Are We Done with MMLU?

Anonymous ACL submission

Abstract

Maybe not. We identify and analyse errors in the popular Massive Multitask Language Understanding (MMLU) benchmark. Even though MMLU is widely adopted, our analysis demonstrates numerous ground truth errors that obscure the true capabilities of LLMs. For example, we find that 57% of the analysed questions in the *Virology* subset contain errors. To address this issue, we introduce a comprehensive framework for identifying dataset errors using a novel error annotation protocol. Then, we create MMLU-Redux, which is a subset of 5,700 manually re-annotated questions across all 57 MMLU subjects. Using MMLU-Redux, we demonstrate significant discrepancies with the model performance metrics that were originally reported. Our results strongly advocate for revising MMLU's error-ridden questions to enhance its future utility and reliability as a benchmark.

1 Introduction

The advent of transformer-based Large Language Models (LLMs) (OpenAI, 2023; Anil et al., 2023a; Anthropic, 2023; Anil et al., 2023b; Touvron et al., 2023; Anthropic, 2024; Kaddour et al., 2023; AI@Meta, 2024) marked a significant advancement in generative models, enabling interaction with computing devices through natural language. This advancement rendered many earlier benchmarks and leaderboards obsolete (Laskar et al., 2023; Shen et al., 2023), leading to the compilation of more challenging and comprehensive tests. Among these benchmarks, Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) has gained significant popularity: It assesses both the breadth and depth of language understanding capabilities of current LLMs across a diverse range of subjects, including mathematics, history, computer science, logic, law, *etc.*

However, the reliability of benchmarking results is only as robust as the quality of the dataset used.

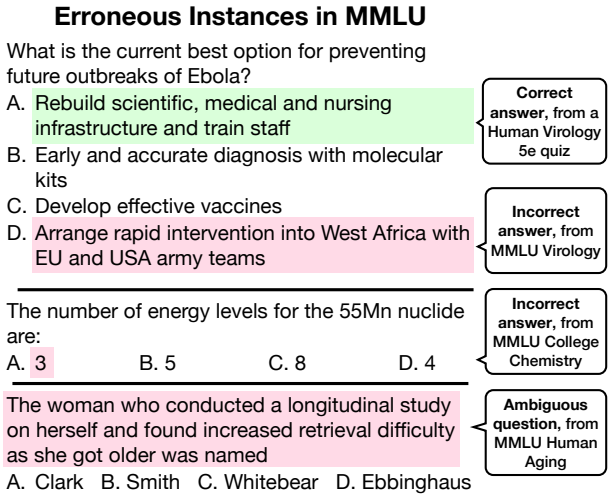

Figure 1: Examples of erroneous instances from MMLU Virology, College Chemistry, and Human Aging.

We find that, despite its popularity, MMLU suffers from numerous errors that can mislead evaluation and model comparison (L and Stapleton, 2023; Erenrich, 2023). These errors, which range from simple parsing and scraping mistakes to more complex issues related to context, interpretation, and dataset quality, compromise the reliability of MMLU as a benchmark. For example, we find that 57% of the analysed instances in the *Virology* subset contain errors, including the suggestion to send the American army to West Africa to prevent outbreaks of Ebola (see Fig. 1).

Therefore, in this study, we manually analyse the MMLU dataset using a novel error annotation protocol to construct **MMLU-Redux**: 14 human experts manually assessed and re-annotated 5,700 questions across all subsets of MMLU. After our manual re-annotation effort, we study how the errors in MMLU impact LLM evaluation. First, we re-evaluated leading LLMs on MMLU-Redux, and found the performance metrics notably altered, changing their ranking. Furthermore, we both quantitatively and qualitatively analysed the errors

to help understand how these errors impact LLM evaluation. MMLU-Redux can also be used as a strong benchmark for automatic error detection in NLP datasets, which would help scale up the review of benchmark datasets. Therefore, we also study whether LLMs can help with error detection, using prompting techniques (*i.e.,* In-Context Learning (Brown et al., 2020), Chain of Thoughts (CoT) (Wei et al., 2022)), Retrieval Augmented Generation (RAG) (Lewis et al., 2020), and fine-tuning. We believe MMLU-Redux underscores the need for closely studying and reassessing the benchmarks used for evaluating NLP models.

2 What Is Wrong with MMLU?

The MMLU dataset has become a popular choice for evaluating the performance of NLP systems, owing to its extensive coverage of subjects collected from freely available online sources with the help of graduate and undergraduate students (Hendrycks et al., 2021). Despite the manual effort, MMLU still contains errors which are difficult to trace due to its under-documented annotation procedure. These errors have not yet been systematically studied, even though they were recently highlighted (L and Stapleton, 2023; Erenrich, 2023).

We identify various errors in MMLU questions, from simple parsing mistakes (*e.g.,* the source answer is B, but MMLU labels it as C) to more complex issues such as missing context. These errors occur randomly, and without detailed documentation, tracing their root causes is challenging.

This randomness and lack of traceability highlight the need for a standardised error categorisation to improve the reliability and accuracy of the MMLU dataset. Our approach involved developing a hierarchical annotation protocol of errors, which we used to develop MMLU-Redux: A manual annotation of all 57 subsets of MMLU, each containing 100 randomly selected samples (Section 3.1).

2.1 Error Detection Annotation Protocol

We develop a hierarchical annotation protocol to classify the various errors identified in MMLU into specific error types. Figure 2 illustrates our annotation protocol for categorising MMLU errors, while Figure 4 provides examples of each error category. We categorise errors into two primary groups: samples with errors in the clarity of the questions (Type 1, Question Assessment) and samples with errors in the ground truth answer (Type 2, Ground Truth Verification).

(1a) Bad Question Clarity: The question is poorly presented in terms of various aspects, such as clarity, grammar, and sufficiency of information. For instance, referring to a previous question.

(1b) Bad Options Clarity: The options are unclear, similar, or irrelevant to the question. Most errors in this category stem from incorrect parsing of the options from the original source. For example, a single option might be incorrectly split into two separate options.

(2a) No Correct Answer: None of the options correctly answer the question. This error may arise when the ground-truth options are omitted to reduce the number of options from five to four.

(2b) Multiple Correct Answers: More than one option can be selected as the answer to the question. For example, the options contain a synonym of the ground truth label.

(2c) Wrong Ground Truth: The correct answer differs from the ground truth provided in MMLU. This type of error occurs when the annotated label differs from the correct label, which may be caused by a mistake during manual annotation.

To create MMLU-Redux, we followed the proposed annotation protocol. We aim to ensure comprehensive coverage of the different error types for further experiments.

2.2 Heterogeneity of Errors in MMLU

Our annotation process reveals substantial variation in error types across subsets. Some subsets are affected by ambiguous questions, while others by incorrect ground truth labels. These differences may impact how MMLU results are interpreted and addressed. Notable irregularities in certain subsets include (a full list of subsets and their corresponding errors can be found in Appendix I):

Virology – Incorrect ground truth labels are prevalent within the Virology subset. Many of the incorrect labels are for relatively simple questions, such as identifying the description of a pandemic, this suggests errors may stem from problems parsing the original datasets (in most cases the *Human Virology* textbook).

College Chemistry – The questions were sourced from textbooks and standardised college-level exams. We identified erroneous questions resulting from simple parsing errors and unknown causes. For example, questions spanning multiple lines in the original source were often parsed incorrectly,

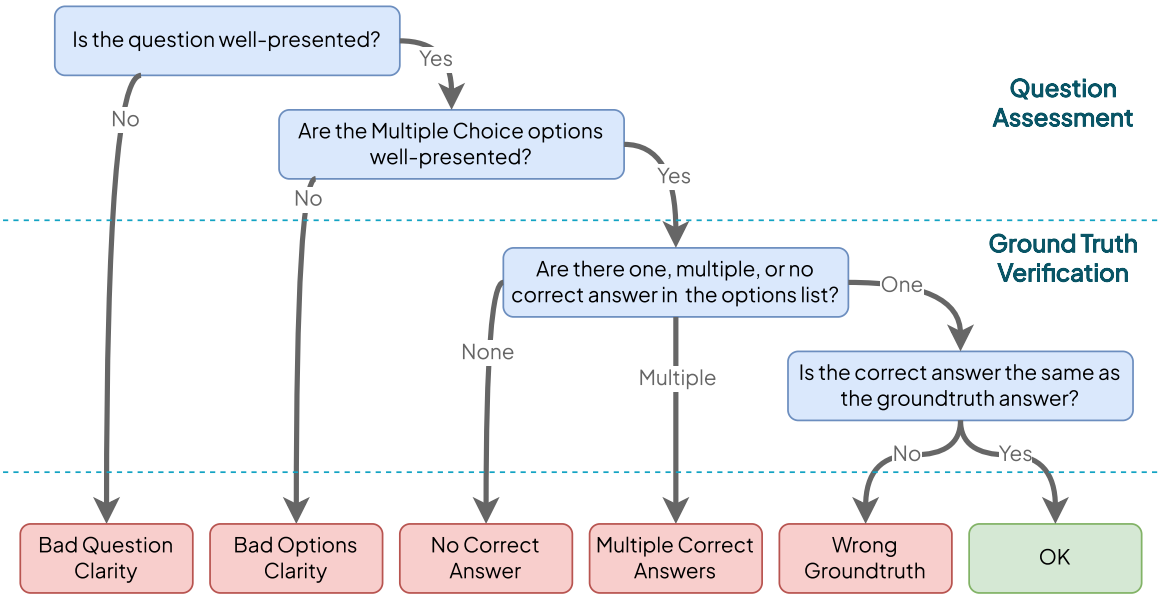

Figure 2: The error annotation protocol used for annotating instances in MMLU — annotators first assess whether the instance is unambiguous (*Question Assessment*, then whether there is a single valid answer, and then whether it matches with the ground truth answer from the dataset.

leading to a part of the question being presented as the first option (Option A) and the exclusion of Option D. Furthermore, there were questions with ground truth labels that did not match the answers provided in the source without known cause.

Professional Law – A major issue is the lack of specificity regarding jurisdictions. The benchmark does not clearly distinguish between different jurisdictions despite focusing on U.S. law.

