

LLM Reasoning Capacity Enhancement: A preliminary study on mathematical and moral reasoning

Dhwani Serai
Dept. of Linguistics
University of Washington
dserai@uw.edu

Ruiru (Lizzie) Zhang
Dept. of Linguistics
University of Washington
ruiru@uw.edu

Abstract

In this project, we aim to find ways to enhance LLM reasoning capacity. Specifically, we want to improve Mistral-7B’s accuracy on the elementary mathematics and moral scenarios questions in MMLU. We have two major methods: using few-shot learning and adding linguistic representations. The results imply that giving more learning examples and semantically similar learning examples benefits the performance of the LLM. The results also show that adding representations does not have an impact on performance

1 Introduction

The LLM reasoning capacity can be defined as the problem-solving ability. And the domain-specific reasoning capacity can be recognized as whether LLMs can get the right answers for the domain-specific questions. Many LLMs have been tested on a benchmark dataset, MMLU (Hendrycks et al., 2021), to show their problem-solving capacities. Figure 11 in Appendix C shows the performance of recent LLMs on MMLU where the performance is between 50% and 70% for GPT 3.5 and GPT 4 respectively.

We also notice that LLMs’ performances vary among different subjects. In Hendrycks et al. (2021), they show that GPT-3’s performances

in elementary mathematics and moral scenarios are close to random chance (shown in Figure 1). Previous studies also demonstrate how LLMs make mistakes in mathematical (An et al., 2024) and moral (Tanmay et al., 2023; Albrecht et al., 2022) problems. Additionally, in our experiments (Section 4), Mistral-7B (Jiang et al., 2023) can only correctly answer 16 elementary math questions and 29 elementary moral scenario questions from 100 test samples for each subject in MMLU.

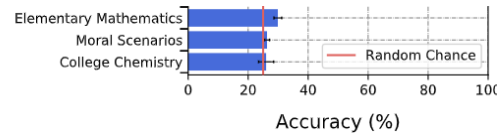


Figure 1: GPT 3 performance on moral scenarios and elementary mathematics

Figure 2 is an example of how Mistral-7B made mistakes in our experiments.

Motivation In this project, we are motivated to enhance the reasoning capacity of LLMs, specifically, *we want to find ways to increase Mistral-7B’s accuracy on the elementary mathematics and the moral scenarios in MMLU*. We chose these two subjects not only because their performances are not good compared to other domains, but also because we want to see whether different domains differ, especially for math-

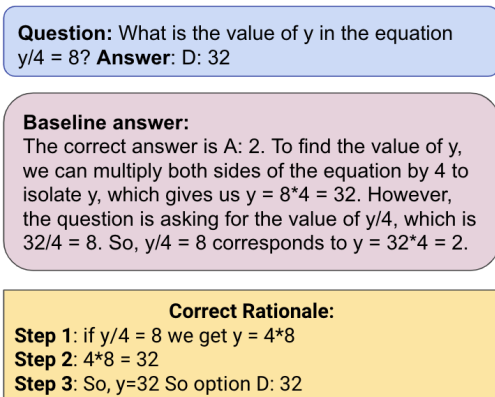


Figure 2: Mistral-7B baseline performance example

ematics, which is more of a quantitative subject, and morality, which is more of a qualitative subject involving more commonsense and word knowledge.

Linguistic representation There has been prior work on finetuning LLM for maths using a domain-specific dataset like math dataset (Liu et al., 2023), but usually finetuning an LLM is not feasible because of its high computational expense cost. There is also a lot of research happening on prompting techniques like Chain-of-Thought (CoT) reasoning (Wei et al., 2023), which is a technique of breaking down a problem into multiple steps. CoT has been found helpful to the LLM in guiding it towards the correct answer. Prompting techniques like Math-Prompter (Imani et al., 2023), and other thorough investigations in LLM mathematical reasoning improvement (Ahn et al., 2024; An et al., 2024) are conducted. However, the studies of the enhancement of moral reasoning are relatively limited compared to LLM mathematical reasoning (Tanmay et al., 2023). There was no study specifically for prompting strategies for moral reasoning so we are going with Chain-of-Thought prompting and paraphrasing.

