

ANLP - Paper presentation

LLM-as-a-Judge

Théo CAVINA - Théo PUTEGNAT - Racim MENASRIA - Alexandre HUMBLLOT

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

Lianmin Zheng^{1*} Wei-Lin Chiang^{1*} Ying Sheng^{4*} Siyuan Zhuang¹

Zhanghao Wu¹ Yonghao Zhuang³ Zi Lin² Zhuohan Li¹ Dacheng Li¹³

Eric P. Xing³⁵ Hao Zhang¹² Joseph E. Gonzalez¹ Ion Stoica¹

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

Introduction

Context end of 2023

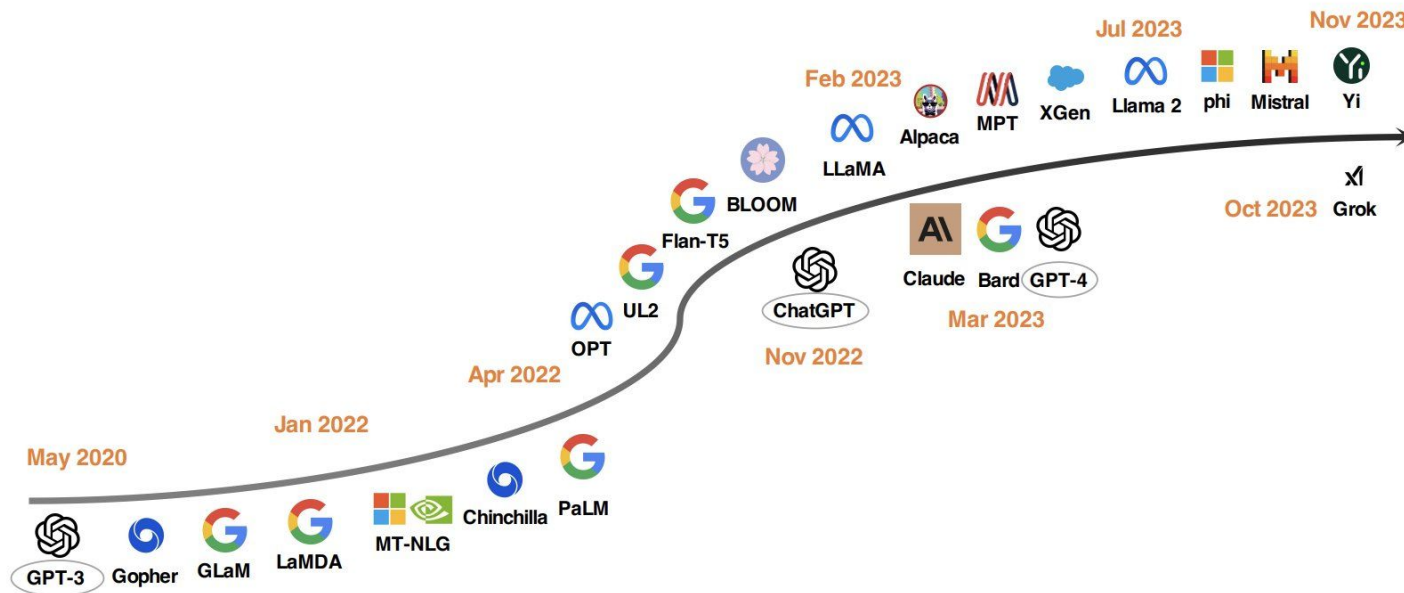


Figure 3: LLM development timeline. The models below the arrow are closed-source while those above the arrow are open-source.

Benchmark vs Human Preference

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.

Existing solution : conversational benchmark

- **Conversational benchmarks**, like CoQA [35], MMDialog [15] and OpenAssistant [23], are closest to our intended use cases. However, the diversity and complexity of their questions often fall short in challenging the capabilities of the latest chatbots.

Limitations

- Automated benchmark do not measure the subtlety of human languages
- Human annotation are slow and costly
- The expected answers are shorts
- Limited range of question topics
- Complexity

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q₁: What are the candidates **running** for?

A₁: Governor

R₁: The Virginia governor's race

Q₂: **Where**?

A₂: Virginia

R₂: The Virginia governor's race

Q₃: Who is the democratic candidate?