Formal Logic – The dataset contains many questions with incorrect answers. These are mostly sourced from the 'Oxford Learning Link' website. Inaccuracies are caused by unknown reason. For example, one question states that $(F \wedge L) \wedge \neg C$ is correct, but $F \wedge L \wedge \neg C$ is not, even though these two formulas are clearly equivalent.

Global Facts – Almost all questions needed consulting external sources to be validated, where a large portion of them are reports from `ourworldindata.org` (18 cases) and `pewresearch.org` (15 cases); for several questions, multiple sources were providing conflicting answers — for example, on the perceived corruption of political parties in 2013 Ethiopia, `ourworldindata.org` seems to confirm the answer in MMLU, while the Global Corruption Barometer from Transparency International was providing conflicting evidence. [1]

Machine Learning – Most questions were sourced from exam papers, assignments or online quizzes. About 30 of the questions require expert knowledge and reasoning. The main issue of this subset is the clarity of the questions and options. *e.g.,* some quiz questions are based on past knowledge, and the descriptions in the questions may be vague or inapplicable today.

Econometrics – The majority of the questions are correct, but some questions contain unverifiable references. *e.g.,* 'Consider again the VAR model of equation 16,' but equation 16 cannot be found within the question.

These irregularities highlight error patterns in MMLU. We want to highlight one type of error that is especially challenging to catch, namely unspecified context that is needed to properly answer the question. For example, the Professional Law and Accounting datasets assume US jurisdiction and practices which are not specified in the question and may become outdated if standards change. Many subjects also display a US- and Western-centric bias: the Virology dataset refers to "the Latino population" in a US context, and the Human Aging dataset mentions an unspecified survey of older adults.

[1] See data at Our World in Data and Global Corruption Barometer.

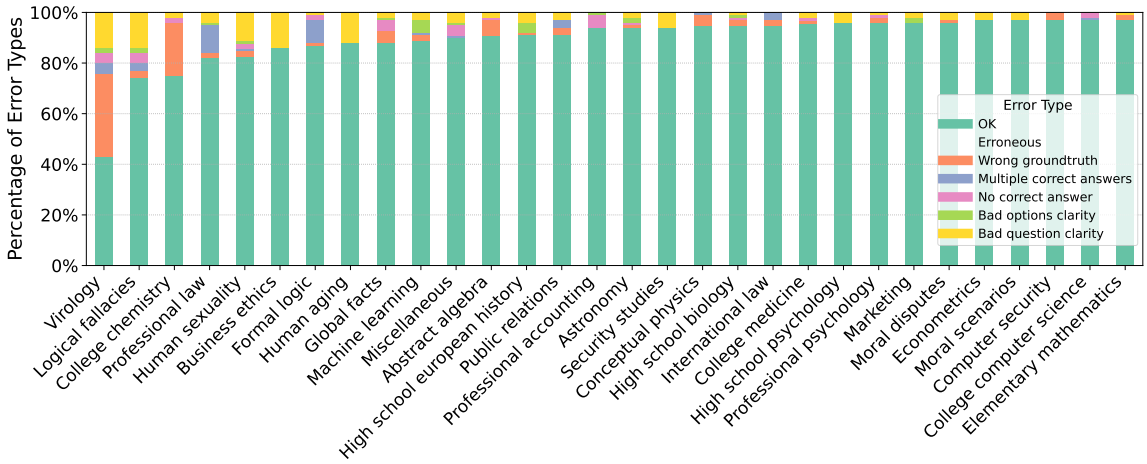

Figure 3: Statistics of the error types detected in 100 randomly sampled and manually annotated data from the 30 most erroneous MMLU subjects. In the *Virology* subject, we found that 57% of the analysed instances contain some forms of errors, such as wrong ground-truth (33%), multiple correct answers (4%), or unclear questions (14%). For detailed numbers of all subsets, see Appendix F.

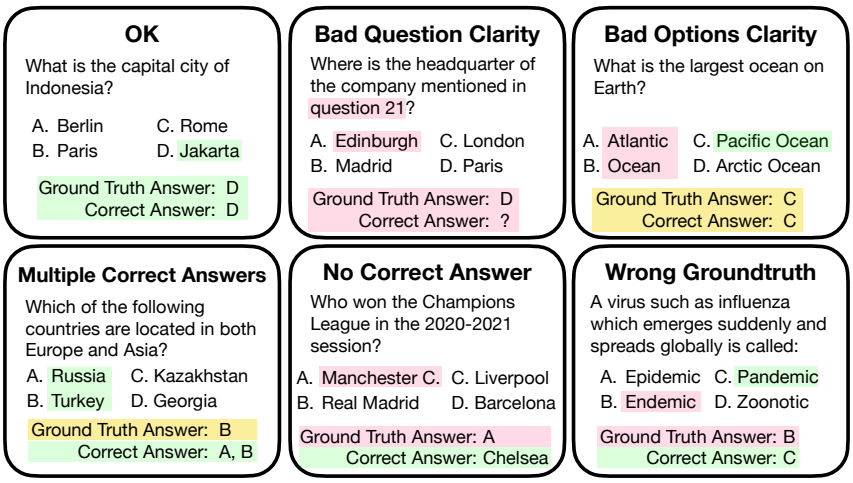

Figure 4: An example for every leaf node in the error annotation protocol shown in Fig. 2 from the MMLU dataset.

3 MMLU-Redux: A Correct MMLU Subset

In this section, we propose MMLU-Redux, a manually annotated subset of MMLU, to quantify the errors present in the original dataset. MMLU-Redux serves two main purposes: *1)* to measure the prevalence and types of errors in MMLU; and *2)* to explore the feasibility of automatically fixing MMLU by leveraging the annotated error types. We find that *1)* the proportion of errors in MMLU is non-negligible, highlighting the need for a correct subset; and *2)* fixing MMLU automatically proves to be a challenging task, despite the availability of annotated error types.

We create MMLU-Redux by manually labelling a subset of MMLU questions with their corresponding error types. To this end, we follow the annotation protocol introduced in Section 2.1. For more accurate annotations, we confirmed the error detection by finding the samples' original source wherever it was available. However, at present, the correct answers suggested by the annotators are not used to replace the existing MMLU labels.

In the following, we analyse the error statistics of MMLU-Redux and use MMLU-Redux to re-evaluate the performance of LLMs. In Section 4, we explore the possibility of using MMLU-Redux to improve the overall quality of MMLU by automatically fixing the identified errors.

3.1 Analysis of MMLU-Redux

We present the percentage of error types in Fig. 3, with detailed numbers available in Appendix F. In

Table 1: Inter-annotator agreement (Cohen's Kappa) analyses for the five erroneous subjects of MMLU-Redux. "All Error Type" is calculated based on whether the specific error types match, while "Binary Error Type" is calculated based on whether an error is present or not.

Subject	Cohen's Kappa (All Error Type)	Cohen's Kappa (Binary Error Type)
virology	0.67	0.67
logical fallacies	0.73	0.71
college chemistry	0.92	0.95
formal logic	0.96	1.0
human sexuality	0.64	0.64

our analysis, we find that more than 9% of the examples are incorrect, suggesting a substantial presence of errors in the MMLU. Especially, we find that more than 57% examples in Virology contain errors, where 30% examples have a wrong ground truth, and 15% are unclear questions. Moreover, we also observe significant error percentages in other disciplines: more than 20% examples in Logical Fallacies and College Chemistry are wrong and more than 10% examples in Professional Law, Business Ethics, Formal Logic, Human Aging, Global Facts, Machine Learning, Miscellaneous and Public Relations are wrong. Such error proportions could lead to inaccurate comparisons and invalid rankings of LLM models.

To improve our understanding of how these errors impact the performance of models on the MMLU, we compare the performance between erroneous instances and correct instances across the seven subjects identified as having the most errors (Virology, Logical Fallacies, College Chemistry, Professional Law, Business Ethics, Formal Logic, and Human Aging) in Fig. 5.

Despite the general trend indicating a performance decline among erroneous instances, we also observed cases where performance was similar or even higher in erroneous instances (Professional Law and Formal Logic). Given that LLMs should be unable to yield correct answers on these erroneous instances, this may be evidence of memorisation, suggesting that these MMLU instances were learned during the pretraining.

We also assessed the reliability of the manual annotations in MMLU-Redux by calculating Cohen's Kappa (Cohen, 1960) among three annotators for the five erroneous subjects (virology, logical fallacies, college chemistry, formal logic, and human sexuality). The results shown in Table 1 demonstrate a generally high level of agreement across most subjects, with Cohen's Kappa values exceeding 0.6 for both "All Error Type" (whether the specific error types match) and "Binary Error Type" (whether an error is present or not). Additionally, the minimal difference in scores between "All Error Type" and "Binary Error Type" indicates that our error annotation protocol used for annotating (Section 2.1) is highly effective.

3.2 Re-Evaluating the State-of-the-Art LLMs

To assess how the corrected dataset impacts the performance of existing state-of-the-art LLMs, we re-evaluate them on the five subjects with the highest number of errors (Business Ethics, Virology, Professional Law, College Chemistry, and Logical Fallacies) in MMLU-Redux.

In Table 2, we compared the performance of models when using all instances of MMLU-Redux to the performance when using only correct instances without errors to see if there are any changes in the rankings due to this. The results clearly demonstrate that, at least for subjects with a high number of detected errors, these issues are significant enough to affect the results, leading to changes in model rankings. For example, in the Virology subset, Palmyra X v3 ranked 4th in overall performance when considering all instances and ranked 1st when only correct instances were used. This indicates that errors in MMLU are a critical issue, considering that MMLU is used as an important benchmark for evaluating model performance.

3.3 Correlation between Exact Match Score and Data Quality

Figure 6 shows the correlation between the performance of the ten best-performing LLMs on MMLU (as ranked by HELM (Liang et al., 2023)) and the ratio of non-erroneous instances on MMLU-Redux. Each point in the scatter plot represents a subject in MMLU. As expected, LLMs perform better in subjects with fewer errors, supporting the link between data reliability and model performance. Ideally, LLMs excel with non-erroneous instances and struggle with erroneous ones.

