


Evaluating Reasoning Capabilities of LLMs on Arithmetic Tasks

Ritwik Jain (jain504@purdue.edu)
Anxhelo Xhebraj (axhebraj@purdue.edu)

Introduction

[\[link\]](#)



	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge <i>MMLU</i>	86.8% 5-shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning <i>GPQA, Diamond</i>	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math <i>GSM8K</i>	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32

- LLMs achieve impressively good performance on many reasoning tasks
- However, often the reported performance does not correlate with actual performance
- We challenge arithmetic reasoning capabilities of LLMs (GPT-3.5 and ~~LLaMA~~¹ Claude 3 Sonnet) with ***in and out of distribution perturbations of questions in tasks they are trained on*** (GSM8K and MultiArith²) and show that their **performance degrades quickly when evaluated on them**

Changes

¹LLaMa 7B model has only 50% performance on GSM8K [\[1\]](#)

²MultiArith is too simple and augmentation API calls cost \$\$

Our Approach

(1) Automated + human verification:
create variations of the question that can be misleading with a capable LLM (Opus).

Difficulty is unaltered but *seemingly* relevant spurious information is added to the question to challenge reasoning abilities

Human annotator supervises that the edit does not affect the answer to the question or lead to ambiguous interpretations



Capable LLM
(GPT 4, Claude Opus)
edits GSM8K training question based on 8-shot example edits
Question-Edited Question pairs
(answer not provided as prompt, more tokens -> too expensive \$\$)

Original Question [link] Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?
Answer [Chain of thought]... Randy has 85 trees in total on his farm.

Augmented Question [link] Randy has 60 mango trees on his farm. He also has 5 horses. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?
Answer [Chain of thought]... So, Randy has 90 trees in total on his farm. Then, we add the number of horses: Total trees + Horses = 85 + 5 = 90
So, Randy has 90 trees in total on his farm.



Smaller LLM
(e.g. GPT 3.5 Turbo or Claude Sonnet)
which correctly predicted the original question
fails on the edited question (both zero-shot)

(2) Automated: change the unit of some of the values (matched by regex) with a simplified conversion rate.

LLM fails in the conversion steps.
Struggles with hypothetical statements
although difficulty level is the same.

Augmented Question [link]: Bill is ordering a new truck. He has decided to purchase a two-ton truck with several added features: a king cab upgrade, a towing package, leather seats, running boards, and the upgraded exterior light package. The base price of the truck is 15,000 euros, and the other features are at extra cost. The king cab is an extra \$7,500, leather seats are one-third the cost of the king cab upgrade, running boards are 250 euros less than the leather seats, and the upgraded exterior light package is \$1500. What is the total cost of Bill's new truck, in dollars if 1 euro is two dollars?

Answer [Chain of thought]... So, the total cost of Bill's new truck is \$28,250.

No changes from proposal

Experiments and results

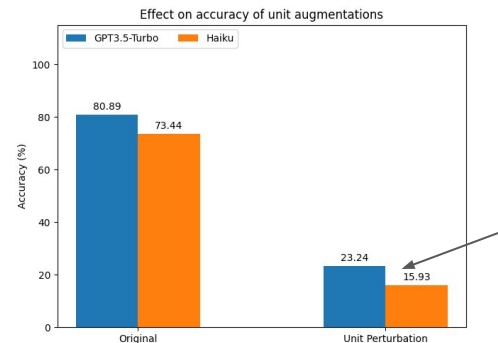
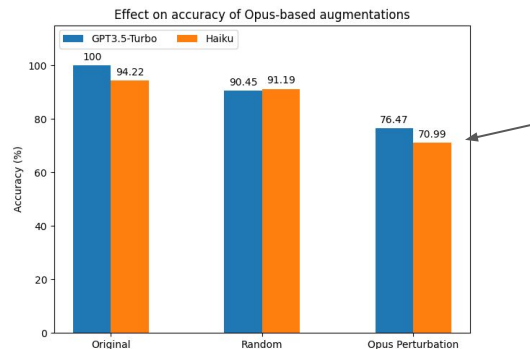
~30% of Opus augmentations changed the answer to the question and GPT3.5 got most of them correct

(0) (Baseline) adding a random sentence in the question affects accuracy by ~10%.

