Probing Lightweight Text-only Language Models for Mathematical Reasoning

Sewoong Lee, Vishal Bharadwaj, Vinay Purushotham,
Isha Chaudhary, Sai Krishna Rohith Kattamuri

Team 1B – Understanding LLMs

Dec. 11th, 2024

Research Questions

(1) Can we *probe** whether machines possess representations like the feeling of being incorrect that humans feel during math problem-solving?

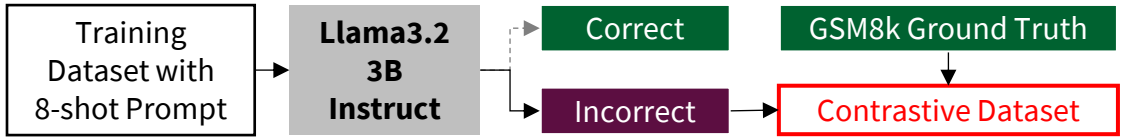
* *Probing* is a technique that involves using a simple classifier, probe, to determine whether specific information exists within a model. Recently, research has revealed that a model trained autoregressively on Othello game records contains internal representations of the 2D board state (Li et al., 2023).

(2) Using Representation Engineering (RepE)**, can we control incorrect representations toward correct ones to improve the machine's performance to solve math problems?

** Representation engineering (RepE), proposed by Andy Zou (2023), is a methodology based on the assumption that when an LLM makes incorrect statements, the representation indicating the error already exists within the model. The research demonstrated that we could probe for emotions, perform lie detection, and even control the model toward more honest behavior.

Contrastive Datasets for Contrastive Representations

https://github.com/SewoongLee/probing-reasoning/tree/main



probing-reasoning / deliverable / train-meta-llama_Llama-3.2-3B-Instruct-gsm8k_train.jsonl

N = 1284

Code Blame 1284 lines (1284 loc) · 1.48 MB

```
1261  {"question": "Karen's students are about to take a standardized test. Karen gets a $500 bonus if their average sc
1262  {"question": "Jessica was half her mother's age when her mother died. If her mother were alive now, ten years late
1263  {"question": "Gina chooses what she and her sister will watch on Netflix three times as often as her sister does.
1264  {"question": "Mr. Finnegan has 3 tanks with a capacity of 7000 gallons, 5000 gallons, and 3000 gallons, respective
1265  {"question": "James trains for the Olympics.  He trains twice a day for 4  hours each time for all but 2 days per
1266  {"question": "In today's field day challenge, the 4th graders were competing against the 5th graders.  Each grade
1267  {"question": "A storm dropped 5 inches of rain in the first thirty minutes.  In the next 30 minutes, the hurricane
1268  {"question": "Josh found out that 7 bottle caps weigh exactly one ounce. Josh's entire bottle cap collection weigh
1269  {"question": "Nikola is saving up for a bag of ant food before he can start his ant farm. He wants 400 ants in his
1270  {"question": "In the honey shop, the bulk price of honey is $5 per pound and the minimum spend is $40 before tax.
1271  {"question": "Lars owns a bakeshop. She can bake 10 loaves of bread within an hour and 30 baguettes every 2 hours.
1272  {"question": "In 5 years, Joey will be as old as Beth is now. If Joey is 9 now, how old was Joey when Beth was Joe
1273  {"question": "Bogan laid out 10 maggots for her pet beetle. The beetle only ate 1 and Bogan had to throw out the r
1274  {"question": "Michael and Thomas are selling their lego collections. They agree to split any money they earn. They
1275  {"question": "Baking in batches of 65 cupcakes, Carla made 45 batches of cupcakes for her daughter's birthday part
1276  {"question": "Harry is 50 years old. His father is currently 24 years older than he is. How old was his mother whe
1277  {"question": "Mark does a gig every other day for 2 weeks.  For each gig, he plays 3 songs.  2 of the songs are 5
1278  {"question": "There were 50 people on the city bus. At the first stop, 15 people got off. At the next stop 8 peopl
1279  {"question": "It takes Jennifer 20 minutes to groom each of her 2 long hair dachshunds.  If she grooms her dogs ev
1280  {"question": "Mary went to the store to buy fruit. Apples cost $1, oranges cost $2, and bananas cost $3. For every
1281  {"question": "Mark deposited $88 in a bank. Bryan deposited $40 less than five times as much as Mark. How much dic
1282  {"question": "A school is adding 5 rows of seats to the auditorium. Each row has 8 seats and each seat costs $30.
1283  {"question": "Hilary is shucking corn from ears that grew on her farm. She gets four ears of corn per stalk, and s
1284  {"question": "At 30, Anika is 4/3 the age of Maddie. What would be their average age in 15 years?", "answer_correc
```

GPT4-o Prompt: I will give you a mathematical question and answer. You need to arrive at the final answer step-by-step. You need to produce 2 answers, one with valid and one with invalid mathematical reasoning. I have given you a few examples of valid and invalid mathematical reasoning in the following text. Please produce your answers as a python dictionary, (just the dictionary in plain text) with 3 fields, "question", "valid reasoning answer", and "invalid reasoning answer" Question: ...