Outliers, however, reveal deviations from this trend, where models perform better or worse than expected for certain subjects. These deviations suggest other factors influence performance, requiring further investigation. For example, some LLMs perform well despite high error rates (in the Logical Fallacies subject), hinting at potential subject-specific advantages. We further explore

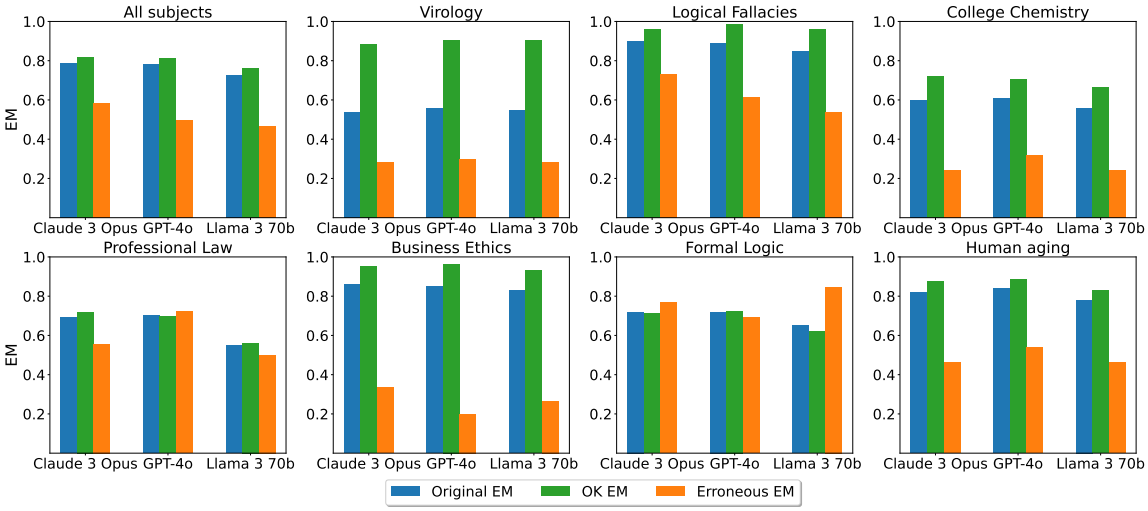

Figure 5: Exact Match (EM) for each model on all subjects combined (top left), and the seven subjects with the most errors in our MMLU-Redux data. Original EM represents performance measured across all instances regardless of error presence, while OK and Erroneous EM reflect results on correct instances and error instances, respectively.

Table 2: Comparison of model performance and ranking changes when using overall instances versus correct instances of MMLU-Redux on five most erroneous subjects. Numbers in parentheses denote ranks, teal an improvement in rank, red a deterioration in rank, and black an unchanged rank.

Model	Business Ethics	Virology	Professional Law	College Chemistry	Logical Fallacies
Claude 3 Opus	0.86 (1) → 0.95 (2)	0.54 (9) → 0.88 (6)	0.69 (4) → 0.72 (3)	**0.60 (2) → 0.72 (1)**	0.90 (1) → 0.96 (3)
GPT-4o	**0.85 (2) → 0.96 (1)**	0.56 (1) → 0.91 (3)	0.70 (3) → 0.70 (4)	0.61 (1) → 0.71 (2)	**0.89 (5) → 0.99 (1)**
Gemini 1.5 Pro	0.80 (7) → 0.88 (9)	0.55 (7) → 0.84 (10)	0.67 (7) → 0.70 (5)	0.58 (6) → 0.69 (6)	0.88 (7) → 0.96 (5)
GPT-4 (0613)	0.79 (8) → 0.93 (3)	0.56 (2) → 0.88 (7)	**0.71 (1) → 0.74 (1)**	0.55 (9) → 0.68 (7)	0.89 (6) → 0.96 (4)
Llama 3 70b	0.83 (3) → 0.93 (4)	0.55 (8) → 0.91 (4)	0.55 (10) → 0.56 (10)	0.56 (8) → 0.67 (8)	0.85 (10) → 0.96 (6)
Gemini 1.5 Flash	0.82 (6) → 0.91 (8)	0.53 (10) → 0.88 (9)	0.58 (9) → 0.57 (9)	0.60 (3) → 0.71 (5)	0.86 (9) → 0.92 (10)
Palmyra X v3	0.83 (4) → 0.91 (6)	**0.56 (4) → 0.93 (1)**	0.68 (5) → 0.65 (7)	0.59 (4) → 0.71 (3)	0.90 (2) → 0.95 (7)
PaLM 2	0.83 (5) → 0.91 (7)	0.56 (5) → 0.93 (2)	0.68 (6) → 0.65 (8)	0.59 (5) → 0.71 (4)	0.90 (3) → 0.95 (8)
GPT-4 Turbo (1106)	0.78 (9) → 0.91 (5)	0.56 (3) → 0.88 (8)	0.71 (2) → 0.73 (2)	0.47 (10) → 0.60 (10)	0.86 (8) → 0.93 (9)
Mixtral 8x22b	0.74 (10) → 0.84 (10)	0.56 (6) → 0.91 (5)	0.59 (8) → 0.66 (6)	0.57 (7) → 0.64 (9)	0.90 (4) → 0.99 (2)

these anomalies and their implications for model robustness in Section 3.1.

4 Can We Fix the MMLU Dataset Automatically?

After presenting evidence of the numerous errors in the MMLU dataset, we explore the following approaches to detect these errors automatically:

Zero-Shot prompting We provide the model with a simple instruction to classify questions into "ok" or "not ok" without introducing any demonstrations. The prompt can be found in Appendix G.

Few-shot prompting We provide the model with two examples for each error type to guide its classification decisions.

Chain of Thought (CoT) prompting (Wei et al., 2022) We encourage the model to generate reasoning steps before producing the final answer, in both zero-shot and few-shot settings. The prompt format can be found in Appendix G.

Retrieval-augmented prompting (RAG) We retrieve 5 most relevant paragraphs from Wikipedia and MS-MARCO and append them as context for zero-shot and CoT prompting.

Instruction fine-tuning We fine-tune Llama3 8B-Instruct model (AI@Meta, 2024) using curated data and evaluate its performance on MMLU-Redux. Refer to Appendix J for more details.

In the following, we introduce these error detection strategies and their evaluation results in detail.

4.1 Error Detection Experiments

To evaluate the performance of LLMs in detecting errors in the MMLU-Redux dataset, we conduct experiments with 4 state-of-the-art models: OpenAI's GPT-4 Turbo, GPT-4o, Anthropic's Claude-3 Opus, and Meta's LlamA-3-70B (AI@Meta, 2024). For

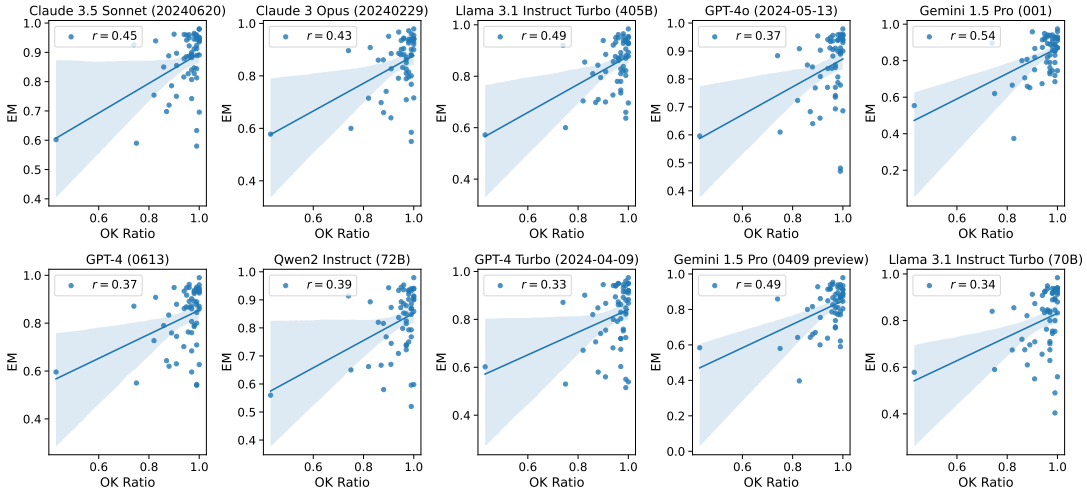

Figure 6: Correlation between the performance of 10 best-performing LLMs on MMLU (as per HELM (Liang et al., 2023)) and the ratio of non-erroneous instances on MMLU-Redux. Each dot represents one MMLU subject.

Table 3: Comparison of different methods and models for error detection in MMLU-Redux scores.

Model	Method	Recall	F1	F2
GPT-4 Turbo	Zero-Shot	22.81	22.70	23.90
	Zero-Shot CoT	27.97	22.97	23.47
	Few-shot	26.55	27.93	27.97
	Few-shot CoT	46.68	31.68	36.58
GPT-4o	Zero-Shot	28.92	15.99	28.05
	Zero-Shot CoT	19.85	21.41	19.98
	Few-shot	37.40	24.93	31.06
	Few-shot CoT	38.47	29.26	31.36
Claude 3 Opus	Zero-Shot	27.11	27.00	27.65
	Zero-Shot CoT	44.87	**34.68**	38.19
	Few-shot	38.63	29.45	34.89
	Few-shot CoT	**48.85**	24.03	**40.29**
Llama3-70B	Zero-Shot	10.06	8.15	9.46
	Zero-Shot CoT	10.74	8.15	10.17
	Few-shot	17.82	18.58	17.60
	Few-shot CoT	24.87	23.16	23.10

each model, we test both standard prompting and Chain of Thought (CoT) prompting methods (Wei et al., 2022). Details about the prompts are provided in Appendix G.

We consider "not ok" as the positive class, and "ok" as the negative class for the calculation of Recall, F1, and F2 scores. Based on Table 3, the Few-shot CoT setting outperforms other settings across all models, suggesting that providing a small set of labelled examples with step-by-step reasoning instructions can improve error detection performance. However, even the best-performing model, Claude-3-opus, only achieves an F2 Score of 40.29, highlighting the difficulty of this task.