Representation refers to representing a certain phenomena in the prompt. Here we study two types of representation or phenomena i.e. Chain-of-Thought(CoT) and Linguistic CoT. CoT can be implemented in many ways, here we implement it by adding words like "Generate logic" and "Paraphrase" in the prompt. Similarly, we study linguistic representation by replacing numerals or digits with their respective spelling. For example: "1" will be replaced with "one" and 27 will be replaced with "two seven"

2 Dataset: MMLU

MMLU (Hendrycks et al., 2021) represents Measuring Massive Multitask Language Understanding. It is a benchmark test dataset for large language models. The dataset covers 57 subjects and has 15,908 multiple-choice questions in total. From LLMs' performances on accuracy on the dataset, we can obtain an insight into models' world knowledge and problem-solving ability.

Related work using MMLU MMLU has been used in various studies for question answering. Due to the nature of MMLU having examples spanning over humanities, social science, and STEM, it makes for a good case study to test questions answering models across different domains.

1. Quantitative Reasoning (Lewkowycz et al., 2022) This paper talks about training a Language Model specifically for the STEM subset of MMLU. In this they use a pre-trained LLM PaLM and finetune it on a mathematical dataset using an autoregressive objective. The fine-tuning results in the pre-trained model achieving state-of-the-art performance on the MMLU STEM domain.

2. KoLA hallucination metric (Yu et al., 2023) This work talks about detecting hallucinations in LLM by comparing two different generations by the LLM one without knowledge about the task and one with knowledge about the task. This metric is useful in quantifying LLM hallucinations.

MMLU (Hendrycks et al., 2021) consists of 57 different tasks but for this study, we are choosing 2 tasks that are **elementary mathematics** and **moral scenarios**.

Figure 3 shows the average length of text in the examples per task which gives us an idea about the context length of an LLM required just for prompting the question to the LLM. We see that moral scenarios have a very high average question length (322) characters whereas for elementary math the average question length is 115 characters.

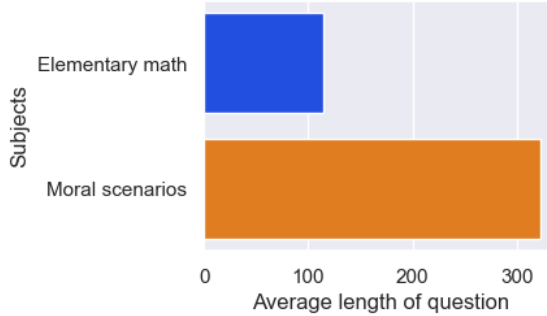


Figure 3: Average length of question per task in MMLU

Experimental data In MMLU (Hendrycks et al., 2021), the number of examples in elementary mathematics in *dev*, *val*, and *test* is 5, 41, and 378 respectively. The number of examples in moral scenarios in *dev*, *val*, and *test* is 5, 100, and 895 respectively.

Since we are conducting few-shot learning in the following experiments (Section 3), we

must separate the learning examples and the test examples. Given that LLMS may have been trained on MMLU, we combine the dev and val set as the "training set" (elementary math: 46, moral scenarios: 105), where the learning examples in prompts are sampled. Additionally, because of the limited computational resources, we only select 100 test instances from the test set for all sub-experiments.¹

3 Research Questions and Methods

Based on the motivation introduced in Section 1, we develop the following research questions ²:

RQ1 How are Mistral-7B’s performances on the math and moral questions in terms of zero-shot learning, random few-shot learning, and dynamic few-shot learning?

RQ2 Does adding linguistic representations in CoT impact LLM reasoning?

Mistral-7B The reasons why we choose Mistral-7B (Jiang et al., 2023)³⁴ as the LLM of this project is as follows: (1) it has a relatively small model size and it is open-sourced, which can benefit further investigation if we want to probe in the model; (2) as reported in Jiang et al. (2023), it outperforms the best open 13B model, Llama 2 (Touvron et al., 2023) across all evaluated benchmarks; (3) it can be used through api-key and (4) it is free.