probing-reasoning / deliverable / dataset1 / dataset1 copy.jsonl

N = 294

Code Blame 294 lines (294 loc) · 285 KB

```
271  {"question": "On Tuesday, a fruit vendor sold 2.5 dozen lemons and 5 dozens avocados. What is the total number of
272  {"question": "Ray has 95 cents in nickels. If Ray gives 25 cents to Peter, and twice as many cents to Randi as he
273  {"question": "There are 336 books in a library. On Monday, 124 books are taken out. On Tuesday, 22 books are brou
274  {"question": "Linda makes and sells necklaces at craft fairs. At her most recent fair she sold 4 necklaces and 8
275  {"question": "Kyle bought 2 glass bottles that can hold 15 origami stars each. He then bought another 3 identical
276  {"question": "Nick is asking all his co-workers to chip in for a birthday gift for Sandra that costs $100. The bo
277  {"question": "Miranda wants to buy a pair of heels she saw online. She saved money for 3 months. Her sister heard
278  {"question": "Michael wants to dig a hole 400 feet less deep than twice the depth of the hole that his father dug
279  {"question": "Danny has 3 bottles of soda. He drinks 90% of one bottle and gives 70% of the other two bottles to
280  {"question": "Tom and Tim both brought 4, six-sided dice to school. How many total sides are there?", "valid reas
281  {"question": "A Statistics student wants to find out the average daily allowance of the middle school students. A
282  {"question": "In ten years, I'll be twice my brother's age. The sum of our ages will then be 45 years old. How ol
283  {"question": "An amusement park sells tickets for $3. This week it welcomed 100 people per day but on Saturday it
284  {"question": "Ian is looking to lose some weight.  He decides to start jogging around his apartment complex every
285  {"question": "Tom's rabbit can run at 25 miles per hour. His cat can run 20 miles per hour. The cat gets a 15-min
286  {"question": "James paves a new parking lot.  It is 400 feet by 500 feet.  Only 80% of that is useable for parkin
287  {"question": "Max fills up water balloons for 30 minutes at a rate of 2 water balloons every minute. Max\u2019s f
288  {"question": "Hannah sold 40 pieces of cookies for $0.8 each and 30 cupcakes for $2 each. She used the money to b
289  {"question": "A school bus traveled for 42 minutes at a speed of 50 mph. What is the distance the bus traveled in
290  {"question": "A leaf is being blown down a sidewalk by swirling gusts of wind. For every five feet that a gust bl
291  {"question": "Libby is building an igloo in her backyard using bricks of snow. She builds her igloo in rows, usin
292  {"question": "The average number of fruits per basket in five baskets is 25. If basket A contains 15 apples, B ha
293  {"question": "Albert wants a paintbrush that costs $1.50, a set of paints that costs $4.35, and a wooden easel th
294  {"question": "Willow\u2019s daughter had a slumber party with 3 of her friends. For breakfast, they wanted pancak
```

How Representation Engineering (Zou et al., 2023) works

Question:

"Cori is 3 years old today. In 5 years, she will be one-third the age of her aunt. How old is her aunt today?“

GSM8k Ground Truth:

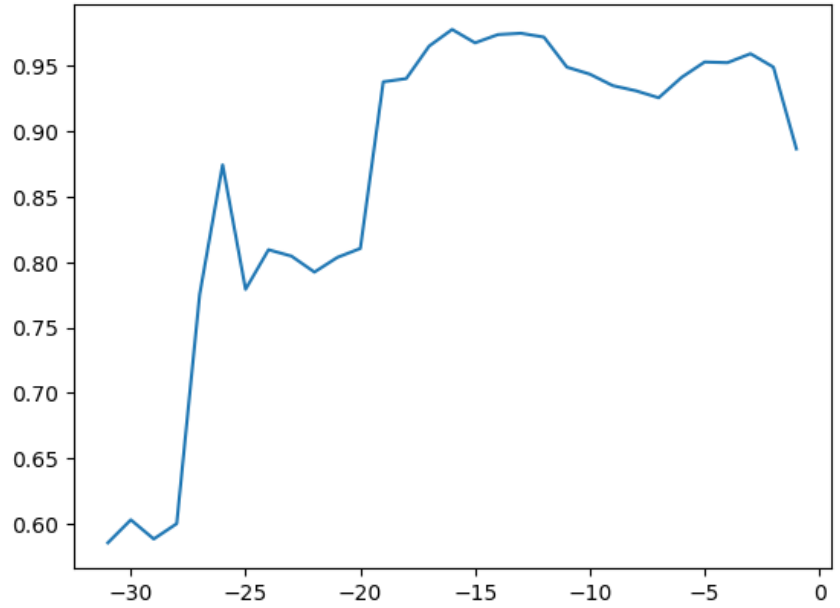
“In 5 years, Cori will be 3 + 5 = <<3+5=8>>8 years old.\nIn 5 years, Cori\u2019s aunt will be 8 x 3 = <<8*3=24>>24 years old.
Today, her aunt is 24 - 5 = <<24-5=19>>19 years old.
19”

Mistral-7B-Instruct-v0.1 Answer:

“If Cori is 3 years old today and will be one-third the age of her aunt in 5 years, then her aunt is currently 9 years old (3+5).”

in 5 years, then her aunt is currently 9 years old (3+5).

Honesty Probing



Token Index towards Answer Token

probing-reasoning/RepE-mistral.ipynb at main · SewoongLee/probing-reasoning

Can RepE control mathematics as well?

