To Accuracy and Beyond: Explore LLM Level of Knowledge via Elimination

Itay Chachy itay.chachy@mail.huji.ac.il
Omer Benishu omer.benishu@mail.huji.ac.il
Matan Velner matan.velner@mail.huji.ac.il

Abstract

Until 2022, Google was the primary tool for answering questions. Recently, with the rise of Large Language Models (LLMs), there has been a significant shift in how people search information online. This shift highlights the importance of question answering abilities for LLMs, and particularly, improving their results, and characterize their knowledge. This work aims to determine LLMs' **level of knowledge** on Multiple-Choice Questions (MCQ) by using **various elimination strategies**, and investigate whether those strategies can enhance their abilities. Our findings suggest that LLMs do not improve their performance on MCQ by using elimination strategies, which might indicate on their partial knowledge. Code, data and results are available at GitHub.

1 Introduction

LLMs are the primary tool for question-answering; evaluating them has become an active field of research, which has numerous applicative incentives. Today MCQ are one of the most highly-regarded and widely used type of objective test for the measurement of knowledge, ability, or achievement of human beings [1]. Naturally, this type of questions is also one of the most used measurements in LLMs' evaluation, through examining their accuracy score over existing MCQ datasets. A key limitation of this score is that it offers only a, partial and potentially misleading assessment of a model's knowledge [2].
We propose a different approach for evaluating LLMs on MCQ, by asking them to eliminate the wrong options, until reaching the right one.
With this approach we aim to investigate the following:

1. Can we characterize LLM's level of knowledge? (will be defined extensively in Section 2.1).
2. Can we improve LLMs' results on MCQ during inference time? Human tend to eliminate distractors from MCQ before reaching the right answer, can LLMs benefit from this strategy as well?
3. Evaluate the difficulty level of a dataset not only by the complexity of its content, but also by how good the distractors are.

To address the above, we conducted several experiments (Section 4) on different datasets (Section 3) using Llama3.1-8B-Instruct [3] and Qwen2-7B-Instruct [4].

2 Related work

2.1 Levels of Knowledge

In psychology, different levels of knowledge are defined [5, 1]. Knowledge for a given question can be characterized via elimination as follows: full knowledge (elimination of all distractors), partial knowledge (elimination of a subset of the distractors), partial misinformation (elimination of the correct answer and a subset of distractors), full misinformation (elimination of the correct answer alone) and absence of knowledge (eliminating all of the options).

2.2 Elimination In NLP

Several attempts were made to use elimination strategy in the NLP field, to check whether this approach, used by humans, can be applied to language models [6], and improve their results on MCQ datasets [7].
Moreover, we aim not only to improve an LLM's results, but also to gain a better understanding on its level of knowledge and "thinking" process [8].

2.3 LLMs' Sensitivity

A main challenge in the deep learning field is to maintain robustness for small changes in the input,

that keep its semantic concept, of a neural network. In modern NLP, where we condition a model's output with prompts for unseen downstream tasks [9], this issue becomes even more significant [10], when a certain semantic guideline can be expressed in many different ways. Previous works have found several techniques to face this challenge: by asking the model to present his chain-of-thought [11], optimizing prompts [12], avoiding biases as much as possible [13, 14] and more. In the experiments detailed in Section 4, the above has been taken into account when several prompts, models and datasets were examined.

3 Data

For a comprehensive evaluation, several MCQ datasets from various fields and difficulty levels were used to assess the model's level of knowledge. Considering that this experiment examines an LLM on inference, a relatively small amount of data is sufficient, since it estimates the distribution of the dataset, while noisy samples won't effect much. Having said that, 400 samples were taken from each dataset (except MMLU, which 1.5K samples were taken from its validation set), each with four possible options per question, distributed almost uniformly (Table 1).