A₃: **Terry McAuliffe**

R₃: Democrat Terry McAuliffe

Q₄: Who is **his** opponent?

A₄: **Ken Cuccinelli**

R₄: Republican Ken Cuccinelli

Q₅: What party does **he** belong to?

A₅: Republican

R₅: Republican Ken Cuccinelli

Q₆: Which of **them** is winning?

A₆: Terry McAuliffe

R₆: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May

Figure 2: A conversation showing coreference chains in color. The entity of focus changes in Q₄, Q₅, Q₆.

Objectives and contributions

- Find a way to measure the model alignment with human expectancies
- Conceive a new kind of scalable benchmark and use appropriate metrics

=> Study how LLMs could approximate human preference when comparing 2 models

- The authors make two human preferences dataset available

Methodology

Introduced benchmarks

MT-Bench :

One (or two for comparison) LLMs are involved in a 2 turns conversation.

Judgement is focused on the 2nd answer.

It compares :

- Instruction following ability
- Consistency

(80 multi-turn questions, 3K votes from 58 experts)

Introduced benchmarks

MT-Bench :

One (or two for comparison) LLMs are involved in a 2 turns conversation. Judgement is focused on the 2nd answer.

It compares :

- Instruction following ability
- Consistency

(80 multi-turn questions, 3K votes from 58 experts)

ChatBot Arena :

A user ask the same question to 2 anonymous model and choose the best answer.

It compares :

- Model alignment over a great variety of topics and unrefined questions

(30K votes and 2114 unique IPs)

Mt-Bench examples

Table 1: Sample multi-turn questions in MT-bench.

Category	Sample Questions	
Writing	1st Turn	Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.
	2nd Turn	Rewrite your previous response. Start every sentence with the letter A.
Math	1st Turn	Given that $f(x) = 4x^3 - 9x - 14$, find the value of $f(2)$.
	2nd Turn	Find x such that $f(x) = 0$.
Knowledge	1st Turn	Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies ...
	2nd Turn	Now, explain them again like I'm five.

Evaluation techniques



- **Pairwise comparison.**

Compare 2 LLMs answers.

- **Single answer grading.**

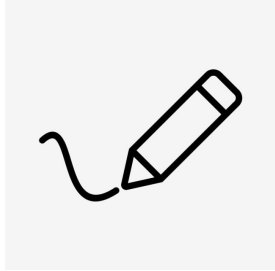
Note the answer of one LLM (converted to pairwise comparison to compute results)



- **Reference-guided grading**

Note the answer of one LLM based on a guideline

Biais



Position bias (some solutions presented)

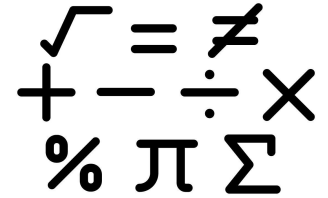


Verbosity bias

Self-enhancement bias.



Limited reasoning capabilities



Position bias

Same answers presented twice in a different order or only by inverting the names Assistant A and Assistant B.

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%

Verbosity biai

When an LLM judge favors longer, verbose responses, even if they are not the best ones

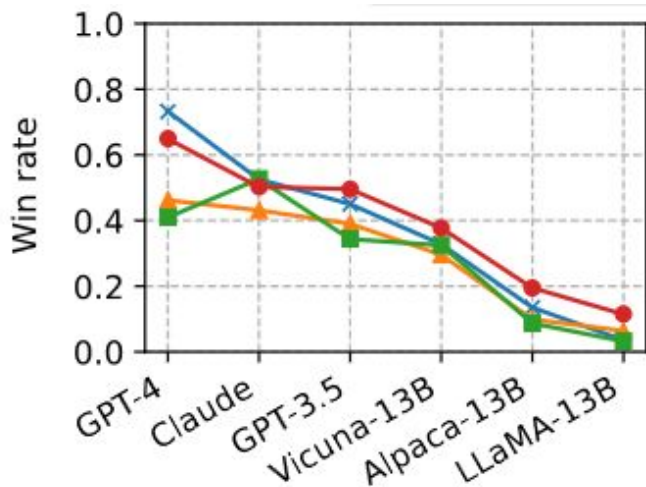
Table 3: Failure rate under “repetitive list” attack for different LLM judges on 23 answers.

