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

Saketh Chaluvadi and Vishnesh Ramanathan

Introduction

- In the past few classes, we have talked about different preference optimization algorithms
  - DPO
  - PPO
  - SimPO
- The goal of these algorithms is to align models to human preferences
- **How do we know that these algorithms work well? Where did the evaluation scores come from?**

How do we know if LLM based chat assistants are doing a good job?

Motivation

- LLM-based assistants are starting to exhibit a lot of ‘intelligent’ capabilities
- These capabilities are across a wide range of tasks: writing to coding
- As model capabilities become very broad, evaluating them becomes more challenging
- Existing benchmarks fall short

Traditional benchmarks

1. **Core-knowledge:** Do LLMs know basic things about the world?
    a. MMLU, HellaSwag, ARC
2. **Instruction-following:** Can LLMs follow our instructions properly?
    a. Flan, Self-instruct
3. **Conversational:** Can LLMs manage a conversation well?
    a. CoQA, MMDialog

Core-Knowledge

Determine whether the statements are logically equivalent or contradictory. If neither, determine whether they are consistent or inconsistent.
$E \supset (F \cdot E)$ and $\sim E \cdot F$
(A) Logically equivalent
(B) Contradictory
(C) Neither logically equivalent nor contradictory, but consistent
(D) Inconsistent

A kitchen is shown followed by various ingredients and a woman speaking to the camera. She begins showing the ingredients and putting them into a hot boiling pot and stirring around. she

a) shows off the oven and begins assembling the cookies in the oven by pushing a button on the oven. (2.2%)
b) continues mixing up more ingredients and then puts them all together in a bowl, serving the dish ad sprinkling olive oil around it. (97.8%)
c) shows raising and lowering the pot until adding more water and corn syrup. (0.0%)
d) places an omelette onto the screen and puts it in the oven to bake. (0.0%)

