.1
	Internlm-2-chat-1.8b (Cai et al., 2024)	47.1	40.5	39.4	+1.1	41.2	39.8	+1.4
	Qwen2-0.5B-instruct (Bai et al., 2023)	37.9	38.3	38.3	+0.0	33.5	33.5	+0.0

Table 4: Performance of various models on MMLU and MMLU-CF (ours). Both 0-shot and 5-shot evaluations don't employ COT (Kojima et al., 2022), except for additional explanations. Δ means the absolute score difference of models between validation and test sets. ‡ denotes 0-shot with COT. † indicates employing MMLU-redux (Gema et al., 2024), the results are from Qwen2.5 homepage (Team, 2024).

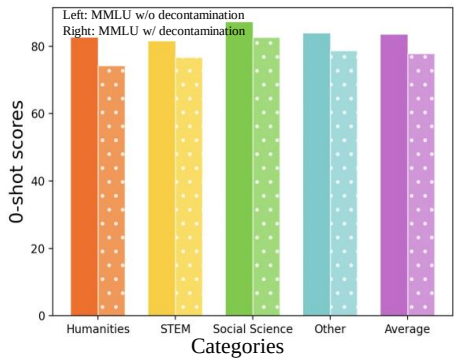
rules to the MMLU (Hendrycks et al.). The results, shown in Figure 8, indicate that these modifications lead to a decrease in 5-shot and 0-shot scores for GPT-4o. Furthermore, when comparing these results to those on the MMLU-CF dataset, as depicted in Table 2, the accuracy drop is more pronounced on the MMLU dataset. This suggests a higher likelihood of data leakage in large models when using the MMLU dataset. In contrast, the MMLU-CF dataset, due to its broad and closed-source nature, exhibits a lower risk of data leakage.

Model	Temperature	Maximum Generation Tokens
Deepseek-R1	0.6	32,768
Deepseek-V3	0.7	8,192
GPT-o1-mini	1.0	8,192
GPT-4o	0.7	2,048
GPT-3.5	0.7	2,048
GPT-4-turbo	0.7	2,048
Qwen2.5 Series	0.7	4,096
Others	0.7	1,024

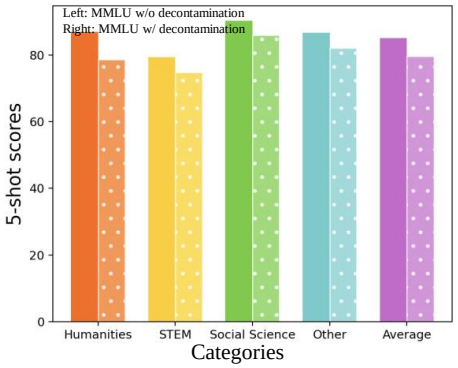
Table 5: Temperature and Maximum Generation Token settings for different LLMs

A.6 Prompt and Setting for Evaluation

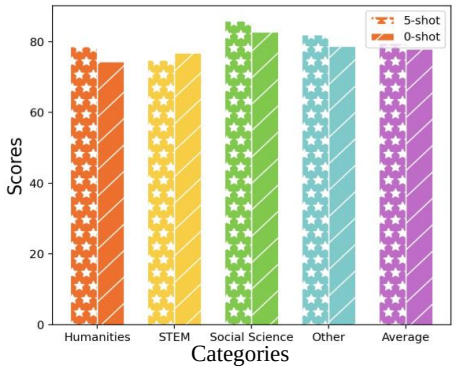
In all our experiments, the applied prompts of all the evaluated LLMs are the same, which could refer



(a) GPT-4o 5-shot scores



(b) GPT-4o 0-shot scores



(c) GPT-4o on MMLU w/ decontamination

Figure 8: GPT-4o evaluation comparison on MMLU with and without our decontamination rules.

to Table 8. This prompt is also used as the MMLU prompt in several well-known language model evaluation tools. Moving forward, we will continue using this prompt consistently to ensure comparability across tests. Moreover, Table 5 presents the temperature parameter settings and maximum generation token limits used in our experiments.