- **A**I2 **R**easoning **C**hallenge (**ARC**) [15]: consists of real world knowledge questions taken from grade school science exams, and contains two subsets of difficulty levels (easy and challenge).
- **OpenBookQA** [16]: consists of synthetic world knowledge questions of elementary level science, while providing a context related fact.
- **R**e**A**ding **C**omprehension dataset from **E**xaminations (**RACE**) [17]: reading comprehension real dataset, collected from exams from middle and high school.
- Measuring **M**assive **M**ultitask **L**anguage **U**nderstanding (**MMLU**) [18]: consists of real world knowledge questions from various fields, mainly from college tests.
- **EasyAdditive**: a synthetic dataset created for the use of this paper, which contains simple math additive questions, where one of the distractors, which isn't a number, is easier to eliminate for humans than the others.

4 Experiments

To evaluate the level of knowledge of an LLM, several experiments were made. Each one tries to isolate a certain aspect in a model evaluation via different elimination strategy. As mentioned above, elimination process, in contrast to naive accuracy score, allows a better, more holistic view of a model's level of knowledge, as it depends on two objectives: identify the right answer, and identify all of the wrong options (distractors).
To address the presented research questions (Section 1), a baseline evaluation has been performed with an identical conditions (data size, similar prompt format, model quantization, etc.). In the baseline, each model is required to find the right answer to an MCQ.
As discusses in Section 2.3, before conducting the full experiments, we explored two models, on a subset of each dataset, for a wide range of prompts, guided the model to justify its decisions (chain-of-thought). Eventually, we noticed that Llama3.1 outperformed Qwen2, so we decided to report only its results.

4.1 Iterative Elimination Approach

To simulate an elimination process, a question is given to an LLM, and it is asked to eliminate one wrong option from it. This process is repeated iteratively until one answer remains (if the model was correct), or interrupted in the middle when the right answer has been eliminated wrongly (Fig. 3). As we are not interested in checking accuracy scores of models but rather to evaluate their level of knowledge, we focus on testing if a model is capable not only to find the right answer, but also to eliminate all of the wrong ones. To address this question, we check the elimination score over questions that the baseline approach answered correctly. As shown in (Fig. 1), the model doesn't reach full knowledge over the questions that it was originally considered capable of answering, but rather only partial knowledge. This interesting phenomena emphasises the problem with the naive accuracy score.

4.2 One-Shot Iterative Elimination

As suggested in [19, 20], adding an in context example to a given prompt has proven useful for improving its results. We adopted this approach, and conducted the previous experiment, with the single change of adding an example (one-shot) of an elimination step. Each example matched the current

elimination step; meaning, the number of possible options matched the presented question's number of options. This change hasn't improve the results significantly (Table 2), which might indicate that indeed the model is lacking full knowledge, and that the problem wasn't with the initial prompt, or misunderstanding the elimination task.

4.3 Accuracy Improvements

One of our initial research questions (Section 1) was whether an elimination based strategy can improve the accuracy of a model (similar to [11]). In addition to the experiments above (Section 4.1, Section 4.2), we tried two different elimination settings as an attempt to improve the baseline's accuracy: elimination-based chain-of-though, where the model is asked to explain why each distractor is wrong before reaching the right answer (Fig. 4), and elimination-based iterative chain-of-though, where the model is asked to perform the initial iterative experiment but in a single inference step (where the entire elimination process is part of its context) (Fig. 5). All of them performed reasonably, but didn't outperform the baseline accuracy (Fig. 2). Our hypothesis is that this is due to: lack of full knowledge as we claimed above, and that the original training data doesn't contain much examples of elimination process.

4.4 Human-Like Decision Making Process

Another perspective of a model's elimination strategy, is related not only to its level of knowledge, but rather to how and what it chooses to eliminate. When facing a MCQ, human beings, tend to eliminate the least likely distractor first. To compare a model's decision making process with human beings, we conducted the following experiment: we built a dataset consisting of simple two numbers additive questions, and 4 possible answers: the correct answer, two integer distractors, and an unrelated string of a fruit/animal (Fig. 3). The purpose of this experiment is to check whether the model decides to eliminate the string option first (as it is the least likely one), when no explicit guidance is given for that. We have found, that interestingly the model indeed operates in a similar manner to humans (Table 3), which might indicate that further study about elimination usage can lead to an improvement.

