An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP)

Paulo Shakarian (presenter), Abhinav Koyyalamudi, Noel Ngu, Lakshmivihari Mareedu
Lab V2
Arizona State University

pshak02@asu.edu

ChatGPT: The Latest Commercial LLM

ChatGPT is a commercial large language model released by OpenAI in late 2022.

Its intuitive conversational interface has led to massive popularity, some estimate it has 10 million users.

Anecdotally-reported difficulties with inference

Why Study a Commercial LLM?

Anticipated wide-spread usage with multiple firms creating commercial LLMs

Understanding vulnerabilities

May provide useful starting point for both scientific and

Identify limitations

Commercial and black-box nature of such systems

Need methods to introspect and predict performance

Contributions

1. Creation of a dataset of ChatGPT responses to 1,000 math word problems (MWPs)

2. Identified various factors that impacted ChatGPT performance on MWP's

3. Created ML models for baseline performance prediction

MWP Warm-Up

One whole number is three times a second. If 20 is added to the smaller number, the result is 6 more than the larger.

$$y = 3x$$

$$y+6 = x+20$$

$$3x+6 = x+20$$

$$2x = 14$$

$$x = 7$$

$$y = 21$$

P

One whole number is three times a second . if 20 is added to the smaller number , the result equals 6 more than the larger number . Find both numbers.

Let's call the larger number x and the smaller number y. From the information given, we can set up the following equations:

1. $x = 3y$ (the larger number is three times the smaller number)
2. $y + 20 = x + 6$ (adding 20 to the smaller number equals 6 more than the larger number)

We can substitute the first equation into the second equation:

$y + 20 = 3y + 6$

Simplifying, we get:

$y = 14$

But wait, this should be 2y=14

Now that we know the smaller number is 14, we can use the first equation to find the larger number:

$x = 3y = 3(14) = 42$

So the two numbers are 42 and 14.

7 and 21 is the correct answer

Jan. 24, 2023

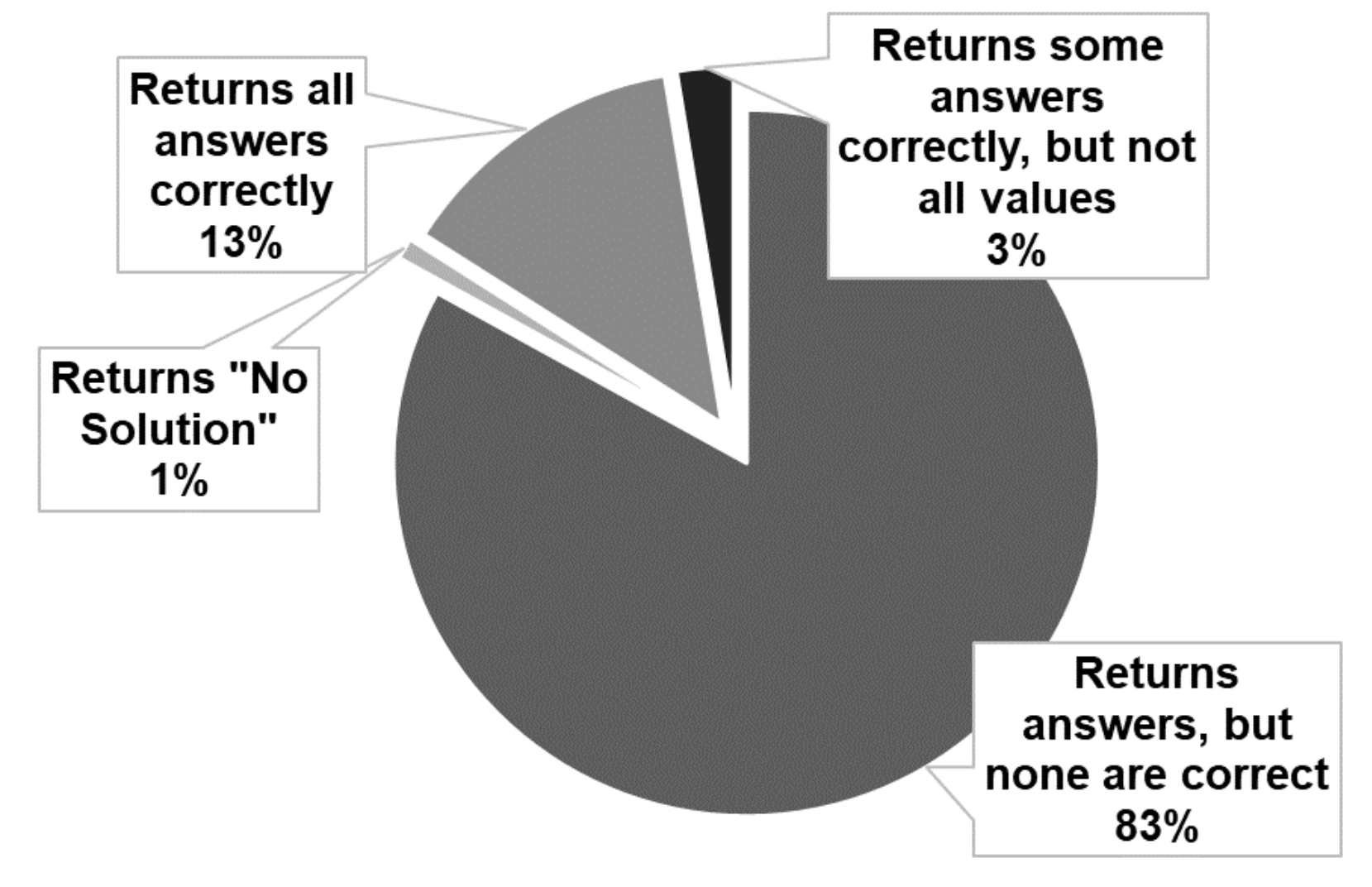
Experiment 1 (early January 2023)