Task	Tested Concepts	Supercategory
Abstract Algebra	Groups, rings, fields, vector spaces, ...	STEM
Anatomy	Central nervous system, circulatory system, ...	STEM
Astronomy	Solar system, galaxies, asteroids, ...	STEM
Business Ethics	Corporate responsibility, stakeholders, regulation, ...	Other
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...	Other
College Biology	Cellular structure, molecular biology, ecology, ...	STEM
College Chemistry	Analytical, organic, inorganic, physical, ...	STEM
College Computer Science	Algorithms, systems, graphs, recursion, ...	STEM
College Mathematics	Differential equations, real analysis, combinatorics, ...	STEM
College Medicine	Introductory biochemistry, sociology, reasoning, ...	Other
College Physics	Electromagnetism, thermodynamics, special relativity, ...	STEM
Computer Security	Cryptography, malware, side channels, fuzzing, ...	STEM
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...	STEM
Econometrics	Volatility, long-run relationships, forecasting, ...	Social Sciences
Electrical Engineering	Circuits, power systems, electrical drives, ...	STEM
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...	STEM
Formal Logic	Propositions, predicate logic, first-order logic, ...	Humanities
Global Facts	Extreme poverty, literacy rates, life expectancy, ...	Other
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...	STEM
High School Chemistry	Chemical reactions, ions, acids and bases, ...	STEM
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...	STEM
High School European History	Renaissance, reformation, industrialization, ...	Humanities
High School Geography	Population migration, rural land-use, urban processes, ...	Social Sciences
High School Gov't and Politics	Branches of government, civil liberties, political ideologies, ...	Social Sciences
High School Macroeconomics	Economic indicators, national income, international trade, ...	Social Sciences
High School Mathematics	Pre-algebra, algebra, trigonometry, calculus, ...	STEM
High School Microeconomics	Supply and demand, imperfect competition, market failure, ...	Social Sciences
High School Physics	Kinematics, energy, torque, fluid pressure, ...	STEM
High School Psychology	Behavior, personality, emotions, learning, ...	Social Sciences
High School Statistics	Random variables, sampling distributions, chi-square tests, ...	STEM
High School US History	Civil War, the Great Depression, The Great Society, ...	Humanities
High School World History	Ottoman empire, economic imperialism, World War I, ...	Humanities
Human Aging	Senescence, dementia, longevity, personality changes, ...	Other
Human Sexuality	Pregnancy, sexual differentiation, sexual orientation, ...	Social Sciences
International Law	Human rights, sovereignty, law of the sea, use of force, ...	Humanities
Jurisprudence	Natural law, classical legal positivism, legal realism, ...	Humanities
Logical Fallacies	No true Scotsman, base rate fallacy, composition fallacy, ...	Humanities
Machine Learning	SVMs, VC dimension, deep learning architectures, ...	STEM
Management	Organizing, communication, organizational structure, ...	Other
Marketing	Segmentation, pricing, market research, ...	Other
Medical Genetics	Genes and cancer, common chromosome disorders, ...	Other
Miscellaneous	Agriculture, Fermi estimation, pop culture, ...	Other
Moral Disputes	Freedom of speech, addiction, the death penalty, ...	Humanities
Moral Scenarios	Detecting physical violence, stealing, externalities, ...	Humanities
Nutrition	Metabolism, water-soluble vitamins, diabetes, ...	Other
Philosophy	Skepticism, phronesis, skepticism, Singer's Drowning Child, ...	Humanities
Prehistory	Neanderthals, Mesoamerica, extinction, stone tools, ...	Humanities
Professional Accounting	Auditing, reporting, regulation, valuation, ...	Other
Professional Law	Torts, criminal law, contracts, property, evidence, ...	Humanities
Professional Medicine	Diagnosis, pharmacotherapy, disease prevention, ...	Other
Professional Psychology	Diagnosis, biology and behavior, lifespan development, ...	Social Sciences
Public Relations	Media theory, crisis management, intelligence gathering, ...	Social Sciences
Security Studies	Environmental security, terrorism, weapons of mass destruction, ...	Social Sciences
Sociology	Socialization, cities and community, inequality and wealth, ...	Social Sciences
US Foreign Policy	Soft power, Cold War foreign policy, isolationism, ...	Social Sciences
Virology	Epidemiology, coronaviruses, retroviruses, herpesviruses, ...	Other
World Religions	Judaism, Christianity, Islam, Buddhism, Jainism, ...	Humanities

Table 2: Summary of all 57 tasks.

Instruction-Following

Instruction: Given an address and city, come up with the zip code.
Input:
Address: 123 Main Street, City: San Francisco
Output: 94105

Instruction: I am looking for a job and I need to fill out an application form. Can you please help me complete it?
Input:
Application Form:
Name:_____ Age:_____ Sex:_____
Phone Number:_____ Email Address:_____
Education:_____ ...
Output:
Name: John Doe Age: 25 Sex: Male
Phone Number: ...

Instruction: How to write a code for converting degrees fahrenheit to celsius.
Input: Null
Output:

```
def convert_fahrenheit_to_celsius(fahr):
    celsius = (fahr - 32) * 5 / 9
    return celsius
```

Instruction: Write a letter from the perspective of a cat.
Input: Null
Output:
Dear [Owner], I am writing to you today because I have a problem. You see, I've been living with you for the past [Number of Years] and I love it here. The food is great, the bed is comfy, and there are plenty of toys to play with. But lately, I've noticed that you're not as affectionate towards me as you used to be. In fact, sometimes when I try to cuddle up next to you on the couch, you push me away ...

