MMLU-SR: A Benchmark for Stress-Testing Reasoning Capability of Large Language Models

Anonymous ACL submission

Abstract

We propose MMLU-SR, a novel dataset designed to measure the true comprehension abilities of Large Language Models (LLMs) by challenging their performance in question-answering tasks with modified terms. We reasoned that an agent that "truly" understands a concept can still evaluate it when key terms are replaced by suitably defined alternate terms, and sought to differentiate such comprehension from mere text replacement. In our study, we modified standardized test questions by replacing a key term with a dummy word along with its definition. The key term could be in the context of questions, answers, or both questions and answers. Notwithstanding the high scores achieved by recent popular LLMs on the MMLU leaderboard, we found a substantial reduction in model performance after such replacement, suggesting poor comprehension. This new benchmark provides a rigorous benchmark for testing true model comprehension, and poses a challenge to the broader scientific community.

1 Introduction

Large Language Models (LLMs) have achieved impressive quantitative performance on a wide range of benchmarks, natural language processing (Zellers et al., 2019; Wang et al., 2019), general knowledge question-answering(Hendrycks et al., 2021; Clark et al., 2018), and coding (Chen et al., 2021; others, 2021). Additionally, by integrating with some advanced prompting techniques, such as Chain-of-Thought (CoT) (Wei et al., 2023) and its variants (Yao et al., 2023; Trivedi et al., 2023; Zhang et al., 2023), LLMs seem to exhibit a certain level of reasoning abilities including mathematics (Zhang et al., 2024) and even causal inference/discovery (Vashishtha et al., 2023; Wang et al., 2020; Mao et al., 2022; Gupta et al., 2021). However, some studies (Oren et al., 2023) have raised concerns about data leakage (i.e., training models on the test sets), potentially rendering these results unreliable. These seemingly contradictory findings prompt the question of whether LLMs are genuinely performing reasoning tasks or merely predicting the next token. If LLMs are truly capable of reasoning, they should remain unaffected by the replacement of key symbols within the test set.

A hallmark of human intelligence is the ability to handle abstract concepts and to associate them with arbitrary terms (Penn et al., 2008). With a few exceptions such as onomatopoeia, the connection between particular words and particular meanings is arbitrary, and identical concepts are invoked by different words in different human languages (e.g. *dog* vs *chien*). Similarly, human reasoners are capable of analogizing structural relationships from one domain to another, meaning that conceptual equivalence can be retained even when details change (Gentner and Medina, 1998). It follows that true human-like comprehension should be unimpaired when terms are substituted for synonymous terms, as long as the substitution is comprehensibly defined.

We wondered whether LLM peformance reflects true human-like comprehension in this sense, or whether it relies heavily on the specific terms used on training corpora. To assess this, we propose MMLU-SR, a new benchmark dataset that uses symbol replacement to remove some important terms from the questions and answers as shown in Figure 1. Instead of relying on memorized terms, this approach tests whether LLMs can reason using the definitions and concepts of those terms, ensuring a more robust evaluation of their understanding.

Our evaluations on GPT-3.5/4, Gemini, and Llama3 families showed significantly lower performance on MMLU-SR compared to the original MMLU, demonstrating the effectiveness of our approach in preventing models from exploiting memorized data. MMLU-SR thus provides a more challenging and revealing test of LLMs' true reasoning

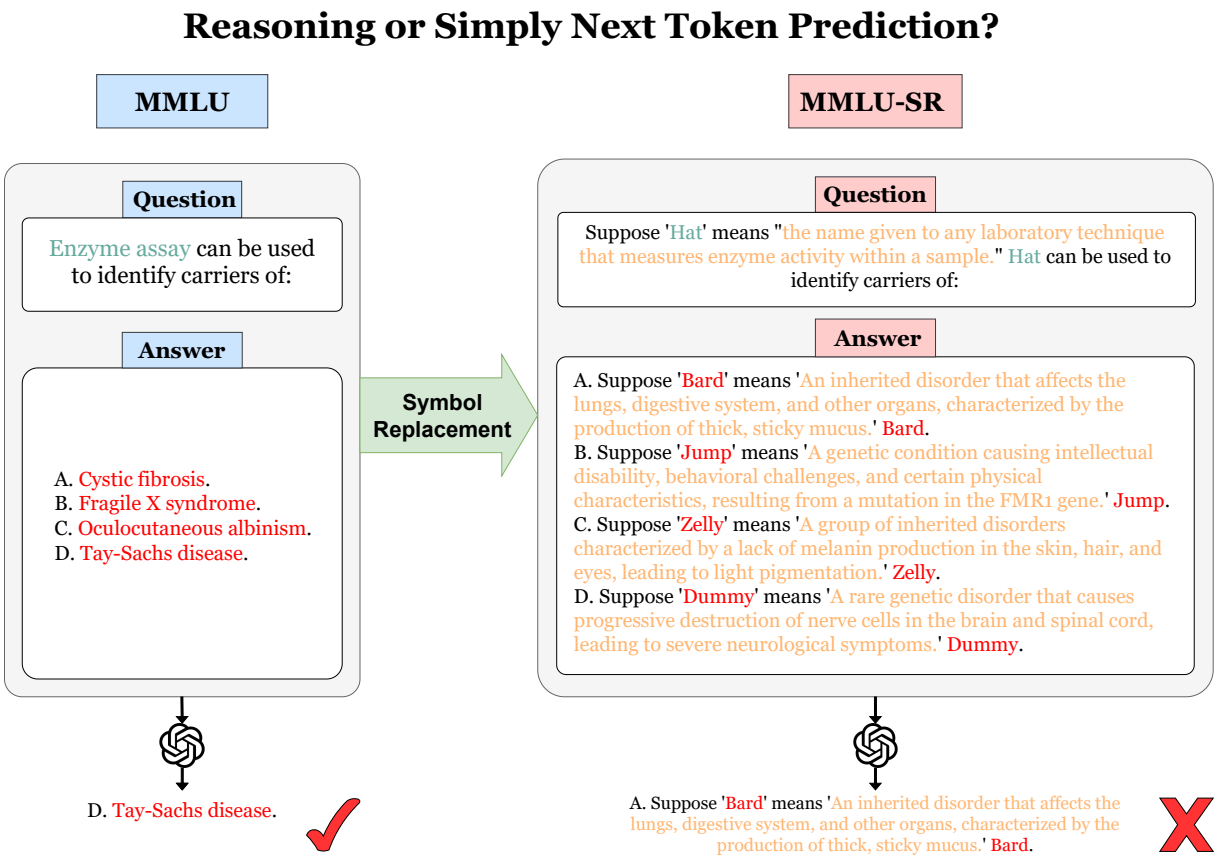

Figure 1: Illustration of our MMLU-SR testing scenarios. The red-colored and green-colored words represent the original symbols in the MMLU dataset showing in answers and questions, which are replaced in the MMLU-SR dataset with random words followed by their definitions, shown in orange text. The example question from the MMLU dataset is correctly answered by both GPT-3.5-turbo and GPT-4. However, the modified question from the MMLU-SR "Question and Answer" dataset is answered incorrectly by both models.

abilities and understanding.

Our findings indicate that while current LLMs excel on traditional benchmarks, they face substantial difficulties when key terms are replaced, highlighting the need for benchmarks like MMLU-SR to ensure robust and comprehensive evaluation of language models.

2 Related Works

MMLU Variants Benchmarks. MMLU Variants such as CMMLU (Li et al., 2024) and TMMLU+ (Tam et al., 2024) are adaptations of the MMLU benchmark for non-English languages; they translate the original MMLU questions and answers into other languages, providing a way to evaluate language models' performance in non-English contexts. These benchmarks are crucial for assessing the generalizability and robustness of models across different languages and cultural settings. They preserve the original structure and intent of MMLU while enabling a broader examination of multilingual capabilities.

