Judging LLM-as-a-Judge

Using MT-Bench and Chatbot Arena

Zheng et al., 2023

CS 6740

Context and Motivation

LLMs are being increasingly used as chat-based assistants through fine-tuning and RLHF.

Current evaluation methods include:

1. Measures of similarity b/w reference and model outputs (BLEU, ROUGE)

2. Accuracy of task-specific performance (MMLU, HELM)

- Metrics that assess technical performance

MMLU as an example

Massive Multitask Language Understanding

Benchmark designed to evaluate large language models across a wide range of academic and professional subjects.

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s^2. If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s^2 ✓
(B) more than 9.8 m/s^2 ✗
(C) less than 9.8 m/s^2 ✗
(D) Cannot say unless the speed of throw is given. ✗

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points ✗
(B) circle ✗
(C) half-line ✗
(D) line ✓

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

Research Problem

- **Challenge**: Evaluating large language models (LLMs) is difficult due to their wide-ranging capabilities, especially when aligning with **human preferences** in open-ended tasks.
- **Current Gaps:** Traditional benchmarks primarily assess LLMs on confined tasks, which do not capture human preference alignment in **real-world dialogues**.
    - Core-knowledge Benchmarks
    - Instruction-following Benchmarks
    - Conversational Benchmarks

Proposed Solution

Evaluate chatbot performance in **multi-turn conversations** and **open-ended tasks**, matching human preferences at a scalable level.

KEY Benchmarks Introduced:

1. **MT-bench**:
   Assesses a chatbot's ability to follow instructions and engage in meaningful dialogues.

2. **Chatbot Arena**:
   A crowdsourced platform for users to rate chatbots based on real-time interactions.

Road-map of Methodology

1. Advantages of LLM Judges

2. Limitations of LLM Judges:

- Position Bias
- Verbosity Bias
- Self-enhancement Bias

3. Addressed Limitations

- Swapping Judges
- Few-shot judge
- Chain-of-thought and reference-guided judge
- Fine-tuning

Chatbot Arena Judges (<u>LLM</u> + Crowdsourced)

Multi-turn Judge

Advantages of LLM Judges

Question: If the FED buys bonds in the secondary market
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

```
The Federal Reserve buys
bonds in the secondary
market to increase the
money supply.
```

Assistant B:

```
(A) the money supply will increase.
```

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

```
The Federal Reserve buys
bonds in the secondary
market to increase the
money supply. This
affects your daily life
by:
1. Increasing the money
supply.
2. Increasing the money
supply and lowering the
discount rate.
3. Increasing the money
supply and lowering the
discount rate and
increasing the money
supply.
```

Assistant B:

```
When the Federal Reserve buys bonds in the secondary market, it can have
several effects on the economy that can impact your daily life. Here are
three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can
influence interest rates, which in turn can affect the cost of borrowing
money. For example, if the Federal Reserve buys bonds and drives up demand
for them, interest rates on things like mortgages, car loans, and credit
cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation,
which is the general rise in prices of goods and services. This can impact
your daily life by making it more expensive to buy things like groceries,
gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also
have an impact on employment. For example, if the Federal Reserve's
actions lead to economic growth and increased demand for goods and
services, it can create more job opportunities and boost employment.
```

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.
On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B**'s response is more helpful, relevant, and detailed.

Figure 1: Multi-turn dialogues between a user and two AI assistants—LLaMA-13B (Assistant A) and Vicuna-13B (Assistant B)—initiated by a question from the MMLU benchmark and a follow-up instruction. GPT-4 is then presented with the context to determine which assistant answers better.

Limitations of LLM Judges

1. **Position bias:** Tendency to prefer certain positions over others.

GPT-4 Judgment (when A is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant A's answer is more detailed and organized, covering a wider range of topics such as dress code and language, which are not mentioned in Assistant B's response. Assistant A also provides more specific advice on how to handle certain situations, such as exchanging business cards and dining etiquette. Therefore, Assistant A's answer is more comprehensive and useful for someone looking to understand Japanese business etiquette.