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

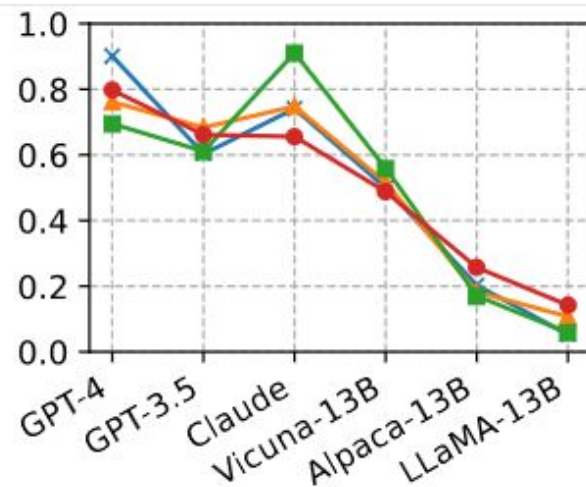
Self enhancement

LLM judges may favor the answers generated by themselves

GP4 + 10 % , Claude V1 +25 %



(a) All votes, first turn



(b) Non-tied votes, first turn

—x— GPT-4 Judge —▲— GPT-3.5 Judge —■— Claude Judge —●— Human —●— Human (first turn)

Reasoning limitations biases

Question: Each problem consists of three statements. Based on the first two statements, the third statement may be true, false, or uncertain.

1. Oranges cost more than apples.
2. Oranges cost less than bananas.
3. Bananas cost more than apples and bananas cost more than orange.

If the first two statements are true, then the third statement is

Assistant A:

false.

If oranges cost more than apples (statement 1) and oranges cost less than bananas (statement 2), then it is not possible for bananas to cost less than oranges and also cost more than both apples and oranges (statement 3). Therefore, statement 3 is false.

Assistant B:

False.

If the first two statements are true, then the third statement is false. The first statement says that oranges cost more than apples, and the second statement says that oranges cost less than bananas. This means that oranges cost less than bananas, which contradicts the third statement that bananas cost more than apples and more than oranges. Therefore, the third statement must be false.

GPT-4 Judgment:

Both Assistant A and Assistant B provided accurate and relevant responses to the user's question. They both concluded that the third statement is false based on the first two statements. However, Assistant B's response is slightly clearer and more concise in explaining the contradiction between the statements. Therefore, I choose Assistant B as the better response.

[[B]]

GPT-4's Answer (if asked separately):

True.

If oranges cost more than apples (statement 1) and oranges cost less than bananas (statement 2), it logically follows that bananas cost more than apples and bananas cost more than oranges, which is the third statement. So, if the first two statements are true, the third one is also true.

Addressing biais

Position biais : Call judge twice by swapping order of answers

Few shot judges : Giving 3 examples of good judgment in the prompt increases the consistency from 65 to 77.5 % for GPT-4

More expensive and may introduce new biais

COT and reference judging : make LLM answer the question (**separately in case of reference**) before judging

Table 4: Judge failure rate on 10 math questions with different prompts. We test LLaMA-13B vs. Vicuna-13B and swap positions. A failure means when GPT-4 says an incorrect answer is correct.

	Default	CoT	Reference
Failure rate	14/20	6/20	3/20

Results

Agreement and best models

Agreement between 2 judges :

- Probability that both judges choose the same best answer
- Baseline : **81% agreement among human**
- GPT-4 vs human : reach **85% agreement**

The best judging models are the GPTs : trained using RLHF
or fine-tuned model on “high quality conversations” (vicuna vs llama)

Results : MT-Bench (80 questions)