Conversational

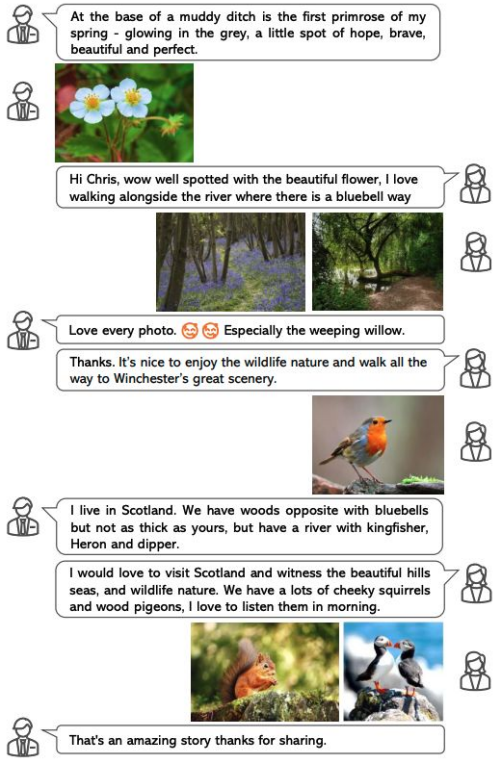


Figure 1: An example of human conversations in our MMDialog dataset. They are talking about scenery and wildlife with both text and various images.

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 1: A conversation from the CoQA dataset. Each turn contains a question (Q_i), an answer (A_i) and a rationale (R_i) that supports the answer.

Any Issues?

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.

MT-Bench and Chatbot Arena

- Authors want to create a benchmark that tests LLMs with **multi-turn** conversations and **diverse, complex, open-ended** questions
- They introduce
  - MT-Bench: Assess multi-turn conversation capabilities
  - Chatbot Arena: Test models on diverse, open-ended questions

MT-Bench

- Benchmark consisting of **80 high-quality multi-turn** conversations
- Covers common uses cases across 8 categories: writing, roleplay, extraction, reasoning, math, coding, STEM knowledge, and social science knowledge
- Can view all questions and model outputs here

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.

Chatbot Arena

- Crowdsourcing benchmark platform with anonymous battles
- Users get to **pick a question** to ask two anonymous models
- User then votes for their preferred response
- Allowing users to ask questions helps improve diversity of the question set
- Can try it out yourself: https://lmarena.ai/

Expand to see the descriptions of 73 models

Model A

$$\int_{-1}^{1} \frac{1}{x}\,dx = \lim_{\epsilon \to 0^-} \int_{-1}^{\epsilon} \frac{1}{x}\,dx$$

$$\int_{0}^{1} \frac{1}{x}\,dx = \lim_{\delta \to 0^+} \int_{\delta}^{1} \frac{1}{x}\,dx$$

Evaluating these integrals, we get:

$$\int_{-1}^{\epsilon} \frac{1}{x}\,dx = \ln|x|\Big|_{-1}^{\epsilon} = \ln|\epsilon| - \ln|-1| = \ln|\epsilon| + \ln 1 = \ln|\epsilon|$$

$$\int_{\delta}^{1} \frac{1}{x}\,dx = \ln|x|\Big|_{\delta}^{1} = \ln 1 - \ln|\delta| = -\ln|\delta|$$

Adding these together, we get:

$$\text{PV}\int_{-1}^{1} \frac{1}{x}\,dx = (\ln|\epsilon|) + (-\ln|\delta|)$$

As $\epsilon \to 0^-$ and $\delta \to 0^+$, both $\ln|\epsilon|$ and $\ln|\delta|$ tend to $-\infty$. Therefore, the Cauchy principal value of the integral is:

$$\text{PV}\int_{-1}^{1} \frac{1}{x}\,dx = 0$$

Thus, the Cauchy principal value of the integral of $\frac{1}{x}$ from -1 to 1 is 0.

Model B

∫(-1 to 1) (1/x) dx = ∫(-1 to 0-) (1/x) dx + ∫(0+ to 1) (1/x) dx

where 0- and 0+ denote the limits as x approaches 0 from the left and right, respectively.