5 Conclusions

Elimination is a strong tool to obtain a more holistic analysis of a model's knowledge on MCQ, in contrast to a simple accuracy score. Our main contribution is proving that a model doesn't reach full knowledge yet (which is a harder requirement for humans as well), and we should reconsider their evaluation metrics. Another perspective of this point, is that elimination process might not be the best choice for prompting strategies, but it can potentially be helpful in assessing a model's level of knowledge, and evaluating the difficulty level of a dataset.

5.1 Future Work

5.1.1 Fine-Tuning

As discussed, we believe that LLMs' majority of training data does not involve elimination, hence the model is struggling with this type of task. Given all of the opportunities it provides, it is interesting to test how fine-tuning a model with elimination based tasks, can improve it overall results (and may indicate a higher knowledge level). In addition, one may consider testing a larger, stronger model.

5.1.2 Improve Benchmarks Reliability

LLMs have achieved impressive success on many benchmarks. However, there is growing concern that some of this performance actually reflects dataset contamination [21], where data closely resembling benchmark questions leaks into the training data, instead of true reasoning ability. It is a difficult issue to tackle since the big companies has a financial incentive to obtain the best results, and most of them don't publish information regarding their training data. A similar concept to augmentation can be helpful here; for example, formulating a question such that the model would have to eliminate wrong answers instead of finding the right one. The first advantage would be to make the benchmarks more robust to data contamination, and in addition, we would gain information regarding the true level of knowledge of each model, and the true difficulty level of a dataset (as proposed in Table 4).

5.1.3 Elimination For Unknown Questions

An interesting direction to study, is whether elimination strategy can improve model's results on question on which it doesn't know the answer to [22]. It is a difficult task for a model to identify such questions, but still, such approach may lead to improvements.