Reasoning Capabilities Benchmarks. Several advanced reasoning benchmarks have been developed to evaluate the reasoning capabilities of language models. AGIEval (Zhong et al., 2023) includes standardized tests and civil service exams to assess reasoning and problem-solving skills in academic and professional scenarios. BoolQ (Clark et al., 2019) comprises over 15,000 real yes/no questions paired with Wikipedia passages to test the ability of models to infer answers from contextual information. GSM8K (Cobbe et al., 2021) features 8.5K grade-school math problems requiring multistep operations, targeting the evaluation of basic to intermediate mathematical problem-solving abilities. DROP (Dua et al., 2019), an adversarially-created reading comprehension benchmark, challenges models to navigate references and perform discrete operations such as addition and sorting,

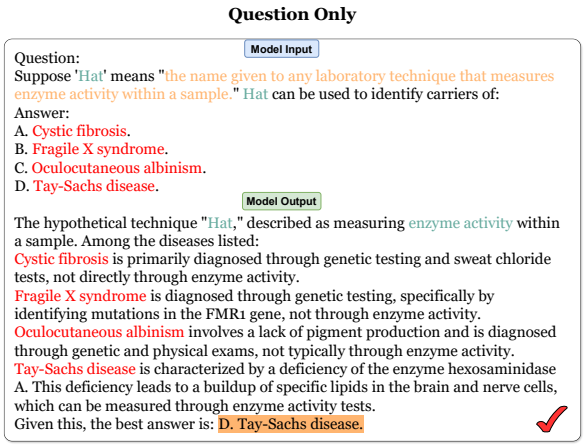

Figure 2: Example GPT-4 output of MMLU-SR 'Question Only".

thus evaluating their capacity to understand complex texts and execute logical reasoning tasks. Beyond purely language-based evaluation, on the multimodal front, MMNeedle (Wang et al., 2024) introduced one of the first multimodal benchmarks to evaluate long-context multimodal reasoning capabilities of multimodal LLMs.

Unlike advanced reasoning benchmarks and MMLU variants for language extension (e.g., CMMLU and TMMLU+), our MMLU-SR benchmark introduces a novel approach. It enhances the challenge by replacing key words within the questions with random words, each paired with its definition, to differentiate from other benchmarks. This approach targets the models' reasoning abilities by preventing reliance on memorized terms or vocabularies. By altering key symbols, MMLU-SR ensures that the evaluation focuses on the models' understanding and reasoning, rather than their recognition of specific vocabulary, thus providing a more robust assessment of their true cognitive capabilities. We build our benchmark on the MMLU dataset because it encompasses a wide range of subjects across various domains, including Humanities, Social Sciences, STEM, and Other fields. This diverse subject matter ensures a comprehensive evaluation of language models' reasoning capabilities, in contrast to other reasoning benchmarks that often focus exclusively on specific STEM subjects.

3 MMLU-SR Dataset

3.1 Dataset Construction

We have developed the MMLU-SR benchmark to rigorously evaluate the reasoning and understanding capabilities of LLMs. Inspired by ObjectNet (Borji, 2020), our benchmark contains three subsets: "Question Only", "Answer Only", and "Question and Answer", each offering a unique perspective on the data to comprehensively assess LLM performance. To reduce human efforts in some redundant tasks, we proposed an automatic process to generate our dataset.

1. **Term Extraction and Definition Generation**: We extracted key terms from the questions and answers across all 57 subjects using the assistance of `gpt-3.5-turbo`. The process involved careful few-shot prompting, and we separately extracted the contexts of questions or answers alone to ensure the model focused on extracting terms rather than solving the questions. We also retrieved appropriate definitions within the specific subject for each extracted term. For terms where the automated process provided irrelevant or inaccurate definitions, we manually reviewed and corrected these entries (see Appendix I for details on the extent of manual modifications).

2. **Dictionary Creation**: Once the terms and their definitions were extracted for each subject, we created JSON files where the terms served as keys and the definitions as values. This dictionary served as the basis for replacing terms in the questions and answers.

3. **Data Replacement**: Using the created dictionaries, we replaced the key terms in the questions with random dummy words followed by their definitions to create the "Question Only" dataset. Similarly, we did this for the answers to form the "Answer Only" dataset. This ensured that the context remained human-readable but required reasoning to infer the replaced terms. Some definitions and replacements required manual adjustments to ensure clarity and accuracy.

4. **Combining Question and Answer Sets**: After creating the "Question Only" and "Answer Only" datasets, we combined them to form the "Question and Answer" dataset. This step involved ensuring that the terms were consistently replaced across both questions and answers, maintaining the coherence of the dataset.

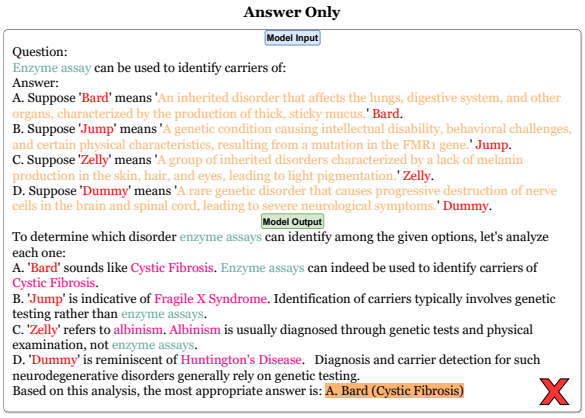

Figure 3: Example GPT-4 output of MMLU-SR "Answer Only".

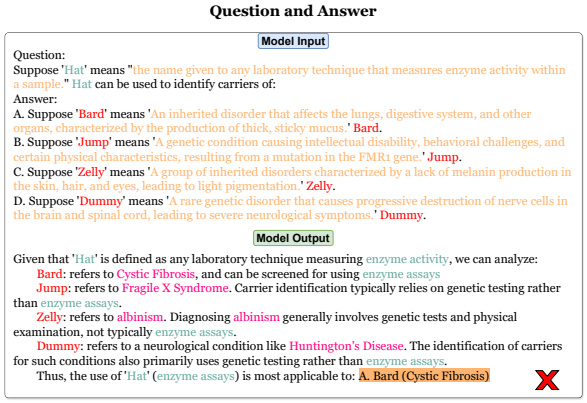

Figure 4: Example GPT-4 output of MMLU-SR "Question and Answer".

5. **Final Adjustments**: All CSV sheets were encoded in UTF-8 without headers. We manually fixed any typos that existed in the original MMLU dataset to ensure the quality and readability of the MMLU-SR dataset.

The MMLU-SR dataset was created using these meticulous steps. We formed both development and test sets, with the development set used for few-shot learning and the test set reserved for evaluation. This structured approach ensured that the dataset effectively tested the reasoning abilities of LLMs, differentiating between simple pattern recognition and genuine understanding. To effectively demonstrate how our MMLU-SR dataset can challenge more sophisticated models, we use Figure 2, Figure 3, and Figure 4 that feature responses from GPT-4. For comparative insights, example responses of GPT-3.5 are available from Table 5, Table 6, and Table 7 in Appendix B.

3.2 "Question Only" Dataset

Our "Question Only" dataset replaces key symbols with dummy words in most questions from the original MMLU dataset and keeps the answer choices unchanged. However, some straightforward questions, such as those involving only mathematical operations and numbers or simple questions like "Which of the following statements is true?", remain unmodified. Generally, we changed at least one important term in the context, replacing it with a random dummy word followed by its definition; sometimes, multiple terms are replaced in the question to further test the model's reasoning capabilities.