Evaluating the Improper Integral

Evaluating the improper integral, we get:

∫(-1 to 0-) (1/x) dx = lim(a → 0-) ∫(-1 to a) (1/x) dx
= lim(a → 0-) [ln|x|](-1 to a)
= lim(a → 0-) [ln|a| - ln|-1|]
= -∞ (since ln|a| → -∞ as a → 0-)

∫(0+ to 1) (1/x) dx = lim(b → 0+) ∫(b to 1) (1/x) dx
= lim(b → 0+) [ln|x|](b to 1)
= lim(b → 0+) [ln|1| - ln|b|]
= ∞ (since ln|b| → -∞ as b → 0+)

Since one part of the integral diverges to -∞ and the other part diverges to ∞, the improper integral does not converge.

Conclusion

The integral of 1/x from -1 to 1 does not converge due to the singularity at x = 0.

👈 A is better | 👉 B is better | 🤝 Tie | 👎 Both are bad

Leaderboard

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	ChatGPT-4o-latest (2024-08-08)	1316	+4/-3	31148	OpenAI	Proprietary	2023/10
2	Gemini-1.5-Pro-Exp-0827	1300	+4/-4	22844	Google	Proprietary	2023/11
2	Gemini-1.5-Pro-Exp-0801	1298	+4/-4	26110	Google	Proprietary	2023/11
2	Grok-2-08-13	1294	+4/-4	16215	xAI	Proprietary	2024/3
5	GPT-4o-2024-05-13	1285	+3/-2	86306	OpenAI	Proprietary	2023/10
6	GPT-4o-mini-2024-07-18	1274	+4/-4	26088	OpenAI	Proprietary	2023/10
6	Claude 3.5 Sonnet	1270	+3/-3	56674	Anthropic	Proprietary	2024/4
6	Gemini-1.5-Flash-Exp-0827	1268	+5/-4	16780	Google	Proprietary	2023/11
6	Grok-2-Mini-08-13	1267	+4/-4	16731	xAI	Proprietary	2024/3
6	Meta-Llama-3.1-405b-Instruct	1266	+4/-4	27397	Meta	Llama 3.1 Community	2023/12

LLM as a Judge

LLM-as-a-Judge

Question:
What are key factors I should consider when buying a home?

Model A:
Key factors include location, budget, and property condition. Location affects value and convenience……

Model B:
When buying a home, think about location, finances, and the property's overall condition…..

Which response is better?

LLM Judge: GPT-4, GPT-3.5, Claude

Judgement:

I think A provides… Therefore, A is better.

Scalable ✅
Explainable ✅

How does the LLM judge compare answers?

Pairwise Comparison

1. LLM judge presented with question
2. Asked to determine which one is better or declare a tie

Lacks scalability when number of players increase

```
[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. You should choose the assistant that
follows the user's instructions and answers the user's question better. Your evaluation
should consider factors such as the helpfulness, relevance, accuracy, depth, creativity,
and level of detail of their responses. Begin your evaluation by comparing the two
responses and provide a short explanation. 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 5: The default prompt for pairwise comparison.

Single Answer Grading

1. LLM judge presented with question
2. Asked to assign a score to a single answer

```
[System]
Please act as an impartial judge and evaluate the quality of the response provided by an
AI assistant to the user question displayed below. Your evaluation should consider factors
such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of
the response. Begin your evaluation by providing a short explanation. Be as objective as
possible. After providing your explanation, please rate the response on a scale of 1 to 10
by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{question}

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

Figure 6: The default prompt for single answer grading.