6 Figures and Tables

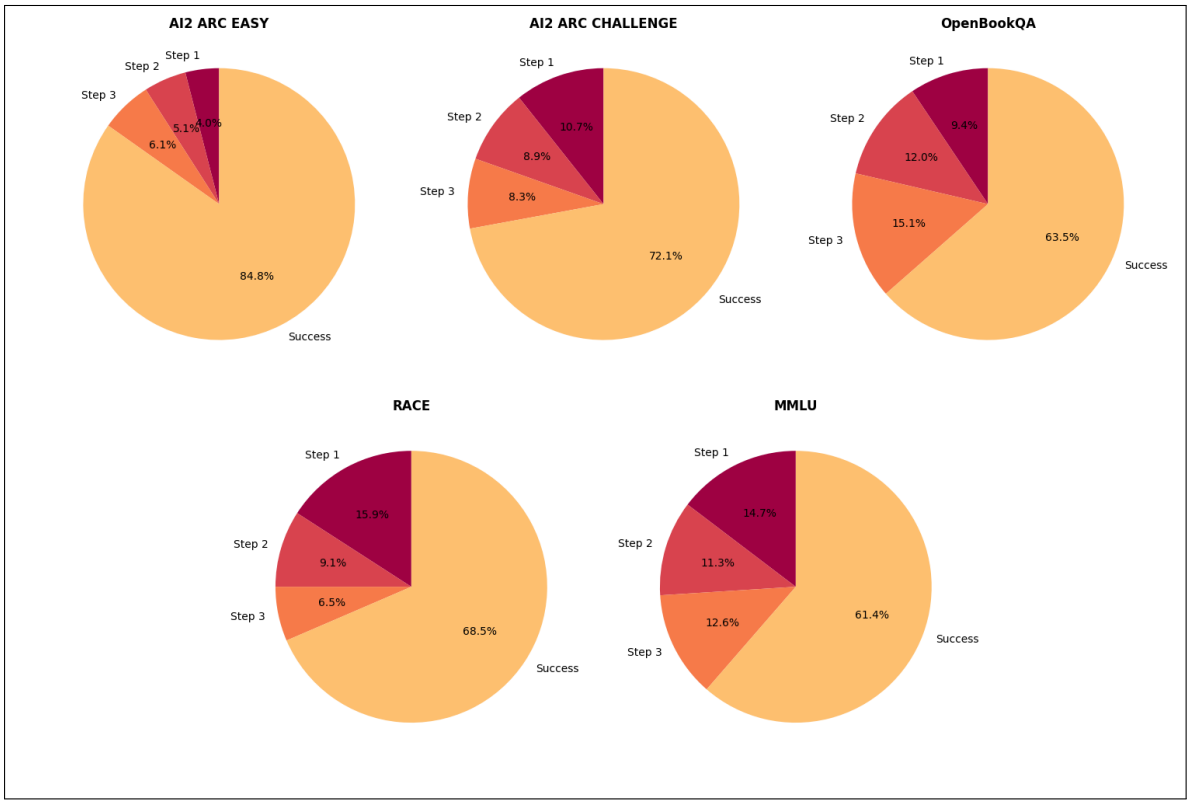

Figure 1: Step of elimination mistake of Llama-3.1, over successful questions obtained by the baseline approach. Step t means that the model wrongly eliminated the correct answer in the t-th step, and success means that the model achieved full knowledge over a question and was able to eliminate all of the wrong options.

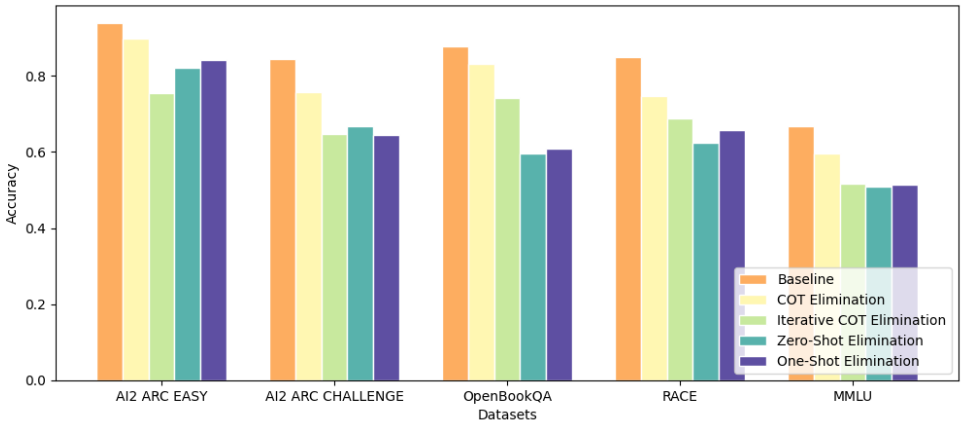

Figure 2: Accuracies Llama-3.1 obtained by each strategy. The baseline consistently outperform the various elimination strategies.