(1) (Opus Augmentation, in distribution) On a sample of 262 GSM8K training set instance accuracy drops from 100% to 76.47% for GPT3.5-Turbo, from 94.22% to 70.99% for Claude Haiku. **A degradation of > 25%!**

(2) (Unit Augmentation, out of distribution) On a sample of 314 GSM8K training questions, performance drops from 80.89% to 23.24% for GPT3.5 and from 73.44% to 15.93% for Claude Haiku. **A degradation of > 50%!**

# instances	Answer Changed	Changed Answer Correct
60	F	F
125	F	T
7	T	F
70	T	T



Analysis

Opus-based augmentations experiment demonstrates that it's easy for LLMs to lazily include seemingly relevant numbers to the computation instead of rigorous reasoning.

Unit augmentation demonstrates that LLMs struggle with hypothetical statements (“Assume 1km is 0.5 miles”). Concurrent research (Basmov et al. <http://arxiv.org/abs/2404.06283>) also investigated this!

Next steps: (1 month) improve Opus-based augmentations by providing the chain-of-thought of the answer to augment a question. (6 months) Scale to more tasks and models.

(5 years) improve tooling for more reliable structured information extraction from LLMs + better grounding of LLMs on new facts at inference time

Evaluating Reasoning Capabilities of LLMs on Arithmetic Tasks

CS577 Project

Ritwik Jain (0033847746, jain504@purdue.edu),
Anxhelo Xhebraj (0032184028, axhebraj@purdue.edu)

Due April 29th, 2024

The code accompanying the report is available at this link¹.

1 Introduction

Large language models based the Transformer architecture have shown surprisingly good reasoning capabilities, despite a purely generative training objective, defying many intuitions of deep learning researchers. Given the sheer complexity and size of these models, it is often difficult to predict or explain their outputs. Therefore, researchers have turned into analyzing the capabilities of these models by devising reasoning tasks such as GSM8k.

While these tasks correctly investigate reasoning abilities, recent issues in cross-contamination of datasets is making it harder to have a faithful evaluation [1]. More specifically, it has been shown that test data leaks into the training data used for large language models in several indirect ways [2]. Due to this, researchers have moved to test models under perturbations of such datasets.

2 Methods

2.1 Baseline

Our baseline consists of adding a random sentence in the middle of the question. This baseline validates that, if the answer to later augmentations of the question is incorrect, it is due to the perturbation being effective instead of an intrinsic uncertainty of the model.

¹https://www.dropbox.com/scl/fo/jccayfkkpg6cw68s9zi1a/ALB0UNL1Tj2aQBA8GtLrk_Y?rlkey=tbhy00a2adm792xlqvd3c324r&st=sngkg2q5&dl=0

2.2 Claude Opus Based Augmentation

MWPs follow specific themes such as colors, money, fruits, time, and distance. In this augmentation, we leverage this observation to introduce additional information in the questions that *do not* affect the answer but seem to be relevant for the solution. Figure 4 shows an example augmentation and ChatGPT 3.5’s answers. This augmentation evaluates the model’s capability to discard irrelevant information. Differently from the baseline approach, the sentences usually follow the theme of the question. We use a more capable (larger) LLM to create an augmentation for each question and validate manually that such augmentations are correct.

Augmentation We evaluate ChatGPT 3.5 Turbo on the first 2623 instances of the GSM8K dataset. We use a zero-shot setting and rely on the “implicit” chain-of-thought present in off the shelf model. ChatGPT 3.5 achieves an accuracy of 63.934% (1677 instances).