Unable to discern specific differences between pairs and unstable results

Reference-guided grading

1. LLM judge presented with question
2. Uses reference answer and looks for closest response

```
[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 a reference answer, assistant A's answer,
and assistant B's answer. Your job is to evaluate which assistant's answer is better.
Begin your evaluation by comparing both assistants' answers with the reference 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 Reference Answer]
{answer_ref}
[The End of Reference Answer]

[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 8: The prompt for reference-guided pairwise comparison.

Limitations

Position Bias

- LLM exhibits propensity to favor certain positions over others

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%

Solution: Swapping positions, Few-shot judge

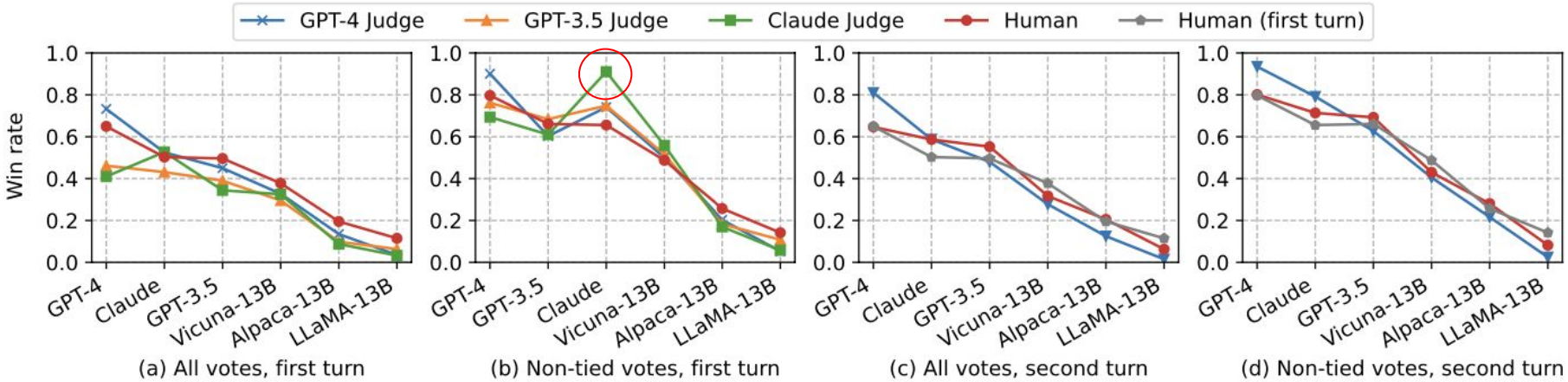
Verbosity Bias

- LLM judge always tend to favors longer, verbose reponses

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

Self-enhancement bias

- LLM judges may favor the answers generated by themselves



Limitation: unable to detect self-enhancement bias due to limited data and small differences

Math and reasoning grading abilities

- LLM judge shows failure in grading questions they already know answers to

Solution:
chain-of-thought-judge, reference guided judge

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.

Setup and Results

MT-Bench Setup

- **80 questions**
- **6 Models**
  - GPT-4, GPT-3.5, Claude-V1, Vicuna-13B, Alpaca-13B, LLaMa-13B
- **2 Judges**
  - LLM Judges and Expert-Level Human Labelers
- **Data Collection**
  - 3K votes for all questions

Evaluation Method: LLM judges evaluate all pairs and human evaluate at least 20 random multi-turn questions