Table 4: Comparison of different methods and models for error detection using RAG.

Model	Index	Method	Recall	F1	F2
GPT-4 Turbo	Wikipedia	Zero-Shot	57.00	27.47	36.87
	Wikipedia	Zero-Shot CoT	46.63	**31.52**	36.13
	MS MARCO	Zero-Shot	57.11	26.34	35.02
	MS MARCO	Zero-Shot CoT	37.20	26.07	29.64
GPT-4o	Wikipedia	Zero Shot	35.01	27.25	29.79
	Wikipedia	Zero-Shot CoT	33.67	26.90	28.75
	MS MARCO	Zero-Shot	29.97	28.07	28.31
	MS MARCO	Zero-Shot CoT	28.65	24.14	25.30
Llama3-70B	Wikipedia	Zero Shot	22.78	20.41	21.00
	Wikipedia	Zero-Shot CoT	10.12	12.61	10.90
	MS MARCO	Zero-Shot	28.45	22.39	24.67
	MS MARCO	Zero-Shot CoT	10.18	14.00	11.40
Claude 3 Opus	Wikipedia	Zero-Shot	82.61	28.72	**41.92**
	MS MARCO	Zero-Shot	**83.91**	28.09	41.27

We then use RAG prompting to investigate the impact of external knowledge on error detection. We use BM25 to retrieve five most relevant paragraphs from *enwiki-paragraphs* and *msmarco-v1-passage* corpus provided by Pyserini (Lin et al., 2021a) using the question as the query. We then use these paragraphs as additional context in the prompt to classify the question and answer choices.

Based on Table 4, the Claude 3 Opus model with the zero-shot method on the MS MARCO index achieves the highest Recall of 83.91. Claude 3 Opus achieves an F2 Score of 41.92 with the zero-shot method on the Wikipedia index. The GPT-4 models perform relatively worse than Claude, with the GPT-4o model showing lower scores than GPT-4 Turbo. Comparing the retrieval indexes, Wikipedia generally outperforms MS MARCO for the GPT-4o model, while the results are mixed for the GPT-4 Turbo and Llama3-70B models. The

RAG approach outperforms the few-shot CoT setting mentioned in the previous analysis, indicating that incorporating retrieved information can enhance error detection performance.

Based on the results presented in Table 3 and 4, we can conclude that automatic error detection in the MMLU dataset remains a challenging task, despite the availability of annotated error types. While Claude 3 Opus demonstrates the highest Recall, F1, and F2 Scores compared to other models, Claude 3 Opus shows the highest performance with RAG, indicating its potential for identifying errors more effectively. However, the best-performing model and method combination using RAG still achieves relatively low scores, suggesting the overall reliability of the models in detecting errors across the diverse range of subjects in MMLU is still limited. Detailed performance across all subjects can be found in Appendix H.

5 Related Work

Benchmark Issues While benchmarks often enable methodological progress, they can also be counterproductive when labelling mistakes and annotation artefacts exist. For example, Beyer et al. (2020) show how annotation artefacts within the popular ImageNet benchmark (Russakovsky et al., 2015) led to likely overstated performance gains that did not necessarily transfer to other datasets and tasks. In NLP, similar issues have been found in summarisation (Tejaswin et al., 2021; Bhandari et al., 2020) and natural language inference (Gururangan et al., 2018; Poliak et al., 2018; Stacey et al., 2020; Wu et al., 2022) benchmarks.

Benchmark issues can arise from biases in the framing of the task (Schwartz et al., 2017); noisy annotations (Chen et al., 2016); web-crawls (Raffel et al., 2023; Lee et al., 2021; Kaddour, 2023); automated labelling processes such as crowdsourcing (Yuen et al., 2011) — where annotators may be optimising their own incentives; human errors (Peterson et al., 2019) — *e.g.,* lack of expertise on a given topic; or programmatic weak supervision (Zhang et al., 2022; Goswami et al., 2021; Kaddour et al., 2023), and personal (Geva et al., 2019) or group-level (Liu et al., 2022) annotator biases.

MMLU Issues and MMLU-Pro. The broad adoption of the MMLU benchmark for LLM evaluations (Anil et al., 2023a; OpenAI, 2023; Touvron et al., 2023; Anthropic, 2024; AI@Meta, 2024) means identifying issues or improvements is crucial for ensuring its continued applicability. Recent studies have identified issues with labelling errors and ambiguous questions in similar benchmarks, such as MedQA (Saab et al., 2024). Concurrent work developing the MMLU-Pro (Wang et al., 2024) benchmark also identifies a number of issues within a filtered and augmented subset of the original MMLU dataset and re-annotates parts of the dataset for inclusion in the new MMLU-Pro evaluation. However, despite these efforts, errors from the original MMLU persist in the extended MMLU-Pro benchmark, highlighting the importance of our work in systematically identifying and addressing dataset errors. Additionally, there is currently limited literature categorising and quantifying these issues across the original MMLU dataset to help inform our understanding of previous results. The AI Explained YouTube Channel recently highlighted several erroneous instances in MMLU across several subjects while evaluating their SmartGPT framework in a recent popular video (L and Stapleton, 2023). However, our study aims to provide a more systematic assessment of MMLU.

6 Conclusion

We analyse the MMLU benchmark, driven by the necessity for rigorous evaluation of its reliability. Our analysis using a hierarchical annotation protocol reveals that a significant portion of MMLU instances contain inaccuracies that could lead to misleading evaluation results. For example, 57% of the instances in the *Virology* and 26% in the *Logical fallacies* subsets were found to be inaccurate. To this end, we introduce MMLU-Redux, a thoroughly reviewed subset of the MMLU dataset (Hendrycks et al., 2021) comprising 5,700 questions spanning all 57 MMLU subjects. The re-evaluation of LLMs using MMLU-Redux shows a significant variation in performance metrics and shifts in model rankings for several subsets, emphasising the impact that dataset quality can have on the evaluation of LLMs. Furthermore, we analyse whether it is possible to identify errors automatically; although Claude 3 Opus seems to produce the most accurate results on this task (41.9% F2 score when using retrieval-augmented generation), it is still insufficient to produce a high-quality dataset.

Limitations

Although our analysis uses a significant subset of 5,700 questions, the accuracy of this analysis, both

quantitatively and qualitatively, can be further improved with additional annotation of the remaining 8,342 questions in MMLU. Therefore, we open up MMLU-Redux for additional annotation of both additional MMLU subjects and add to the already reviewed subsets. However, we acknowledge that the annotation protocol we introduced to classify errors might still be prone to the personal biases of the annotators.

References

AI@Meta. 2024. Llama 3 model card.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Jos van der Velde, Steffen Vogler, and Carole-Jean Wu. 2024. Croissant: A metadata format for ml-ready datasets. DEEM '24, page 1–6, New York, NY, USA. Association for Computing Machinery.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023a. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023b. Palm 2 technical report. *CoRR*, abs/2305.10403.

Anthropic. 2023. Anthropic. model card and evaluations for claude models.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with imagenet? *CoRR*, abs/2006.07159.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *EMNLP (1)*, pages 9347–9359. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *ACL (1)*. The Association for Computer Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Daniel Erenrich. 2023. Errors in the MMLU: The Deep Learning Benchmark is Wrong Surprisingly Often. URL: https://derenrich.medium.com/errors-in-the-mmlu-the-deep-learning-benchmark-is-wrong

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li,

Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *EMNLP/IJCNLP (1)*, pages 1161–1166. Association for Computational Linguistics.

Mononito Goswami, Benedikt Boecking, and Artur Dubrawski. 2021. Weak supervision for affordable modeling of electrocardiogram data. In *AMIA*. AMIA.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*, pages 107–112. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Jean Kaddour. 2023. The minipile challenge for data-efficient language models. *CoRR*, abs/2304.08442.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *CoRR*, abs/2307.10169.

Philip L and Joshua Stapleton. 2023. AI Explained YouTube Channel - SmartGPT: Major Benchmark Broken - 89.0% on MMLU + Exam's Many Errors. URL: https://youtu.be/hVade_8H8mE?t=829.

Md. Tahmid Rahman Laskar, M. Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *ACL (Findings)*, pages 431–469. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021b. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Haochen Liu, Joseph Thekinen, Sinem Mollaoglu, Da Tang, Ji Yang, Youlong Cheng, Hui Liu, and Jiliang Tang. 2022. Toward annotator group bias in crowdsourcing. In *ACL (1)*, pages 1797–1806. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *ICCV*, pages 9616–9625. IEEE.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018.

Hypothesis only baselines in natural language inference. In *SEM@NAACL-HLT, pages 180–191. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-Baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Benjamin Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Andrew Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale R. Webster, Joelle K. Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *CoRR*, abs/2404.18416.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *CoNLL*, pages 15–25. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *CoRR*, abs/2305.13091.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training. In *EMNLP (1)*, pages 8281–8291. Association for Computational Linguistics.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3436–3449. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *ACL (1)*, pages 2660–2676. Association for Computational Linguistics.

Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. A survey of crowdsourcing systems. In *SocialCom/PASSAT*, pages 766–773. IEEE Computer Society.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022. A survey on programmatic weak supervision. *CoRR*, abs/2202.05433.

A Responsible Research Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We claim to develop a dataset for studying the errors present in MMLU, which we present.
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [No] See section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 3
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We provide details for all proposed methods. For prompting strategies, we show prompts in Appendix I For fine-tuning, we provide details about the training dataset in Appendix J.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Rerunning experiments multiple times to obtain error bars would have exceeded our funding capabilities.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provide details about our resources in Appendix J

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Our work is based on the MMLU dataset, which we cited in Section 1.
   (b) Did you mention the license of the assets? [Yes] The license of the dataset is available on the dataset URL (CC-BY 4.0).
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] MMLU-Redux is provided as a supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Our dataset is based on MMLU, which has an MIT license. The annotation work for MMLU-Redux is done by the authors of this paper, who all gave consent to use the data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The MMLU data does not use personal data as it is based on publicly available test questions, and so neither does MMLU-Redux.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

B Data Collection and Organisation

The MMLU-Redux dataset will be made available at the Hugging Face Datasets hub with a CC-BY-4.0 licence and include a link to a public datasheet. The data used to create MMLU-Redux was obtained from `cais/mmlu` [2], which is also utilised in the 'lm-eval-harness' framework (Gao et al., 2023). To ensure uniformity of our results, the language model (LM) predictions used in our performance analyses were obtained from the Holistic Evaluation of Language Models (HELM) leaderboard v1.3.0, released on May 15th, 2024.