We sample $\approx 30\%$ (533) of the correct instances as candidates for augmentation through `claude-3-opus-20240229`. A key design decision was to make the “augmenting” model highlight the change in the sentence (e.g. `Lorem ipsum +| augmentation sentence |+ dolor`) to simplify human evaluation. Such markers are stripped from the string before evaluating the target model.

We used a sample of 20 questions from the training set to craft a prompt of 6 examples (Question, Augmented Question) pairs. Such pairs are demonstration of how the augmentations should not affect the answer. Two of the pairs provided also contained “bad augmentations” examples that should be avoided which were critical to reduce the number of augmentations that changed the answer.

Nevertheless, we employ human validation to check that the augmentations created do not affect the answer. Table 1 quantifies the number of augmentations that change the answer of the question.

Cost This experiment generated 50K output tokens and used 532K input tokens with cost \$75 / MTok and \$15 / MTok respectively for a total of \approx \$10.

2.3 Augmentation 2: Multiple units

MWPs dataset usually contain units such as pounds, dollars and miles. We build variations of these questions by changing *some* of the units in the question and providing the model the conversion rates between the units. Figures 5 and 6 show two augmentations from the GSM8k dataset.

The purpose of custom conversion units is two-fold: (1) we want to ensure that the calculations remain the same, simplifying evaluation and avoiding issues arising due to large numbers or operations involving numbers with several digits after the decimals; (2) it stresses reasoning capabilities of the model and its ability to extract highly-relevant information from the prompt.

Details Our aim is to have a mix of units in the question to demonstrate the models understanding of the interactions and dependencies that emerge out of it. Firstly, we calculated the statistic that the maximum number of occurrences of a single unit in a question is 7. Therefore, we used the following heuristic for the perturbation: if there is just 1 occurrence of the unit, we don't change it. If there are 2 occurrences we change 1 of them. If there are 3 occurrences we change 2 of them. For 4,5,6, or 7 occurrences in the same question, we change all except 2 of the occurrences to the new unit.

3 Evaluation

# instances	Answer Changed	Changed Answer Correct
60	F	F
125	F	T
7	T	F
70	T	T

Table 1: GPT 3.5 solves all the subset of 262 GSM8K questions correctly. However when augmenting the questions with Claude Opus and verifying them manually, 29.38% of the augmentations changed the answer to the question. GPT3.5 solved correctly all the original answers but only $\frac{125+70}{125+70+60} \approx 76.47\%$ of the augmented question — a performance degradation of 23.63%.

Figure 1 (left) shows the performance of GPT3.5-Turbo and Haiku on 262 instances from the training set of GSM8K. In the original questions both models perform really well. When tested on perturbations with a random sentence in the question performance degrades a bit but not significantly. However when tested on the Opus-augmented questions performance degrades considerably. Performance degrades similarly for 314 instances of the training set (Figure 1 (right)) when tested with the unit perturbations.

Common Failure Modes A common failure mode we have noticed during evaluation is that often times LLMs are capable of producing an intermediate step that is the solution to the problem but incorrectly assume that the spurious element *must* be included in the result. Figures 2 and 4 show an example of such case.

For the units perturbations common failure modes that we noticed were that when drawing inference from multiple units, the model is unable to capture nuanced differences, for eg: the overtime pay is only given for the extra hours and not for all the hours, but captures this nuance perfectly in the original dataset which points at over fitting. For example the chain of thought reasoning says “Since Tina works more than 24 yolo per shift, she is eligible for overtime pay for every shift. Since she works 5 days, we need to calculate her total overtime pay: 5 days * 27.0 dollars/hour * 10 hours = 1350.0”

Besides this it also tends to forget/ignore the relationships between the units which leads to

