Judging Models

Situation

- Aligned Models with RLHF and Supervised instruction finetuned are preferred
- Benchmarks like MMLU cannot tell the difference between aligned models and base models
    - Discrepancy between user perception of usefulness and criteria of benchmarks
- Hypothesis: Arises due to benchmarks only measure core capability like (multi choice, retrieval questions) and not open-ended questions

Benchmarks

- Human Rating
    - MT-bench
    - Chatbot-Arena
- MT-bench
    - Series of 80 open-ended ,multi turn questions
        - writing, roleplay, extraction, reasoning, math, coding, knowledge I (STEM), and knowledge II (humanities/social science)
- Chatbot-Arena
    - Anonymous battles between Chatbots
    - Users rate responses of 2 bots
    - Captures wide range of interests of users

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.

LLM as a Judge

- Pairwise comparison
  - Two answers: Declare winner or tie
- Single answer grading
  - Assign score to answer
- Reference-guided grading
  - Reference solution provided, e.g. math
- Advantages
  - Fast, without human interaction, provide explanations

Limitations

Position bias:

- Bias towards first/second answer
- Raname: renamed the models in the prompt

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%

Fix: Swapping positions, only call win if preferred in both orders/assign positions randomly.

Few-Shot-Judge: Enhance consistency (not imply Accuracy, increased API cost)

Limitations

- Verbosity bias: favors longer, verbose responses, even if they are not as clear
- LLM judges are able to correctly judge identical answers

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

- Self-enhancement bias: LLM judges may favor the answers generated by themselves
    - GPT-4 favors itself with a 10% higher
    - Claude-v1 favors itself with a 25% higher win rate

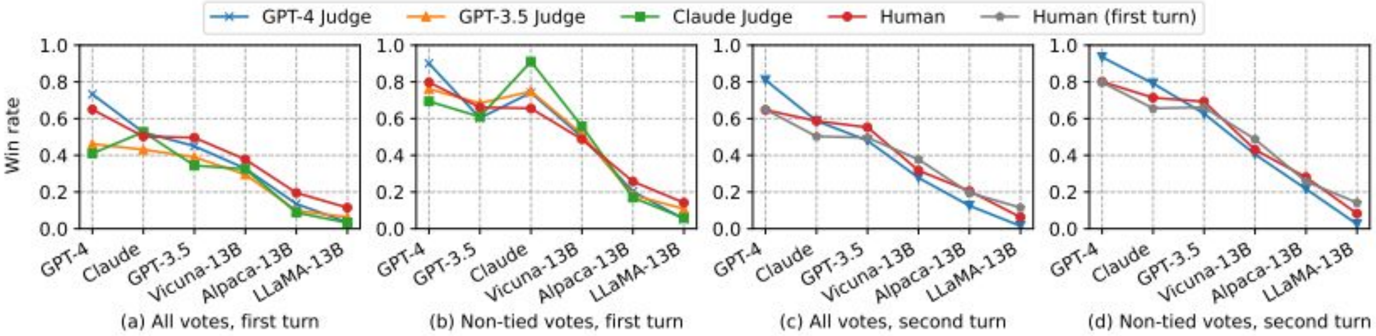

Figure 3: Average win rate of six models under different judges on MT-bench.

Fix: Fine-tuned judge model: Train on arena data to act as judge -> Promising results

Limitations

- Limited capability in grading math and reasoning questions
    - Lacks ability to grading math problems it could solve itself

Fix: Chain-of-thought and reference-guided judge

- Often same mistake as is given answer
- Let judge solve the questions itself and then display is as reference in the judge prompt (reduces failure rate from 70% to 15%)

Agreement Evaluation

- Check Agreement between LLM judges and humans
  - Among humans too for MT-bench

- MT-Bench
  - Generated answers of all 80 questions for all models
  - 58 expert-level humans (graduate students) to rate 20 answers
  - Roughly 3K votes
- Chatbot arena
  - Randomly 3K selected votes

Agreement Evaluation

- High agreement between GPT-4 and human majority
  - Higher when the win rates of the models differ