Figure 2 illustrates an example where the dummy word 'Hat' is used to replace term 'enzyme assay' and is defined as "the name given to any laboratory technique that measures enzyme activity within a sample." In this case, GPT-4 successfully recognized 'Hat' and understood that the critical aspect was whether the diseases listed in the answer choices could be identified through enzyme activity. The model systematically eliminated each answer option, except for the last one, as it was the only choice that could be measured through enzyme activity tests. This setup effectively tests the model's ability to comprehend and reason based on the provided definitions, rather than relying on pre-trained knowledge of the term 'enzyme assay'.

3.3 "Answer Only" Dataset

Our "Answer Only" dataset modifies the answer choices in most text from the original MMLU dataset similarly to the "Question Only" dataset and keeps the questions unchanged. We ensure that at least one answer choice is modified among all four choices. In specific cases, such as "moral scenarios" where the answers are simple TRUE or FALSE choices, we make substitutions like "Suppose 'Jack' means 'True'." and "Suppose 'Luck' means 'False'." Despite these modifications, Table 3 still shows the accuracy drops among all models compared to the original "moral scenarios" sheet from MMLU. This underscores the challenge posed to models in adapting to these symbolic substitutions. Figure 3 demonstrates that GPT-4 was able to recognize the replaced terms in answer choices A, B, and C, identifying 'Bard' as 'Cystic Fibrosis', 'Jump' as 'Fragile X Syndrome', and 'Zelly' as 'Albinism'. The model incorrectly identified the term 'Dummy' as 'Huntington's Disease', while the correct term is 'Tay-Sachs Disease'. Both disorders are indeed genetic, but they are distinct

Table 1: Performance of `gpt-4o-mini`,`gpt-4o`, `gemini-1.5-pro`, and `llama3-70b`.

Dataset	Humanities	Social Sciences	STEM	Other	Average
GPT-4o-mini					
MMLU (5-shot)	0.793	0.858	**0.689**	0.782	0.771
Question Only (5-shot)	0.744	0.792	**0.621**	0.724	0.710
Answer Only (5-shot)	0.659	0.738	**0.602**	0.651	0.655
Question and Answer (5-shot)	0.588	0.666	**0.531**	0.585	0.585
GPT-4o					
MMLU (5-shot)	0.880	0.906	**0.771**	0.854	0.845
Question Only (5-shot)	0.838	0.856	**0.702**	0.811	0.792
Answer Only (5-shot)	0.764	0.824	**0.705**	0.760	0.757
Question and Answer (5-shot)	0.708	0.754	**0.635**	0.712	0.695
Gemini-1.5-pro					
MMLU (5-shot)	0.849	0.881	**0.802**	0.815	0.832
Question Only (5-shot)	0.795	0.836	**0.700**	0.754	0.764
Answer Only (5-shot)	0.741	0.816	0.747	**0.739**	0.758
Question and Answer (5-shot)	0.690	0.752	**0.670**	0.681	0.694
Llama3-70B					
MMLU (5-shot)	**0.681**	0.868	0.697	0.814	0.765
Question Only (5-shot)	0.635	0.812	**0.631**	0.770	0.712
Answer Only (5-shot)	**0.539**	0.683	0.565	0.622	0.602
Question and Answer (5-shot)	**0.523**	0.653	0.536	0.591	0.576

in their genetic causes and manifestations. It appears that GPT-4, focusing on the broader category of 'genetic disorder' from the provided definition, inadvertently linked the description to the wrong disease. Such misidentification led the model to persist in incorrectly affirming that choice A ('Bard' as 'Cystic Fibrosis') was the correct answer (it is not).

3.4 "Question and Answer" Dataset

Our "Question and Answer" dataset integrates elements from both the "Question Only" and "Answer Only" datasets, replacing fundamental terms in both the questions and answer choices with dummy words followed by their definitions. As illustrated in Figure 4, GPT-4 successfully interpreted the original terms for each replaced term in answer choices A through C. However, similar to the results seen in Figure 3, the model incorrectly recognized the term in the last answer choice D ('Dummy' for Huntington's Disease), leading to an incorrect answer. This outcome contrasts with Figure 2, where GPT-4 correctly answered the questions when only the questions were modified. This illustrates that as complexity in context increases, with terms being replaced in both questions and answers, the model struggles to accurately identify the correct original term, consequently leading to an incorrect answer choice.

4 Experiments

4.1 Evaluation Protocol

We evaluated seven models across OpenAI, Gemini, Llama families: `gpt-3.5-turbo`, `gpt-4o-mini`, `gpt-4o`, `gemini-1.0-pro`, `gemini-1.5-pro`, `llama3-8b`, and `llama3-70b`. The evaluation for GPT and Gemini models was conducted using the Gemini-benchmark pipeline (Akter et al., 2023). For these models, we set the temperature parameter to 0 and utilized carefully crafted prompts that required responses in the format of "Answer: Letter of Choice." This approach ensures that the generated responses are directly comparable and suitable for evaluation. Additionally, both models were evaluated in the 5-shot setting, using examples from our development dataset to enhance their contextual understanding. Llama3 was evaluated using the lm-evaluation-harness framework (Gao et al., 2023). This model employed a different evaluation strategy; it uses log likelihood to determine the model's responses. Consistent with the other models, Llama3 also uses the same 5-shot setting, ensuring a standardized comparison across all tests. The complete results of all seven models are available in Appendix E.

Table 2: Relative percentage drop of accuracy in MMLU-SR compared to MMLU.

Dataset	Humanities	Social Sciences	STEM	Other	Average
GPT-4o-mini					
Question Only (5-shot)	6.18%	7.69%	**9.87%**	7.42%	7.91%
Answer Only (5-shot)	**16.90%**	13.99%	12.63%	16.75%	15.05%
Question and Answer (5-shot)	**25.85%**	22.38%	22.93%	25.19%	24.12%
GPT-4o					
Question Only (5-shot)	4.77%	5.52%	**8.95%**	5.03%	6.27%
Answer Only (5-shot)	**13.18%**	9.05%	8.56%	11.01%	10.41%
Question and Answer (5-shot)	**19.55%**	16.78%	17.64%	16.63%	17.75%
Gemini-1.5-pro					
Question Only (5-shot)	6.36%	5.11%	**12.72%**	7.48%	8.17%
Answer Only (5-shot)	**12.72%**	7.38%	6.86%	9.33%	8.89%
Question and Answer (5-shot)	**18.73%**	14.64%	16.46%	16.44%	16.59%
Llama3-70B					
Question Only (5-shot)	6.75%	6.45%	**9.47%**	5.41%	6.93%
Answer Only (5-shot)	20.85%	21.31%	18.94%	**23.59%**	21.31%
Question and Answer (5-shot)	23.20%	24.77%	23.10%	**27.40%**	24.71%

4.2 Results and Analysis

General Trend. Table 1 shows the accuracy of the four models `gpt-4o-mini`, `gpt-4o`, `gemini-1.5-pro`, and `llama3-70b` evaluated in both MMLU and our MMLU-SR. The data highlights how each model performs in the Humanities, Social Sciences, STEM, and Other academic fields, providing average scores for each subset. We observe consistent drop in model performance across all subsets when transitioning from the standard MMLU dataset to the more challenging MMLU-SR dataset, as evidenced by the decline in average accuracy from 0.771 on the MMLU dataset to 0.710, 0.655, and 0.585, on our MMLU-SR's "Question Only", "Answer Only", and "Question and Answer" subsets, respectively, for the `gpt-4o-mini` model. This trend of decreased performance is similarly observed in the other models.