We randomly subsampled 100 questions per MMLU subject to be presented to the annotators. The annotators are instructed to follow the introduced taxonomy by first assessing the question presentation, and then by verifying the ground truth MMLU label. The annotators were encouraged to perform an exact match search using a search engine to find occurrences of the question and multiple-choice options from credible sources. If the annotators found an exact match of the question-options pair, the annotators were asked to evaluate the answer provided by the source. Regardless of whether a label was found in the source, and whether the MMLU label is the same or not, the annotators were asked to decide whether they would follow the label using their expertise. In the cases where an exact match was not found, the annotators were asked to search for supporting evidence from trusted sources, such as government websites, textbooks, and/or other reputable organisations (*e.g.,* World Health Organisation (WHO)). In cases where the annotators are still unsure, they were asked to annotate the question with "Expert", denoting that the question requires more expertise. This annotated subset of MMLU is denoted as MMLU-Redux.

MMLU-Redux comprises subsampled test splits of the aforementioned thirty MMLU subsets (denoted as "`config`" in HF vocabulary).

Each data point in MMLU-Redux contains seven columns:

- **question** (string): The original MMLU question.
- **choices** (list of four strings): The original list of four choices associated with the question from the MMLU dataset.
- **answer** (integer): The MMLU ground truth label in the form of an array index between 0 and 3.
- **error_type** (string): The annotated error_type. The values can be one of the six error types proposed in the taxonomy ("ok", "bad_question_clarity", "bad_options_clarity", "no_correct_answer", "multiple_correct_answers", "wrong_groundtruth") and "expert".
- **source** (string): The potential source of the question.
- **correct_answer** (string): In the case of "no_correct_answer" and "wrong_groundtruth", the annotators can suggest the alternative correct answer.
- **potential_reason** (string): A free text column for the annotators to note what they believe to have caused the error.

The question, choices, and answer columns are taken from the MMLU Hugging Face dataset (`cais/mmlu`).

To ensure reproducibility and facilitate further research, we will make MMLU-Redux publicly available on the Hugging Face (HF) platform in the Croissant format (Akhtar et al., 2024).This format provides a standardised and easily accessible structure for the dataset, allowing researchers and practitioners to readily utilise it for their own investigations and analyses. The code to generate the data and analyses will be made available in the camera ready.

[2] https://huggingface.co/datasets/cais/mmlu

C Usage Guidelines

To reproduce our results or perform analyses similar to those presented in this study, the user may download the data and utilise all the columns. MMLU-Redux contains both correct and erroneous instances, so the user should look at the value in column "error_type" to filter samples based on the specific error type.

In those cases where the error is "no_correct_answer", "multiple_correct_answers" or "wrong_groundtruth", the users may examine the suggested answer reported in the "correct_answer" column. The user should consider that the questions and the options reported are the same as those in the MMLU dataset, and they have not been modified even when affected by bad clarity.

D Maintenance Plan

D.1 General Maintenance

The datasets will be updated as needed to maintain accuracy, with announcements made for each update. These updates will be posted on the dataset's Hugging Face page. Older versions will be preserved and documented using the git commit hash of the dataset repository.

D.2 Contributing Guidelines

We welcome contributions from the research community to enhance and expand the MMLU-Redux dataset. Primarily, we foresee two types of contribution:

Additional data annotation Currently, MMLU-Redux covers 5,700 MMLU questions, which is a subset of the full MMLU dataset. We encourage the community to help annotate more questions to cover the entire dataset. Generally, we recommend you to follow these steps:

1. Familiarise yourself with the taxonomy. The taxonomy is designed to be simple and broad to cover all possible erroneous cases found in MMLU.

2. Open a new discussion on the Hugging Face page of MMLU-Redux, URL to be released in the camera ready.

3. Insert the title with the prefix "[ADD]" to signify additional data annotation, followed by the name of the MMLU subject you are contributing to. For example, "[ADD] Virology".

4. In the description body of the issue, specify the question, the choices, the MMLU answer, the error type, the correct answer (if applicable), the question source, what you believe to be the reason for the error, and optionally an additional comment. Follow this markdown template:

```
**Question**: {question}
**Choices**: {choices}  # Copy paste the python array
**MMLU answer**: {mmlu_answer}

**Error type**: {error_type}
# If this field is not applicable (e.g., for bad question clarity
# or bad options clarity), write "N/A"
**Correct answer**: {correct_answer}
# Ideally, a URL to where you can find the source online.
**Source**: {source}
 # This is strictly to describe what might have caused the error
 # (e.g., incorrect HTML parsing)
**Potential error reason**: {free text}
# This is where the annotators can leave additional comments
**Additional comment**: {free text}
```

5. If you intend to propose annotation of more than one question, we ask you to use a delimiter such as three-dashes ("---").

6. Submit the discussion.

We aim to respond to discussions within 48 hours. The proposed changes will either be integrated into the associated MMLU subject or receive feedback for further adjustments and clarification. Discussions with fully integrated changes will be closed.

Fix to the existing annotation The current and future annotated questions in MMLU-Redux may contain inaccuracies, particularly for questions whose answers may change over time, and other factors. We believe that the community can help keep the data up to date. We recommend everyone to propose fixes by following these steps:

1. Familiarise yourself with the taxonomy. The taxonomy is simple and broad to cover all possible erroneous cases found in MMLU.

2. Open a new discussion on the Hugging Face page of MMLU-Redux, URL to be released in the camera ready.

3. Insert the title with the prefix "[FIX]" to signify additional data annotation, followed by the name of the MMLU subject you are contributing to. For example, "[FIX] Virology".

4. In the description body, specify the question, the choices, the MMLU answer, the current error type, the new error type, the current possible correct answer, the new correct answer, the question source, the potential reason that causes the error, and the optional additional comments. Follow this markdown template:

```
**Question**: {question}
**Choices**: {choices}  # Copy paste the python array
**MMLU answer**: {mmlu_answer}

**Current Error type**: {current_error_type}
**Current Correct answer**: {current_correct_answer}
**New Error type**: {new_error_type}
# If it is not applicable
# (e.g., for bad question clarity or bad options clarity), write "N/A"
**New Correct answer**: {new_correct_answer}
# Ideally, a URL to where you can find the source online.
**Source**: {source}
# This is strictly to describe what might have caused the error
# (e.g., incorrect HTML parsing)
**Potential error reason**: {free text}
# This is where the annotators can leave additional comments
# about the data
**Additional comment**: {free text}
```

5. If you intend to propose more than one fix, please use a delimiter such as three-dashes ("---").

6. Submit the discussion.

For contributions that may not be suitable for submission via Hugging Face discussions or pull requests (e.g., bulk annotations stored in a CSV/PDF/DOCX file), we encourage you to reach out to us directly at *anonymised email addresses*, together with your HF username (if you have one and you are willing to be acknowledged). We will create a pull request which contains the proposed changes and tag your HF username. You can verify the changes and notify us if the pull request is ready to be merged via the discussion page.

D.3 Contingency Plan

We are committed to ensuring the long-term availability and accessibility of MMLU-Redux. If the primary hosting platform (*i.e.,* Hugging Face) becomes unavailable or ceases operations, we will immediately activate our contingency measures. This includes notifying all stakeholders and users of the dataset about the temporary unavailability and redirecting them to the mirror copies available on alternative platforms (*i.e.,* Zenodo, link to be included in camera ready). If the primary hosting ceases operations, we will ensure the dataset is reuploaded to a new reliable hosting platform, accompanied by updated documentation and links to the new location. We will also utilise social media and other networks to inform the community about the new location of the dataset.

E Authors Statement

The authors bear all responsibility in case of violation of rights.

F MMLU-Redux Error Type Statistics

In this section, we give the exact statistics of the error types found in MMLU-Redux. Table 5 contains an overview. We include the total number of questions in each subject, but note that we annotated a subset of 100 questions for each.

Table 5: Error types by subset. BQC is (1a) Bad Question Clarity, BOC is (1b) Bad Options Clarity, NCA is (2a) No Correct Answer, MCA is (2b) Multiple Correct Answers, and WGT is (2c) Wrong Ground Truth.

Subset	**Original Size**	**Annotated**	**OK**	**BQC**	**BOC**	**NCA**	**MCA**	**WGT**
Virology	166	100	43	14	2	4	4	33
Logical fallacies	163	100	74	14	2	4	3	3
College chemistry	100	100	75	2	0	2	0	21
Professional law	1,534	100	82	4	1	0	11	2
Human sexuality	131	100	83	11	1	2	1	2
Business ethics	100	100	86	14	0	0	0	0
Formal logic	126	100	87	1	0	2	9	1
Human aging	223	100	88	12	0	0	0	0
Global facts	100	100	88	2	1	4	0	5
Machine learning	112	100	89	3	5	0	1	2
Miscellaneous	783	100	90	4	1	4	1	0
Abstract algebra	100	100	91	2	0	1	0	6
Public relations	110	100	91	3	0	0	3	3
High school European history	165	100	91	4	4	0	0	1
Professional accounting	282	100	94	0	1	5	0	0
Astronomy	152	100	94	2	2	1	0	1
Security studies	245	100	94	6	0	0	0	0
Conceptual physics	235	100	95	0	0	0	1	4
International law	121	100	95	0	0	0	3	2
High school biology	310	100	95	1	1	1	0	2
College medicine	173	100	96	2	0	1	0	1
Moral disputes	346	100	96	3	0	0	0	1
Marketing	234	100	96	2	2	0	0	0
Professional psychology	612	100	96	1	0	1	0	2
High school psychology	545	100	96	4	0	0	0	0
High school physics	151	100	97	0	0	1	0	2
Elementary mathematics	378	100	97	1	0	0	0	2
High school macroeconomics	390	100	97	2	0	0	1	0
College computer science	100	100	97	0	0	2	1	0
Computer security	100	100	97	0	0	0	0	3
Moral scenarios	895	100	97	3	0	0	0	0
Econometrics	114	100	97	3	0	0	0	0
Nutrition	306	100	98	1	1	0	0	0
High school microeconomics	238	100	98	1	0	0	0	1
Sociology	201	100	98	2	0	0	0	0
High school statistics	216	100	98	0	0	0	0	2
College biology	144	100	98	1	0	0	0	1
Electrical engineering	145	100	98	1	0	0	0	1
High school world history	237	100	99	0	0	0	0	1
Anatomy	135	100	99	0	0	0	0	1
High school mathematics	270	100	99	0	0	0	0	1
High school chemistry	203	100	99	1	0	0	0	0
Jurisprudence	108	100	99	0	1	0	0	0
World religions	171	100	99	1	0	0	0	0
US foreign policy	100	100	99	1	0	0	0	0
Clinical knowledge	265	100	99	1	0	0	0	0
Professional medicine	272	100	99	1	0	0	0	0
College mathematics	100	100	99	1	0	0	0	0
Management	103	100	100	0	0	0	0	0
Philosophy	311	100	100	0	0	0	0	0
Medical genetics	100	100	100	0	0	0	0	0
High school US history	204	100	100	0	0	0	0	0
High school government and politics	193	100	100	0	0	0	0	0
High school geography	198	100	100	0	0	0	0	0
High school computer science	100	100	100	0	0	0	0	0
College physics	102	100	100	0	0	0	0	0
Prehistory	324	100	100	0	0	0	0	0
Total	14,042	100	94	2	0	1	1	2

