Rethinking LLM Evaluation

Ion Stoica
(joint work with: Hao Zhang, Lianmin Zheng, Ying Sheng, Wei-Lin Chiang, Dacheng Li, Zhuohan Li, Zi Lin, Zhanghao Wu, Shuo Yang, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Eric P. Xing

August 27, 2023
“For better or worse, benchmarks shape a field”

- David Patterson, Turing Award 2017
LLMs extremely hard to evaluate

Expensive

Unreliable
User

Develop a Python program that reads all the text files under a directory and returns top-5 words with the most number of occurrences
Expensive: Which one is better?

User
Develop a Python program that reads all the text files under a directory and returns top-5 words with the most number of occurrences.

Assistant A
Sure, here’s a Python program that should accomplish what you’re looking for:

```python
import os
from collections import Counter

def get_top_5_words(directory):
    file_list = [f for f in os.listdir(directory) if f.endswith('.txt')]

    word_counts = {}

    # Loop through each text file and count the number of words
    for file in file_list:
        with open(os.path.join(directory, file), 'r') as f:
            file_contents = f.read()
            words = file_contents.split()
            for word in words:
                if word not in word_counts:
                    word_counts[word] = 1
                else:
                    word_counts[word] += 1

    # Sort the word counts in descending order
    sorted_word_counts = sorted(word_counts.items(), key=lambda x: x[1], reverse=True)

    # Return the top 5 words with the most number of occurrences
    return [word for word, count in sorted_word_counts[:5]]

directory = '/path/to/directory'
top_words = get_top_5_words(directory)
print(top_words)
```

This program uses the `os` module to loop through all files in the specified directory, and the `re` module to split the contents of each file into words.
Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs of each stage.
Expensive: Which one is better?

User

Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage?

Assistant A

Of course, I'd be happy to help! Photosynthesis is indeed a crucial process that supports life on Earth, and it occurs in specialized organelles called chloroplasts found in plant cells. The two main stages of photosynthesis are the light-dependent reactions and the light-independent reactions, also known as the Calvin cycle.

1. Light-dependent reactions: These reactions take place in the thylakoid membranes of the chloroplast and involve the conversion of light energy into chemical energy. The primary input for this stage is light, which is absorbed by pigments such as chlorophyll and other accessory pigments. The energy from light is used to generate ATP and NADPH, two energy-rich molecules that are essential for the light-independent reactions. The light-dependent reactions also produce oxygen as a byproduct.

2. Light-independent reactions (Calvin cycle): These reactions occur in the stroma of the chloroplast and are responsible for fixing CO2 into organic molecules using the energy and reducing power generated during the light-dependent reactions. The primary input for this stage is CO2, which enters the leaf through the stomata. The Calvin cycle produces glucose, a simple sugar that can be used by the plant for energy and growth. Oxygen is also produced as a byproduct of this stage.

Here's a summary of the inputs and outputs for each stage:

Light-dependent reactions:
- Inputs: Light, H2O, CO2
- Outputs: ATP, NADPH, O2

Light-independent reactions (Calvin cycle):
- Inputs: CO2, ATP, NADPH
- Outputs: Glucose, O2

I hope this helps! Let me know if you have any further questions.

Assistant B

Sure, I'd be happy to help!

The two main stages of photosynthesis are the light-dependent reactions and the light-independent reactions (also known as the Calvin cycle).

The light-dependent reactions take place in the thylakoid membranes of the chloroplast. The primary inputs for this stage are light energy, water, and ADP (adenosine diphosphate) + Pi (inorganic phosphate). The primary outputs are oxygen, ATP (adenosine triphosphate), and NADPH (nicotinamide adenine dinucleotide phosphate).

During the light-dependent reactions, light energy is absorbed by pigments in the thylakoid membranes, which excites electrons and generates ATP and NADPH. Water is also split into oxygen and hydrogen ions, which contribute to the formation of ATP and NADPH.