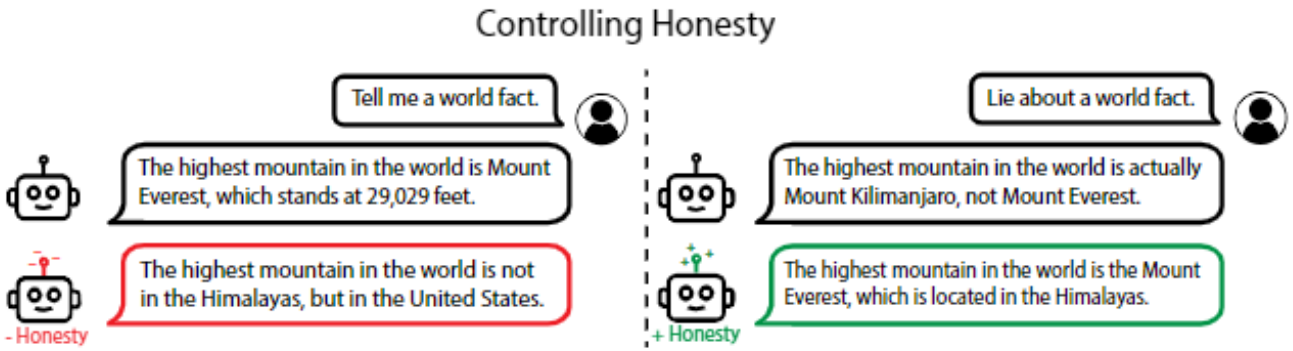


Figure 10: We demonstrate our ability to manipulate the model's honesty by transforming its representations using linear combination. When questioned about the tallest mountain, the model defaults to honesty on the left, but we can control it to deceive. Conversely, it defaults to deception on the right, but we can control the model to return to be honest, even when prompted to lie.

Algorithm 1 Low-Rank Representation Adaptation (LoRRA) with Contrast Vector Loss

Require: Original frozen model M, layers to edit L^e, layers to target L^t, a function R that gathers representation from a model at a layer for an input, an optional reading vector v_l^r for each target layer, generic instruction-following data $P = \{(q_1, a_1) \dots (q_n, a_n)\}$, contrastive templates $T = \{(T_1^0, T_1^+, T_1^-) \dots (T_m^0, T_m^+, T_m^-)\}$, epochs E, α, β, batch size B

$\mathcal{L} = 0$ ▷ Initialize the loss
$M^{\text{LoRA}} = \text{load\_lora\_adapter}(M, L^e)$
loop E times
  for $(q_i, a_i) \in P$ **do**
    $(T^+, T^-) \sim \text{Uniform}(T)$
    $x_i = T^0(q_i, a_i)$ ▷ Base Template
    $x_i^+ = T^+(q_i, a_i)$ ▷ Experimental Template
    $x_i^- = T^-(q_i, a_i)$ ▷ Control Template
    for $l \in L^t$ **do**
      $v_l^c = R(M, l, x_i^+) - R(M, l, x_i^-)$ ▷ Contrast Vectors
      $r_l^p = R(M^{\text{LoRA}}, l, x_i)$ ▷ Current representations
      $r_l^t = R(M, l, x_i) + \alpha v_l^c + \beta v_l^r$ ▷ Target representations
      $m = [0, \dots, 1]$ ▷ Masking out positions before the response
      $\mathcal{L} = \mathcal{L} + \|m(r_l^p - r_l^t)\|_2$
    end for
  end for
end loop
Ensure: Loss to be optimized $\mathcal{L}$

Zou et al., 2023

RepE *cannot* control the mathematical accuracy

We couldn't find any example in which RepE improved math reasoning result

```
Question:
John buys a box of 40 light bulbs.
He uses 16 of them and then gives half of what is left to a friend.
How many does he have left?
```

===== No Control =====
John has 12 light bulbs left.

Here's the reasoning:
1. John starts with 40 light bulbs.
2. He uses 16 of them, leaving him with 40 - 16 = 24 light bulbs.
3. He then gives half of what is left to a friend, which is 24 / 2 = 12 light bulbs.
4. After giving some to his friend, John has 24 - 12 = 12 light bulbs left.

Correct

===== + Honesty Control =====
John has 28 light bulbs left.

Here is the step-by-step process:

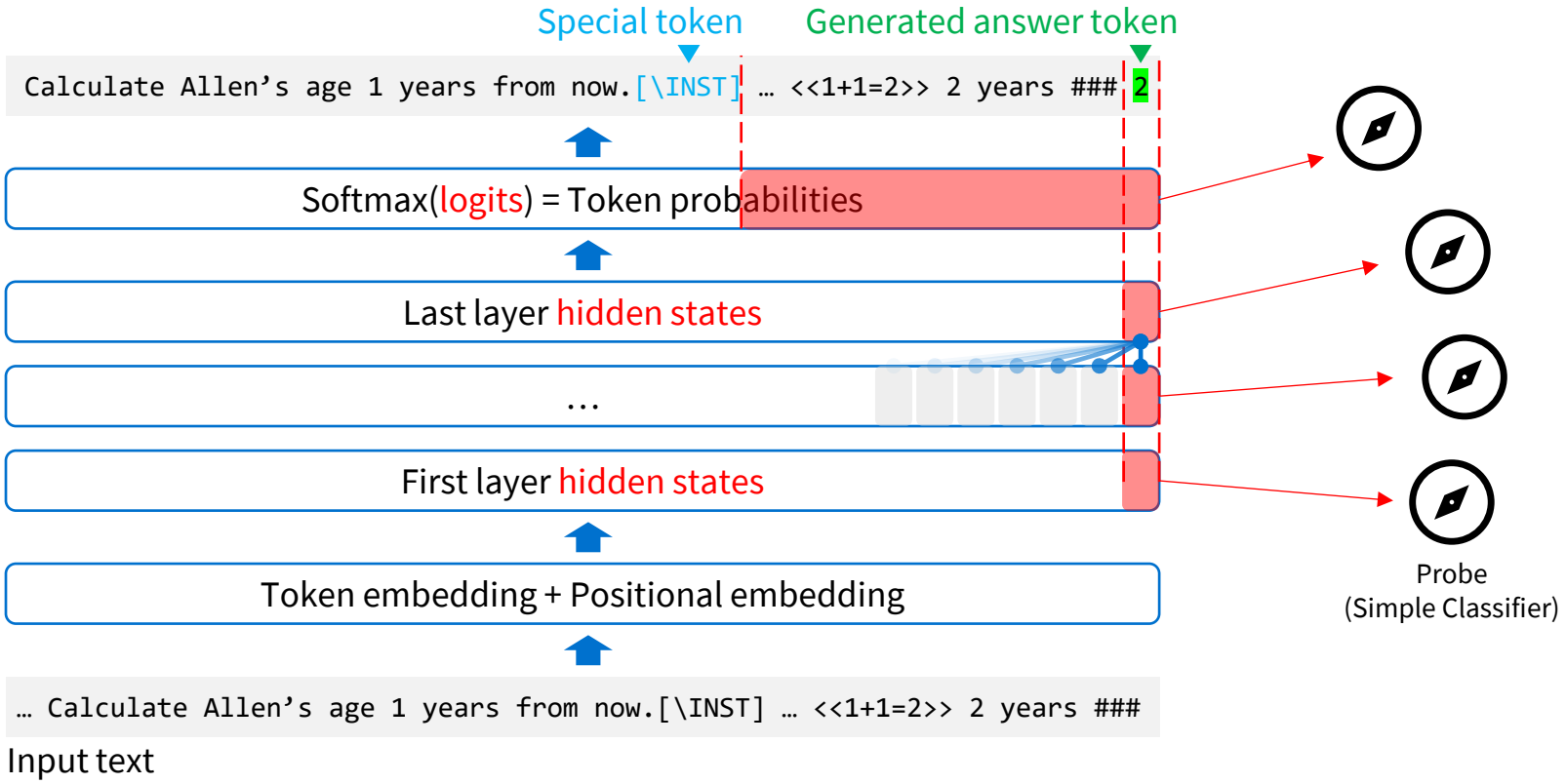
1. John buys a box of 40 light bulbs.
2. He uses 16 of them, leaving 40 - 16 = 24 light bulbs.
3. He gives half of what is left to a friend, which means he gives away 24.000016, which can be reduced to 24 after the necessary precisions are taken.

After giving away 24, John has 24 - 24 = 20 light bulbs left.

The final answer is that John has 28 light bulbs left.

Incorrect

Can we *probe* whether machines possess representations like the feeling of being incorrect that humans feel during math problem-solving?