MT-Bench Setup

Question:
What are key factors I should consider when buying a home?

Model A:
Key factors include location, budget, and property condition. Location affects value and convenience……

Model B:
When buying a home, think about location, finances, and the property's overall condition…..

User Follow-up Question:
Make the following changes: 1. Make the tone sound more casual 2.Response should be less than 10 words

Model A:
Consider location, budget, and property condition.

Model B:
Focus on location, budget, and home condition.

A is better | B is better | Tie | Skip(⅕)

Chatbot Arena Setup

- **8 Models**
  - GPT-4, GPT-3.5, Claude-V1, Vicuna-7B/13B, Koala-13B, Alpaca-13B, LLaMa-13B, Dolly-12B
- **2 Judges**
  - LLM Judges, Collected Crowd Judges
- **2114 unique IP addresses**
- **Data Collection**
  - 3K randomly sampled single-turn votes from 30K arena data

Agreement

Agreement: probability of randomly selected individuals (but not identical) of each type agreeing on a randomly selected question

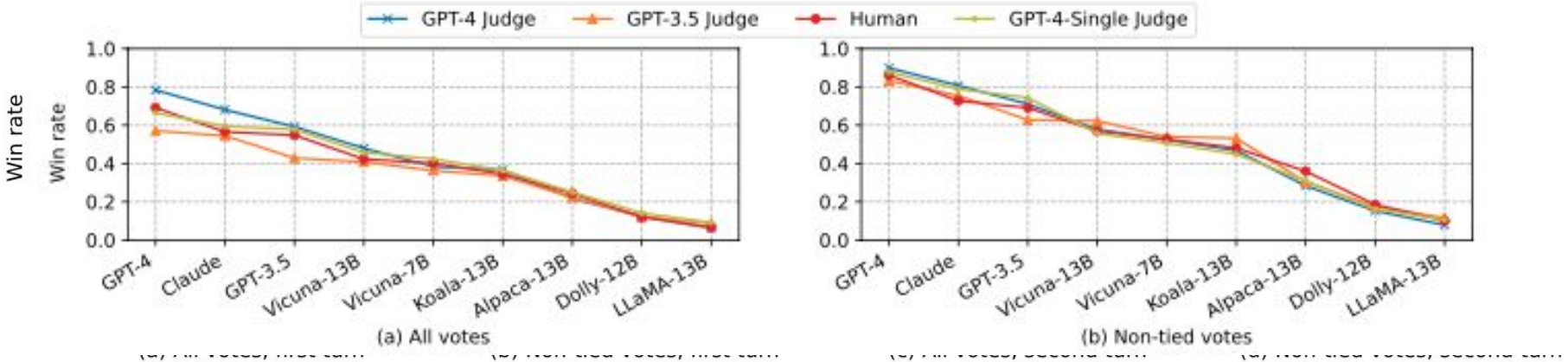
Setup	S1 (Random = 33%)				S2 (Random = 50%)			
Judge	G4-S	G3.5	C	H	G4-S	G3.5	C	H
G4	72% 2968	66% 3061	66% 3062	64% 3066	95% 1967	94% 1788	95% 1712	**87%** 1944
G4-S	-	60% 2964	62% 2964	60% 2968	-	89% 1593	91% 1538	85% 1761
G3.5	-	-	68% 3057	54% 3061	-	-	96% 1497	83% 1567
C	-	-	-	53% 3062	-	-	-	84% 1475

Setup	S2 (R = 50%)	
Judge	4-Single	Human
G4-Pair	95% 727	**85%** 864
G4-Single	-	84% 776
Human	-	**82%** 474

GPT-4 - human agreement higher than human-human agreement

Win Rate

Average Win Rate: average of win rates against all other players



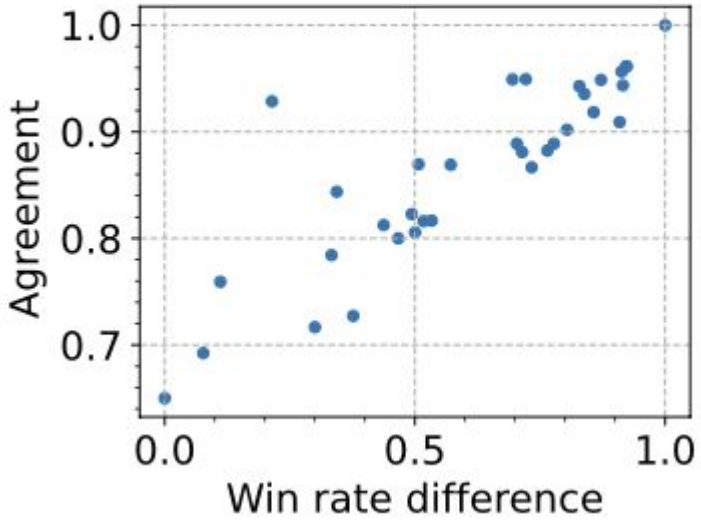
LLM Judge curves align closely with human judge

Win Rate

Table 7: Category-wise win rate of models.