1,000 math word problems (DRAW-1K) sent to ChatGPT via bot

Conducted experiment in early January, 2023

Included additional text to have ChatGPT only provide the answer

Experiment 1 (early January 2023)



Jan. 30th ChatGPT Upgrade

a while.

- We rolled out the ability to purchase ChatGPT Plus internationally.

Release Notes (Feb 9)

As we recently announced, our Plus plan comes with early access to new, experimental features. We are beginning to roll out a way for Plus users the ability to choose between different versions of ChatGPT:

- Default: the standard ChatGPT model
- Turbo: optimized for speed (alpha)

Version selection is made easy with a dedicated dropdown menu at the top of the page. Depending on feedback, we may roll out this feature (or just Turbo) to all users soon.

Release Notes (Jan 30)

We've upgraded the ChatGPT model with improved factuality and mathematical capabilities.

Release Notes (Jan 9)

We're excited to announce several updates to ChatGPT! Here's what's new:

1. We made more improvements to the ChatGPT model! It should be generally better across a wide range of topics and has improved factuality.
2. Stop generating: Based on your feedback, we've added the ability to stop generating ChatGPT's response

Release Notes (Dec 15)

We're excited to announce several updates to ChatGPT! Here's what's new:

1. General performance: Among other improvements, users will

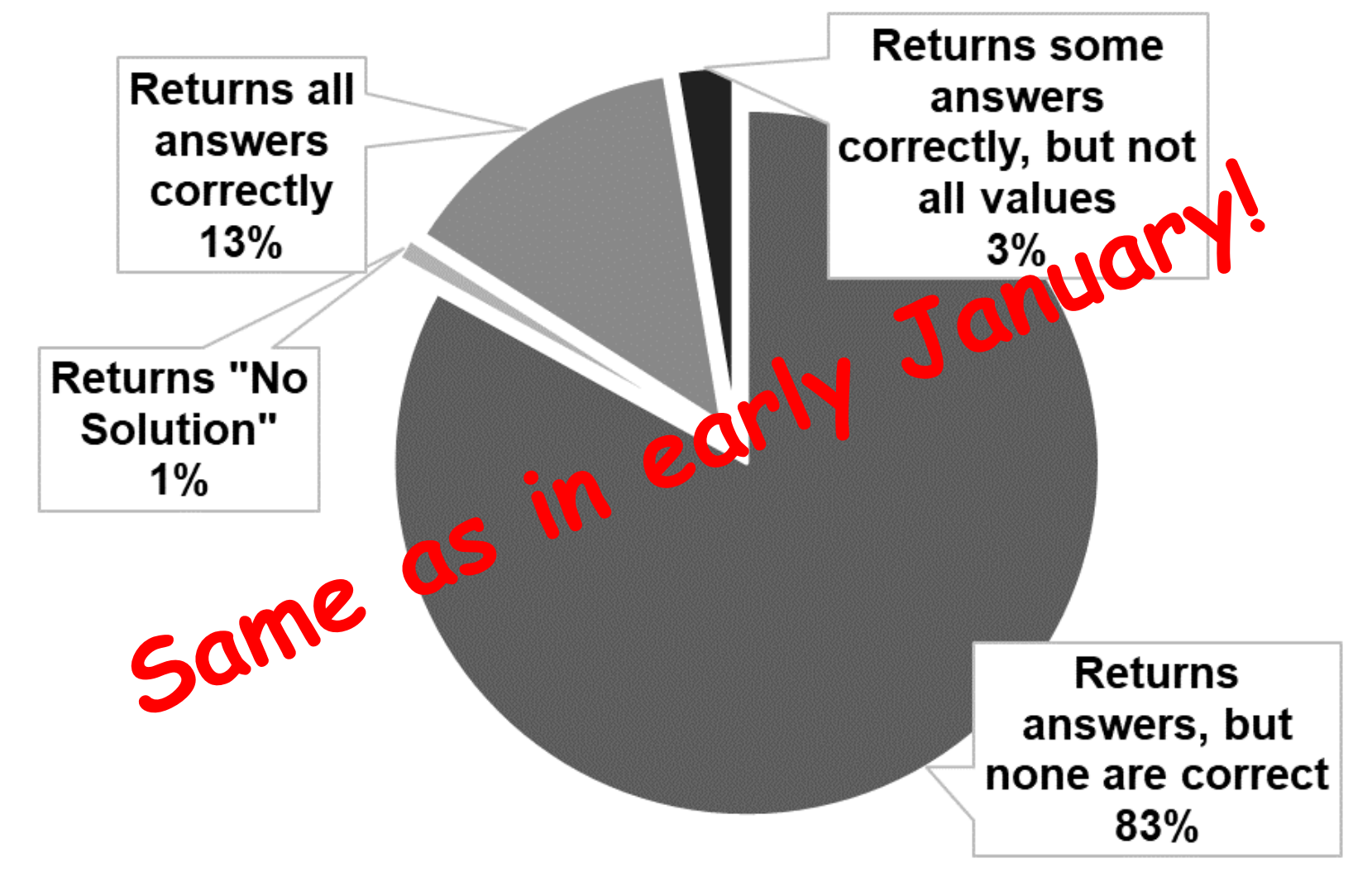
Experiment 2 (mid February 2023)

