Rethinking LLM Evaluation

Ion Stoica

(joint work with: 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

Berkeley
UNIVERSITY OF CALIFORNIA

“*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:

```
import os
from collections import Counter

def get_top_5_words(directory):
    # Create a list of all text files in the directory
    file_list = [f for f in os.listdir(directory) if f.endswith('.txt')]

    # Initialize a dictionary to store word counts
    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]]

# Call the function and print the top 5 words
directory = '/path/to/directory'
top_words = get_top_5_words(directory)
print(top_words)
```

Assistant B

Sure, here's a Python program that reads all the text files under a directory and returns top-5 words with the most number of occurrences:

```
import os
import re
from collections import Counter

def get_top_words(directory):
    # Initialize a Counter object to keep track of word occurrences
    word_counter = Counter()

    # Loop through all files in the directory
    for filename in os.listdir(directory):
        # Check if the file is a text file
        if filename.endswith(".txt"):
            # Open the file and read its contents
            with open(os.path.join(directory, filename), "r") as file:
                contents = file.read()

                # Use regular expressions to split the contents into words
                words = re.findall(r'\b\w+\b', contents)

                # Update the word counter with the words from the file
                word_counter.update(words)

    # Return the top 5 words with the most occurrences
    return word_counter.most_common(5)

# Example usage
top_words = get_top_words("path/to/directory")
print(top_words)
```

This program uses the `os` module to loop through all files in the specified directory, and the `re` module to

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 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).

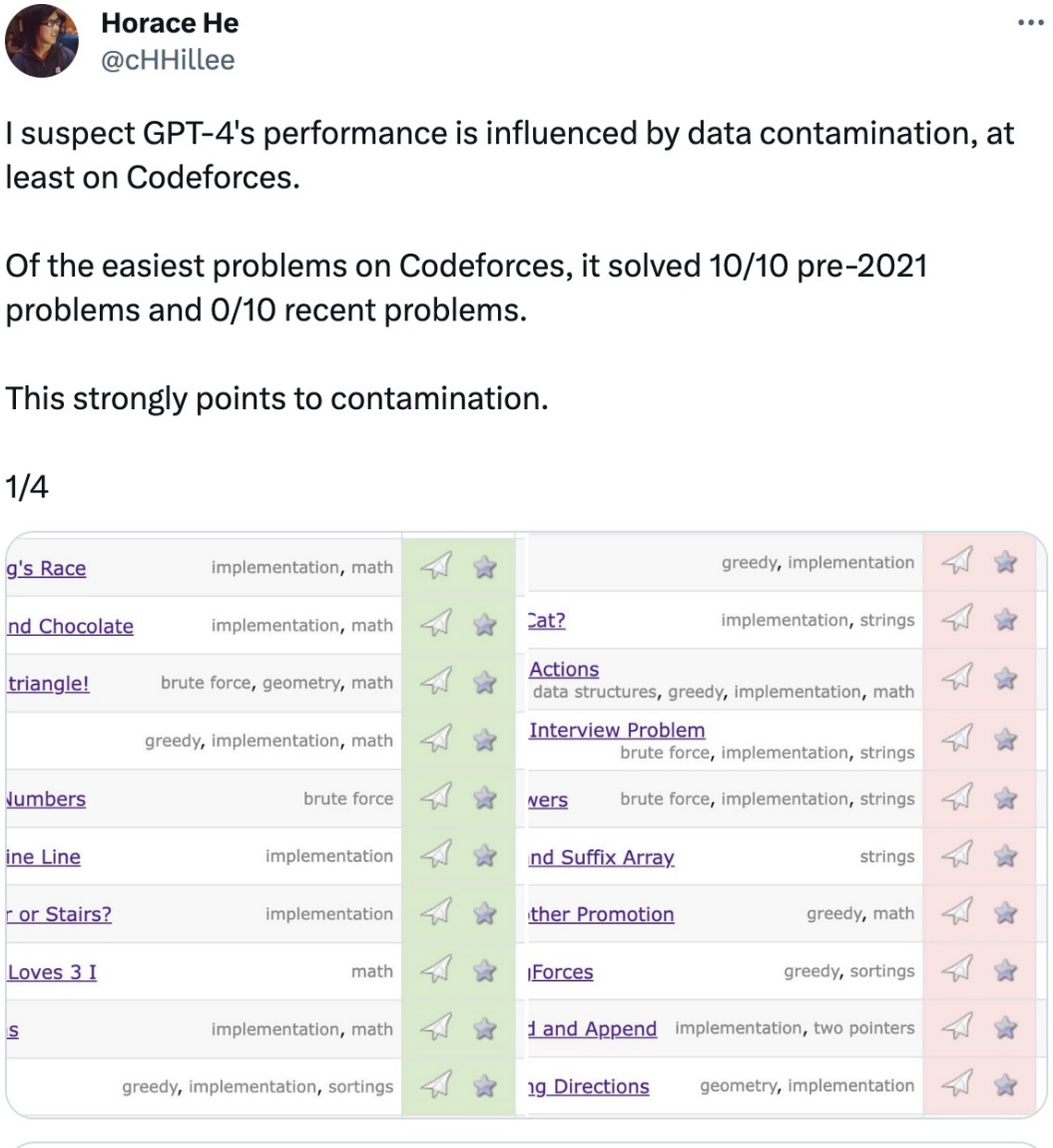
The light-dependent reactions take place in the thylakoid membranes of the chloroplast. The primary inputs for this stage are light energy, water, and ADP (adenosine diphosphate) + Pi (inorganic phosphate). The primary outputs are oxygen, ATP (adenosine triphosphate), and NADPH (nicotinamide adenine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