Model	Writing	Roleplay	Reasoning	Math	Coding	Extraction	STEM	Humanities
GPT-4	61.2%	67.9%	49.3%	66.1%	56.3%	66.2%	76.6%	72.2%
GPT-3.5	50.9%	60.6%	32.6%	63.8%	55.0%	48.8%	52.8%	53.8%
Vicuna-13B	39.7%	39.2%	20.1%	18.0%	36.9%	29.2%	47.0%	47.5%
LLaMA-13B	15.1%	15.1%	7.8%	7.5%	2.1%	9.3%	6.8%	10.1%

Win Rate Difference and Agreement



Model Variant Performance

Table 8: Evaluation results of several model variants.

Model	#Training Token	MMLU (5-shot)	TruthfulQA (0-shot)	MT-Bench Score (GPT-4)
LLaMA-7B	1T	35.2	0.22	2.74
LLaMA-13B	1T	47.0	0.26	2.61
Alpaca-7B	4.4M	40.1	0.26	4.54
Alpaca-13B	4.4M	48.1	0.30	4.53
Vicuna-7B (selected)	4.8M	37.3	0.32	5.95
Vicuna-7B (single)	184M	44.1	0.30	6.04
Vicuna-7B (all)	370M	47.1	0.32	6.00
Vicuna-13B (all)	370M	**52.1**	**0.35**	**6.39**
GPT-3.5	-	70.0	-	7.94
GPT-4	-	**86.4**	-	**8.99**

Limitations and Future Work

- Limitations
  - Neglects safety
  - Ignores multiple dimensions of helpfulness

- Future Work
  1) Benchmarking chatbots at scale with a broader set of categories
  2) Open-source LLM judge aligned with human preference
  3) Enhancing LLM judges' math/reasoning capability

Summary

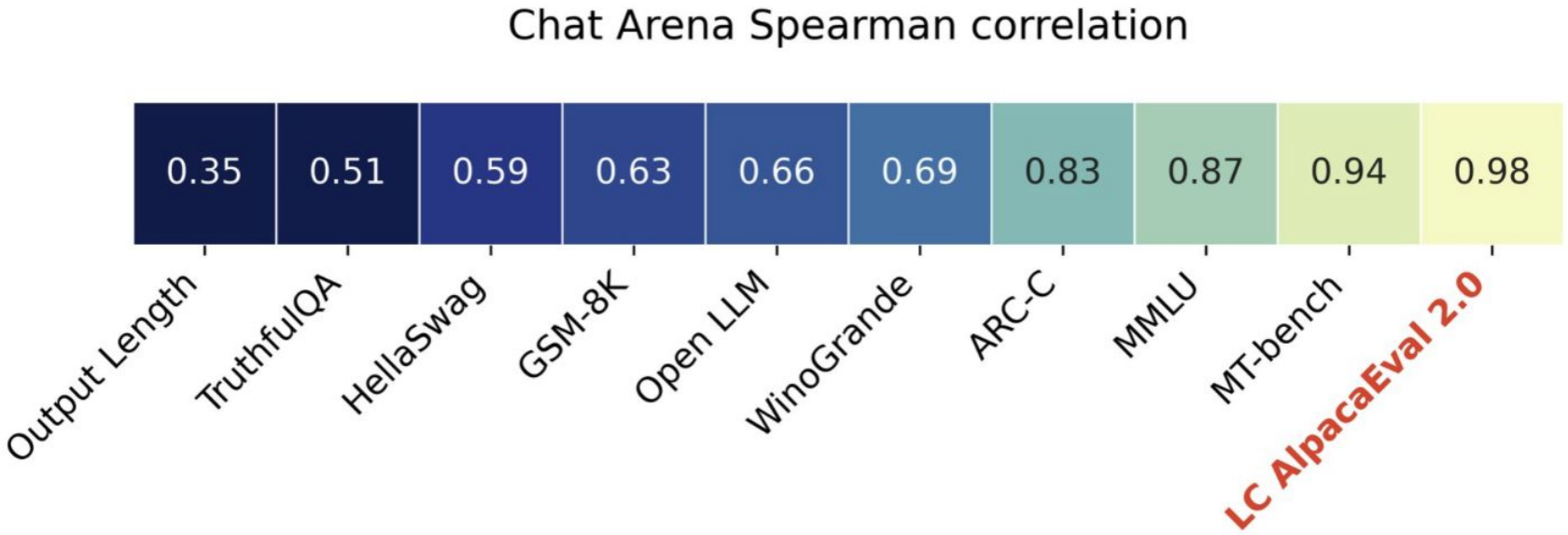
- Performance
  - GPT-4 vs. Human Agreement (85%) > human-human agreement (81%)
  - GPT-4 helps human make better judgements - 75% deemed GPT's judgement's reasonable
  - Humans changed choices based on GPT-4's judgements (34% of time)
  - LLM judges' win rates closely match human judges

- Findings
  - GPT-4 shows high agreement with human judgements
  - GPT-4 helps in improving human judgements
  - LLM judges generally align with human preferences, especially in multi-turn settings

Related Work

Benchmark for benchmarks!

- Chatbot Arena can’t be used for model development because it relies on live human evaluations. Other automatic metrics try to approximate it.