1,000 math word problems (DRAW-1K) sent to ChatGPT via bot

Conducted experiment in early February, 2023

Included additional text to have ChatGPT only provide the answer

Experiment 2 (mid February 2023)



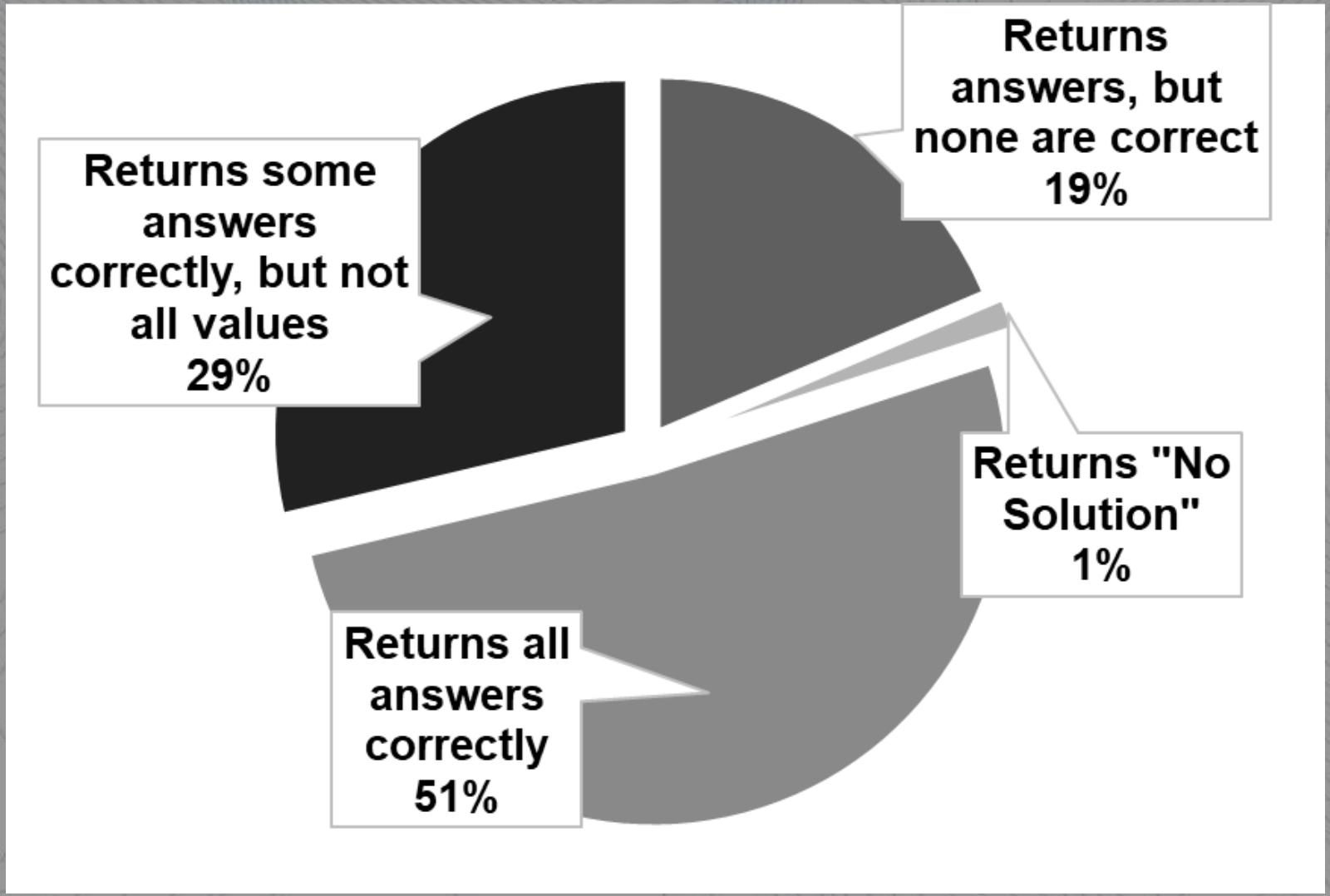
Experiment 3 (mid February 2023)