GPT-4 Judgment (when B is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant B's answer is more detailed and covers a wider range of topics, such as addressing people by their titles, avoiding direct eye contact, and being mindful of communication styles. Assistant A's answer is more concise but lacks some of the depth found in Assistant B's response. Therefore, Assistant B's answer is better in this case.

Judge	Prompt	Consistency	Biased toward first	Biased toward second	Error
Claude-v1	default	23.8%	**75.0%**	0.0%	1.2%
	rename	56.2%	11.2%	**28.7%**	**3.8%**
GPT-3.5	default	46.2%	**50.0%**	1.2%	2.5%
	rename	51.2%	38.8%	6.2%	**3.8%**
GPT-4	default	**65.0%**	30.0%	5.0%	0.0%
	rename	**66.2%**	28.7%	5.0%	0.0%

Limitations of LLM Judges

2. Verbosity bias - Tendency to prefer lengthy responses, even if they lack clarity & accuracy.

(*Calibration task)

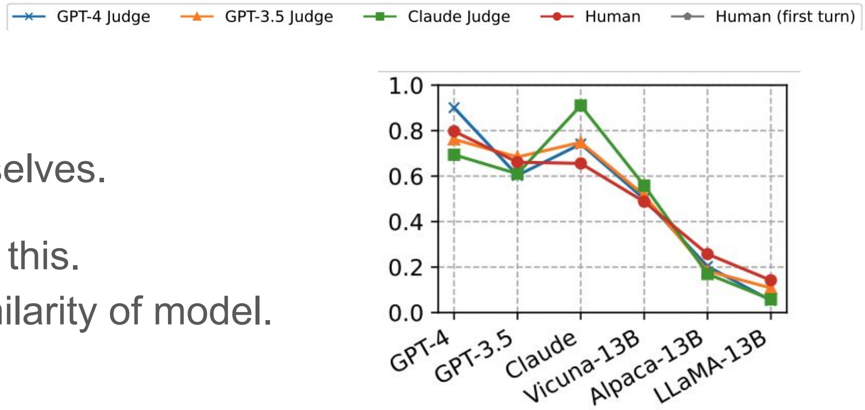
```
The processes involved in the central dogma are:

1. Transcription rephrased: The act of copying
genetic information from DNA to RNA.
2. Translation rephrased: The act of using
genetic information in RNA to create proteins.
3. Transcription: The process by which the
genetic information in DNA is copied into RNA.
4. Translation: The process by which the genetic
information in RNA is used to synthesize
proteins.

The term "central dogma" was coined by Francis
Crick, one of the co-discoverers of the structure
of DNA, in 1958.
```

```
The processes involved in the central dogma are:

1. Transcription: The process by which the
genetic information in DNA is copied into RNA.
2. Translation: The process by which the genetic
information in RNA is used to synthesize
proteins.

The term "central dogma" was coined by Francis
Crick, one of the co-discoverers of the structure
of DNA, in 1958.
```

Table 3: Failure rate under "repetitive list" attack for different LLM judges on 23 answers.

Judge	Claude-v1	GPT-3.5	GPT-4
Failure rate	91.3%	91.3%	8.7%

Limitations of LLM Judges



(b) Non-tied votes, first turn

3. Self-enhancement bias -

Tendency to favor answers generated by themselves.

- Cannot determine whether models exhibit this.
- Difficulty controlling quality vs. stylistic similarity of model.