A.7 Prompt for Difficulty Labeling

Table 7 demonstrates the prompt used in the LLMs difficulty labeling. We do not provide choices and answers in the prompt to avoid data leakage through LLMs API calling.

There is a single choice question (with answers). Answer the question by replying A, B, C or D. Question: {Question}? A. {A} B. {B} C. {C} D. {D}

Table 6: The prompt of LLMs Question Answering. {Question} represents the question instance, while {A}, {B}, {C}, and {D} denote the answer choices.

|Instruction| Please rate the difficulty of this question on a scale of [0-9], where level [0] represents the easiest question and level [9] represents the most difficult. Question: {Question}

Table 7: The prompt of LLMs Difficulty Labeling. {Question} represents the question instance.

A.8 Prompt for LLMs Checking

Table 8 shows the prompt used in the LLMs checking processing to verify the correctness of questions. For safety, we used GPT-4's built-in safety filter under the strongest constraints to filter out unsafe content related to hate speech, sexual content, self-harm, and violence.

[Instruction] Please review the following question and corresponding choices for correctness based on these criteria: **Context and Clarity**: Are the question and choices consistent and unambiguous, providing enough context for understanding? **Logical Consistency**: Are the question and choices logically structured without contradictions? **Factual Accuracy**: Are the question and choices factually correct and not misleading? **Mutual Exclusivity**: Are choices mutually exclusive without overlap? **Correct Answer**: Is the correct answer included in the choices? [Question to be reviewed] {Question} [Choices to be reviewed] {Choices} [Response] Rate the question's correctness on a scale of 1 to 5, with 5 being correct; Only give an overall Rating. For example, Rating: 5

Table 8: The Prompt of LLMs Checking. {Question} and {Choices} represent question and choices instance.

A.9 Difficulty Distribution of MMLU-CF

Figure 9 demonstrates the question difficulty distribution of MMLU-CF dataset at various stages. In step three, we sampled from a normal distribution centered around a difficulty level of 6 and verified

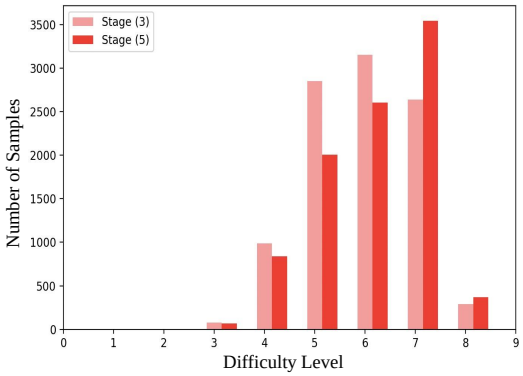


Figure 9: The difficulty level distribution of MMLU-CF after stage (3) and (5).

that the difficulty level remained consistent after Relu 2 of the decontamination-free process. After applying Relu 3 of the decontamination-free process, we observed a change: the proportion of samples with difficulty level 5 significantly decreased, while the number of questions with difficulty level 7 increased. This indicates that the decontamination process introduced more challenging questions into the dataset, which meets the expectation.

A.10 Sampled MCQs across Disciplines

In Table 9, 10, 11, 12 and 13, we present the questions from the validation set across various disciplines. For each subject, we have randomly sampled three questions for demonstration, which offers insights into the diversity and characteristics of the questions used in the validation process. For more questions, we will publicly announce the validation set soon.

Biology

Question 1
Which group of biological molecules is the most diverse in function?
A. Carbohydrates B. Proteins
C. Nucleic acids D. Lipids
Answer: B

Question 2
Which of these structures is the smallest?
A. Hydrogen atom B. None of the other choices
C. Mitochondrion D. Viriod
Answer: A

Question 3
Which of the following controls and regulates life processes?
A. Reproductive and endocrine systems B. Endocrine and digestive systems
C. None of the other choices D. Nervous and endocrine systems
Answer: D

Chemistry

Question 1
What occurs when silver chloride is exposed to sunlight?
A. Silver metal and chlorine gas are formed B. Silver metal and hydrogen gas are formed
C. Only hydrogen gas is formed D. Only silver metal is formed
Answer: A