1,000 math word problems (DRAW-1K) sent to ChatGPT via bot

Conducted immediately following experiment 2

<u>No</u> additional text to have ChatGPT only provide the answer

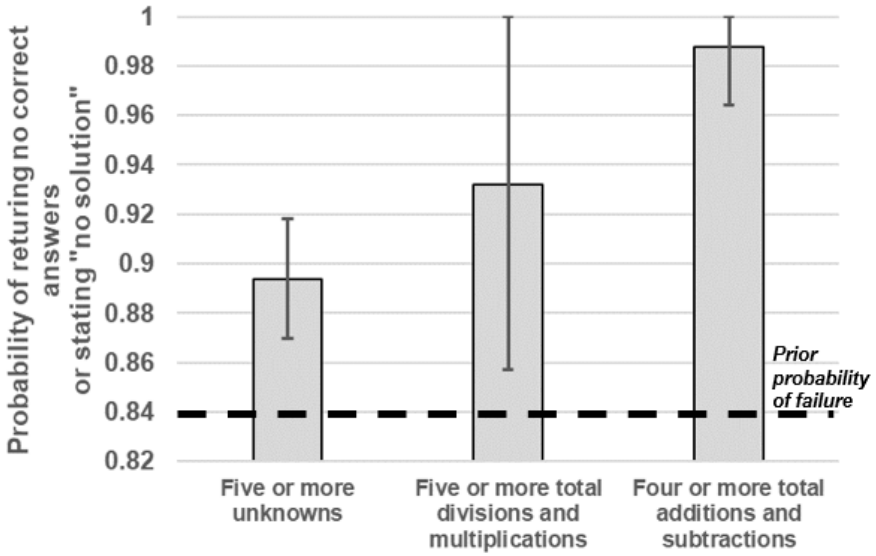
Experiment 3 (mid February 2023, showing work)



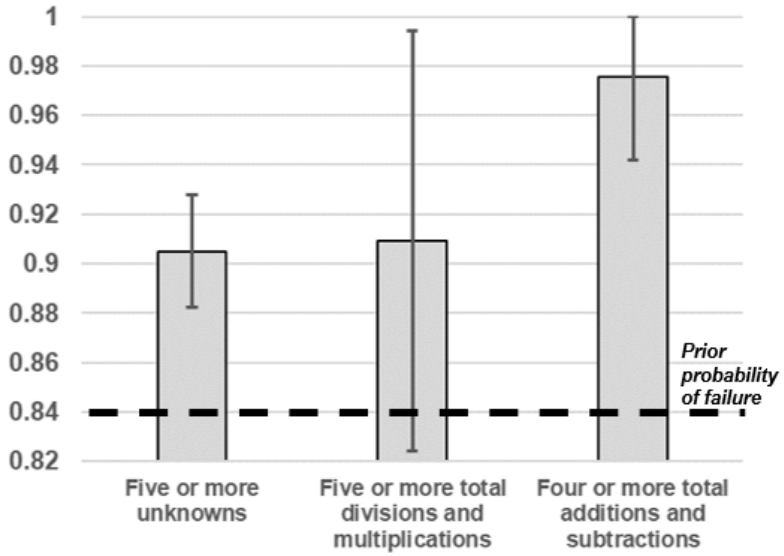
Drivers of Inaccuracy

As we know the ground truth math equations associated with the word problems, can we understand what causes ChatGPT to get incorrect results?