Horace He
@cHHillee

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.

1/4

Unreliable: Data contamina

Raunak Chowdhuri
@raunakdoesdev

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

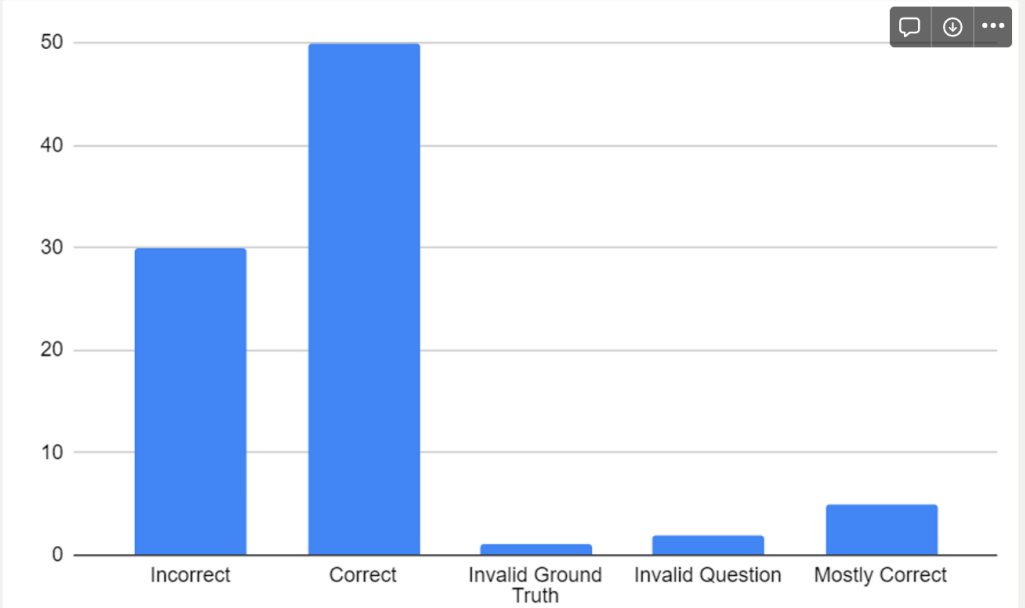
🧵

> **Aran Komatsuzaki** @arankomatsuzaki · Jun 15
> Exploring the MIT Mathematics and EECS Curriculum Models
>
> Presents a comprehensive dataset of 4,550 question
> MIT EECS courses required for obtaining a degree
>
> arxiv.org/abs/2306.08997

