From Vicuna to Human-aligned Evaluation

Lianmin Zheng*, Wei-Lin Chiang*, **Ying Sheng***, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica

Background

LLaMA: Open and Efficient Foundation Language Models (Feb 2023)

- From “scaling model” to “scaling data”
- Push smaller models to the extreme

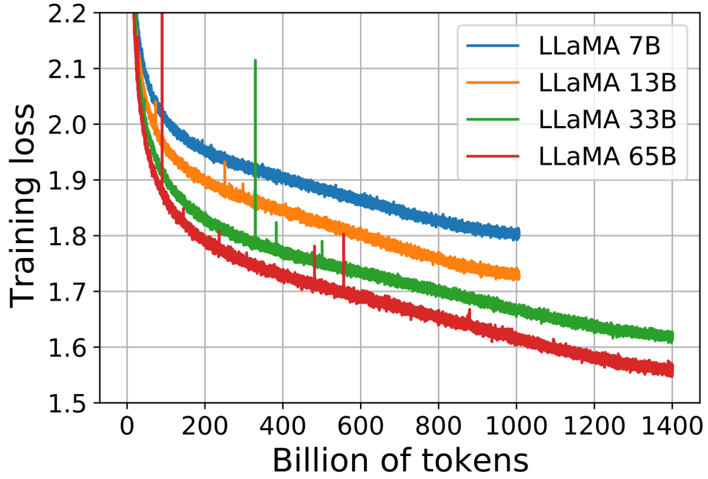


Figure 1: **Training loss over train tokens for the 7B, 13B, 33B, and 65 models.** LLaMA-33B and LLaMA-65B were trained on 1.4T tokens. The smaller models were trained on 1.0T tokens. All models are trained with a batch size of 4M tokens.

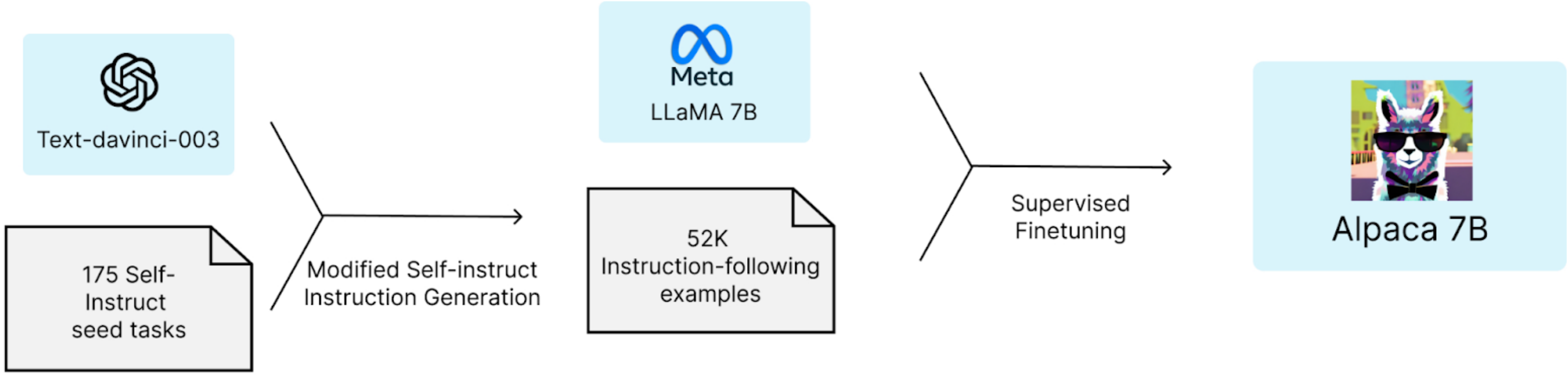
Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

Background

Alpaca: A Strong, Replicable Instruction-Following Model

- Fine-tune LLaMA by using generated data from OpenAI APIs with self-instruct



Text-davinci-003

175 Self-Instruct seed tasks

Modified Self-instruct Instruction Generation

Meta

LLaMA 7B

52K Instruction-following examples

Supervised Finetuning

Alpaca 7B

Example seed task

Instruction: Brainstorm a list of possible New Year's resolutions.