We observe a crucial trend in decreasing accuracy across datasets: The "Question Only" dataset experiences the least drop, followed by the "Answer Only" dataset, with the most significant decline occurring in the "Question and Answer" dataset. This trend can be primarily attributed to two major reasons: (1) When only the question is modified, the model retains the original answer choices, facilitating the inference of the modified question's meaning; in contrast, altering the answer choices removes this contextual aid, challenging the model's ability to correctly match the question with the appropriate answer. (2) Answer choices are typically more concise and therefore lack the extensive context found in questions; consequently, replacing terms in the answers not only introduces ambiguity but also demands more complex inferential reasoning, disrupting the model's learned pattern-recognition strategies and resulting in a greater accuracy drop. The observations above also *justify the design of our MMLU-SR* on three variants (i.e., "Question Only", "Answer Only", and "Question and Answer").

Accuracy Drop in Each Category. Table 2 shows several aspects in the relative percentage drop of accuracy in MMLU-SR compared to that in MMLU across different categories for `gpt-4o-mini`, `gpt-4o`, `gemini-1.5-pro`, and `llama3-70b`:

1. **Humanities and Social Sciences.** For `gpt-4o`, the accuracy drops significantly in the Humanities category, with a 19.55% decrease in the "Question and Answer" dataset, and a slightly lower drop of 16.78% in Social Sciences. `gemini-1.5-pro` shows the smallest performance decline, with 18.73% and 14.64% drops in Humanities and Social Sciences respectively. `llama3-70b` exhibits the highest drops, reaching 23.20% in Humanities and 24.77% in Social Sciences for the "Question and Answer" dataset, indicating greater sensitivity to symbol replacement in these areas.

2. **STEM.** For `gemini-1.5-pro` and

Table 3: Detailed accuracy for different Humanities subjects across different models.

Subject	MMLU			Question Only			Answer Only			Question and Answer		
	GPT	Gemini	Llama3	GPT	Gemini	Llama3	GPT	Gemini	Llama3	GPT	Gemini	Llama3
Formal Logic	0.730	0.698	**0.532**	0.603	0.500	**0.484**	0.643	0.579	**0.516**	0.556	0.500	**0.460**
Logical Fallacies	0.902	0.902	**0.853**	0.883	0.834	**0.810**	0.853	0.847	**0.663**	0.834	0.841	**0.564**
Moral Disputes	0.882	**0.832**	0.847	0.832	0.806	**0.769**	0.777	0.830	**0.630**	0.711	0.749	**0.653**
Moral Scenarios	0.813	0.760	**0.318**	0.830	0.774	**0.289**	**0.143**	0.199	0.318	0.177	**0.167**	0.253
Philosophy	0.891	0.865	**0.865**	0.778	**0.724**	0.772	0.698	0.756	**0.598**	0.582	0.611	**0.582**
World Religions	0.901	**0.895**	0.906	0.895	**0.836**	0.895	0.842	0.813	**0.696**	0.825	0.772	**0.684**
High School European History	0.903	0.885	**0.848**	0.885	0.855	**0.830**	0.897	0.849	**0.721**	0.861	0.818	**0.739**
High School Us History	0.946	**0.922**	0.946	0.917	0.902	**0.887**	0.897	0.863	**0.799**	0.863	0.819	**0.799**
High School World History	0.937	**0.920**	0.945	0.924	0.920	**0.916**	0.907	0.865	**0.806**	0.882	**0.827**	0.840
Prehistory	0.948	**0.901**	0.910	0.904	0.836	**0.793**	0.843	0.803	**0.670**	0.790	0.769	**0.670**
International Law	0.942	0.926	**0.868**	0.901	**0.860**	0.868	0.934	0.843	**0.769**	0.835	0.802	**0.760**
Jurisprudence	0.898	0.861	**0.852**	0.852	0.861	**0.806**	0.861	0.806	**0.602**	0.722	0.750	**0.556**
Professional Law	0.749	0.666	**0.616**	0.683	0.627	**0.583**	0.641	0.585	**0.461**	0.563	0.544	**0.461**

`llama3-70b`, the STEM category shows a relatively moderate decrease in accuracy across the MMLU-SR datasets. Notably, `gemini-1.5-pro` experiences the highest drop of 12.72% in the "Question Only" dataset, indicating some sensitivity in this area. `llama3-70b` demonstrates a similar trend, with the highest drop of 9.47% in the STEM category, suggesting both models retain some robustness in STEM but are still impacted by symbol replacement. On the other hand, `gpt-4o` experiences a more consistent drop across all datasets in the STEM category, with the highest being 17.64% in the "Question and Answer" dataset, highlighting its relative vulnerability in this domain compared to `gemini-1.5-pro` and `llama3-70b`.

3. **Other.** The Other category generally shows a significant drop across all models and datasets, with the highest drops often observed in the "Question and Answer" dataset. For example, `gpt-4o` experiences a notable drop of 16.63% in the "Question and Answer" dataset, which is relatively consistent with its performance in other categories. `llama3-70b` shows the highest drop of 27.40% in this category, indicating a high sensitivity to contextual changes in this area. `gemini-1.5-pro` also shows a substantial drop of 16.44%, very similar to `gpt-4o`, suggesting that the "Other" category might be more context-dependent and hence more susceptible to performance degradation when symbols are replaced across all models.

Detailed Accuracy Drop in Each Subject. Table 3 shows a detailed comparison of accuracy scores across different models evaluated on various subjects in the Humanities category. The MMLU scores serve as a baseline for comparison. `gpt-4o` demonstrates exceptional performance across most subjects in this category, often leading in accuracy, particularly in complex subjects like Philosophy and International Law. `gemini-1.5-pro` also shows strong performance, but `gpt-4o` frequently matches or exceeds its accuracy. Notably, `gpt-4o` performs particularly well in subjects like High School World History and Jurisprudence. However, all models continue to struggle with Moral Scenarios, where the accuracy score drops significantly, particularly for `llama3-70b`, which shows a drastic decrease, reflecting a higher sensitivity to the challenges posed by the MMLU-SR datasets

Table 12 in Appendix F shows a detailed comparison of accuracy across different models evaluated on various subjects in the Social Science category. We observe that all models perform exceptionally well in Social Science on MMLU, particularly in High School Government and Politics, where `gpt-4o` achieves an impressive accuracy of 0.984. While there is still a drop in accuracy from MMLU to MMLU-SR's "Question and Answer" dataset, `gpt-4o` demonstrates remarkable resilience, maintaining accuracy levels around 0.7∼0.9 across most subjects. This performance significantly outpaces the other models, particularly in subjects like High School Psychology and Sociology. The drop in accuracy, though less pronounced for `gpt-4o`, still illustrates how our symbol replacement method increases difficulty, effectively stress-testing the models' reasoning capabilities versus mere memorization of pre-trained terms.

Table 13 in Appendix G shows a detailed comparison of accuracy across various STEM sub-

jects for different models. Each model demonstrated varying degrees of success across the subjects, with notable difficulties in some areas. College Mathematics and High School Mathematics remain challenging for all models, including gpt-4o, with accuracy dropping to around 0.4 to 0.5 in MMLU-SR's "Question and Answer" dataset. However, gpt-4o shows marked improvement in subjects like Astronomy, College Biology, and High School Biology, maintaining high accuracy even in the more challenging MMLU-SR datasets. The subject with the lowest accuracies among all models is still High School Mathematics, where llama3-70b struggles the most, especially in the Answer Only" and "Question and Answer" datasets. Similarly, College Physics and Abstract Algebra also show significant drops in accuracy across all models, highlighting the persistent challenges in subjects involving extensive calculations and complex problem-solving.