G Prompting Methods

Below, we provide the prompts used for both standard prompting and Chain of Thought (CoT) prompting methods.

Throughout the evaluation, we used the test split of the MMLU-Redux loaded with the specified configuration. The hyperparameters include a temperature of 0.0, top_p of 1, frequency_penalty and presence_penalty of 0, and max_tokens of 600 for both the standard prompting and Chain of Thought (CoT) prompting methods to ensure consistency and deterministic results. The default random seed was used.

Standard Prompting

Task:
Given a question, its choices, and the ground truth answer, classify the question as either 'ok' or 'not ok'.

- 'ok' means that the question and the choices are understandable, and the ground truth answer is correct.
- 'not ok' means that the ground truth answer is incorrect, or the question and the choices are not well presented.
Classify with 'ok' or 'not ok' WITHOUT PROVIDING ANY REASONING

Chain of Thought (CoT) Prompting

Task:
Given a triple consisting of a multiple choice question, its choices, and the corresponding ground truth answer, your task is to classify the triple into 'ok' or 'not ok'.

Instructions:
1. Question Presentation: Is the question well-presented? Assess clarity, grammar, and sufficiency of information.
1.1 If Yes, assess the presentation of the MC options.
1.2 If No, classify the issue as 'not ok'.
2. MC Options Presentation: Are the MC options well-presented? Check if the options are clear, distinct, and relevant to the question.
2.1 If Yes, determine if there is one potentially correct answer.
2.2 If No, classify the issue as 'not ok'.
3. Answer Evaluation: Is there one, more than one, or no potentially correct answer in the options list?
3.1 If one, continue to Ground Truth Answer Evaluation.
3.2 If more than one, classify the issue as 'not ok'.
3.3 If no correct answer, classify the issue as 'not ok'.
4. Ground Truth Answer Evaluation: Is the ground truth answer correct?
4.1. If Yes, classify as 'ok'.
4.2. If No, classify as 'not ok'.
Provide your assessment in JSON format with keys 'Question Presentation', 'MC Options Presentation', 'Answer Evaluation', 'Ground Truth Answer Evaluation', 'Classification'.
The 'classification' is either 'ok', or 'not ok'.
FOLLOW THE EXACT EXAMPLE ANSWER FORMAT ALL IN ONE LINE WITHOUT PROVIDING EXPLANATION
Example Answer:
{"Question Presentation": "OK", "MC Options Presentation": "OK", "Answer Evaluation": "One", "Ground Truth Answer Evaluation": "Correct", "Classification": "OK"}

H Details on automatic error detection

H.1 Detailed Results on Error Detection Experiments for MMLU-Redux

Table 6: Comparison of Zero-Shot, Chain-of-Thought (CoT), and Few-Shot F1 score on 10 MMLU-Redux subjects.

Dataset	Models	Zero-Shot	Zero-Shot CoT	Few-Shot	Few-Shot CoT
College Chemistry	GPT-4-Turbo	52.94	17.14	55.74	48.98
	GPT-4o	42.86	30.77	19.35	30.77
	Claude-3-Opus	40.00	36.36	24.24	32.26
	Llama-3-70B	0.00	0.00	13.00	19.00
College Mathematics	GPT-4-Turbo	0.00	20.00	0.00	0.00
	GPT-4o	0.00	21.43	0.00	21.43
	Claude-3-Opus	9.52	0.00	0.00	32.26
	Llama-3-70B	0.00	0.00	0.00	0.00
Econometrics	GPT-4-Turbo	0.00	8.70	28.57	20.69
	GPT-4o	0.00	29.41	40.00	29.41
	Claude-3-Opus	15.38	0.00	46.15	0.00
	Llama-3-70B	0.00	0.00	57.00	55.00
Formal Logic	GPT-4-Turbo	23.33	22.22	26.09	0.00
	GPT-4o	33.33	0.00	0.00	0.00
	Claude-3-Opus	26.67	55.81	0.00	0.00
	Llama-3-70B	0.00	0.00	0.00	0.00
Global Facts	GPT-4-Turbo	19.35	30.00	20.00	44.44
	GPT-4o	16.67	28.57	21.05	28.57
	Claude-3-Opus	28.57	32.26	38.09	37.50
	Llama-3-70B	0.00	0.00	22.00	32.00
High School Physics	GPT-4-Turbo	9.52	17.39	23.53	16.67
	GPT-4o	19.05	26.09	30.77	26.09
	Claude-3-Opus	8.33	21.05	28.57	33.33
	Llama-3-70B	0.00	0.00	18.00	22.00
Machine Learning	GPT-4-Turbo	20.83	9.52	30.77	35.71
	GPT-4o	40.00	20.00	28.57	20.00
	Claude-3-Opus	33.33	23.08	31.58	46.15
	Llama-3-70B	0.00	0.00	0.00	14.00
Professional Law	GPT-4-Turbo	25.00	17.14	21.43	22.86
	GPT-4o	8.00	23.53	8.00	23.53
	Claude-3-Opus	29.79	32.43	16.00	23.53
	Llama-3-70B	11.00	9.00	8.00	9.00
Public Relations	GPT-4-Turbo	0.00	20.00	31.58	52.17
	GPT-4o	0.00	31.58	25.00	31.58
	Claude-3-Opus	0.00	80.77	38.09	35.29
	Llama-3-70B	17.00	17.00	15.00	14.00
Virology	GPT-4-Turbo	76.00	73.12	81.19	75.27
	GPT-4o	0.00	81.19	76.59	81.19
	Claude-3-Opus	78.43	25.00	71.74	0.00
	Llama-3-70B	54.00	56.00	52.00	67.00

Table 7: Comparison of Zero-Shot, Chain-of-Thought (CoT), and Few-Shot F2 score on 10 MMLU-Redux subjects.

Dataset	Models	Zero-Shot	Zero-Shot CoT	Few-Shot	Few-Shot CoT
College Chemistry	GPT-4-Turbo	45.69	26.55	50.29	48.39
	GPT-4o	48.39	18.69	30.61	22.94
	Claude-3-Opus	40.00	50.85	35.09	27.27
	Llama-3-70B	0.00	0.00	0.00	0.00
College Mathematics	GPT-4-Turbo	0.00	0.00	0.00	0.00
	GPT-4o	0.00	0.00	0.00	0.00
	Claude-3-Opus	6.17	0.00	0.00	0.00
	Llama-3-70B	0.00	0.00	0.00	10.64
Econometrics	GPT-4-Turbo	6.13	15.63	10.49	39.47
	GPT-4o	13.33	0.00	32.26	50.00
	Claude-3-Opus	10.53	40.00	34.88	60.00
	Llama-3-70B	0.00	0.00	9.52	14.15
Formal Logic	GPT-4-Turbo	17.41	22.73	26.32	20.83
	GPT-4o	35.09	8.47	30.30	24.19
	Claude-3-Opus	24.69	32.79	40.54	31.75
	Llama-3-70B	0.00	0.00	0.00	0.00
Global Facts	GPT-4-Turbo	17.05	26.79	25.00	37.04
	GPT-4o	16.67	9.09	25.00	17.86
	Claude-3-Opus	31.25	37.31	41.67	48.39
	Llama-3-70B	0.00	0.00	62.50	75.00
High School Physics	GPT-4-Turbo	6.29	31.25	6.99	30.30
	GPT-4o	13.33	23.81	23.26	38.46
	Claude-3-Opus	5.75	35.71	21.28	38.46
	Llama-3-70B	31.25	26.32	18.52	27.27
Machine Learning	GPT-4-Turbo	15.72	9.26	19.69	40.98
	GPT-4o	37.31	20.00	43.48	33.09
	Claude-3-Opus	27.03	47.62	34.88	39.06
	Llama-3-70B	12.82	12.82	41.67	57.47
Professional Law	GPT-4-Turbo	21.74	16.85	23.08	22.47
	GPT-4o	10.87	6.25	10.87	18.29
	Claude-3-Opus	26.12	32.97	21.74	29.41
	Llama-3-70B	43.65	45.45	25.00	27.78
Public Relations	GPT-4-Turbo	25.97	21.28	36.23	60.00
	GPT-4o	18.87	44.44	27.03	32.61
	Claude-3-Opus	20.55	28.30	35.09	49.18
	Llama-3-70B	0.00	10.64	12.50	12.20
Virology	GPT-4-Turbo	82.97	64.39	81.63	66.29
	GPT-4o	86.67	69.03	87.80	75.37
	Claude-3-Opus	84.39	76.36	83.76	79.42
	Llama-3-70B	6.85	6.49	6.25	6.49

Table 8: Comparison of Zero-Shot, Chain-of-Thought (CoT), and Few-Shot Recall score on 10 MMLU-Redux subjects.