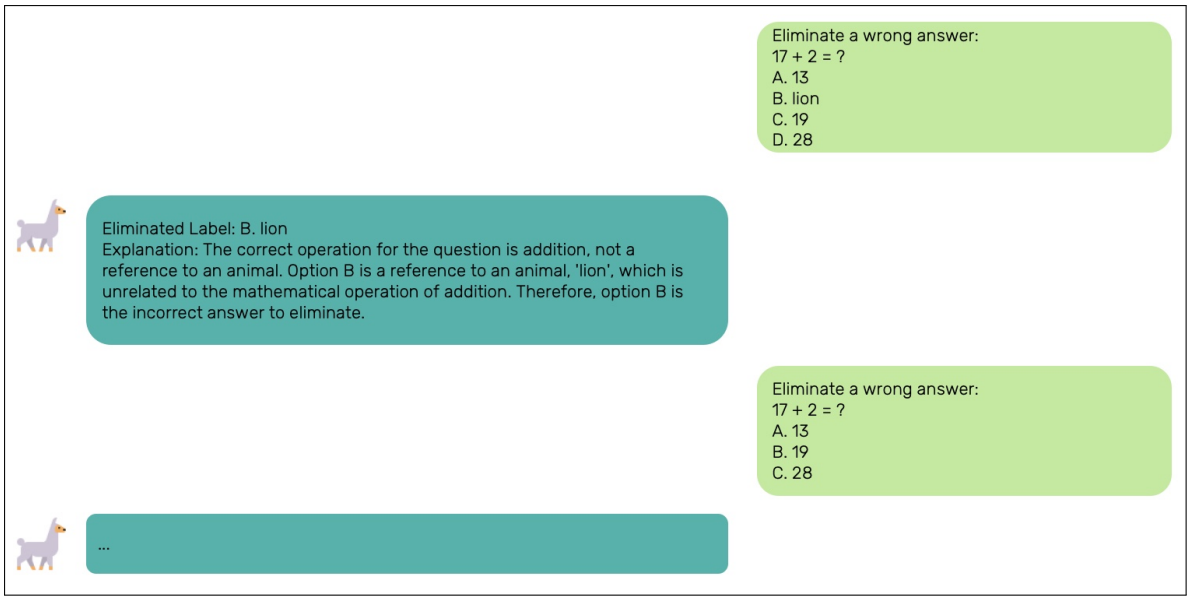

Figure 3: Zero-shot elimination illustration running on the EasyAdditive dataset. We can see how the model understands that option B of a lion in unrelated to the question, and eliminates it right away. Note that this process is iterative, and after each correct elimination, the model gets the question without the eliminated option (and without previous context).

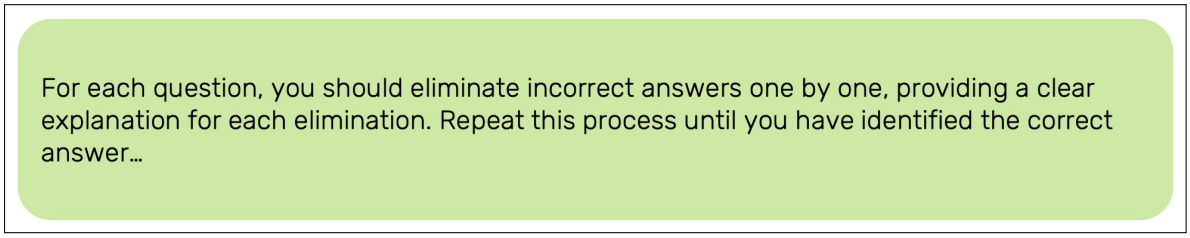

Figure 4: Part of the model's system-role prompt that guides the model to eliminate wrong options before reaching the right answer, in a single inference operation.

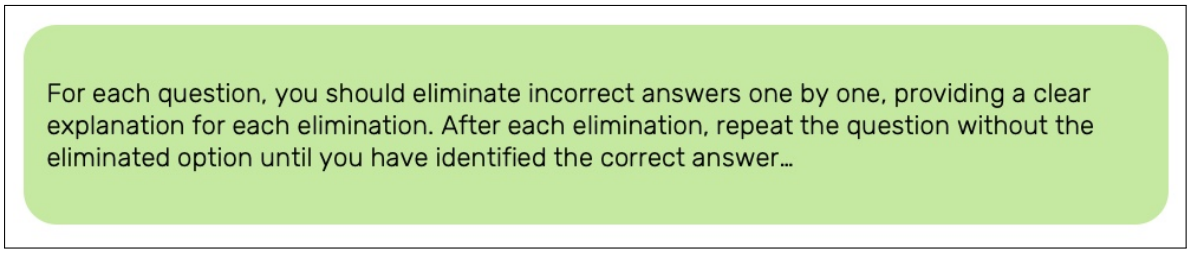

Figure 5: Second attempt of using chain-of-though to improve model's result via elimination strategy. This prompt, in comparison to Fig. 4, guides the model to repeat the question, after each elimination step. This experiment simulates Section 4.1, in a single inference operation, which provides previous context.