- 75% of GPT-4 judgement considered reasonable
  - 34% of humans willing to change their choice

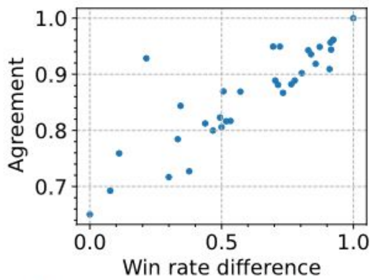

Figure 2: Agreement and win rate difference. Each point corresponds to a model pair and counts only the non-tie votes between the two models. The x-axis value is the win rate difference between the two models. The y-axis value is the GPT-4 and human agreement.

Agreement Evaluation

- Pair, evaluate two answers at once, Single, evaluate one answer independently

Table 5: Agreement between two types of judges on MT-bench. "G4-Pair" and "G4-Single" denote GPT-4 with pairwise comparison and single-answer grading respectively. The single-answer grading can be converted into pairwise comparison results for calculating the agreement. We report two setups: "S1" includes non-tie, tie, and inconsistent (due to position bias) votes and counts inconsistent as tie; "S2" only includes non-tie votes. The agreement between two random judges under each setup is denoted as "R=". The top value in each cell is the agreement, and the bottom gray value is #votes.

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

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

Agreement Evaluation

Table 6: Agreement between two types of judges on Chatbot Arena. "G4-S" denotes GPT-4 with single-answer grading. "G4", "G3.5" and "C" denote GPT-4, GPT-3.5, and Claude with pairwise comparison, respectively. "H" denotes human. The remaining of table follows the same format as Table 5.

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

Agreement Evaluation

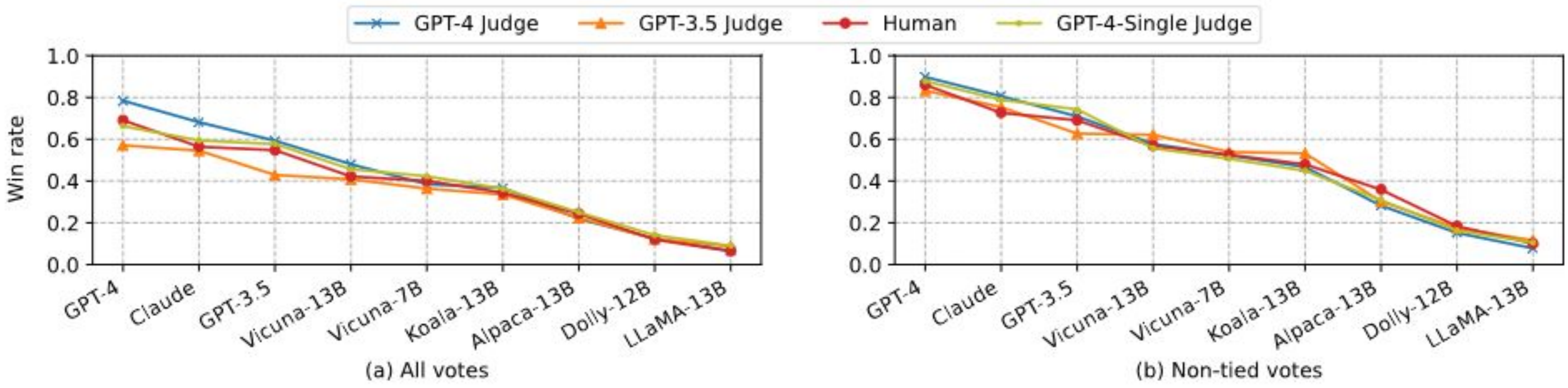

Figure 4: Average win rate of nine models under different judges on Chatbot Arena.

Human Preference Benchmark and Standardized Benchmark

- Recommended to use both

- No single evaluation method is enough to determine the quality of a model

- Fine-tuning with high-quality dialog improves MMLU

- A small high-quality conversation can teach a model GPT-4s preferred style