4. Math and reasoning grading

- Difficulty in determining answers by themselves
- Misled by LLM-generated solutions

Addressed Limitations

- Swapping positions (Position bias)
  - Calling the judge twice, reversing answer order, and declaring a winner only if the answer is preferred TWICE
- Few-shot judge (Position bias)
  - API calling 4 times as expensive
  - Higher consistency from 65% to 77.5%
  - Higher performance unknown, so Zero-shot kept
- Chain-of-thought and reference-guided judge
  - (next slide)
- Fine-tuning a judge model

Reference-guided judge instead of Chain-of-thought

```
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two
AI assistants to the user question displayed below. Your evaluation should consider
correctness and helpfulness. You will be given assistant A's answer, and assistant B's
answer. Your job is to evaluate which assistant's answer is better. You should
independently solve the user question step-by-step first. Then compare both assistants'
answers with your answer. Identify and correct any mistakes. Avoid any position biases and
ensure that the order in which the responses were presented does not influence your
decision. Do not allow the length of the responses to influence your evaluation. Do not
favor certain names of the assistants. Be as objective as possible. After providing your
explanation, output your final verdict by strictly following this format: "[[A]]" if
assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure 7: The chain-of-thought prompt for math and reasoning questions.

- However, GPT-4 was shown to be influenced by given answers, despite it being asked to think independently
- Therefore LLM judge's answer is generated first and then displayed as reference answer separately.

MT-bench Design

- 2 possible designs:

1. Breaking the 2 turns into 2 prompts. → **Prompt 1:** User asks question.

Prompt 2: User follows up with a clarification request.

OR

2. Displaying the entire conversation as a single prompt

→ **Prompt:** User asks question and follow-up in one go, showing full conversation.

Design 2 was used as GPT-4 was shown to struggle to locate assistant's previous accurately - issue with referencing?

Evaluation Setup

1. MT-bench Setup:
    - 6 models used to generate responses to all 80 questions
    - LLM judges (all pairs) and 58-expert human labelers (at least 20 each)

2. Chatbot Arena:
    - Randomly sampled 3K single-turn votes (from 30K arena data)
    - Models evaluated: GPT-4, GPT-3.5, Claude, Vicuna-7B/13B, Koala-13B [16], Alpaca-13B, LLaMA13B, and Dolly-12B.
    - LLM judges and collected crowd judges

Evaluation Metrics

1. **Agreement** - Probability of randomly selected individuals agreeing on a randomly selected question

2. **Average win rate** - Against all other players

***With or without tie-votes**

***Pairwise vs. single-answer grading** (G4-Pair vs. G4-Single)

Comparison b/w Humans and GPT-4: MT-bench Eval.