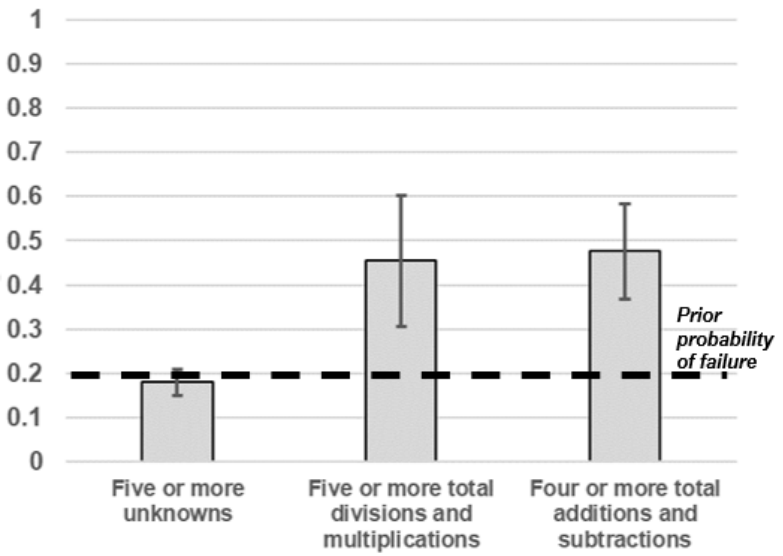
Factors Contributing to Inaccuracy



Performance of ChatGPT on DRAW-1K when not showing work (Jan. 2023)



Performance of ChatGPT on DRAW-1K when not showing work (Feb. 2023)

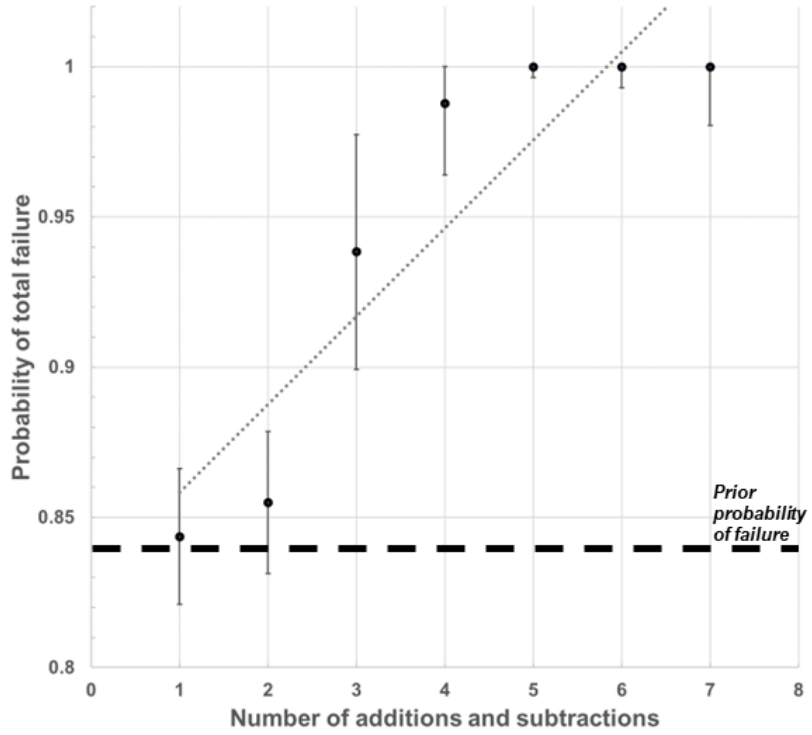


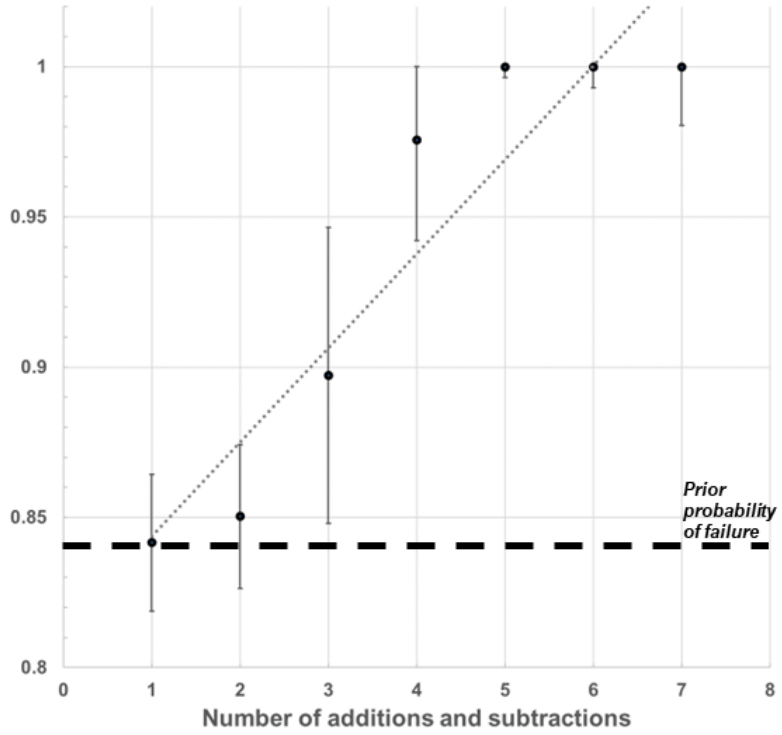
Performance of ChatGPT on DRAW-1K when showing work (Feb. 2023)