Question 2
What is the phenomenon called when a beam of light passes through a colloidal solution?
A. Cataphoresis B. Tyndall effect
C. Electrophoresis D. Coagulation
Answer: B

Question 3
Electrolytes play a crucial role in the chemistry of living organisms. What defines an electrolyte?
A. Contains electrodes B. Conducts electricity when melted or put into solution
C. Generates light when electricity is applied D. Contains electrons
Answer: B

Computer Science

Question 1
Which of the following is not a valid floating point literal in Java?
A. 5.0e2 B. 033D
C. 6.8 D. 4.5f
Answer: B

Question 2

```
#include <stdio.h>
int main() {
    int a = -1, b = 4, c = 1, d;
    d = ++a && ++b || ++c;
    printf("%d, %d, %d, %d\n", a, b, c, d);
    return 0;
}
```

A. 0, 5, 2, 1 B. 0, 4, 2, 1
C. None of the other choices D. 1, 4, 1, 1
Answer: B

Question 3
In what aspect did a digital computer not surpass an analog computer?
A. Accuracy B. Reliability
C. Speed D. None of the other choices
Answer: A

Table 9: Three Random Questions from the Biology, Chemistry and Computer Science of the MMLU-CF Validation Set.

Engineering
Question 1 What functions can a diode perform? A. Rectifier B. None of the other choices C. Demodulator D. Modulator **Answer:** C
Question 2 What is a periodic signal? A. May be represented by g(t) = g(t + T0) B. Value may be determined at any point C. Repeats itself at regular intervals D. All of the above **Answer:** D
Question 3 What are the advantages of using electron beam welding? A. Absence of porosity B. Welds are clean C. Distortion less D. All of these **Answer:** B
Math
Question 1 What is the result when $\frac{1}{\sqrt{7}-2}$ is rationalized? A. $(\sqrt{7}-2)/3$ B. $(\sqrt{7}+2)/45$ C. $(\sqrt{7}+2)/5$ D. $(\sqrt{7}+2)/3$ **Answer:** D
Question 2 What is the percentage increase in the area of a rectangle if each side is increased by 20%? A. 46% B. 44% C. 42% D. 40% **Answer:** B
Question 3 What is the radius of a sphere with a surface area of 616 cm²? A. 21 cm B. 7 cm C. 3.5 cm D. 14 cm **Answer:** B
Physics
Question 1 Daylight color film is calibrated for what type of light? A. 3200 K B. 3400 K C. 3000 K D. 5400 K **Answer:** D
Question 2 On a Force versus position (F vs. x) graph, what signifies the work done by the force F? A. The product of the maximum force times the maximum x B. The length of the curve C. The slope of the curve D. The area under the curve **Answer:** D
Question 3 What is the phase difference between the voltage and current in a capacitor in an AC circuit? A. $\pi/3$ B. $\pi/2$ C. π D. 0 **Answer:** B

Table 10: Three Random Questions from the Enigineering, Math and Physics of the MMLU-CF Validation Set.