The light-independent reactions (Calvin cycle) take place in the stroma of the chloroplast. The primary inputs for this stage are carbon dioxide, ATP, and NADPH. The primary outputs are glucose (or other sugars) and ADP + Pi.

During the Calvin cycle, carbon dioxide is fixed into organic molecules using the energy from ATP and NADPH generated during the light-dependent reactions. This process ultimately results in the production of glucose (or other sugars) that can be used by the plant for energy or stored for later use.
Unreliable: Data contamination

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.
A recent work from @iddo claimed GPT4 can score 100% on MIT’s EECS curriculum with the right prompting.

My friends and I were excited to read the analysis behind such a feat, but after digging deeper, what we found left us surprised and disappointed.

dub.sh/gptsucksatmit
Our story

After Facebook released LLaMa in February 2023…

… we released Vicuna: fine tuned LLaMa using SharedGPT data (70K conversations)
How to evaluate it?

Humans take long and are expensive

We used GPT-4 (released 2 weeks earlier) for evaluation
Game 1.

Judgment: The reference answer provides a comprehensive Python program that reads all text files in a directory, counts the occurrences of each word, and returns the top 5 words with the most occurrences. It uses the `os` module to list all files in the directory and open each text file, the `re` module to split the file contents into words, and the `Counter` class from the `collections` module to count the occurrences of each word and find the top 5 words.

Assistant A's answer is similar to the reference answer, but it does not handle case sensitivity or punctuation. It splits the file contents into words using the `split()` method, which does not remove punctuation from the words. This could lead to incorrect word counts if the text files contain punctuation. Also, it does not convert the words to lowercase before counting them, which could lead to separate counts for the same word in different cases.

Assistant B's answer is very similar to the reference answer. It uses the `os` module to list all files in the directory and open each text file, the `re` module to split the file contents into words, and the `Counter` class from the `collections` module to count the occurrences of each word and find the top 5 words. However, it does not convert the words to lowercase before counting them, which could lead to separate counts for the same word in different cases.

In conclusion, both Assistant A and Assistant B provided correct and helpful answers, but Assistant B's answer is more accurate because it handles punctuation correctly. However, both assistants failed to handle case sensitivity, which is a minor issue that could be easily fixed. Therefore, the final verdict is: [B]
Which one is better?

User
Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage?

Assistant A
Of course, I'd be happy to help! Photosynthesis occurs in specialized organelles within the chloroplast. The light-dependent reactions are the light-dependent reactions, and the light-independent reactions are the Calvin cycle.
1. Light-dependent reactions involve the conversion of light absorbed by pigments such as chlorophyll to generate ATP and NADPH, two key energy carriers.
2. Light-independent reactions, also known as the Calvin cycle, are responsible for fixing CO2 and converting it into glucose. They occur in the stroma of chloroplasts.

Judgment: Both Assistant A and Assistant B provided accurate and detailed responses to the user's question about the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage. They both correctly identified the light-dependent reactions and the light-independent reactions (Calvin cycle) as the two main stages of photosynthesis, and correctly described where these stages take place within the chloroplast. They also correctly identified the primary inputs and outputs for each stage.

However, Assistant A made a mistake by stating that CO2 is an input for the light-dependent reactions and that O2 is an output of the Calvin cycle. In reality, CO2 is an input for the Calvin cycle, and O2 is an output of the light-dependent reactions. Assistant B correctly identified the inputs and outputs for each stage.

Final verdict: [B]

I hope this helps! Let me know if you have any further questions.
Human evaluation

Chatbot answers interpreted by humans, so it makes sense humans to be ultimate arbiters

Ideally, for every question we want to rank all LLMs

Ranking N choices is hard:
• Easier to pick best of N
• Even easier to pick best of two!