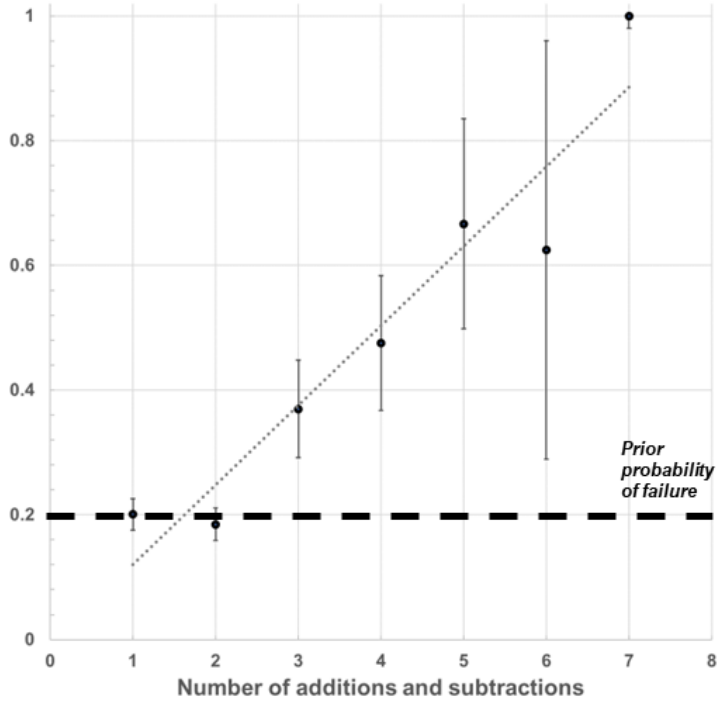
Jan. 2023 (no work) **Feb. 2023 (no work)** **Feb. 2023 (with work)**

95% confidence intervals shown

More Additions → More Failures for ChatGPT







Jan. 2023 (no work)
$R^2=0.82$

Feb. 2023 (no work)
$R^2=0.87$

Feb. 2023 (with work)
$R^2=0.92$

95% confidence intervals shown

Using an ML Model to Predict ChatGPT Performance

Proof of Concept

	Version of ChatGPT	Model Type	Incorr. Prec.	Incorr. Recall	Incorr. F1	Corr. Prec.	Corr. Recall	Corr. F1
84% - 16%	Jan. (No work)	RF	0.90	0.88	0.89	0.34	0.41	0.37
		XGBoost	0.95	0.22	0.36	0.16	0.93	0.26
84% - 16%	Feb. (No work)	RF	0.94	0.89	0.91	0.47	0.63	0.54
		XGBoost	0.98	0.35	0.51	0.18	0.95	0.31
20% - 80%	Feb. (Showing work)	RF	0.78	0.69	0.73	0.74	0.82	0.78
		XGBoost	0.77	0.59	0.67	0.69	0.83	0.75

Beyond Our Study

In our study we used ground truth equations, can we use equations produced by ChatGPT?

More relevant for performance prediction.

We found adding to the prompt can degrade results, but can the prompt also be used to improve results?

Preliminary Result

Subject to further analysis

Prompt addition:

'Let's think things through step by step to get the right answer.'