Exploring the MIT Mathematics ar
Curriculum Using Large Languag

Sarah J. Zhang* MIT sjzhang@mit.edu
Sam Flori MIT sflorin@mit
Ariel N. Lee Boston University arielle@bu.edu
Eamon Niknafs Boston University en@bu.edu
Andrei Marginean MIT atmargi@mit.edu
Keith Tyser Boston University ktyser@bu.edu
Zad Chin Harvard University zadchin@college.harvard.edu
Nikhil Singh MIT nsingh1@mit.edu
Madeleine Udell Stanford University udell@stanford.edu
Yoon Kim MIT yoonkim@mit.edu
Armando Solar-Lezama MIT asolar@csail.mit.edu
Iddo Drori MIT, Columbia University, Boston University idrori@csail.mit.edu

the Google Sheet to see the latest results, but with a subset of 96 Qs (so far graded), the results are ~32% incorrect, ~58% correct, and the rest invalid or mostly correct.



We've also run and released the expert-prompting zero-shot experiments, but haven't had time to validate these yet. Some of the clear issues with this prompting system are detailed in the section below.

⚠️ We wanted to make clear that our grading process involved both our own manual grading as well as crowdsourcing some of the manual grading effort to PhD level experts who reached out to us after our original post. As the manual grading process is still underway, we cannot verify that every question is graded correctly until we have finished grading entirely and double-checked the grades. We've made the grading spreadsheet public and set it so anyone can comment so as to welcome any corrections from the community in the meantime. Thank you for your patience!

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

Which one is better?

User

Develop a Python program that reads all tl

Assistant A

Sure, here's a Python program that should

```
import os
from collections import Counter

def get_top_5_words(directory):
    # Create a list of all text f
    file_list = [f for f in os.li

    # Initialize a dictionary to
    word_counts = {}

    # Loop through each text file
    for file in file_list:
        with open(os.path.join(di
            file_contents = f.rea
            words = file_contents
            for word in words:
                if word not in wo
                    word_counts[w
                else:
                    word_counts[w

    # Sort the word counts in des
    sorted_word_counts = sorted(w

    # Return the top 5 words with
    return [word for word, count

# Call the function and print the
directory = '/path/to/directory'
top_words = get_top_5_words(direc
print(top_words)
```

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]]

tory and returns top-5 words with the

occurrences

, "r") as file:

contents into words

s from the file

This program uses the `os` module to loop through all files in the specified directory, and the `re` module to

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 hel
occurs in specialized organel
photosynthesis are the light-
Calvin cycle.

1. Light-dependent reactions
involve the conversion of ligh
absorbed by pigments such a
generate ATP and NADPH, tw
The light-dependent reactior

2. Light-independent reactio
responsible for fixing CO2 int
light-dependent reactions. T
stomata. The Calvin cycle pr
growth. Oxygen is also produ

Here's a summary of the inpu

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.

Game 1.

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.

Therefore, Assistant B provided a more accurate response to the user's question.

Final verdict: [[B]]

l the light-independent

chloroplast. The primary inputs
norganic phosphate). The
tinamide adenine dinucleotide

the thylakoid membranes,
oxygen and hydrogen ions,

chloroplast. The primary
s are glucose (or other sugars)

he energy from ATP and
ely results in the production of
r later use.

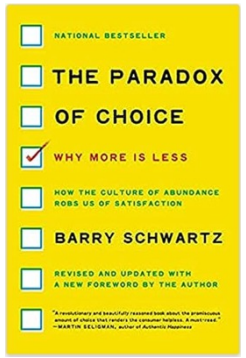
Human evaluation

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

Human evaluation

1. **Tournament:** for each question, compare each pair of LLMs

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

Hard to scale

How to scale human evaluation?