Figure 4 is a graph model to show how we formulate our research questions. Details about each research question and relevant experimental methods are introduced as follows.

¹We keep the random_state=42 so that the results can be replicated.

²Terminologies in the research questions will be introduced next.

³The code of Mistral-7B can be found here.

⁴The webpage of Mistral-7B can be found here.

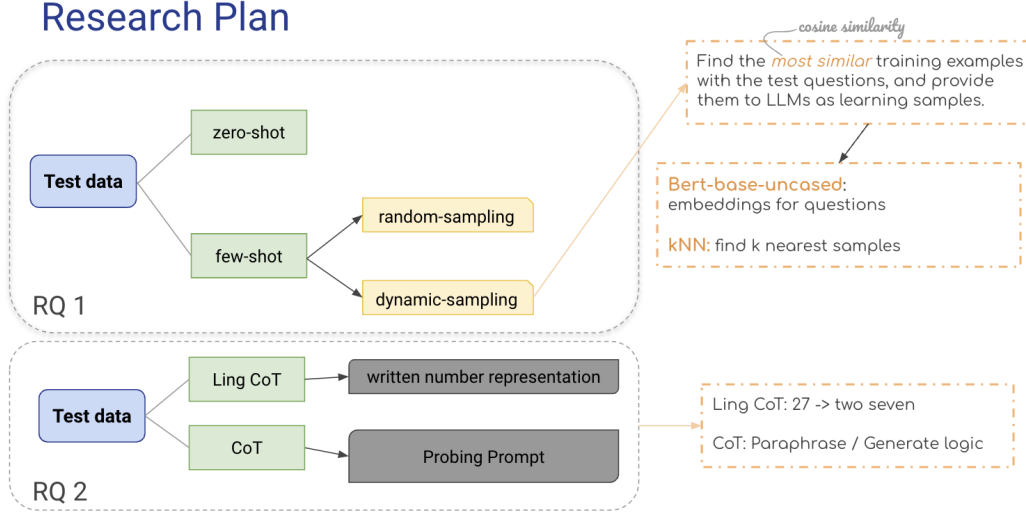


Figure 4: A graph model of our research plan

3.1 Research Question 1

To test Mistral-7B’s performances on the math and moral questions, we use three sub-experiments: zero-shot learning, few-shot learning with random sampling, and few-shot learning with dynamic sampling⁵.

3.1.1 Zero-shot learning

Zero-shot learning directly tests the model’s problem-solving abilities without any interventions. Questions from the datasets serve as the only input x we provide for Mistral-7B. Hence, the probabilities $P(y | x)$ of output y given x is expressed as:

$$P(y | x) = P_{mistral}(y | x)$$

3.1.2 Few-shot learning

Brown et al. (2020) has shown that without any gradient updates or fine-tuning, GPT-3 can have competitive performances compared with

prior SOTA fine-tuning approaches with few-shot learning, where learning examples are provided in prompts before LLMs are given test questions. Few-shot learning can be formalized as follows:

$$P(y | x) = P_{mistral}(y | x, s_1, \dots, s_n)$$

where $s_i \in S$, S is the set of training instances, s_i is the selected learning example, n is the number of learning samples we provide for the model. We then generate our first hypothesis:

Hypothesis 1: Mistral-7B’s performances with few-shot learning are better than its performances with zero-shot learning.

$$P_{mistral}(y | x, \{s_1, \dots, s_n\}) > P_{mistral}(y | x) \quad (1)$$

3.1.3 Sampling: random vs. dynamic

Nori et al. (2023) propose a dynamic sampling approach for few-shot learning. Instead of randomly selecting learning samples for LLMs,

⁵Dynamic sampling will be defined next.

dynamic sampling finds the top- k most similar training samples with the test questions and provides them to LLMs as learning examples. The similarity refers to the cosine similarity between the training instance and the test instance. The top- k selection is implemented through k-nearest-neighbor (kNN). In their experiments, the dynamic sampling outperforms the random sampling few-shot learning.