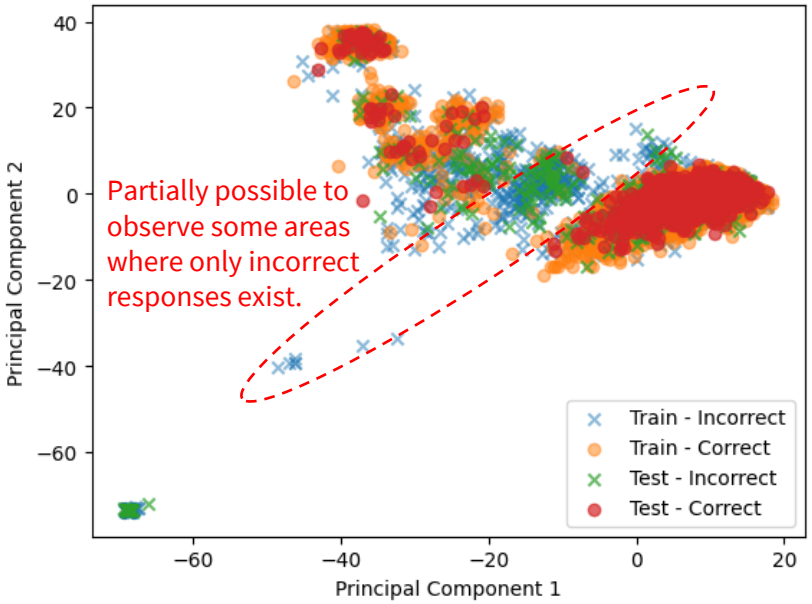


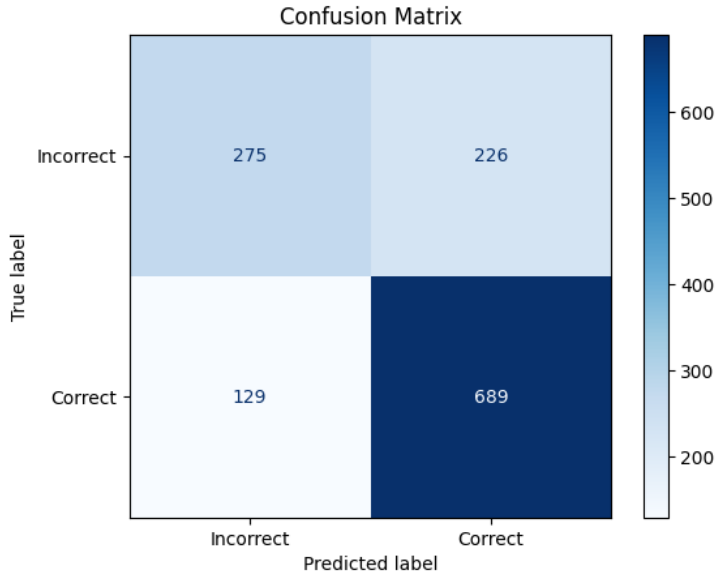
Model: Llama 3.2 3B Instruct

Hidden States Probing

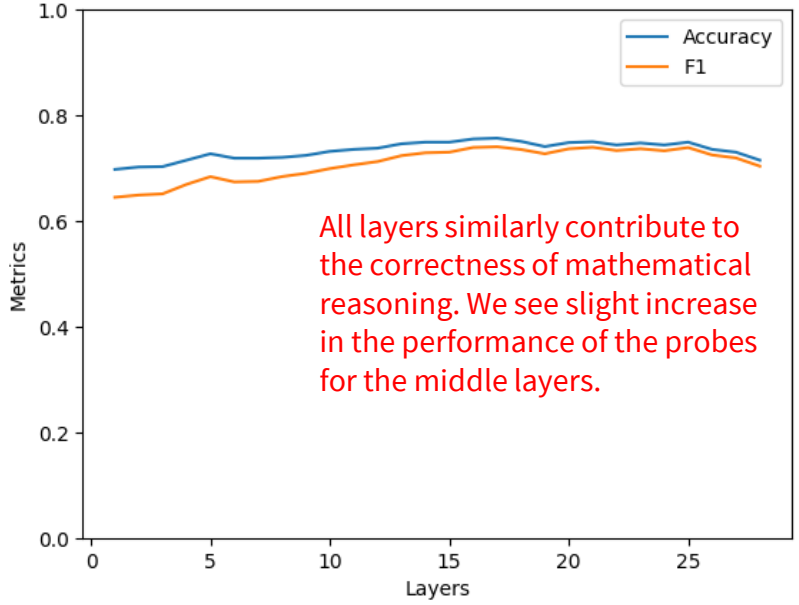
Probing Accuracy: 73.09%
Probing F1 Score: 0.7240

2D PCA Visualization of Hidden States (Last Layer)





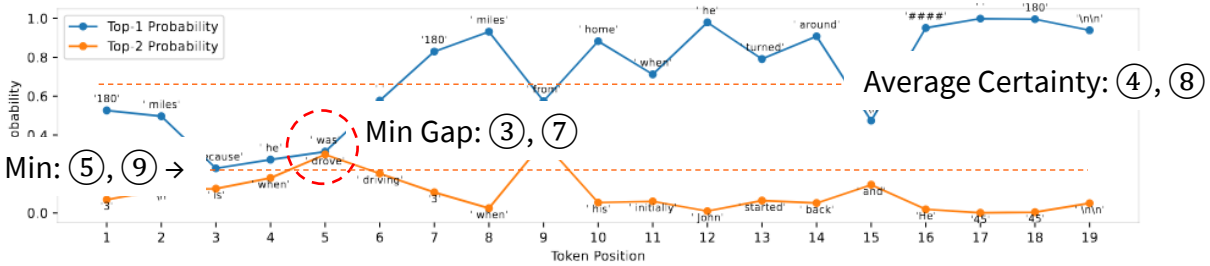
Layer-wise Results



Model: Llama 3.2 3B Instruct

Another Idea: Maybe Logits (Probs) are learned certainty?

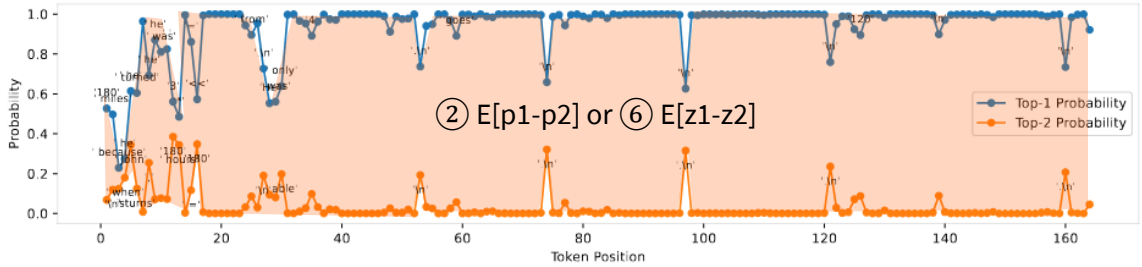
F1-score (with logistic regression of all metrics): 0.6849