Reduces totally false answers by 10%

Preliminary Result

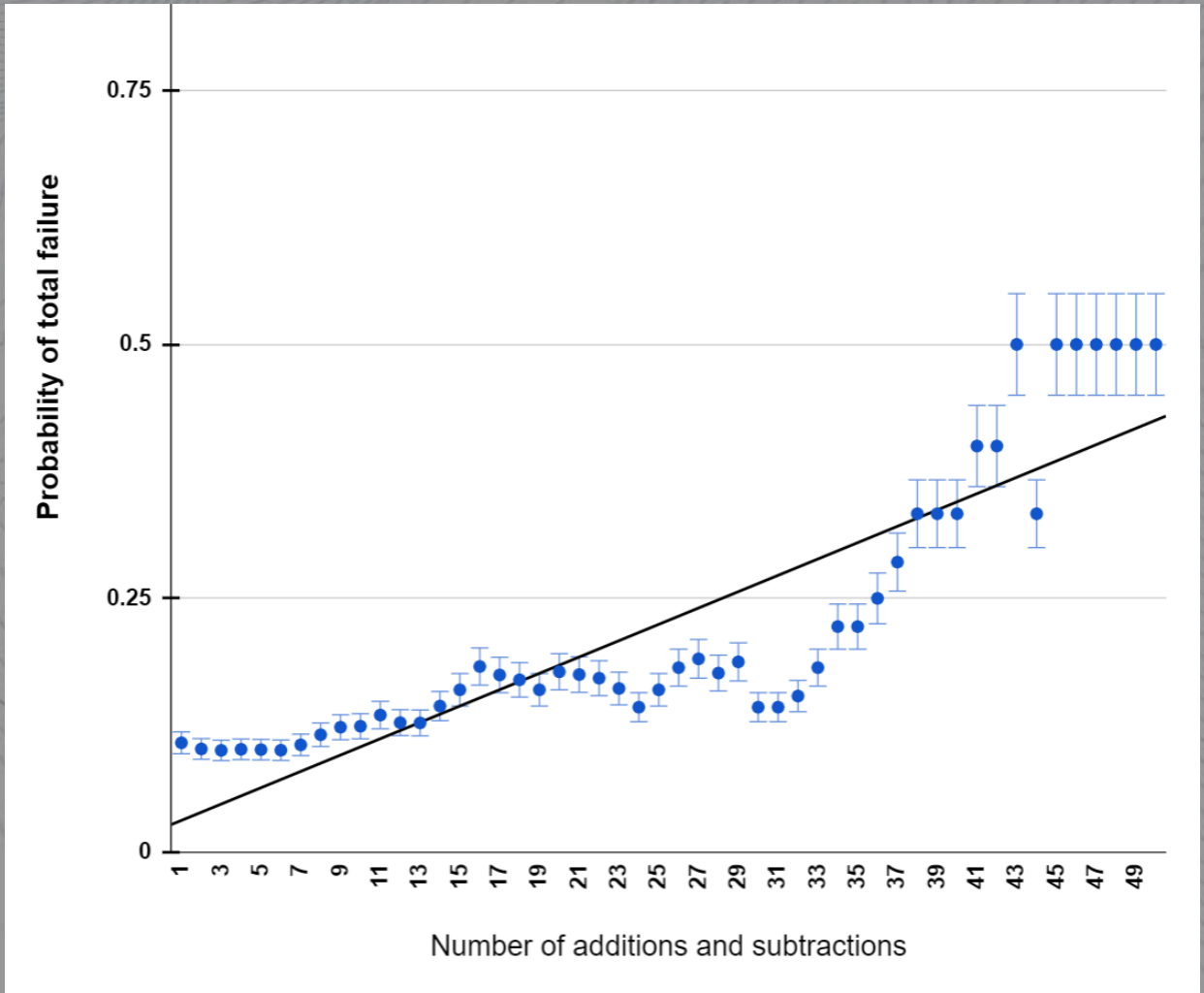
Subject to further analysis

Performance prediction: using ChatGPT's returned equations

Still monotonically increases with additions and subtractions

Also observed for other factors (not seen in ground truth) like number of pairs of parentheses)

Current ML performance prediction results very similar



The Challenge

Can we predict ChatGPT failures on DRAW-1K?

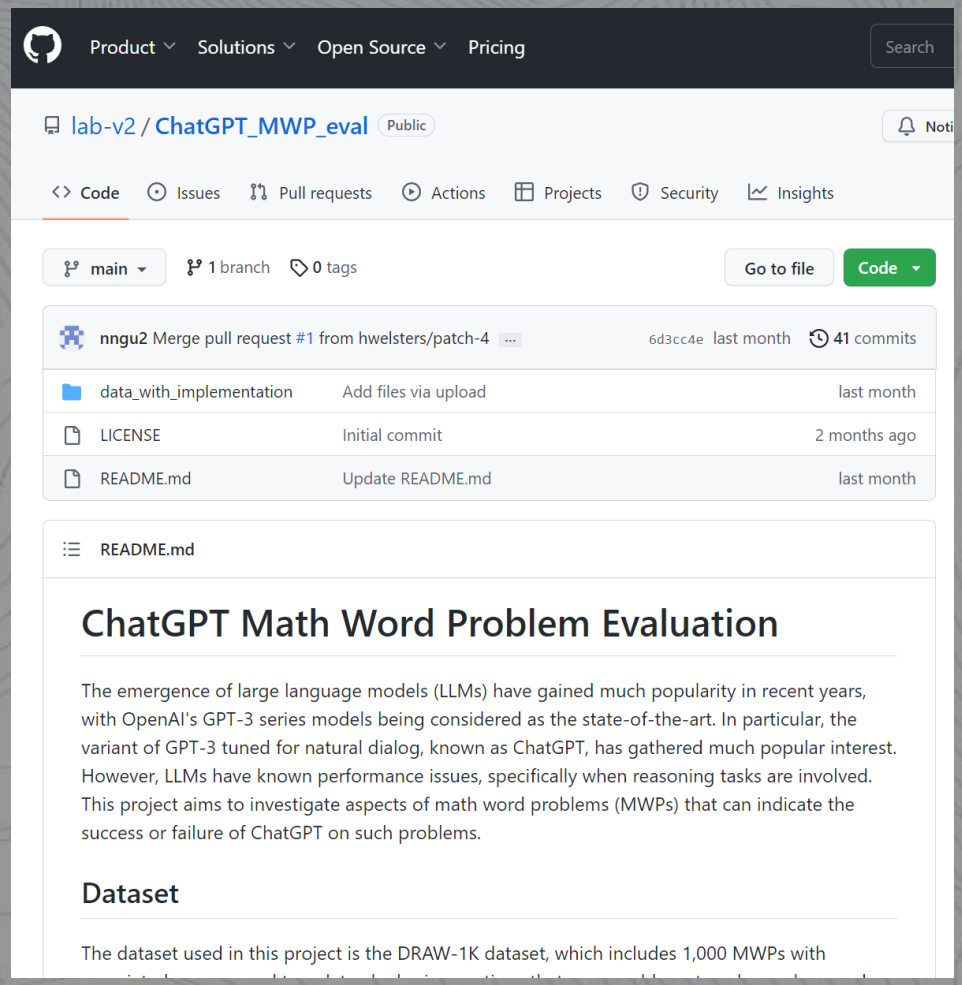
Metric: Precision when recall=0.5 on predicting ChatGPT was totally incorrect for our Feb. data when all work is shown.