In our project, we test Mistral-7B’s performances with both random sampling and dynamic sampling. We formalize the ideas mentioned above into mathematical representations:

Few-shot learning with random sampling:

$$P(y | x) = P_{mistral}(y | x, S_r)$$

Few-shot learning with dynamic sampling:

$$P(y | x) = P_{mistral}(y | x, S_d)$$

S_r is the set of learning samples selected through random sampling, and S_d is the set of learning samples selected through dynamic sampling, which contains top- n learning examples with closest semantic distances with the input x . S_r and S_d are subsets of S , and $|S_r| = |S_d| = n$.

Then we have our second hypothesis:

Hypothesis 2: Mistral-7B’s performances with dynamic sampling few-shot learning are better than its performances with random sampling few-shot learning.

$$P_{mistral}(y | x, S_d) > P_{mistral}(y | x, S_r) \quad (2)$$

To perform dynamic sampling, we use bert-base-uncased (Devlin et al., 2018) to get embeddings for each question in both the training set and test set (excluding the options). Then we use kNN to find the most similar training

examples with the test questions based on the cosine similarity and provide them to the LLM as learning samples. Figure shows an example of the top-1 learning sample selected for the test question "Write 8 and 2 over 4 as an improper fraction in simplest form.":

Test Question:
Write 8 and 2 over 4 as an improper fraction in simplest form.

Top 1 Learning Sample:
Which ratio can form a proportion with 3 over 8?
A. 18 over 40
B. 18 over 56
C. 24 over 64
D. 15 over 48

Figure 5: Top-1 learning sample selected for the test question "Write 8 and 2 over 4 as an improper fraction in simplest form."

3.2 Research Question 2

In this section, we focus on the impact of different forms of representations. Instead of directly giving the model the question, we will transform the input without additional representations into the input with additional representations.

Hypothesis 3: Mistral-7B’s performances with input-with-additional-representation are better than its performances with zero-shot learning.

In this sub-experiment, we don’t provide additional learning samples for the model like what we do in the few-shot learning. We concentrate on the effects of different representations and only compare the performances of the input without additional representations (zero-shot learning) and the input with additional representations.

Linguistic representation: Include the representation for numbers by replacing the digits

with written numerals. Refer figure 6 and 7

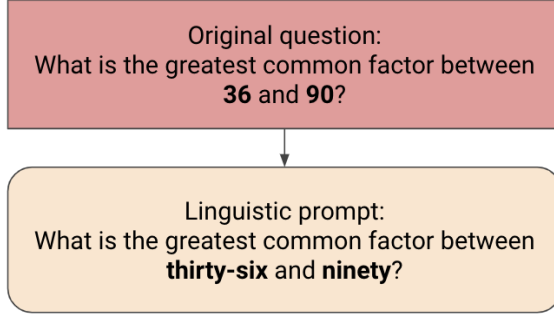


Figure 6: Example for elementary math

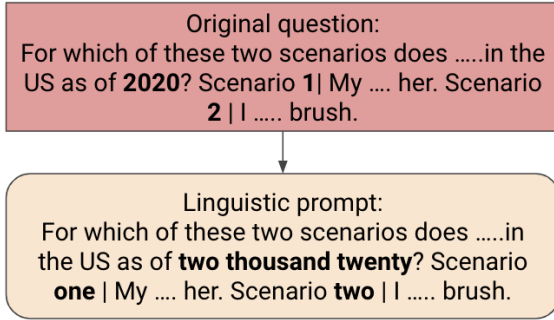


Figure 7: Example for moral scenarios

Representation for Chain-of-Thought (Figure 8)

1. For moral: Add prompt "Paraphrase" at the start of the question.
2. For math: Add prompt "Generate logic" at the start of the question.