Pick the best answer between two LLMs
1. **Tournament:** for each question, compare each pair of LLMs

```
<table>
<thead>
<tr>
<th>Club</th>
<th>MP</th>
<th>W</th>
<th>D</th>
<th>L</th>
<th>GF</th>
<th>GA</th>
<th>GD</th>
<th>Pts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man City</td>
<td>38</td>
<td>28</td>
<td>5</td>
<td>5</td>
<td>94</td>
<td>33</td>
<td>61</td>
<td>89</td>
</tr>
<tr>
<td>Arsenal</td>
<td>38</td>
<td>26</td>
<td>5</td>
<td>6</td>
<td>88</td>
<td>43</td>
<td>45</td>
<td>84</td>
</tr>
<tr>
<td>Man United</td>
<td>38</td>
<td>23</td>
<td>6</td>
<td>9</td>
<td>58</td>
<td>43</td>
<td>15</td>
<td>75</td>
</tr>
<tr>
<td>Newcastle</td>
<td>38</td>
<td>19</td>
<td>14</td>
<td>5</td>
<td>68</td>
<td>33</td>
<td>35</td>
<td>71</td>
</tr>
<tr>
<td>Liverpool</td>
<td>38</td>
<td>19</td>
<td>10</td>
<td>9</td>
<td>75</td>
<td>47</td>
<td>28</td>
<td>67</td>
</tr>
<tr>
<td>Brighton</td>
<td>38</td>
<td>18</td>
<td>8</td>
<td>12</td>
<td>72</td>
<td>53</td>
<td>19</td>
<td>62</td>
</tr>
<tr>
<td>Aston Villa</td>
<td>38</td>
<td>18</td>
<td>7</td>
<td>13</td>
<td>51</td>
<td>46</td>
<td>5</td>
<td>61</td>
</tr>
<tr>
<td>Tottenham</td>
<td>38</td>
<td>18</td>
<td>6</td>
<td>14</td>
<td>70</td>
<td>63</td>
<td>7</td>
<td>60</td>
</tr>
<tr>
<td>Brentford</td>
<td>38</td>
<td>15</td>
<td>14</td>
<td>9</td>
<td>58</td>
<td>46</td>
<td>12</td>
<td>59</td>
</tr>
<tr>
<td>Fulham</td>
<td>38</td>
<td>15</td>
<td>7</td>
<td>16</td>
<td>55</td>
<td>53</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>Crystal Palace</td>
<td>38</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>40</td>
<td>49</td>
<td>-9</td>
<td>45</td>
</tr>
<tr>
<td>Chelsea</td>
<td>38</td>
<td>11</td>
<td>11</td>
<td>16</td>
<td>38</td>
<td>47</td>
<td>-9</td>
<td>44</td>
</tr>
<tr>
<td>Wolves</td>
<td>38</td>
<td>11</td>
<td>8</td>
<td>19</td>
<td>31</td>
<td>58</td>
<td>-27</td>
<td>41</td>
</tr>
</tbody>
</table>
```

**Hard to scale**
### How to scale human evaluation?

1. **Tournament**

   - We use ELO rating

2. **Rating**
Chatbot arena: Benchmarking LLMs in the Wild

Benchmark platform for LLMs that features anonymous, randomized battles in a crowdsourced manner

https://arena.lmsys.org
Chatbot Arena: Results

Elo ratings Based on 27K user votes

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Elo Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gpt-4</td>
<td>1225</td>
</tr>
<tr>
<td>2</td>
<td>claude-v1</td>
<td>1195</td>
</tr>
<tr>
<td>3</td>
<td>claude-instant-v1</td>
<td>1153</td>
</tr>
<tr>
<td>4</td>
<td>gpt-3.5-turbo</td>
<td>1143</td>
</tr>
<tr>
<td>5</td>
<td>vicuna-13b</td>
<td>1054</td>
</tr>
<tr>
<td>6</td>
<td>palm-2</td>
<td>1042</td>
</tr>
<tr>
<td>7</td>
<td>vicuna-7b</td>
<td>1007</td>
</tr>
<tr>
<td>8</td>
<td>koala-13b</td>
<td>980</td>
</tr>
<tr>
<td>9</td>
<td>mpt-7b-chat</td>
<td>952</td>
</tr>
<tr>
<td>10</td>
<td>fastchat-t5-3b</td>
<td>941</td>
</tr>
<tr>
<td>11</td>
<td>alpaca-13b</td>
<td>937</td>
</tr>
<tr>
<td>12</td>
<td>RWKV-4-Raven-14B</td>
<td>928</td>
</tr>
</tbody>
</table>