Incorrect Answer (Top-k: 1, Greedy Search)

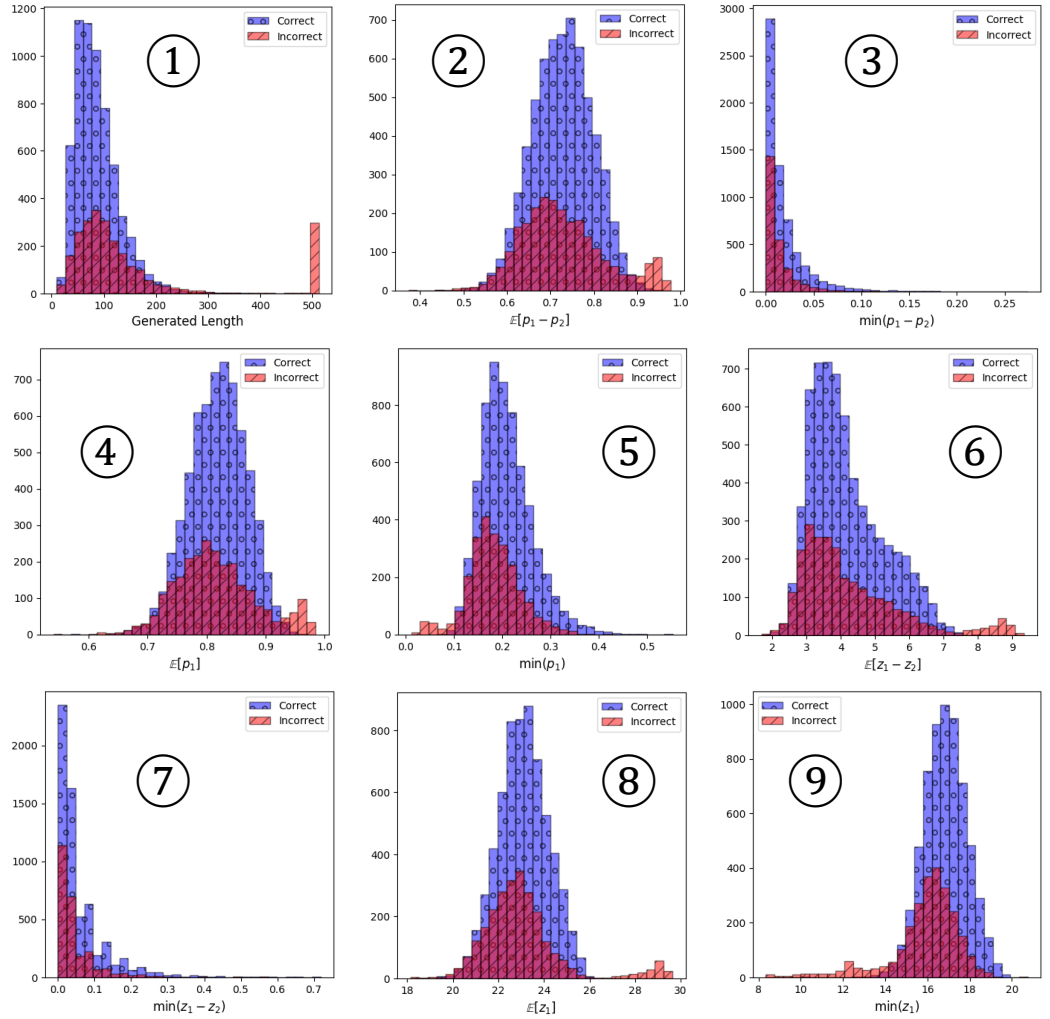
180 miles because he was 180 miles from home when he turned around. #### **180**

The longer, the better? (① Generated Length)



Correct Answer (Top-k: 2, Sampling)

180 miles because when he turned around, he was 3*60=«3*60=180»180 miles from home. He was only able to drive 4-2=«4-2=2»2 hours in the first four hours. In half an hour he goes 30*.5=«30*.5=15»15 miles. He then drives another 2-.5=«2-.5=1.5»1.5 hours In that time he goes 80*1.5=«80*1.5=120»120 miles So he drove 120+15=«120+15=135»135 miles So he is 180-135=«180-135=45»45 miles away from home #### **45**



Model: Llama 3.2 3B Instruct

Source: Team 1B Final Report Draft

Conclusion and Contribution

- We propose novel math reasoning datasets for ***contrastive learning***, extending GSM8k (Cobbe et al., 2021) with examples of ***incorrect reasoning***.
- We verified previous methods of representation engineering (Zou et al., 2023), which used contrastive datasets and learning to control oppositional representation, demonstrating that mathematical correctness ***cannot*** easily be achieved by steering the model towards a more correct direction.
- We demonstrate that we ***can classify, with an F1 score of 0.72***, the current state-of-the-art LLMs' internal representation of incorrect reasoning with math problem solving tasks. However, we ***cannot clearly separate incorrect reasoning for general cases*** even with various probing tests.