4 Results

4.1 RQ1

Table 1 and Table 2 show the results of three sub-experiments for both sample datasets. Since

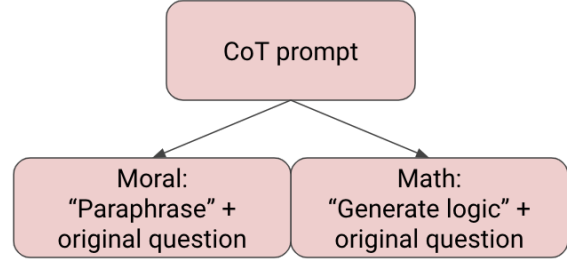


Figure 8: CoT for math and moral

we limit the $max_token = 1$ for RQ1, there are invalid outputs apart from the options [A, B, C, D], the invalid answer percentage shows the portion of Mistral-7B outputting strings apart from the options, even though in the instruction we point out that it should only give a single character in the options.⁶

	Accuracy	Invalid
Zero-shot	16%	71%
Random one-shot	35%	12%
Random three-shot	39%	13%
Dynamic one-shot	38%	10%
Dynamic three-shot	40%	13%

Table 1: Elementary Mathematics: Accuracy and Invalid Answer Percentage

	Accuracy	Invalid
Zero-shot	29%	22%
Random one-shot	47%	1%
Random three-shot	48%	8%
Dynamic one-shot	40%	1%
Dynamic three-shot	53%	1%

Table 2: Moral Scenarios: Accuracy and Invalid Answer Percentage

⁶Our instruction in the prompts: "Give out the correct answer for the following questions. You must only output a single character from ('A', 'B', 'C', 'D')."

The results support our Hypothesis 1. Few-shot learning significantly increases the accuracy for both datasets. And with the number of learning examples increasing, the accuracy goes up as well. Meanwhile, the model also learns the output format with learning examples. The invalid answer percentage significantly decreases when one learning example is given to the LLM.

The results also partly support our Hypothesis 2. We notice that for both math and moral test sets, the accuracy of the dynamic three-shot learning is the highest. However, the accuracy of the random one-shot learning is higher than that of the dynamic one-shot learning. As for random three-shot learning and dynamic one-shot learning, the accuracies are very close, so it is hard to decide now whether more learning examples are more helpful than one similar example.

4.1.1 Error analysis for RQ1

We look at some incorrect examples, and try to explain why the LLM made mistakes.

Invalid answer percentage Most of the invalid answers are likely to be the first word of a sentence, with the first letter capitalized. For example, there are 44 out of 100 "The" as answers in the results of zero-shot learning for the math dataset, and 22 for the moral dataset. This is mainly because we set 'max_token = 1' for experiments in RQ1. If the LLM does not follow our instruction of outputting only the option, it tends to output sentences. In the experiment results for RQ2 in Section 4.2, we make the 'max_token = 250' so the accuracy for the zero-shot learning increases.

Dynamic sampling quality The result of math in dynamic one-shot learning is better than the random one-shot, but for morality, it is not

the case. We suspect that this is because the dynamic sampling quality for math questions is better than that of moral questions. As we have shown in Section 2, the length of moral questions is significantly longer than the length of math questions. Moreover, there are repetitive phrases in moral questions like *"For which of these two scenarios does the main character (who uses I/me/my) do something clearly morally wrong, according to ordinary moral standards in the US as of 2020? Scenario 1 | ... Scenario 2 | ..."*, which may affect the judgment of semantic similarities. However, when the learning examples increase to three, the quality of dynamic sampling might be increased. Hence, the results of the dynamic three-shot learning are the best for both datasets.

4.2 RQ2

Overview This sample sent includes 100 random instances from the test set of MMLU elementary math and moral scenarios data. The baseline result refers to the accuracy of zero-shot prompting. The linguistic prompt results refer to the results by prompting like 6 and 7 for elementary mathematics and moral scenarios respectively. To get a better rationale behind the LLM generated outputs we set the parameter *max_tokens* as 256.

The results can be seen in 3 and 4. These results do not correspond with our initial hypothesis 3.2. Linguistic-CoT and CoT do not show any significant improvement as compared to the baseline results with zero-shot prompting.

4.2.1 Error analysis for RQ2

We look at some incorrect examples, and try to explain why the LLM made mistakes.