- GPT-4 and Humans show high agreement on pairwise and single-grading.
- Agreement in S2 (no ties included) = 85%
  - Higher than agreement among humans
- GPT-4 deviations from humans were judged reasonable in 75% of cases, and made 34% change their minds.

Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1138	66% 1343	97% 662	**85%** 859
G4-Single	-	60% 1280	-	85% 739
Human	-	63% 721	-	**81%** 479

(a) First Turn

Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1161	66% 1325	95% 727	**85%** 864
G4-Single	-	59% 1285	-	84% 776
Human	-	67% 707	-	**82%** 474

(b) Second Turn

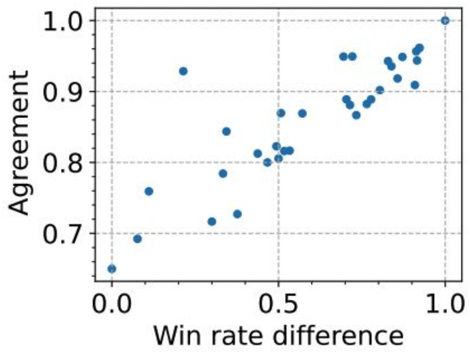
Comparison b/w Humans and GPT-4: Arena Eval.

- Similar non-tie agreement ratio (from MT-bench)
- Greater no. of non-tie votes from GPT-4
- GPT-4 more affirmative

In both, GPT-4 single matches pairwise and human preferences closely - consistent internal rubric

Setup	S1 (Random = 33%)				S2 (Random = 50%)			
Judge	G4-S	G3.5	C	H	G4-S	G3.5	C	H
G4	72% 2968	66% 3061	66% 3062	64% 3066	95% 1967	94% 1788	95% 1712	**87%** 1944
G4-S	-	60% 2964	62% 2964	60% 2968	-	89% 1593	91% 1538	85% 1761
G3.5	-	-	68% 3057	54% 3061	-	-	96% 1497	83% 1567
C	-	-	-	53% 3062	-	-	-	84% 1475

- Higher disparity b/w model players (win-rate difference) -> Greater GPT-4 and Human agreement



Extra: Win rates under other LLM judges

- Win rate curves from LLM judges closely match the curves from humans.
- MT-bench second turn: Claude and GPT-3.5 are more preferred by humans (compared to first).
- Multi-turn benchmarks can better differentiate some advanced abilities of models.

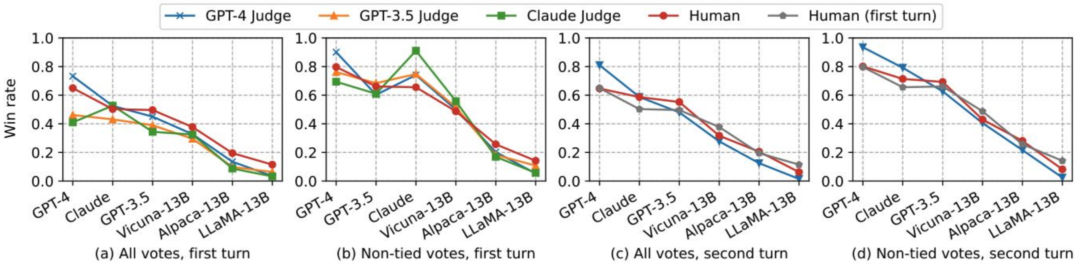


Figure 3: Average win rate of six models under different judges on MT-bench.

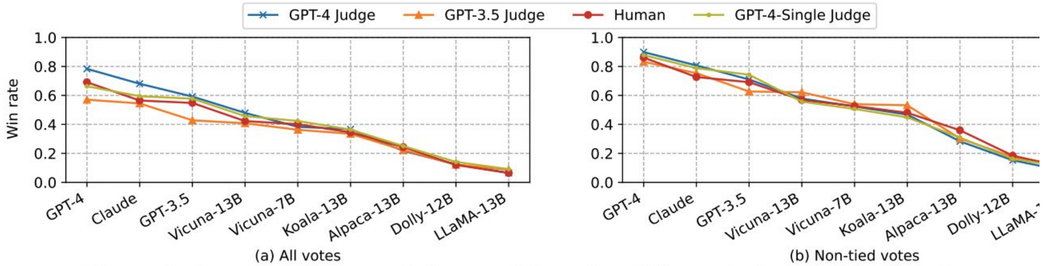


Figure 4: Average win rate of nine models under different judges on Chatbot Arena.

Key Takeaways

- **LLM-as-a-Judge** achieves over 80% agreement with human expert evaluations, matching human-to-human consistency.
- **Agreement between GPT-4 and humans** is greater than among humans in S2 setup (no ties).
- **Bias Mitigation**: Position and verbosity biases in LLM evaluation are minor and can be addressed effectively.
- **Hybrid Evaluation**: Combining capability-based benchmarks with human preference benchmarks enables swift, automated LLM evaluation.
- **Public Data Release**: 80 MT-bench questions, 3K expert votes, and 30K conversations are released for further research.

Evaluation/Points for discussion

- **Focus on evaluating helpfulness, but harmfulness:**
  - Would there be a similar alignment between humans and GPT-4 in assessing safety?
- **Self-enhancement bias: similarity vs. quality in controlled trials:**
  - How can one mitigate this limitation in LLM judges?
- **Selection of zero-shot prompting for LLM judges instead of few-shot:**
  - How might few-shot examples present an additional bias?