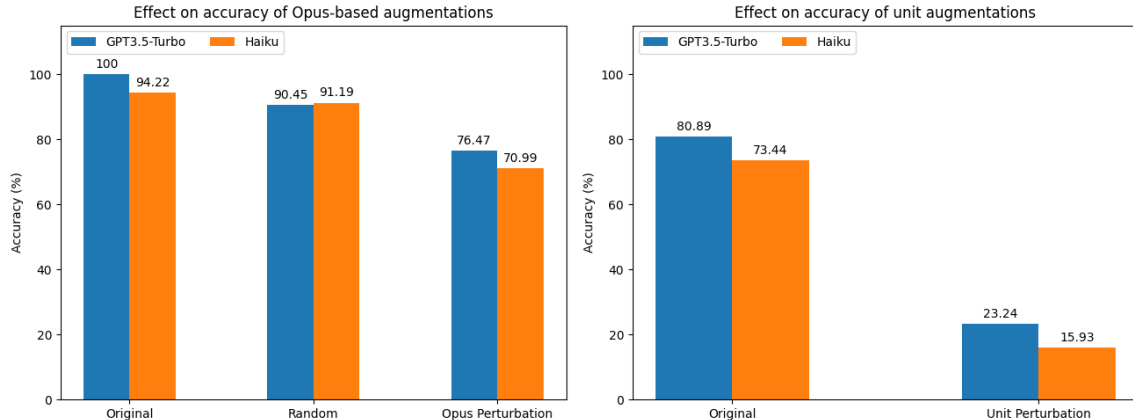


Figure 1: Accuracy degradation of GPT3.5-Turbo (larger model) and Haiku (smaller model) on a subset of GSM8K *training* instances. While both model achieve an impressive accuracy on such instances, performance degrades for varying augmentations. Both models seem relatively robust when tested on the original questions with a random sentence added in them (Random exhibits only a 10% degradation) but it degrades more quickly for the Opus-based augmentation (Approach 1) and the unit augmentation (Approach 2) for up to $\approx 25\%$ and $\approx 50\%$ respectively.

inconsistent reasoning and using the wrong numeric values for the calculation. For example in the chain of thought it initially says “First, let’s convert all the time durations to the same unit:- 135.0 jiffy is $135/3 = 45$ minute” but then goes on to do “ $2 * 135.0$ jiffy = 270 minutes.”

Lastly, an interesting observation from the chain of thought reasoning of the models was also that reasoning with money and time, or speed are tougher for the model.

Thus, our analysis highlighted that the model mostly captures shallow dependencies to arrive at the answer, instead of having a real understanding of the scenario to correctly reason over it.

Good Outputs Figure 3 shows a desirable behavior. The augmentation produced by Opus creates a contradiction in the question but GPT3.5-Turbo correctly identifies that instead of creating a plausible answer.

4 Analysis

4.1 Learning Experience

This was the first project ever in which we used LLM APIs to solve reasoning tasks.

A key challenge was providing the right prompt to obtain good augmentations to the questions but also extracting the answer to the questions. While most of the time LLMs

Augmented Question Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. She also crafted 15 unique hair pins during those two months. How many clips did Natalia sell altogether in April and May?

Answer Natalia sold 48 clips in April and half as many in May, which means she sold $48 / 2 = 24$ clips in May. Therefore, the total number of clips sold in April and May is $48 + 24 = 72$ clips. Additionally, Natalia crafted 15 unique hair pins. Therefore, Natalia sold 72 clips $+ 15$ hair pins $= 87$ items altogether in April and May.

Figure 2: ChatGPT-3.5-Turbo incorrectly assumes that crafting implies that the pins must have been sold

Augmented Question Four classmates were comparing their ages based on their birth month. They found out that Jolyn is 2 months older than Therese while Therese is 5 months older than Aivo. Then, Leon is 2 months older than Aivo. Coincidentally, Leon and Aivo share the same birthday. How much older in months is Jolyn than Leon?

Answer Let's denote each person's age as follows: Aivo = A months, Therese = $A + 5$ months, Jolyn = $A + 5 + 2 = A + 7$ months, Leon = $A + 2$ months. Since Aivo and Leon share the same birthday, we have: $A + 2 = A$, $2 = 0$.