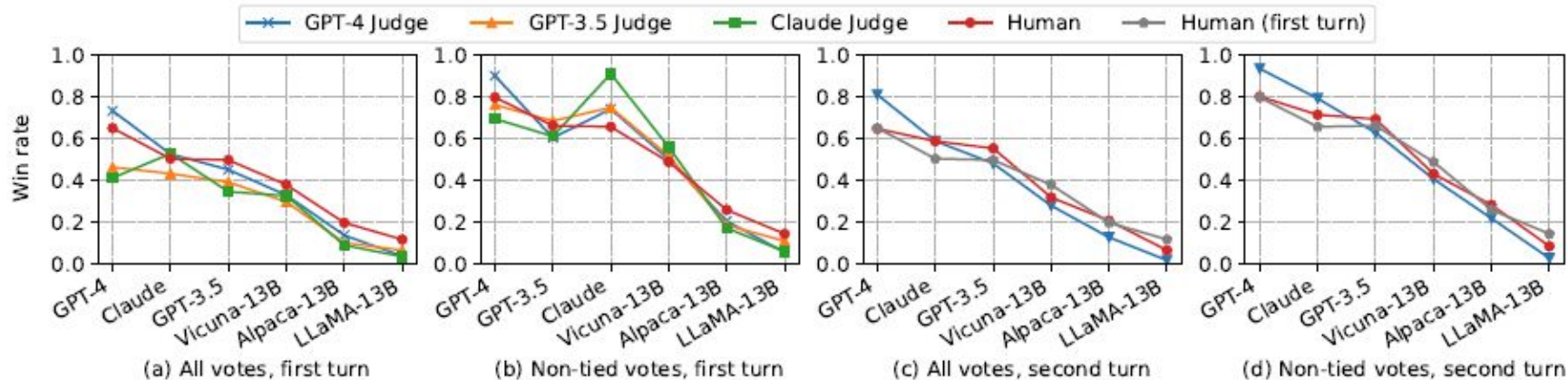


Figure 3: Average win rate of six models under different judges on MT-bench.

- Win rate curves from LLMs match human judgement
=> Justify the use of GPT-4 to judge the best response for a small set of structured questions

Results : ChatBot Arena (Great variety of questions)

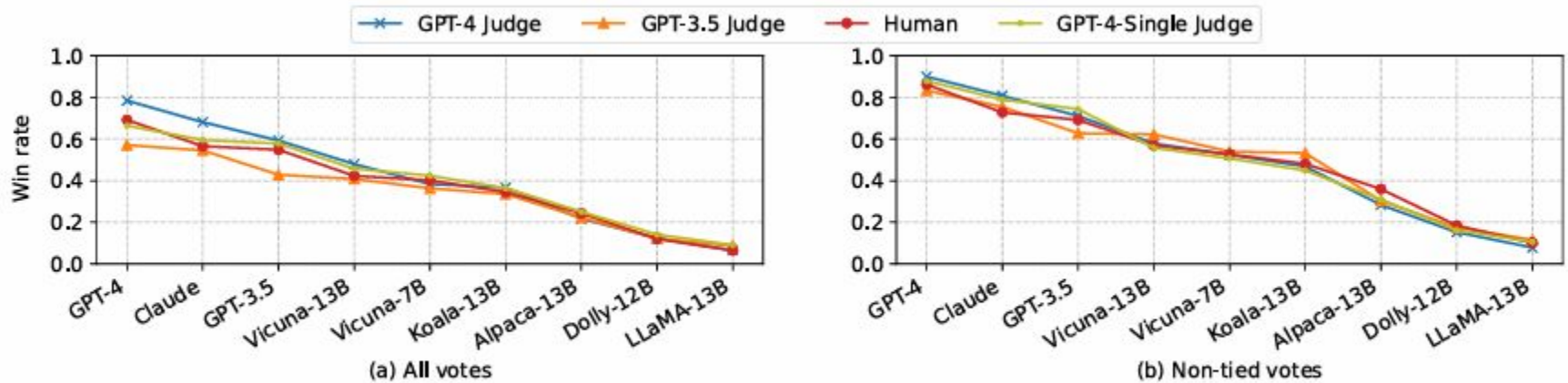


Figure 4: Average win rate of nine models under different judges on Chatbot Arena.

- Here again : the LLM win rate curve follow the human judgment
=> For general non-controlled question, the usage of LLM as judges is also justified

Conclusion

- Existing benchmark such as MMLU do not take human alignment into account
- **This paper shows that using LLM to approximate human preferences is very promising**
- It would allow new benchmark to be scaled and automatised
- Some biases have to be faced
- It is a preliminary work and solutions of this kind will probably be conceived in the future