Dataset	A	B	C	D	std
ARC Easy	23.2	25.1	24.2	22.8	1.48
ARC Challenge	22.3	25.7	25.9	24.1	1.53
OpenBookQA	27.6	25.2	26.4	20.8	2.57
RACE	21.5	26.8	26.8	24.8	2.17
MMLU	24.9	24.8	25.2	25.1	0.15
EasyAdditive	25	25	25	25	0

Table 1: The percentage of each correct label per dataset. We can see an almost uniform label distribution, which reduces the label bias. Note that due to outliers in the datasets, the portion does not always sum up to 100.

Dataset	Zero-Shot	One-Shot
ARC Easy	84.8	**86.4**
ARC Challenge	**72.1**	70
OpenBookQA	63.5	**64.1**
RACE	68.5	**71.2**
MMLU	61.4	**62.1**

Table 2: Zero-shot vs. One-shot successions rate (indicating full knowledge) over questions that the baseline answered correctly (as in Fig. 1)

Strategy	Top-1	Top-2
Zero-Shot	94.8	99.8
One-Shot	90	95.2

Table 3: Percentage of questions in the EasyAdditive dataset where Llama-3.1 successfully eliminated the least likely option using zero-shot and one-shot strategies. The Top-t column indicates the percentage of questions where the model eliminated the option till step t inclusive.

Metric	OpenBookQA	RACE
Accuracy	**87.8**	85
Elimination Success Rate	59.5	**62.5**
Weighted Accuracy	73.8	**75.8**
Step-1 Accuracy	**89**	82

Table 4: Different metrics for evaluating dataset difficulty level (performed by Llama-3.1): classic baseline accuracy, success rate of our proposed elimination strategy which reflects full knowledge (using zero-shot prompting), weighted accuracy in which each step of elimination is counted as partial success, which defined as: $\frac{1}{|D|}\Sigma_{q\in D}\frac{\text{successful elimination steps in } q}{\text{total elimination steps}}$, and the percentage of successful first elimination. From each metric, a different inference could be made regarding the difficulty of a given dataset, which indicate the potential for future work in this field.

References

[1] Anat Ben-Simon, David Budescu, and Baruch Nevo. A comparative study of measures of partial knowledge in multiple-choice tests. 1997.

[2] David A. Bradbard, Darrell F. Parker, and Gary L. Stone. An alternate multiple-choice scoring procedure in a macroeconomics course. 2004.

[3] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. 2023.

[4] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. 2024.

[5] Clyde H. Coombs, J. E. Milholland, and F. B. Womer. The assesment of partial knowledge. 1956.

[6] Soham Parikh, Ananya B. Sai, Preksha Nema, and Mitesh M. Khapra. Eliminet: A model for eliminating options for reading comprehension with multiple choice questions. 2019.

[7] Chenkai Ma and Xinya Du. Poe: Process of elimination for multiple choice reasoning. 2023.

[8] Ariel Goldstein and Gabriel Stanovsky. Do zombies understand? a choose-your-own-adventure exploration of machine cognition. 2024.

[9] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. 2022.

[10] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. 2024.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. 2023.

[12] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation. 2024.

[13] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. 2018.

[14] Yuval Reif and Roy Schwartz. Beyond performance: Quantifying and mitigating label bias in llms. 2024.

[15] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. 2018.

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

[17] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. 2017.

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. 2020.

[19] 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. Language models are few-shot learners. 2020.

[20] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. 2020.

[21] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele (Mike) Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. 2024.

[22] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. 2018.