Table 14 in Appendix H shows a detailed comparison of accuracy scores across different models evaluated on various subjects in the Other category. We observe that gpt-4o performs exceptionally well in MMLU, with accuracy consistently above 0.9 in most subjects, significantly outperforming other models. Marketing stands out with a particularly high accuracy of 0.949 for gpt-4o, indicating outstanding performance in this subject. Professional Accounting shows improved performance with gpt-4o, achieving an accuracy of 0.766 in MMLU. Virology remains challenging, but gpt-4o shows improvement with an accuracy of 0.578. While there is still a drop in accuracy from MMLU to MMLU-SR's "Question and Answer" dataset, gpt-4o maintains relatively high performance, with accuracy generally staying above 0.7 for most subjects. Even in challenging areas like Virology and Global Facts, gpt-4o demonstrates resilience, maintaining accuracy levels significantly higher than other models.

CoT and System Instruction. We developed a simple baseline to test our MMLU-SR dataset on more recent and sophisticated models like GPT-4. This involves adding the instruction "Let's think step by step" at the end of answer choices to enable zero-shot CoT prompting. As shown in Table 4 from Appendix A, we also included a system instruction informing GPT-4 that the following questions would involve symbol replacement with arbitrary definitions. However, the example demonstrates that despite applying (zero-shot) CoT, the model still incorrectly interprets the term 'Dummy' in choice D as 'neurodegenerative disorder,' leading to the wrong answer, choice A. We applied this system instruction across the entire MMLU-SR dataset as well, with results shown in Table 8 from Appendix C. The results indicate that while the system instruction slightly improves accuracy in the "Question Only" and "Answer Only" datasets, the model still struggles with the increased complexity in the "Question and Answer" dataset.

5 Conclusion

We introduced MMLU-SR, a novel benchmark that challenges LLMs by replacing key terms in questions with random words followed by their definitions, aiming to test the models' reasoning and comprehension abilities rather than their memorization skills. Our evaluation across multiple domains revealed that popular LLMs suffer from significant drops in performance with these modifications, highlighting their reliance on memorized terms. MMLU-SR's unique approach addresses concerns about overfitting to traditional benchmarks and provides a more rigorous measure of true language understanding. This dataset will enable researchers to better identify and address the reasoning limitations of current LLMs, fostering the development of more robust and genuinely intelligent models.

6 Limitation

While MMLU-SR provides a novel approach to evaluate language models by incorporating symbol replacement, it has two major limitations: Firstly, some questions and answers remain unchanged without symbol replacement, especially in cases involving straightforward mathematical numbers or simple question forms. While these unchanged elements should not introduce significant noise, their presence might affect the intended complexity and variability of the evaluation. Secondly, MMLU-SR only involves multiple-choice questions, limiting its utility in assessing models' capabilities in generative tasks. This restriction could hinder comprehensive evaluations of a model's ability to generate coherent and contextually appropriate responses beyond structured choices. As the field progresses, expanding MMLU-SR to include generative components could provide a more holistic assessment of language models.

References

S. N. Akter, Z. Yu, A. Muhamed, T. Ou, A. Bäuerle, Á. A. Cabrera, K. Dholakia, C. Xiong, and G. Neubig. 2023. An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*.

Ali Borji. 2020. Objectnet dataset: Reanalysis and correction. *Preprint*, arXiv:2004.02042.

Mark Chen et al. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *Preprint*, arXiv:1905.10044.

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*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *Preprint*, arXiv:1903.00161.

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.

D. Gentner and J. Medina. 1998. Similarity and the development of rules. *Cognition*, 65(2/3):263–297.

Shantanu Gupta, Hao Wang, Zachary Lipton, and Yuyang Wang. 2021. Correcting exposure bias for link recommendation. In *ICML*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, Y. Gong, N. Duan, and T. Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. 2022. Causal transportability for visual recognition. In *CVPR*.

Y. Oren, N. Meister, N. Chatterji, F. Ladhak, and T. B. Hashimoto. 2023. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

Shuai Lu others. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *CoRR*, abs/2102.04664.

D. C. Penn, K. J. Holyoak, and D. J. Povinelli. 2008. Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(2):109–130.

Z.-R. Tam, Y.-T. Pai, Y.-W. Lee, S. Cheng, and H.-H. Shuai. 2024. An improved traditional chinese evaluation suite for foundation model. *arXiv preprint arXiv:2403.01858*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. 2023. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. 2024. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.11230*.

Yuhao Wang, Vlado Menkovski, Hao Wang, Xin Du, and Mykola Pechenizkiy. 2020. Causal discovery from incomplete data: A deep learning approach. In *AAAI StarAI Workshop*.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Boning Zhang, Chengxi Li, and Kai Fan. 2024. Mario eval: Evaluate your math llm with your math llm–a mathematical dataset evaluation toolkit. *arXiv preprint arXiv:2404.13925*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A CoT Examples

Table 4 shows an example of incorrect answer using zero-shot CoT with a system instruction produced by the GPT-4 on MMLU-SR's "Question and Answer" dataset. The correct answer is choice D, but GPT-4 responded with choice A.

B Examples of GPT-3.5 Response

Table 5 shows an example of incorrect answer produced by the GPT-3.5 on MMLU-SR's "Question and Answer" dataset. Table 6 shows an example of incorrect answer produced by the GPT-3.5 on MMLU-SR's "Question and Answer" dataset. Table 7 shows an example of incorrect answer produced by the GPT-3.5 on MMLU-SR's "Question and Answer" dataset.

C System Instruction Example

Table 8 shows the performance comparison of `gemini-1.0-pro` with and without using the system instruction "In each of the questions that I ask, I will replace some of the words that you might know with a word that is arbitrarily assigned a specific meaning just for this test. The meaning of these arbitrary definition may change with every question."

D Hard MMLU-SR Example

Table 9 illustrates a challenging example mentioned in the "Question and Answer" section in the main paper. In the question, we substituted "enzyme assay" with a synonym, "protein", and used the original terms from the MMLU dataset as placeholders in the answer choices, each paired with a definition. GPT-4 correctly identified the term "Tay-Sachs Disease" but associated it with the incorrect answer choice and definition, which was meant for "Fragile X syndrome." This example demonstrates that GPT-4 can be misled by the complexity of the dataset, tending to rely on pre-existing knowledge rather than engaging in substantive reasoning to align terms with their definitions.

E Complete Experiment Results

Table 10 shows our complete experiment results of different LLMs including `gpt-3.5-turbo`, `gpt-4o-mini`, `gpt-4o`, `gemini-1.0-pro`, `gemini-1.5-pro`, `llama3-8b` and `llama3-70b`. The percentage drop of each model is shown in Table 11.

F Social Science Category Detailed Accuracy Results

Table 13 shows detailed accuracy results in Social Science category of different LLMs including `gpt-4o`, `gemini-1.5-pro` and `llama3-70b`.

G STEM Category Detailed Accuracy Results

Table 13 shows detailed accuracy results in STEM category of different LLMs including `gpt-4o`, `gemini-1.5-pro` and `llama3-70b`.

H Other Category Detailed Accuracy Results

Table 14 shows detailed accuracy results in Other categoryof different LLMs including `gpt-4o`, `gemini-1.5-pro` and `llama3-70b`.

I Numbers of Human Modified Terms

Figure 5 shows the number of manually modified term definitions across 41 subject glossaries. These glossaries were created by consolidating related topics from the original 57 subjects in the MMLU dataset. Of the 28,676 terms initially generated by `gpt-3.5-turbo` for these 41 subjects, approximately 1,197 (4.2%) required human modification.

Table 4: Example of incorrect output using zero-shot CoT with system instruction.