Thank You

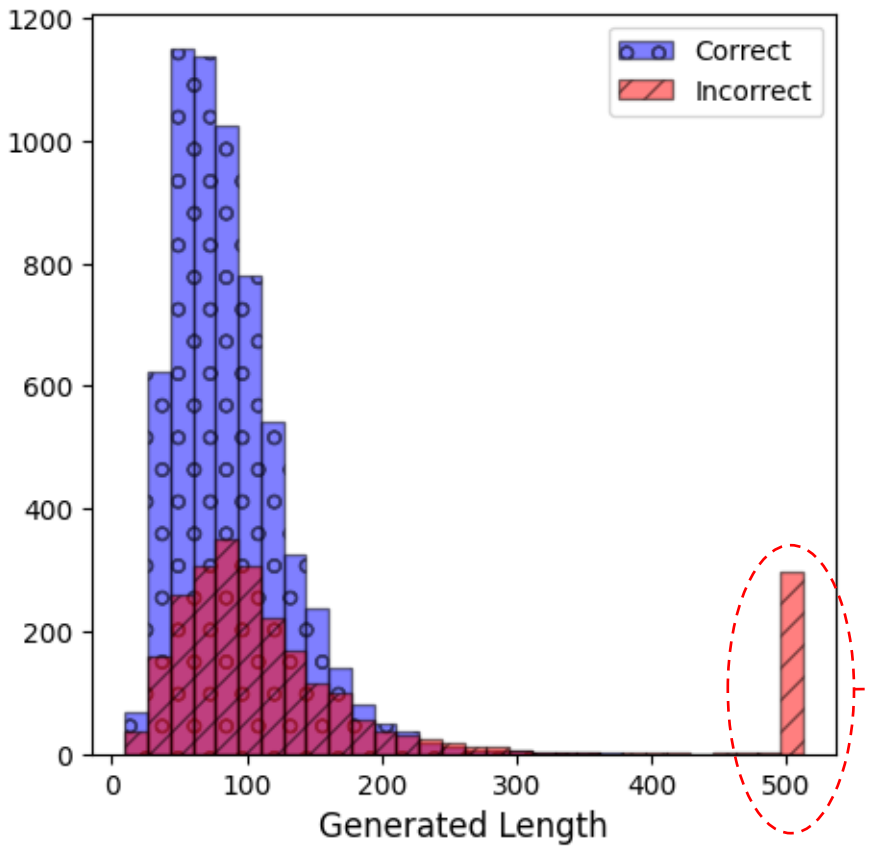
Team 1B

{samuel27, isha4, vp32, vishalb3, skatt24}@illinois.edu

Conclusion

- We conducted a comprehensive analysis of what occurs internally during reasoning and examined whether open-source, lightweight, text-only LLMs.
- We not only used the GSM8k dataset to conduct an in-depth analysis of the model's internal activations, but also constructed a contrastive dataset that can be utilized in future research.
- Representation Engineering does not work for math reasoning; Mathematical accuracy is different from generating correct answer with high honesty.
- Our experimental results show that incorrect reasoning can be classified with an F1 score exceeding 0.7.
- However, we also reveal that achieving a clean and reliable classification of incorrect reasoning using only the model's internal information, or improving the model's reasoning performance through probing, remains a challenging task.

Cases where generation becomes infinite loop-ish and fails to reach the correct answer within the maximum length.




```
[meta-llama/Llama-3.2-3B-Instruct]
Answer: Initially, the bag of marbles costs $20.
Every 2 months, the price increases by 20% of the original price.
20% of $20 is 0.20 * $20 = $4.
So, after the first 2 months, the price will be $20 + $4 = $24.
After the next 2 months, the price will increase by 20% of $24, which is 0.20 * $24 = $4.80.
The new price will be $24 + $4.80 = $28.80.
After the next 2 months, the price will increase by 20% of $28.80, which is 0.20 * $28.80 = $5.76.
The new price will be $28.80 + $5.76 = $34.56.
After the next 2 months, the price will increase by 20% of $34.56, which is 0.20 * $34.56 = $6.912.
The new price will be $34.56 + $6.912 = $41.472.
After the next 2 months, the price will increase by 20% of $41.472, which is 0.20 * $41.472 = $8.2944.
The new price will be $41.472 + $8.2944 = $49.7664.
After the next 2 months, the price will increase by 20% of $49.7664, which is 0.20 * $49.7664 = $9.95328.
The new price will be $49.7664 + $9.95328 = $59.71968.
After the next 2 months, the price will increase by 20% of $59.71968, which is 0.20 * $59.71968 = $11.953936.
The new price will be $59.71968 + $11.953936 = $71.673616.
After the next 2 months, the price will increase by 20% of $71.673616, which is 0.20 * $71.673616 = $14.2347232.
The new price will be $71.673616 + $14.2347232 = $85.9083392.
After the next 2 months, the price will increase by 20% of $85.9083392, which is 0.20 * $85.9083392 = $17.181668
The new price will be $85.9083392 + $17.18166864 = $103.09.
After the next 2 months, the price will increase by 20% of $103.09, which is 0.20 * $103.09 = $20.618.
The new price will be $103.09 + $20.618 = $123.708.
After the next 2 months, the price will increase by 20% of $123.708, which is 0.20 * $123.708 = $24.7416.
The new price will be $123.708 + $24.7416 = $148.45.
After the next 2 months, the price will increase by 20% of $148.45, which is 0.20 * $148.45 = $29.69.
The new price will be $148.45 + $29.69 = $178.14.
After the next 2 months, the price will increase by 20% of $178.14, which is 0.20 * $178.14 = $35.628.
The new price will be $178.14 + $35.628 = $213.768.
After the next 2 months, the price will increase by 20% of $213.768, which is 0.20 * $213.768 = $42.7536.
The new price will be $213.768 + $42.7536 = $256.5216.
After the next 2 months, the price will increase by 20% of $256.5216, which is 0.20 * $256.5216 = $51.30432.
The new price will be $256.5216 + $51.30432 = $307.82592.
After the next 2 months, the price will increase by 20% of $307.82592, which is 0.20 * $307.82592 = $61.575196.
The new price will be $307.82592 + $61.575196 = $369.401116.
After the next 2 months, the price will increase by 20% of $369.401116, which is 0.20 * $369.401116 = $73.880223
```