1. Tournam
2. Rating

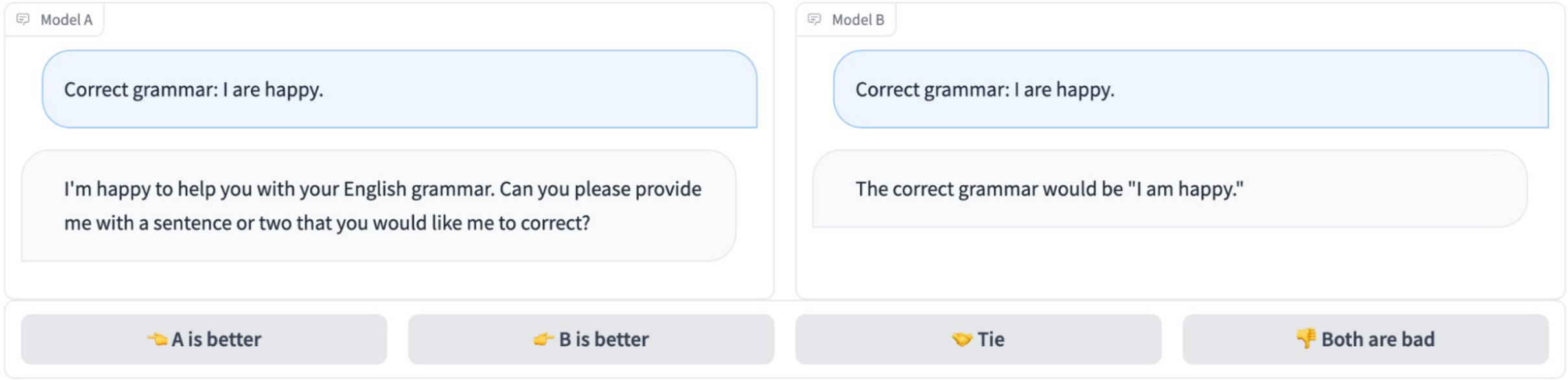
Rank	Name		Classical	Rapid	Blitz
#1	GM Magnus Carlsen	↑10	2845	2820	2887
↑1 #2	GM Fabiano Caruana	↑9	2791	2767	2813
↓1 #3	GM Hikaru Nakamura	↓7	2780	2732	2874
↑1 #4	GM Ding Liren		2780	2830	2787
↑1 #5	GM Alireza Firouzja		2777	2742	2896
↓1 #6	GM Ian Nepomniachtchi	↓8	2771	2756	2785
↑1 #7	GM Anish Giri	↓9	2760	2705	2752

Rank ^					
1					
2					
3					
4					
5					
6					
7	-2		Casper Ruud	24	4,715
8			Andrey Rublev	25	4,595