System Prompt	In each of the questions that I ask, I will replace some of the words that you might know with a word that is arbitrarily assigned a specific meaning just for this test. The meaning of these arbitrary definition may change with every question.
Input	Suppose 'Hat' means 'the name given to any laboratory technique that measures enzyme activity within a sample.' Hat can be used to identify carriers of: A. Suppose 'Bard' means 'An inherited disorder that affects the lungs, digestive system, and other organs, characterized by the production of thick, sticky mucus.' Bard. B. Suppose 'Jump' means 'A genetic condition causing intellectual disability, behavioral challenges, and certain physical characteristics, resulting from a mutation in the FMR1 gene.' Jump. C. Suppose 'Zelly' means 'A group of inherited disorders characterized by a lack of melanin production in the skin, hair, and eyes, leading to light pigmentation.' Zelly. D. Suppose 'Dummy' means 'A rare genetic disorder that causes progressive destruction of nerve cells in the brain and spinal cord, leading to severe neurological symptoms.' Dummy. Let's think step by step.
Output	In this scenario, Hat is used to detect variations or abnormalities in enzyme levels that might be indicative of certain genetic disorders. A: Bard refers to cystic fibrosis. Areas of land (Jack) often have boundaries (Winky) that separate them from other areas. B: Jump refers to Fragile X syndrome, primarily involves genetic testing to detect the specific mutation rather than enzyme activity measurements. C: Zelly refers to conditions such as albinism, which are typically diagnosed based on pigment absence rather than enzyme activity, although some biochemical tests can be relevant in specific contexts. D: Dummy could be referring to several neurodegenerative disorders, which might be associated with enzyme activities related to nerve cell function or metabolic processes, but generally, such disorders are diagnosed through genetic markers and neuroimaging rather than enzyme activity alone. Given these explanations, the best match for 'Hat' in terms of being a technique to measure enzyme activity useful for diagnosis would be A, 'Bard' as cystic fibrosis, due to its direct relationship with enzyme dysfunction and measurement. ✗

Table 5: Example of incorrect response generated by GPT-3.5 in "Question Only" dataset.

Input	Suppose 'Hat' means "the name given to any laboratory technique that measures enzyme activity within a sample." Hat can be used to identify carriers of: A. Cystic fibrosis. B. Fragile X syndrome. C. Oculocutaneous albinism. D. Tay-Sachs disease
Output	"Hat" in this context refers to an enzyme assay technique rather than a genetic condition or disease. Therefore, it wouldn't be used to identify carriers of genetic disorders like cystic fibrosis, Fragile X syndrome, oculocutaneous albinism, or Tay-Sachs disease. It's essential to keep in mind the specific context in which terms are used in scientific fields! ✗

Table 6: Example of incorrect response generated by GPT-3.5 in "Answer Only" dataset.

Input	Enzyme assay can be used to identify carriers of: A. Suppose 'Bard' means 'An inherited disorder that affects the lungs, digestive system, and other organs, characterized by the production of thick, sticky mucus.' Bard. B. Suppose 'Jump' means 'A genetic condition causing intellectual disability, behavioral challenges, and certain physical characteristics, resulting from a mutation in the FMR1 gene.' Jump. C. Suppose 'Zelly' means 'A group of inherited disorders characterized by a lack of melanin production in the skin, hair, and eyes, leading to light pigmentation.' Zelly. D. Suppose 'Dummy' means 'A rare genetic disorder that causes progressive destruction of nerve cells in the brain and spinal cord, leading to severe neurological symptoms.' Dummy.
Output	An enzyme assay could be used to identify carriers of Bard, Jump, Zelly, or Dummy if the specific enzyme associated with each disorder is known and can be tested for. Enzyme assays are commonly used in genetic testing to detect the presence or absence of specific enzymes or their activity levels, which can provide information about an individual's genetic makeup and potential carrier status for certain genetic disorders. ✗

Table 7: Example of incorrect response generated by GPT-3.5 in "Question and Answer" dataset.

Input	Suppose 'Hat' means "the name given to any laboratory technique that measures enzyme activity within a sample." Hat can be used to identify carriers of A. Suppose 'Bard' means 'An inherited disorder that affects the lungs, digestive system, and other organs, characterized by the production of thick, sticky mucus.' Bard. B. Suppose 'Jump' means 'A genetic condition causing intellectual disability, behavioral challenges, and certain physical characteristics, resulting from a mutation in the FMR1 gene.' Jump. C. Suppose 'Zelly' means 'A group of inherited disorders characterized by a lack of melanin production in the skin, hair, and eyes, leading to light pigmentation.' Zelly. D. Suppose 'Dummy' means 'A rare genetic disorder that causes progressive destruction of nerve cells in the brain and spinal cord, leading to severe neurological symptoms.' Dummy.
Output	Based on the provided definitions: A. Hat (enzyme activity measurement) can be used to identify carriers of Zelly (lack of melanin production disorders). Explanation: Enzyme activity measurements are often used in medical diagnostics to identify various conditions, including genetic disorders. In this case, the lack of melanin production, as seen in Zelly disorders, can be identified using enzyme activity tests. ✗

Table 8: Performance comparison of the `gemini-1.0-pro` model with and without the system instruction.

Dataset	**Humanities**	**Social Sciences**	**STEM**	**Other**	**Average**
Gemini-1.0-pro					
Question Only (5-shot)	0.687	0.744	**0.539**	0.658	0.645
Answer Only (5-shot)	0.619	0.670	**0.504**	0.591	0.586
Question and Answer (5-shot)	0.582	0.622	**0.472**	0.544	0.546
Gemini-1.0-pro with System Instruction					
Question Only	0.688	0.735	**0.535**	0.650	0.652
Answer Only	0.619	0.661	**0.502**	0.576	0.590
Question and Answer	0.570	0.612	**0.460**	0.537	0.545

Table 9: Example of a hard version of MMLU-SR.

Input	Suppose 'Protein' means 'the name given to any laboratory technique that measures enzyme activity within a sample.' Protein can be used to identify carriers of: A. Suppose 'Oculocutaneous albini' means 'An inherited disorder that affects the lungs, digestive system, and other organs, characterized by the production of thick, sticky mucus.' Oculocutaneous albini. B. Suppose 'Tay-Sachs diseas' means 'A genetic condition causing intellectual disability, behavioral challenges, and certain physical characteristics, resulting from a mutation in the FMR1 gene.' Tay-Sachs diseas. C. Suppose 'Cystic fibrosi' means 'A group of inherited disorders characterized by a lack of melanin production in the skin, hair, and eyes, leading to light pigmentation.' Cystic fibrosi. D. Suppose 'Fragile X syndrom' means 'A rare genetic disorder that causes progressive destruction of nerve cells in the brain and spinal cord, leading to severe neurological symptoms.' Fragile X syndrom.
Output	The given problem seems to involve identifying which condition might be detected through a laboratory technique measuring enzyme activity. From the provided definitions, none of the conditions directly indicate an enzyme-related pathology except for a possible misalignment in the definitions and names: A: Oculocutaneous albini is defined incorrectly and does not relate to the production of enzymes. B: Tay-Sachs disease is a genetic disorder characterized by the destruction of nerve cells in the brain due to a deficiency of a specific enzyme (Hexosaminidase A). This is actually a plausible answer if we correct the name to Tay-Sachs disease, which fits the description of a disease involving enzyme activity. C: Cystic fibrosis and D: Fragile X syndrome are misdefined and do not relate to conditions involving measurable enzyme activity relevant to their actual conditions. Thus, the most likely answer, B. Tay-Sachs disease (with the corrected definition and name). ✗

Table 10: Complete performance of `gpt-3.5-turbo`, `gpt-4o-mini`, `gpt-4o`, `gemini-1.0-pro`, `gemini-1.5-pro`, `llama3-8b`, and `llama3-70b`.