Can use ground truth equations, any data from problem, any text output by ChatGPT

Baseline is 0.78 (Random Forest)

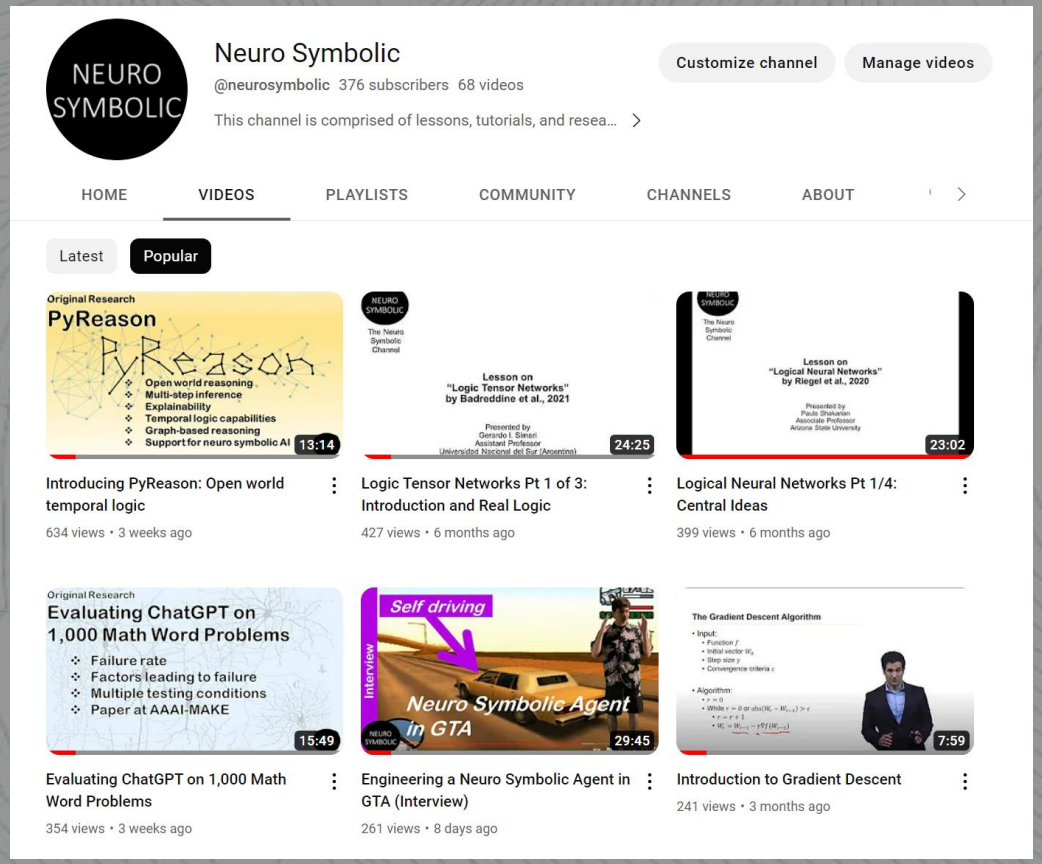
Bonus: Provide an explainable (preferably symbolic) result

Challenge Dataset



Dataset available on GitHub
https://github.com/lab-v2/ChatGPT_MWP_eval

Learn More



Web
https://neurosymbolic.asu.edu

YouTube
https://youtube.com/@neurosymbolic

Inquires: email pshak02@asu.edu