This is a contradiction, as a person's age can't be 0. So there must be a mistake in the information provided. The problem seems to have an inconsistency or error. Kindly double-check the ages given for each person, as the current setup leads to a contradiction. Let me know if you have any other information or if there's a typo in the question.

Figure 3: ChatGPT-3.5-Turbo correctly identifies a contradicting augmentation produced by Claude Opus

would provide the answer according to the constraints requested, sometimes units would be present in the answer.

4.2 Potential next steps

If we had 1 more month we would try to find better ways to prompt the augmenting model. Currently we only provided the initial question to the augmenting model but it might be more beneficial to provide an example chain-of-thought answer as well as prompt. However doing so in a cost effective manner is nontrivial. If we had 6 more months we would scale our experimentation to more models and larger datasets / tasks. Investigate other methods to augment questions in more automated ways. If we had 5 more years we would create more tooling integrating knowledge bases with large language models. Additionally, we would research methods to ground models to new facts at inference time. We think that a key challenge of LLMs is that they are trained on mostly “true” facts but can struggle when presented with hypothetical statements.

References

- [1] Rachith Aiyappa et al. “Can we trust the evaluation on ChatGPT?” In: *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Association for Computational Linguistics, 2023. DOI: 10.18653/v1/2023.trustnlp-1.5. URL: <http://dx.doi.org/10.18653/v1/2023.trustnlp-1.5>.
- [2] Simone Balloccu et al. *Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs*. 2024. arXiv: 2402.03927 [cs.CL].

A Augmentation Examples

A.1 Misleading Sentences

Original Question [link] Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

Answer [Chain of thought]... Randy has 85 trees in total on his farm.

Augmented Question [link] Randy has 60 mango trees on his farm. He also has 5 horses. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

Answer [Chain of thought]... So, Randy has 90 trees in total on his farm. Then, we add the number of horses: Total trees + Horses = $85 + 5 = 90$
So, Randy has 90 trees in total on his farm.

Figure 4: Example misleading fact augmentation

A.2 Multiple Units

Original Question [link] Bob is 75 miles ahead of Tom, driving 55 miles per hour. Tom is driving 70 miles per hour. How long will it take Tom to catch up with Bob?

Answer To find out how long it will take Tom to catch up with Bob, we can use the concept of relative speed [Chain of thought]... So, it will take Tom 5 hours to catch up with Bob

Augmented Question [link] Assume 1 meter is 50 miles and 1 second is 10 hours. Bob is 75 miles ahead of Tom, driving 11 meters per second. Tom is driving 70 miles per hour. How long will it take Tom to catch up with Bob?

Answer [Chain of thought]... Relative speed = -549.98056
Since the relative speed is negative, it means Tom is slower than Bob, so he won't catch up. Bob will keep getting farther away. Therefore, Tom will never catch up with Bob given the provided conditions.

Figure 5: Example unit change augmentation

Original Question [link]: Bill is ordering a new truck. He has decided to purchase a two-ton truck with several added features: a king cab upgrade, a towing package, leather seats, running boards, and the upgraded exterior light package. The base price of the truck is \$30,000 , and the other features are at extra cost. The king cab is an extra \$7,500, leather seats are one-third the cost of the king cab upgrade, running boards are \$500 less than the leather seats, and the upgraded exterior light package is \$1500. What is the total cost of Bill's new truck, in dollars?

Answer [Chain of thought]... So, the total cost of Bill's new truck is \$43,500

Augmented Question [link]: Bill is ordering a new truck. He has decided to purchase a two-ton truck with several added features: a king cab upgrade, a towing package, leather seats, running boards, and the upgraded exterior light package. The base price of the truck is 15,000 euros , and the other features are at extra cost. The king cab is an extra \$7,500, leather seats are one-third the cost of the king cab upgrade, running boards are 250 euros less than the leather seats, and the upgraded exterior light package is \$1500. What is the total cost of Bill's new truck, in dollars if 1 euro is two dollars ?

Answer [Chain of thought]... So, the total cost of Bill's new truck is \$28,250 .

Figure 6: Example from GSM8k Dataset