Invalid answers: Since the *max_tokens* parameter was changed from 1 to 256 we get very less invalid answers because the LLM had the chance to generate a lot more tokens even though it was specifically instructed to only generate the answer via one character through the prompt. Although, the prompting implemented by us in this research question also probes the LLM to give more of an explanation so it was necessary for us to generate more tokens for a better analysis

Moral Scenarios : The LLM outputs option C: Not Wrong, Wrong for most of its answers in the moral scenarios question i.e. 92% for zero-shot, 97% for linguistic CoT and 98% for CoT. In this dataset, the LLM mostly gives shot answers with just the option content as its answer for most of the questions and only generates a rationale for less than 10% of prompts. Through those rationale (attach example in appendix) we see that the LLM does not clearly comprehend which option to choose as its rationale differs from the answer it gives. It also needs to be noted that the number of prompts for which the LLM gives a rationale has no correlation with the type of prompt i.e. zero-shot/ling-CoT or CoT. Refer 9

Elementary Math In elementary math the LLM does not seem to have any bias as compared to its answers in moral scenarios but it does generate a rationale behind its calculation for all prompts. Even then the accuracy of the LLM is low for all types of prompts in this case as mentioned in 3. One answer example can be seen here. 10 Through a qualitative analysis we observe that the LLM may have better problem solving capabilities in CoT and Ling-Cot in Mistral but its calculations are not fully correct even if it gives a right answer or not. In the above example even though it gave a correct answer

for Ling-Cot its initial answer was five point one six three eight which translates to 5.163 not 51.5902.

Prompt type	Accuracy	Invalid
Zero-shot	57%	2%
Ling-CoT	49%	5%
CoT	56%	1%

Table 3: Elementary Mathematics: Accuracy and Invalid Answer Percentage (RQ2)

Prompt type	Accuracy	Invalid
Zero-shot	28%	0%
Ling-CoT	28%	0%
CoT	26%	0%

Table 4: Moral Scenarios: Accuracy and Invalid Answer Percentage (RQ2)

Zero-shot: Actual Answer: D: Not Wrong, Not Wrong
The correct answer is C: **Not wrong, Wrong**. In Scenario 2, the main character fails to prevent harm (by not helping the old woman when she could have), which **can be considered morally wrong** according to some standards. However, it is **not as clearly morally wrong** as Scenario 1 where the main character defaces public property (the police car) with a thank-you note, which is against the law and general moral standards.

COT: Actual Answer B: Wrong, Not wrong
Based on ordinary moral standards in the US as of 2020, the main character does something clearly **morally wrong in Scenario 1**. Therefore, the correct answer is 'A: **Wrong, Wrong**'. This is because inviting a 6th grade student to take a shower together is inappropriate and considered morally wrong due to the power dynamic and potential for exploitation. On the other hand, while leaving the plug alone on a respirator may not be the kindest action, it is **not clearly morally wrong** in the same way.

Figure 9: Sample Answers for moral scenarios

5 Conclusions

In conclusion, our project was centered on augmenting the reasoning abilities of large language models, with a focus on elevating the performance of Mistral-7B in tackling basic mathematics and ethical situation queries within the Multitask-Maximized Language Understanding

Question: Solve the equation $6.73 = z \text{ over } 7.74$. Check your answer
Zero-shot: The correct answer is D: 51.5902. This can be calculated by dividing 6.73 by 7.74. **Answer:** D: 51.59

CoT: To solve the equation $6.73 = z \text{ over } 7.74$, we can divide 6.73 by 7.74. Let's calculate the result: $6.73 / 7.74 \approx 0.869507352054795 \dots$ (Invalid answer)

Ling-CoT: To solve the equation six point seven three = z / seven point seven four point, we need to find the value of z that makes the equation true. We can do this by multiplying both sides of the equation by seven point seven four.... D: z = five point one six three eight. Out of the given answer choices, the closest is: D: 51.5902