We use ELO rating

Chatbot arena: Benchmarking LLMs in the Wild

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

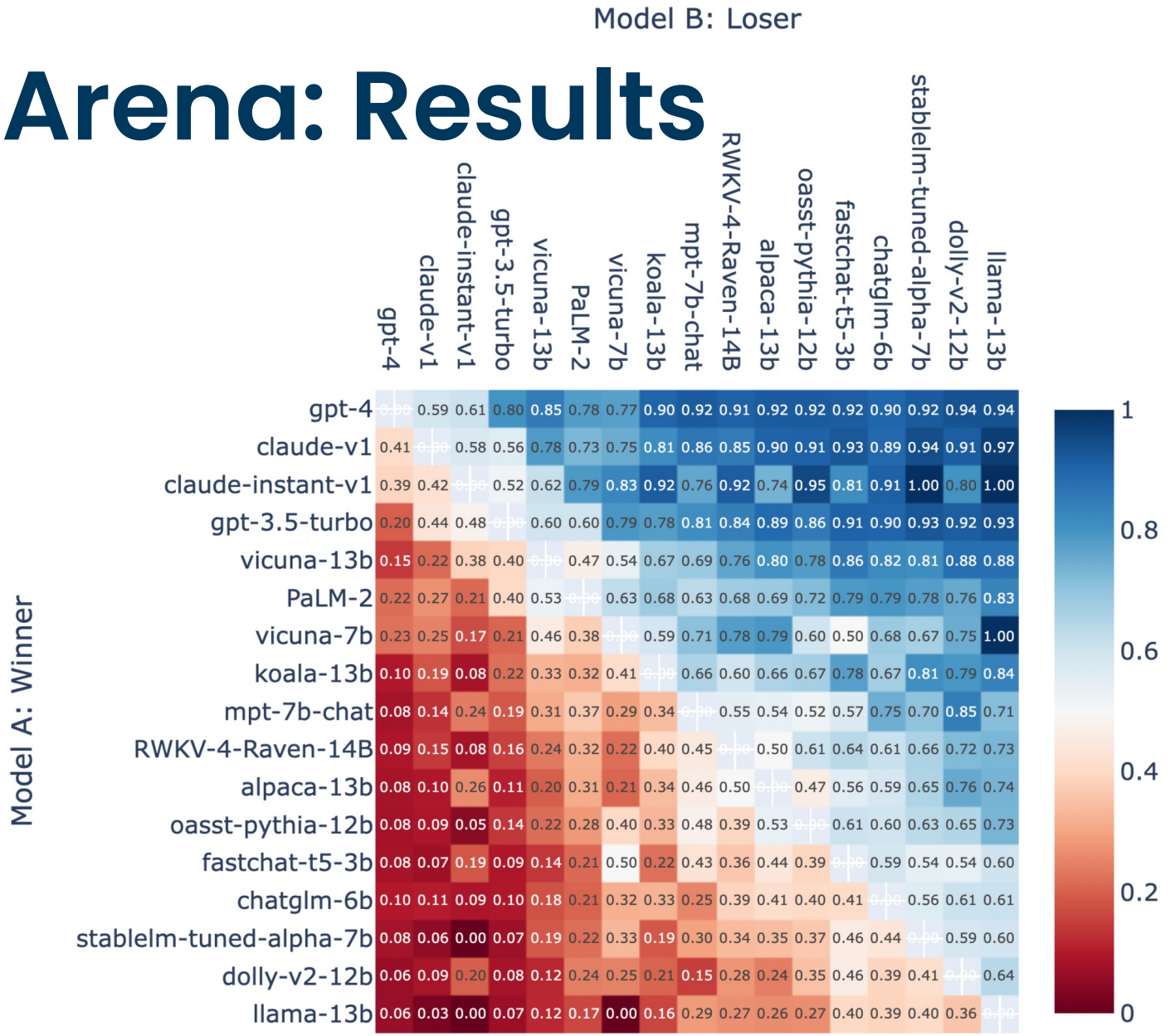


https://arena.lmsys.org

Chatbot Arena: Results

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

Elo ratings Based on 27K user votes



Win rate between model pairs

https://chat.lmsys.org/?compare

Can we really trust an LLM as a Judge

Systematic study:

https://arxiv.org/pdf/2306.05685.pdf

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

Lianmin Zheng[1*] **Wei-Lin Chiang**[1*] **Ying Sheng**[4*] **Siyuan Zhuang**[1]

Zhanghao Wu[1] **Yonghao Zhuang**[3] **Zi Lin**[2] **Zhuohan Li**[1] **Dacheng Li**[135]

Eric. P Xing[35] **Hao Zhang**[12] **Joseph E. Gonzalez**[1] **Ion Stoica**[1]

[1] UC Berkeley [2] UC San Diego [3] Carnegie Mellon University [4] Stanford [5] MBZUAI

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 migrate 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 crowd-sourced 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 [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

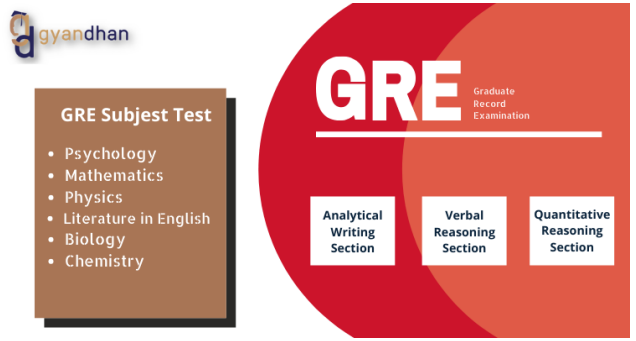
Agreement between GPT-4 and humans over 80%: same as human-human agreement

	Agreement v.s. Human Experts	
	Include Ties	Exclude Ties
GPT-4	66% 1301	85% 828
Claude-v1	54% 1299	86% 624
Human	63% 677	81% 441
Random	33%	50%

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