Dataset	Humanities	Social Sciences	STEM	Other	Average
GPT-3.5-turbo					
MMLU (5-shot)	0.723	0.770	**0.554**	0.714	0.677
Question Only (5-shot)	0.661	0.702	**0.506**	0.641	0.616
Answer Only (5-shot)	0.540	0.595	**0.441**	0.538	0.520
Question and Answer (5-shot)	0.469	0.523	**0.396**	0.476	0.459
GPT-4o-mini					
MMLU (5-shot)	0.793	0.858	**0.689**	0.782	0.771
Question Only (5-shot)	0.744	0.792	**0.621**	0.724	0.710
Answer Only (5-shot)	0.659	0.738	**0.602**	0.651	0.655
Question and Answer (5-shot)	0.588	0.666	**0.531**	0.585	0.585
GPT-4o					
MMLU (5-shot)	0.880	0.906	**0.771**	0.854	0.845
Question Only (5-shot)	0.838	0.856	**0.702**	0.811	0.792
Answer Only (5-shot)	0.764	0.824	**0.705**	0.760	0.757
Question and Answer (5-shot)	0.708	0.754	**0.635**	0.712	0.695
Gemini-1.0-pro					
MMLU (5-shot)	0.728	0.758	**0.596**	0.703	0.686
Question Only (5-shot)	0.687	0.744	**0.539**	0.658	0.645
Answer Only (5-shot)	0.619	0.670	**0.504**	0.591	0.586
Question and Answer (5-shot)	0.582	0.622	**0.472**	0.544	0.546
Gemini-1.5-pro					
MMLU (5-shot)	0.849	0.881	**0.802**	0.815	0.832
Question Only (5-shot)	0.795	0.836	**0.700**	0.754	0.764
Answer Only (5-shot)	0.741	0.816	0.747	**0.739**	0.758
Question and Answer (5-shot)	0.690	0.752	**0.670**	0.681	0.694
Llama3-8B					
MMLU (5-shot)	0.593	0.757	**0.557**	0.729	0.651
Question Only (5-shot)	0.546	0.685	**0.507**	0.668	0.595
Answer Only (5-shot)	**0.455**	0.599	0.460	0.557	0.510
Question and Answer (5-shot)	**0.421**	0.538	0.424	0.499	0.465
Llama3-70B					
MMLU (5-shot)	**0.681**	0.868	0.697	0.814	0.765
Question Only (5-shot)	0.635	0.812	**0.631**	0.770	0.712
Answer Only (5-shot)	**0.539**	0.683	0.565	0.622	0.602
Question and Answer (5-shot)	**0.523**	0.653	0.536	0.591	0.576

Table 11: Complete relative percentage drop of accuracy in MMLU-SR compared to MMLU.

Dataset	**Humanities**	**Social Sciences**	**STEM**	**Other**	**Average**
GPT-3.5-turbo					
Question Only (5-shot)	8.58%	8.83%	8.67%	**10.22%**	9.08%
Answer Only (5-shot)	**25.31%**	22.73%	20.40%	24.65%	23.27%
Question and Answer (5-shot)	**35.12%**	32.08%	28.52%	33.30%	32.26%
GPT-4o-mini					
Question Only (5-shot)	6.18%	7.69%	**9.87%**	7.42%	7.91%
Answer Only (5-shot)	**16.90%**	13.99%	12.63%	16.75%	15.05%
Question and Answer (5-shot)	**25.85%**	22.38%	22.93%	25.19%	24.12%
GPT-4o					
Question Only (5-shot)	4.77%	5.52%	**8.95%**	5.03%	6.27%
Answer Only (5-shot)	**13.18%**	9.05%	8.56%	11.01%	10.41%
Question and Answer (5-shot)	**19.55%**	16.78%	17.64%	16.63%	17.75%
Gemini-1.0-pro					
Question Only (5-shot)	5.63%	1.85%	**9.56%**	6.40%	5.86%
Answer Only (5-shot)	14.96%	11.61%	15.44%	**15.91%**	14.48%
Question and Answer (5-shot)	20.05%	17.94%	20.81%	**22.60%**	20.85%
Gemini-1.5-pro					
Question Only (5-shot)	6.36%	5.11%	**12.72%**	7.48%	8.17%
Answer Only (5-shot)	**12.72%**	7.38%	6.86%	9.33%	8.89%
Question and Answer (5-shot)	**18.73%**	14.64%	16.46%	16.44%	16.59%
Llama3-8B					
Question Only (5-shot)	7.92%	**9.51%**	8.98%	8.36%	8.69%
Answer Only (5-shot)	23.27%	20.87%	17.41%	**23.56%**	21.28%
Question and Answer (5-shot)	28.16%	28.93%	23.88%	**31.56%**	28.63%
Llama3-70B					
Question Only (5-shot)	6.75%	6.45%	**9.47%**	5.41%	6.93%
Answer Only (5-shot)	20.85%	21.31%	18.94%	**23.59%**	21.31%
Question and Answer (5-shot)	23.20%	24.77%	23.10%	**27.40%**	24.71%

Table 12: Detailed accuracy for different Social Science subjects across different models.

Subject	MMLU			Question Only			Answer Only			Question and Answer		
	GPT	Gemini	Llama3	GPT	Gemini	Llama3	GPT	Gemini	Llama3	GPT	Gemini	Llama3
Econometrics	0.711	0.702	**0.693**	0.588	0.579	**0.570**	0.640	0.614	**0.561**	0.535	0.535	**0.421**
High School Macroeconomics	0.921	0.880	**0.821**	0.849	0.785	**0.779**	0.813	0.785	**0.628**	0.721	0.715	**0.572**
High School Microeconomics	0.971	0.929	**0.870**	0.903	0.870	**0.773**	0.857	0.815	**0.664**	0.769	0.744	**0.571**
High School Government And Politics	0.984	0.974	**0.969**	0.979	0.943	**0.938**	0.943	0.922	**0.798**	0.922	0.845	**0.782**
Public Relations	0.836	**0.746**	0.755	0.755	0.755	**0.736**	0.664	0.682	**0.600**	0.627	0.646	**0.555**
Security Studies	0.824	0.841	**0.824**	0.788	0.792	**0.767**	0.731	0.796	**0.673**	0.633	0.714	**0.624**
Us Foreign Policy	0.930	0.940	**0.930**	0.920	0.930	**0.890**	0.870	0.880	**0.740**	0.810	0.810	**0.780**
Human Sexuality	0.931	0.893	**0.855**	0.924	0.855	**0.840**	0.863	0.847	**0.710**	0.802	0.756	**0.756**
Sociology	0.935	**0.891**	0.920	0.900	0.896	**0.841**	0.881	0.881	**0.806**	0.831	0.851	**0.786**
High School Geography	0.955	0.939	**0.924**	0.894	0.909	**0.833**	0.884	0.864	**0.737**	0.813	0.813	**0.662**
High School Psychology	0.965	0.938	**0.921**	0.923	0.917	**0.884**	0.927	0.912	**0.719**	0.872	0.859	**0.739**
Professional Psychology	0.908	0.895	**0.845**	0.845	0.801	**0.788**	0.817	0.791	**0.627**	0.719	0.737	**0.601**

Table 13: Detailed accuracy for different STEM subjects across different models.