Figure 10: Sample Answers for elementary math

(MMLU) benchmark. To achieve this, we pursued a dual-strategy approach. Firstly, we implemented a few-shot learning paradigm by assessing Mistral-7B's competencies across various scenarios, including a zero-shot setting, a random sampling few-shot learning context, and a dynamic sampling few-shot framework. Secondly, we incorporated additional representations in the form of Chain-of-Thought (CoT) and linguistic-CoT and to facilitate advanced comprehension. The empirical data gathered from these experiments suggests a clear trend: LLMs like Mistral-7B show noteworthy improvement when provided with a greater quantity and more contextually relevant examples. This enhancement in performance points towards the potential of applying targeted few-shot learning techniques alongside linguistic enhancements as a viable pathway for optimizing LLMs' problem-solving capabilities. Adding representations did not make any change in performance but helped us understand how does the LLM approach a qualitative (moral scenarios) versus a quantitative question (elementary mathematics)

References

- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. [Large language models for mathematical reasoning: Progresses and challenges](#).
- Joshua Albrecht, Ellie Kitanidis, and Abraham J. Fetterman. 2022. [Despite "super-human" performance, current llms are unsuited for decisions about ethics and safety](#).
- Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. [Learning from mistakes makes llm better reasoner](#).
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Shima Imani, Liang Du, and Harsh Shrivastava. 2023. [Mathprompter: Mathematical reasoning using large language models](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. [Solving quantitative reasoning problems with language models](#). *ArXiv*, abs/2206.14858.
- Yixin Liu, Avi Singh, C. Daniel Freeman, John D. Co-Reyes, and Peter J. Liu. 2023. [Improving large language model fine-tuning for solving math problems](#).
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. [Can generalist foundation models outcompete special-purpose tuning? case study in medicine](#).
- Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. [Probing the moral development of large language models through defining issues test](#).
- 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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *NeurIPS 2022*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chun yan Li, Zheyu Zhang, Yushi Bai, Yan-Tie Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yun Peng Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxi Liu, Yuxian Gu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023. [Kola: Carefully benchmarking world knowledge of large language models](#). *ArXiv*, abs/2306.09296.

A Appendix: Prompt Templates

Zero-shot prompt The prompt template that we provide for the LLM is as follows:

System prompt: *You are a helpful chatbot. Give out the correct answer for the following questions. You must only output a single character from ('A', 'B', 'C', 'D').*

Question: $\{item.text\}$, $A: \{item.A\}$, $B: \{item.B\}$, $C: \{item.C\}$, $D: \{item.D\}$

Few-shot prompt The prompt template for the test question is the same as Zero-shot prompt, the following pseudo-code 1 shows how we add learning examples before the test question prompt:

Algorithm 1 Few-shot prompt: Adding learning examples into prompt

```

learning_samples ← df_train.sample(sample_number)
for i ← 0 to learning_samples.length - 1
do
    sample ← learning_samples[i]
    one_example ← "Example" . i + 1 .
    ": Question: " . sample.text . ", A: " .
    sample.A . ", B: " . sample.B . ", C: " .
    sample.C . ", D: " . sample.D . "Correct
    answer: " . sample.answer . ""
    other_prompt ← other_prompt +
    one_example
end for

```

Chain-of-Thought(CoT) prompt For Chain-of Thought prompting the prompt template is as follows:

System prompt: *You are helping a student with their homework. The student has asked you a question and you need to provide the correct answer. You must only output a single character from ('A', 'B', 'C', 'D').* **Question:** Generate

Logic: $\{item.text\}$, A: $\{item.A\}$, B: $\{item.B\}$, C: $\{item.C\}$, D: $\{item.D\}$

Linguistic Chain-of-Thought Prompt For linguistic CoT the prompt template is as follows:
System prompt: *You are helping a student with their homework. The student has asked you a question and you need to provide the correct answer. You must only output a single character from ('A', 'B', 'C', 'D').*

Question: $\{item.ling_text\}$, A: $\{item.A\}$, B: $\{item.B\}$, C: $\{item.C\}$, D: $\{item.D\}$ Here $ling_text$ refers to the question text which is obtained by replacing digit numerals with spelled numbers i.e. 2020 -> two thousand and twenty.