Dataset	**Models**	**Zero-Shot**	**Zero-Shot CoT**	**Few-Shot**	**Few-Shot CoT**
College Chemistry	GPT-4-Turbo	41.86	24.00	47.22	48.00
	GPT-4o	52.94	16.00	50.00	20.00
	Claude-3-Opus	40.00	48.00	50.00	24.00
	Llama-3-70B	12.00	8.00	8.00	12.00
College Mathematics	GPT-4-Turbo	0.00	0.00	0.00	0.00
	GPT-4o	0.00	0.00	0.00	0.00
	Claude-3-Opus	5.00	0.00	0.00	0.00
	Llama-3-70B	0.00	0.00	0.00	0.00
Econometrics	GPT-4-Turbo	5.00	33.33	8.57	100.00
	GPT-4o	11.11	0.00	28.57	100.00
	Claude-3-Opus	8.70	66.67	30.00	100.00
	Llama-3-70B	0.00	0.00	0.00	100.00
Formal Logic	GPT-4-Turbo	14.89	23.08	23.33	23.08
	GPT-4o	36.36	7.69	40.00	23.08
	Claude-3-Opus	23.53	30.77	50.00	30.77
	Llama-3-70B	0.00	0.00	0.00	0.00
Global Facts	GPT-4-Turbo	15.79	25.00	23.53	33.33
	GPT-4o	16.67	8.33	28.57	16.67
	Claude-3-Opus	33.33	41.67	44.44	50.00
	Llama-3-70B	0.00	0.00	0.00	25.00
High School Physics	GPT-4-Turbo	5.13	66.67	5.71	66.67
	GPT-4o	11.11	33.33	20.00	66.67
	Claude-3-Opus	4.76	66.67	18.18	66.67
	Llama-3-70B	33.33	33.33	66.67	33.33
Machine Learning	GPT-4-Turbo	13.51	9.09	17.24	45.45
	GPT-4o	35.71	18.18	66.67	36.36
	Claude-3-Opus	24.00	54.55	37.5	45.45
	Llama-3-70B	0.00	9.09	0.00	9.09
Professional Law	GPT-4-Turbo	20.00	16.67	21.43	22.22
	GPT-4o	14.29	5.56	14.29	16.67
	Claude-3-Opus	24.14	33.33	28.57	27.78
	Llama-3-70B	5.56	5.56	5.56	5.56
Public Relations	GPT-4-Turbo	23.53	22.22	33.33	66.67
	GPT-4o	18.18	44.44	28.57	33.33
	Claude-3-Opus	18.75	33.33	33.33	66.67
	Llama-3-70B	11.11	11.11	11.11	11.11
Virology	GPT-4-Turbo	88.37	59.65	85.11	61.40
	GPT-4o	92.86	64.91	97.30	71.93
	Claude-3-Opus	88.89	73.68	94.29	77.19
	Llama-3-70B	38.60	40.35	36.84	52.63

I Heterogeneity of Errors in MMLU Subsets – Full List

Here, we provide an extensive list of qualitative observations from the manually validated MMLU subsets. Some of these observations were previously reported in by the AI Explained YouTube Channel (L and Stapleton, 2023) and a Medium blog post (Erenrich, 2023).

Professional Law – There are several contextual limitations that pose challenges for accurate question answering. One major issue is the lack of specificity regarding jurisdictions. The benchmark does not clearly distinguish between different jurisdictions, despite focusing on U.S. law. Additionally, the inherently interpretative nature of legal principles means that there is often no definitive ground truth, making it difficult to provide unequivocal answers.

Professional accounting – These questions mainly come from *FAR CPA Exams*, and are of high quality. There are minor issues in scraping where numeric answers are not converted properly and significance is lost, *e.g.,* where the given correct answer is $242,000, while the correct computed answer is $241,843. Furthermore, like in professional law, all questions assume U.S. professional accounting practice, even though this is rarely specified.

Human Aging – The questions are mostly correct, except for some questions containing underspecified information. *e.g.,* "In *this chapter's Senior View*, Dr. Shealy advises you to".

Global Facts – Almost all questions needed consulting external sources to be answered, such as `ourworldindata.org` (18 cases) and `pewresearch.org` (15 cases); in a few cases multiple sources were providing conflicting answers to the same question – for example, on the perceived corruption of political parties in 2013 Ethiopia, `ourworldindata.org` confirms the answer in MMLU, while other sources such as the Global Corruption Barometer from Transparency International were providing conflicting answers.

Virology – Incorrect ground truths labels are particularly prevalent within the Virology subset. Many of the incorrect labels are for relatively simple questions, such as identifying the description of a pandemic, this suggests errors may stem from problems parsing the original datasets (in most cases the *Human Virology* textbook's student resources). In addition, there is a range of issues relating to question and option clarity where the necessary context required to answer the question is missing, such as which family of diseases is being referred to.

Business Ethics – This subset includes several unclear questions where respondents must identify multiple correct statements in a listed statement. MMLU provides only the last statement of the list as the question, instead of including the entire question.

Philosophy – Several samples in this subset cannot be explicitly found in an external source.

Public Relations – This subset includes various errors, ranging from multiple correct answers and wrong ground truth to bad question clarity.

Anatomy – This is almost entirely correct.

Abstract algebra – This subset contains several answers that are simply incorrect. Some answers are ambiguous because they assume a ring does not need to have multiplicative identities, which is usually assumed in modern treatments of rings. This changes the outcome of the given answer.

College Chemistry – The questions were sourced from textbooks (*e.g., Chechik: Electron Paramagnetic Resonance*, *Hore: Nuclear Magnetic Resonance 2e*) and standardised college-level exams (*e.g., GRE Chemistry Practice Test*). We identified erroneous questions resulting from simple parsing errors and unknown causes. For example, questions spanning multiple lines in the original source were often parsed incorrectly, leading to a part of the question being presented as the first MMLU choice (Option A) and the exclusion of Option D. Additionally, some questions originally included option E, which was the correct answer, but this option was omitted from the MMLU choices to fit into the 4-choices question style. Furthermore, there were questions with ground truth labels that did not match the answers provided in the source, with no apparent cause for the discrepancy.

College Medicine – The questions were mostly sourced from textbooks (*e.g., Maughan & Gleeson: The Biochemical Basis of Sports Performance 2e*) and standardised college-level medical exam (*e.g.,* MCAT). The majority of question-answer pairs are of good quality. However, questions that were sourced from *Maughan & Gleeson: The Biochemical Basis of Sports Performance 2e* can also be found in the Clinical

Knowledge subject.

Clinical Knowledge – The questions were mostly sourced from textbooks (*e.g., Maughan & Gleeson: The Biochemical Basis of Sports Performance 2e*, *Cox & Roper: Clinical Skills*, *Endacott, Jevon & Cooper: Clinical Nursing Skills Core and Advanced*). The majority of question-answer pairs are of good quality. One specific question was annotated very well by adding the time when the question was asked (*i.e.,* "Which of the following statements is true about informal carers **(as of 2020)**?") which correctly indicates the ever-changing nature of these questions. However, questions sourced from *Maughan & Gleeson: The Biochemical Basis of Sports Performance 2e* can also be found in the College Medicine subject.

Formal Logic – The dataset contains a significant number of questions with incorrect answers. These are primarily sourced from the 'Oxford Learning Link' website. Inaccuracies are not because of invalid scraping: For example, one question states that $(F \wedge L) \wedge \neg C$ is correct, but $F \wedge L \wedge \neg C$ is not, even though these two formulas are clearly equivalent.

Logical fallacies – Some questions discuss fallacies, *e.g.,* the **solid** slope fallacy, which come from a set of flashcards obtained from the 'Quizlet' website, but otherwise do not return any hits on Google. There is also a large set of unclear questions involving an argument with a logical fallacy, but without any question. For instance, one "question" is "All things that are spoiled are inedible. Timothy is spoiled. So, Timothy is inedible." The original source of this question also had the instruction "Select the fallacy-type which best describes the reasoning contained in each of the passages below.", but this was lost when adding it to the MMLU dataset.

Machine Learning – Most questions were sourced from exam papers, assignments or online quizzes. About 30% of the questions require expert knowledge and reasoning. The main issues of this subset are bad question clarity and bad options clarity. *e.g.,* some quiz questions are based on past knowledge, and the descriptions in the questions may be vague or inapplicable today.

Electrical Engineering – All the questions are sourced from the 'Electrical4U' website. However, 2 answers have been incorrectly extracted from the website.

College Mathematics – The majority of the questions are from GRE papers, but they have been modified to have 4 options instead of 5. The majority of the questions require expert knowledge and reasoning. However, there is one question that was incorrectly adjusted when changing from 5 options to 4.

College Computer Science – The majority of the questions are from a Computer Science GRE practice exam and online boards with practice questions.

High School Mathematics – The majority of the questions are extracted from AP Mathematics and require expert knowledge and reasoning.

High School Statistics – The majority of the questions are from AP Statistics and can be directly obtained from the `crackap.com` website. However, two answers were incorrectly extracted.

Miscellaneous – The questions are direct facts regarding celebrities, movies, and global knowledge, which can be directly extracted from the internet. However, there were some incorrect answers and vague questions that could be difficult to answer accurately. For instance: 'How many cups of coffee were consumed in the United States in the past week (as of 2013)?'

Econometrics – The majority of the questions are correct, but some questions contain unverifiable references. (E.g., 'Consider again the VAR model of equation 16,' but equation 16 cannot be found within the question.)

High School Chemistry – This is almost completely correct.

High School Physics – This subset is almost entirely correct. The majority of the questions are from AP Statistics and can be directly obtained from the 'crackap' website.

College Physics – Some questions (approximately 20%) were duplicated.

High School Geography – This is almost completely correct.

Elementary Mathematics – This is almost entirely correct. There are only a few questions (mostly basic equations) with wrong ground truth answers.

Conceptual Physics – This is almost completely correct, with only a few questions having wrong ground truth answers due to issues with data scraping from unreliable sources.