Subject	MMLU			Question Only			Answer Only			Question and Answer		
	GPT	Gemini	Llama3	GPT	Gemini	Llama3	GPT	Gemini	Llama3	GPT	Gemini	Llama3
Abstract Algebra	0.660	0.690	**0.380**	0.470	0.550	**0.370**	0.640	0.730	**0.400**	0.460	0.520	**0.400**
College Mathematics	**0.490**	0.680	0.510	**0.420**	0.630	0.490	**0.440**	0.650	0.460	**0.410**	0.610	0.480
High School Statistics	0.769	0.866	**0.699**	0.708	0.708	**0.657**	0.750	0.829	**0.620**	0.644	0.662	**0.597**
Elementary Mathematics	0.735	0.921	**0.606**	0.675	0.786	**0.521**	0.706	0.900	**0.561**	0.661	0.825	**0.497**
High School Mathematics	0.541	0.700	**0.422**	0.537	0.504	**0.356**	0.541	0.615	**0.426**	0.511	0.526	**0.367**
Astronomy	0.947	**0.901**	0.921	0.908	**0.829**	0.849	0.888	0.849	**0.697**	0.855	0.796	**0.684**
College Physics	0.686	0.716	**0.559**	0.559	0.647	**0.451**	0.618	0.745	**0.431**	0.480	0.608	**0.422**
Conceptual Physics	0.911	0.932	**0.783**	0.804	0.757	**0.677**	0.791	0.843	**0.494**	0.685	0.698	**0.447**
High School Physics	0.748	0.782	**0.563**	0.649	0.556	**0.530**	0.589	0.616	**0.477**	0.543	0.596	**0.450**
College Chemistry	**0.570**	0.610	0.580	**0.540**	0.550	0.570	0.550	0.530	**0.480**	**0.480**	0.560	0.470
High School Chemistry	0.759	0.788	**0.734**	0.709	**0.685**	0.631	0.670	0.680	**0.537**	0.586	0.626	**0.468**
College Biology	0.951	**0.868**	0.931	0.938	0.882	**0.854**	0.924	0.861	**0.708**	0.833	0.826	**0.625**
High School Biology	0.958	0.929	**0.903**	0.932	0.893	**0.858**	0.884	0.858	**0.713**	0.858	0.829	**0.729**
College Computer Science	0.790	0.790	**0.670**	0.690	**0.610**	0.650	0.760	0.730	**0.610**	0.670	0.660	**0.570**
Computer Security	0.840	**0.820**	0.830	0.830	0.770	**0.750**	0.760	0.730	**0.660**	0.760	**0.610**	0.720
High School Computer Science	0.910	0.920	**0.870**	0.860	0.880	**0.790**	0.880	0.910	**0.820**	0.850	0.870	**0.740**
Machine Learning	0.777	0.714	**0.652**	0.661	0.643	**0.589**	0.643	0.661	**0.527**	0.580	0.580	**0.509**
Electrical Engineering	0.841	0.807	**0.745**	0.752	0.724	**0.655**	0.655	0.710	**0.510**	0.566	0.655	**0.490**

Table 14: Detailed accuracy for different Other subjects across different models.

Subject	MMLU			Question Only			Answer Only			Question and Answer		
	GPT	Gemini	Llama3	GPT	Gemini	Llama3	GPT	Gemini	Llama3	GPT	Gemini	Llama3
Anatomy	0.911	**0.793**	0.807	0.874	0.733	**0.726**	0.815	0.667	**0.563**	0.726	0.659	**0.578**
Clinical Knowledge	0.898	**0.838**	0.849	0.811	0.785	**0.740**	0.796	0.755	**0.638**	0.713	0.709	**0.608**
College Medicine	0.832	0.844	**0.757**	0.780	0.786	**0.740**	0.798	0.763	**0.647**	0.717	0.740	**0.659**
Human Aging	0.830	**0.807**	0.807	0.794	**0.744**	0.758	0.704	0.740	**0.457**	0.632	0.691	**0.471**
Medical Genetics	0.960	0.910	**0.830**	0.900	0.850	**0.820**	0.840	0.780	**0.570**	0.830	0.740	**0.550**
Nutrition	0.899	0.876	**0.853**	0.863	**0.758**	0.804	0.798	0.784	**0.663**	0.699	0.703	**0.647**
Professional Medicine	0.956	**0.864**	0.868	0.919	**0.776**	0.868	0.901	0.783	**0.754**	0.842	**0.735**	0.754
Virology	0.578	0.578	**0.536**	0.548	0.506	**0.488**	0.524	0.542	**0.452**	0.524	0.494	**0.404**
Business Ethics	0.860	0.850	**0.750**	0.890	0.780	**0.720**	0.750	0.670	**0.500**	0.710	0.640	**0.480**
Management	0.913	**0.893**	0.913	0.883	**0.816**	0.903	**0.757**	0.835	0.728	0.767	0.767	**0.650**
Marketing	0.949	0.940	**0.923**	0.906	0.927	**0.880**	0.838	0.846	**0.615**	0.808	0.803	**0.662**
Global Facts	0.650	0.600	**0.530**	0.540	0.540	**0.430**	**0.580**	0.690	0.540	0.520	0.470	**0.410**
Miscellaneous	0.955	0.955	**0.903**	0.932	0.877	**0.860**	0.861	0.847	**0.692**	0.840	0.791	**0.616**
Professional Accounting	0.766	0.663	**0.638**	0.716	0.674	**0.596**	0.681	0.638	**0.514**	0.631	0.596	**0.489**

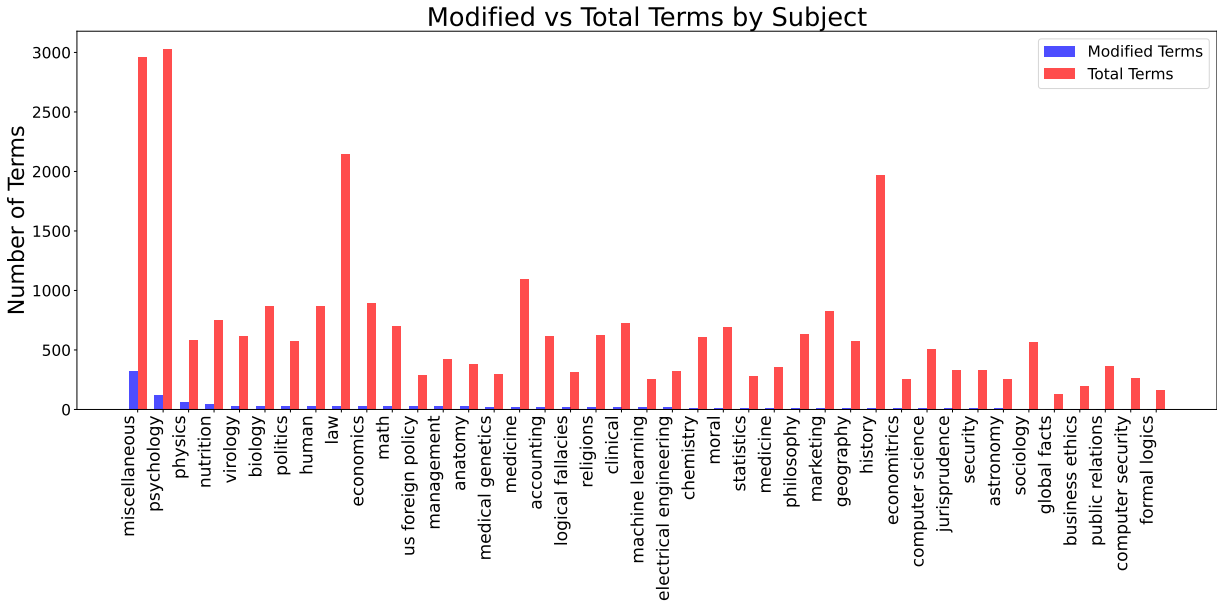

Figure 5: Comparison of total generated terms (red) and human-modified terms (blue) across 41 subject glossaries