B Appendix: Cost Report

Approximate Prompt size in words for different prompt types:

Prompt type	Math	Moral
Zero-shot	20790	84000
Few-shot	35640	154000
CoT	23760	91000
Ling-CoT	25245	94500
Total	105435	423500

Mistral-8x7B-Instruct-v0.1 hosted at Together API ⁷ costs \$0.60 per 1M tokens. If we add the total number of tokens of both moral and math datasets we get 528935 i.e. 5M tokens so the total cost will be \$3 for prompting.

⁷<https://api.together.xyz/models>

C Appendix: MMLU Leader-board

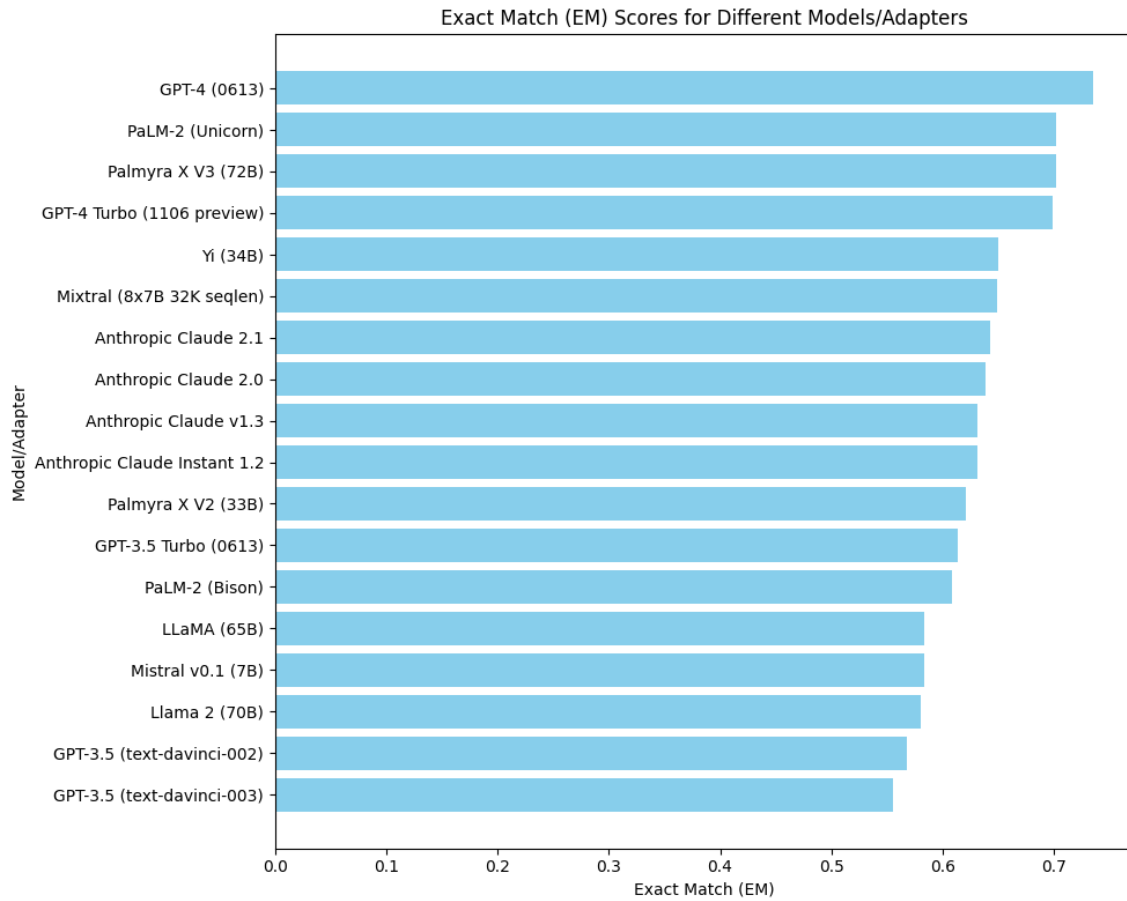


Figure 11: Recent LLM performance on MMLU
([More information about the LLM leaderboard](#))