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

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

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

\[ y = 3x \]
\[ y + 6 = x + 20 \]
\[ 3x + 6 = x + 20 \]
\[ 2x = 14 \]
\[ x = 7 \]
\[ y = 21 \]
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.
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)

- Returns all answers correctly: 13%
- Returns "No Solution": 1%
- Returns some answers correctly, but not all values: 3%
- Returns answers, but none are correct: 83%
Jan. 30th ChatGPT Upgrade

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 and chat commerce

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 19)

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, there will be...
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)

- Returns all answers correctly: 13%
- Returns "No Solution": 1%
- Returns some answers correctly, but not all values: 3%
- Returns answers, but none are correct: 83%

Same as in early January!
Experiment 3 (mid February 2023)

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

Conducted immediately following experiment 2

No additional text to have ChatGPT only provide the answer
Experiment 3 (mid February 2023, showing work)

- Returns some answers correctly, but not all values: 29%
- Returns all answers correctly: 51%
- Returns answers, but none are correct: 19%
- Returns "No Solution": 1%
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

95% confidence intervals shown

Jan. 2023 (no work)

Feb. 2023 (no work)

Feb. 2023 (with work)
More Additions $\rightarrow$ More Failures for ChatGPT

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

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

95% confidence intervals shown

Feb. 2023 (with work)
$R^2=0.92$
Using an ML Model to Predict ChatGPT Performance

**Proof of Concept**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. (No work)</td>
<td>RF</td>
<td>0.90</td>
<td>0.88</td>
<td>0.89</td>
<td>0.34</td>
<td>0.41</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.95</td>
<td>0.22</td>
<td>0.36</td>
<td>0.16</td>
<td>0.93</td>
<td>0.26</td>
</tr>
<tr>
<td>Feb. (No work)</td>
<td>RF</td>
<td>0.94</td>
<td>0.89</td>
<td>0.91</td>
<td>0.47</td>
<td>0.63</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.98</td>
<td>0.35</td>
<td>0.51</td>
<td>0.18</td>
<td>0.95</td>
<td>0.31</td>
</tr>
<tr>
<td>Feb. (Showing work)</td>
<td>RF</td>
<td>0.78</td>
<td>0.69</td>
<td>0.73</td>
<td>0.74</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.77</td>
<td>0.59</td>
<td>0.67</td>
<td>0.69</td>
<td>0.83</td>
<td>0.75</td>
</tr>
</tbody>
</table>
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?
Prompt addition:

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

Reduces totally false answers by 10%
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

Neuro Symbolic AI
Resources website

Web
https://neurosymbolic.asu.edu

YouTube
https://youtube.com/@neurosymbolic

Inquires: email pshak02@asu.edu