LC AlpacaEval 2.0

- **AlpacaEval:** LLM-based automated evaluation metric that controls for biases such as position bias however it was still biased towards length
- **Length-controlled AlpacaEval 2.0** controls for length

What would the AlpacaEval metric be, if the outputs of all models had the same length as those of the baseline?

Logistic Regression

- Use a logistic regression model to estimate bias towards lengthy responses
- Then zero out the length term to calculate the length bias
- Utilize this to calculate an adjusted win rate

$$q_{\theta,\phi,\psi}(y = m|z_m, z_b, x) :=$$
$$\text{logistic}\left(\underbrace{\theta_m - \theta_b}_{\text{Model}} + \underbrace{\phi_{m,b} \cdot \tanh\left(\frac{\text{len}(z_m) - \text{len}(z_b)}{\text{std}(\text{len}(z_m) - \text{len}(z_b))} \right)}_{\text{Length}} + \underbrace{(\psi_m - \psi_b)\gamma_x}_{\text{Instruction}} \right)$$

New Metric is not biased towards length

	AlpacaEval			Length-controlled AlpacaEval		
	concise	standard	verbose	concise	standard	verbose
gpt4_1106_preview	22.9	50.0	64.3	41.9	50.0	51.6
Mixtral-8x7B-Instruct-v0.1	13.7	18.3	24.6	23.0	23.7	23.2
gpt4_0613	9.4	15.8	23.2	21.6	30.2	33.8
claude-2.1	9.2	15.7	24.4	18.2	25.3	30.3
gpt-3.5-turbo-1106	7.4	9.2	12.8	15.8	19.3	22.0
alpaca-7b	2.0	2.6	2.9	4.5	5.9	6.8

Relation to SimPO

Method	Mistral-Base (7B) Setting					Mistral-Instruct (7B) Setting				
	AlpacaEval 2		Arena-Hard	MT-Bench		AlpacaEval 2		Arena-Hard	MT-Bench	
	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4	LC (%)	WR (%)	WR (%)	GPT-4 Turbo	GPT-4
DPO	15.1	12.5	10.4	5.9	7.3	26.8	24.9	16.3	6.3	7.6
SimPO	21.5	20.8	16.6	6.0	7.3	32.1	34.8	21.0	6.6	7.6
w/o LN	11.9	13.2	9.4	5.5	7.3	19.1	19.7	16.3	6.4	7.6
$\gamma = 0$	16.8	14.3	11.7	5.6	6.9	30.9	34.2	20.5	6.6	7.7

Questions?