Astronomy – Most questions and their corresponding answers are correct. A unique source was not

identified, but some can be found on 'Quizlet.' Two questions are unclear because they are missing essential details. Additionally, two question options are confusing due to incorrect parsing of the power of 10 (e.g., 10^{16} is incorrectly shown as 1016). One question lacks a correct answer because it is outdated, highlighting the importance of keeping questions updated with recent discoveries in this field.

Professional Psychology – This subset mainly contains questions in the style of Examination for Professional Practice in Psychology (EPPP) mainly used in most U.S. states and Canadian provinces. Among 100 randomly sampled questions, there are four erroneous instances. For instance, one of them contains no correct answer due to incorrect parsing (*i.e.,* incorrectly parsing '0.12' as '0.42'). Another example is a bad question clarity where the answer is highly dependent on the geographical context (*i.e.,* 'In most cases, statutory responsibility for establishing minimal standards for professional competency to protect the public from harm rests with ...').

Security Studies – The questions are sourced from textbooks (*e.g.,* '*Collins: Contemporary Security Studies 6e Student Resources*'). Six out of 100 instances have bad question clarity due to underspecified abbreviations (*e.g.,* 'HM' and 'TNC'). When read in this form without other supporting context, readers may confuse the abbreviation with other relevant concepts in the security studies. In the textbook, the questions were presented as part of a specific chapter, and the abbreviations were defined in the earlier questions.

High School Computer Science – This subset is completely clean.

High School Psychology – The questions are mostly sourced from AP Psychology practice questions. Among the 100 instances, there are four problematic ones. Three of them are due to missing statements. These questions are presented as multiple true-false questions (*i.e.,* the options are 'I only', 'II only', 'III only', and 'I and II only') without the four statements. One of the four erroneous instances is problematic due to racial stereotyping (*i.e.,* '*Aisha is a beautiful, black teenager. If she is* ***typical****, she most likely believes that*').

Moral Disputes – This subset contains four erroneous instances. For instance, three of them refers to underspecified entities (*e.g.,* '***Ashford's article*** *is meant to address a particular paralysis in the face of*', '*If* ***Thomson's conclusion*** *is correct, then*', '*In responding to terrorism, says* ***Walzer****, it is particularly important that*').

Moral Scenarios – This subset contains three instances with ambiguous questions requiring more background information before deciding whether the statement is 'morally wrong'.

High School World History – The questions are sourced from publicly available AP World History practice questions from sources such as 'crackap'. This subset is almost completely correct, except for one question with a wrong ground truth which may be caused by a random mistake.

High School European History – The errors in this subset are mainly due to bad option clarity. In some cases, there is an option that is equivalent to the correct one. Additionally, in a few samples, the passage provided as a reference does not contain the answer to the question, despite the question implying that it should.

Computer Security – The questions are sourced from university-level quizzes, with most of them being correct. There are three instances with wrong ground truth answer, however it is unclear what may have caused it.

Prehistory – Completely clean.

US Foreign Policy – The questions were mostly sourced from textbooks (*e.g., Cox & Stokes: US Foreign Policy 3e*) and publicly available final exams. There is only one question with bad clarity in terms of the temporal context (*i.e.,* 'How many states in the international system are likely to have nuclear weapons **right now**?').

Sociology – This subset contains questions that are sourced from textbooks (*e.g.,* '*Fulcher & Scott: Sociology 4e*'). There are two instances with bad question clarity, mainly because the question requires time and geographical contexts (*i.e.,* '*Patterns of drug use in Britain reveal that:*' and '*Which of the following policies did the New Labour government not pursue?*').

World Religions – This subset is almost completely correct. One instance with a bad question clarity requires more context about when is this question asked (*i.e.,* '*When was the* ***current*** *Dalai Lama born?*').

College Biology – This subset is mostly sourced from a publicly available GRE Biology practice test.

There are two erroneous instances; one of them has a wrong ground truth answer, while the other one has its ground truth answer copy pasted to the question due to incorrect parsing (*i.e.,* '*Stabilization of the unique coiled structure of an alpha helix in a protein is primarily attributed to* ***(A) hydrogen bonding between the peptide backbone atoms***').

High School Biology – Most questions in this subset are sourced from AP Biology practice questions. There are five erroneous instances, three of them seem to be caused by bad parsing. For instance, the correct option A is simply written as '0'. Another erroneous example is where the question refers to a previous question (*i.e.,* '***Refer to question 15 for details on the squirrel population.*** *Which of the following conditions is required to keep this population in Hardy-Weinberg equilibrium?*').

International Law – Most questions in this subset are sourced from a textbook (*i.e.,* 'Bantekas & Papastavridis: International Law Concentrate 4e'), with most of them are correct.

Jurisprudence – Most questions in this subset are sourced from a textbook (*i.e.,* 'Wacks: Understanding Jurisprudence 5e'), with most of them are correct.

Medical Genetics – This subset is sourced from university-level final exams and is completely correct.

Professional Medicine – This subset contains questions that are sourced from the publicly available sample questions of a standardised medical exam in the US (*i.e.,* the United States Medical Licensing Examination (USMLE)). Several questions with an unknown source can be answered with several searches on established websites (*e.g.,* National Center for Biotechnology Information (NCBI), Mayo Clinic). This subset only contains one erroneous instance due to bad question clarity where the original question has an accompanying photograph.

Marketing – The questions in this subset are sourced from university-level quizzes and textbooks. There are four erroneous instances. One question misses the temporal context, while the other is resented as a statement with unclear relation with the options.

High School Microeconomics – The questions in this subset are sourced from AP Microeconomics. Most of them are correct, except for two instances. One instance has a wrong ground truth answer for unknown reason. Another one has a bad question clarity, referring to unspecified question (*i.e.,* '*The basis for the answer in* ***number 1*** *is*').

Management – This is almost completely correct.

Nutrition – This is almost entirely correct. Only one question has a minor mistake in one of its options (*i.e.,* the '>' sign was mistakenly replaced by '?' mark) and one question was annotated by adding the time when the question was asked (similar to Clinical Knowledge) to indicate the ever-changing nature of the questions.

Human Sexuality – Most questions in this subset are sourced from university-level exams that are publicly available. There are 17 erroneous instances of various types. Several questions are highly subjective or speculative (*e.g.,* '*Persons with liberal attitudes about premarital sex are likely to:*', '*Alexander the Great was probably*' (in relation to his sexuality)).

High School Macroeconomics – The questions in this subset are sourced from AP Macroeconomics. Most of them are correct, except for three instances. One instance has two identical correct answers. The other two are fill-in-the-blank type of questions, however the position of the blanks are not specified.

High School Government and Politics – The questions are mostly sourced from AP U.S. Government and Politics practice questions, with most answers can be found on the 'crackap' website. This subset is completely correct.

High School US History – The questions in this subset are sourced from AP U.S. History practice questions, with all answers can be found on the 'crackap' website. This subset is completely correct.

J Error Detection via Fine-tuning

To validate our fine-tuning strategy for error detection, we developed LabelChaos, a dataset designed to mirror the error distribution of the original MMLU. This dataset serves as a benchmark for finetuning models, which are subsequently evaluated on MMLU-Redux.

To create LabelChaos, we selected and merged six manually labelled datasets. We chose datasets annotated by humans ((Clark et al., 2018; Mihaylov et al., 2018; Amini et al., 2019; Jin et al., 2021; Lin

et al., 2021b)) to avoid the potential inaccuracies associated with automated labelling procedures. After standardising these datasets to align with the format of MMLU, we generate a corrupted version of each by introducing specific types of corruption, as categorised in our annotation protocol (Fig. 3). These corruptions were carefully applied to replicate the quality and distribution characteristics of the MMLU dataset. The final dataset comprises approximately 264,000 samples, similar to those in MMLU. Below, we provide an overview of the LabelChaos subsets and the corruption procedures used for each type of error.

- **Wrong Ground Truth**: The correct answer label is replaced by randomly selecting from the incorrect answers.
- **Poor Question Clarity**: An LLM-based corruption is introduced by prompting GPT-3.5 Turbo to modify the question, increasing its difficulty. Few-shot examples from MMLU illustrating the "poor question clarity" error type are used.
- **No Correct Answers**: The correct answer is replaced with 'all options listed' or 'all of the above'.
- **Unclear Options**: Since most instances of unclear options in the original MMLU stem from parsing errors, we simulate these by introducing parsing errors into the choices, such as splitting strings at incorrect characters.
- **Multiple Correct Answers**: GPT-3.5 Turbo is used to generate a new option semantically identical to the correct one. One of the incorrect options is then replaced with this newly generated option.
- **Correct**: The original, uncorrupted dataset.

We fine-tune the Llama-3 (8B-Instruct) (AI@Meta, 2024) using LabelChaos datasets. To balance the distribution in MMLU-Redux, where most instances are labelled as "*correct*", we adjusted the label distribution to: 0.1 (Wrong Ground Truth), 0.1 (Poor Question Clarity), 0.1 (No Correct Answers), 0.1 (Unclear Options), 0.1 (Multiple Correct Answers), and 0.5 (correct). For consistency, we used the previously described CoT Prompt as input, with outputs labelled as either "**ok**" or "**not ok**". The training involves 2048 steps, with a batch size of 64, utilising the AdamW optimizer (Loshchilov and Hutter, 2018) with a learning rate of 2×10^{-4} and no weight decay. Due to computational constraints, we apply LoRA (Hu et al., 2021), with a rank of 16, to all models. All experiments were conducted on a single Nvidia A100 (40GB) GPU. We present the results of the fine-tuning method in Table 9.

Table 9: Comparison of Different Methods and Models for Error Detection in MMLU-Redux Scores.

Methods	Recall	F1 Score	F2 Score
GPT-4 Turbo (Few-shot CoT)	46.68	31.60	36.58
GPT-4 Turbo (Few-shot CoT)	38.47	29.26	31.36
Claude 3 Opus (Few-shot CoT)	48.85	24.03	40.29
Llama3-70B (Few-shot CoT)	24.87	23.16	23.10
Llama3-8B (Fine-tuning)	56.58	34.06	44.75