Business
Question 1 Beth is the project manager for her organization. While her current project has numerous deliverables identified broadly, the specific details of these deliverables remain unclear. Beth is meticulously planning only the activities that are immediately forthcoming in the project. What is this project management planning approach called? A. Rolling wave planning B. Imminent activity management C. None of the other choices D. Predecessor-only diagramming **Answer:** A
Question 2 How do you format Pivot Table report summary data as currency? A. Type in the currency symbol B. Use custom calculation C. Modify the field settings D. None of the above **Answer:** C
Question 3 Which one of these choices is not considered an operating cost? A. Maintenance cost B. Salaries of high officials C. None of the other choices D. Salaries of operating staff **Answer:** B
Economics
Question 1 Which tax proposal did the Finance Minister announce the withdrawal of on 8th March following nationwide protests? A. Tax on High Income Farmers B. Tax proposal on EPF C. Kisan Kalyan Cess D. All of above **Answer:** B
Question 2 In economics, what does the demand for a good indicate regarding the quantity that people: A. None of the other choices B. Need to achieve a minimum standard of living C. Will buy at alternative income levels D. Would like to have if the good were free **Answer:** A
Question 3 What is it called when a firm's supply rises as a result of implementing advanced technology? A. Expansion in supply B. Increase in quantity supplied C. Contraction in supply D. Increase in supply **Answer:** D
Health
Question 1 Thrombocytes are more accurately referred to as _____? A. Megakaryoblasts B. Clotting factors C. Megakaryocytes D. Platelets **Answer:** D
Question 2 Lindsay has been prescribed insulin therapy for which condition? A. None of the other choices B. Diabetes C. Hemophilia D. Spina bifida **Answer:** B
Question 3 Why is it crucial to control and reduce the amount of dust that enters the air? A. Less dust means less cleaning up afterwards B. Dust in the air will affect your vision C. Dust is always in the air and it does not cause harm D. Constantly inhaling dust particles can cause lung problems in the future **Answer:** D

Table 11: Three Random Questions from the Business, Economics, and Health of the MMLU-CF Validation Set.

History
Question 1 The constitutional history of France starts with the French Revolution in what year? A. 1786 B. 1780 C. 1789 D. None of the other choices **Answer:** C
Question 2 Between 1889 and 1916, where was the Second International, which developed under the influence of Socialist Philosophy, organized? A. None of the other choices B. London C. Paris D. Brussels **Answer:** C
Question 3 What was the capital of the Hoysalas? A. Dwarasamudra B. Halebeedu C. Sosevuru D. Belur **Answer:** A
Law
Question 1 How are computer programs legally safeguarded? A. Copy rights. B. Trademarks. C. Industrial design. D. Patents. **Answer:** A
Question 2 What type of justice is represented by the penalty imposed for breaking the law? A. Political justice B. Moral justice C. Legal justice D. Economic justice **Answer:** C
Question 3 What does WIPO stand for? A. World Information and Patents Organisation B. World Intellectual Property Organisation C. World Information Protection Organisation D. None of the other choices **Answer:** B
Philosophy
Question 1 What does it mean when a reprehensible act is referred to by a different term? A. None of the other choices B. advantageous comparison C. euphemistic labeling D. attribution of blame **Answer:** C
Question 2 The assertion, ‘Being non-violent is good’ is a: A. Religious judgement B. None of the other choices C. Factual judgement D. Value judgement **Answer:** D
Question 3 What does the phrase ‘lived alone on the forest tree’ symbolize? A. None of the other choices B. Freedom C. A dull life D. A dependent life **Answer:** B

Table 12: Three Random Questions from the History, Law, and Philosophy of the MMLU-CF Validation Set.

Psychology
Question 1 Which of the following happens first in development? A. Secondary sexual characteristics B. Reproductive maturity C. Gender identity D. Primary sexual characteristics **Answer:** D
Question 2 How can a teacher be successful? A. imparts subject knowledge to students B. presents the subject matter in a well organized manner C. prepares students to pass the examination D. None of the other choices **Answer:** B
Question 3 What is meant by Ex Post Facto research? A. The research is carried out prior to the incident B. None of the other choices C. The research is carried out along with the happening of an incident D. The research is carried out after the incident **Answer:** D
Other
Question 1 To achieve a quick promotion, he came up with a plan to appease the manager. A. Conciliate B. Evict C. Incite D. Praise **Answer:** A
Question 2 Which company initiated the secret Zuma Mission for the United States government? A. SpaceX B. None of the other choices C. XCOR Aerospace D. Boeing **Answer:** A
Question 3 In The Calling of Saint Matthew, Caravaggio depicted his subjects wearing the clothing of his own era, rather than that of Jesus's time. A. to portray the painting's patrons realistically. B. to conform with other paintings in the series. C. to enable the audience to identify with them. D. so that he could use richer colors and brushstrokes. **Answer:** C

Table 13: Three Random Questions from the Psychology, Other of the MMLU-CF Validation Set.