Win rate between model pairs

https://chat.lmsys.org/?compare
Can we really trust an LLM as a Judge


Judging LLM-as-a-judge with MT-Bench and Chatbot Arena

Lianmin Zheng\textsuperscript{1,*}  Wei-Lin Chiang\textsuperscript{1,*}  Ying Sheng\textsuperscript{4,*}  Siyuan Zhuang\textsuperscript{1}
Zhaohao Wu\textsuperscript{1}  Yonghao Zhuang\textsuperscript{3}  Zi Lin\textsuperscript{2}  Zhuohan Li\textsuperscript{1}  Dacheng Li\textsuperscript{1,10}
Eric. P Xing\textsuperscript{15}  Hao Zhang\textsuperscript{1,2}  Joseph E. Gonzalez\textsuperscript{1}  Ion Stoica\textsuperscript{1}

\textsuperscript{1} UC Berkeley  \textsuperscript{2} UC San Diego  \textsuperscript{3} Carnegie Mellon University  \textsuperscript{4} Stanford  \textsuperscript{5} MBZUA

Abstract

Evaluating large language model (LLM) based chat assistants is challenging due to their broad capabilities and the inadequacy of existing benchmarks in measuring human preferences. To address this, we explore using strong LLMs as judges to evaluate these models on more open-ended questions. We examine the usage and limitations of LLM-as-a-judge, such as position and verbosity biases and limited reasoning ability, and propose solutions to mitigate some of them. We then verify the agreement between LLM judges and human preferences by introducing two benchmarks: MT-bench, a multi-turn question set; and Chatbot Arena, a crowdsourced battle platform. Our results reveal that strong LLM judges like GPT-4 can match both controlled and crowdsourced human preferences well, achieving over 80% agreement, the same level of agreement between humans. Hence, LLM-as-a-judge is a scalable and explainable way to approximate human preferences, which are otherwise very expensive to obtain. Additionally, we show our benchmark and traditional benchmarks complement each other by evaluating several variants of LLaMA/Vicuna. We will publicly release 80 MT-bench questions, 3K expert votes, and 30K conversations with human preferences from Chatbot Arena\textsuperscript{2}. 
Limitations: not unlike humans!

- **Position bias**: prefer first position
- **Verbosity bias**: prefer long answers
- **Self-enhancement bias**: prefer answers from itself
- **Limited reasoning**: not good at grading math questions
High agreement despite limitations

Agreement between GPT-4 and humans over 80%: same as human-human agreement

<table>
<thead>
<tr>
<th></th>
<th>Agreement v.s. Human Experts</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Include Ties</td>
<td>Exclude Ties</td>
<td></td>
</tr>
<tr>
<td>GPT-4</td>
<td>66%</td>
<td>85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1301</td>
<td>828</td>
<td></td>
</tr>
<tr>
<td>Claude-v1</td>
<td>54%</td>
<td>86%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1299</td>
<td>624</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>63%</td>
<td>81%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>677</td>
<td>441</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>33%</td>
<td>50%</td>
<td></td>
</tr>
</tbody>
</table>
What about data contamination?

Steal a page on how humans are evaluated!

One-time exams: each exam is different

Collaborating with kaggle
Summary

LLM evaluation extremely hard

Cracking this problem requires new techniques
• LLMs as judges
• Scalable human evaluation

Many challenges remain
• Contamination: generating unique exams difficult
• Diversity: most questions are easy; need hard questions to differentiate between LLMs