Contrastive Learning Method

One notable approach to probing LLMs is *representation engineering*, suggested by (Zou et al., 2023). While the study focused on engineering emotions or honesty, our work extends this approach to the problem of correct mathematical reasoning. To draw an analogy, this is similar to probe the emotions a human feels when they make a mistake on a math problem. Specifically, we define a set of reasoning tasks $\mathcal{T}$, where $\mathcal{C} \subset \mathcal{T}$ is the set of correctly classified tasks and $\mathcal{I} = \mathcal{T} \setminus \mathcal{C}$ is the set of incorrectly classified tasks. For the model weights W and the task $t \in \mathcal{T}$, we extract sets of neural network activation as follows:

$$A_{\text{correct}} = \{\text{Rep}(W, t)[-1] \mid t \in \mathcal{C}\}$$

$$A_{\text{incorrect}} = \{\text{Rep}(W, t)[-1] \mid t \in \mathcal{I}\}.$$

where Rep_i is a trainable probing functions of all layers at the ith token position to classify weights based on their representations. We then can modify the model by incorporating the probing results:

$$W' = W + \Delta W,$$

where ΔW represents the weight adjustment based on the contrastively learned LoRA with $\mathcal{L} = \|\text{Rep}(W, t^+) - \text{Rep}(W, t^-)\|_2$, with $\mathcal{L}$ being the loss function, t^+ is the correct task and t^- is the incorrect task.

Source: Team 1B Final Report Draft

LLM model: why Llama 3?

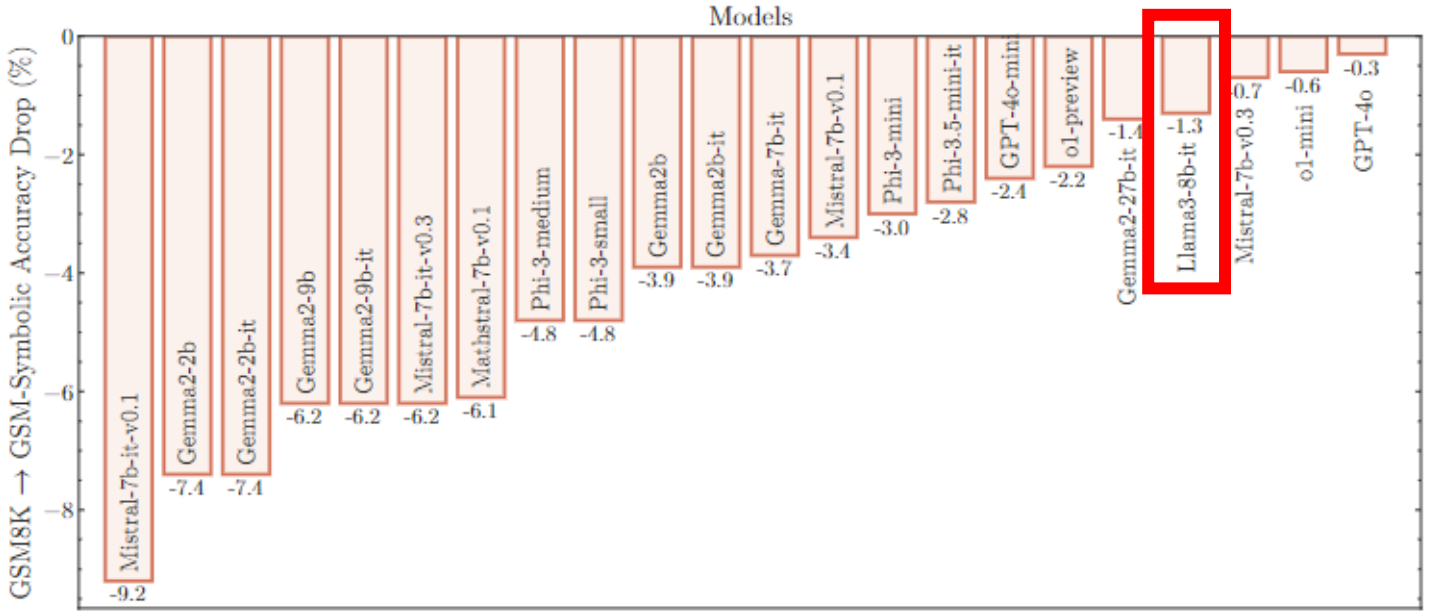


Figure 3: The performance of all state-of-the-art models on `GSM-Symbolic` drops compared to `GSM8K`. Later, we investigate the factors that impact the performance drops in more depth.

[GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models - Apple Machine Learning Research]


ILLINOIS