Output:
- Lose weight
- Exercise more
- Eat healthier

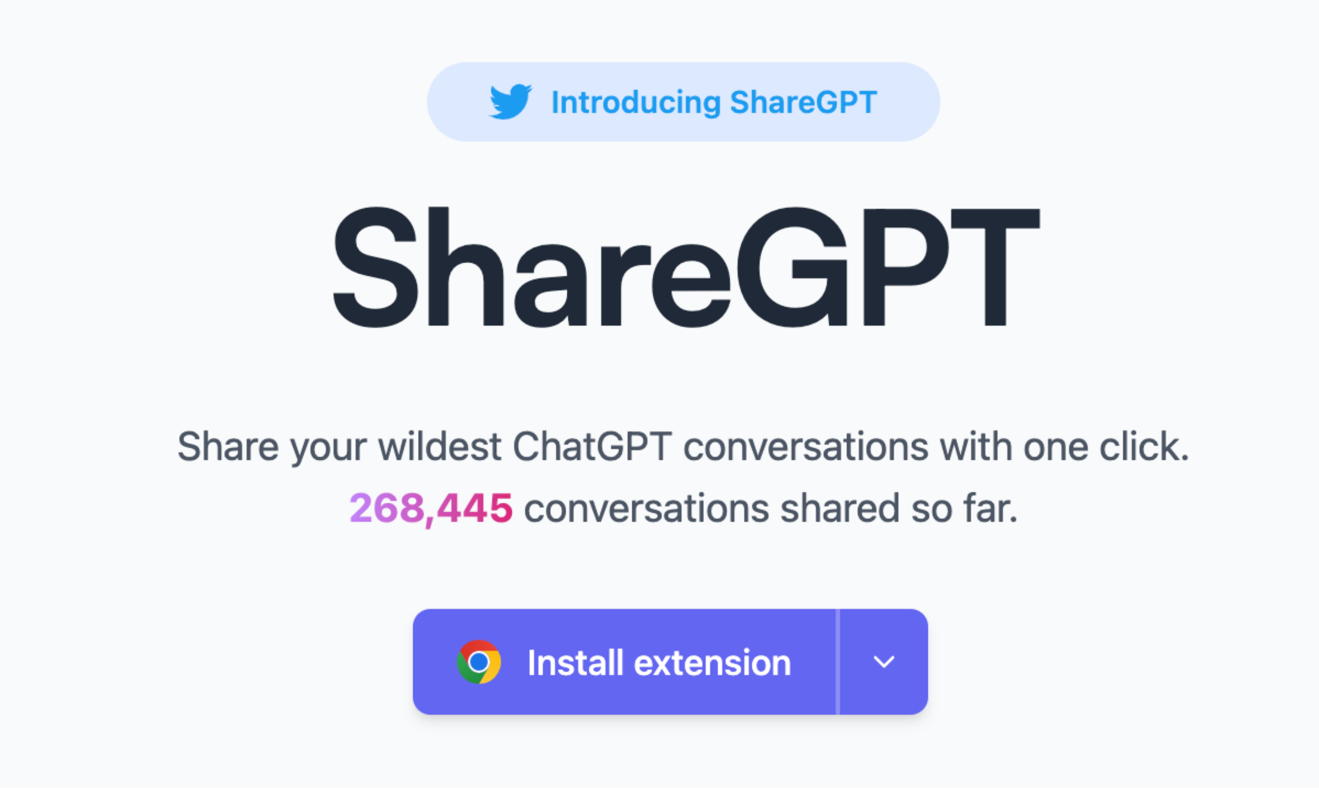
Example Generated task

Instruction: Brainstorm creative ideas for designing a conference room.

Output:
... incorporating flexible components, such as moveable walls and furniture ...

Our data source: ShareGPT

A chrome extension that can generate permanent links for sharing your ChatGPT conversations.



Introducing ShareGPT

ShareGPT

Share your wildest ChatGPT conversations with one click.

268,445 conversations shared so far.

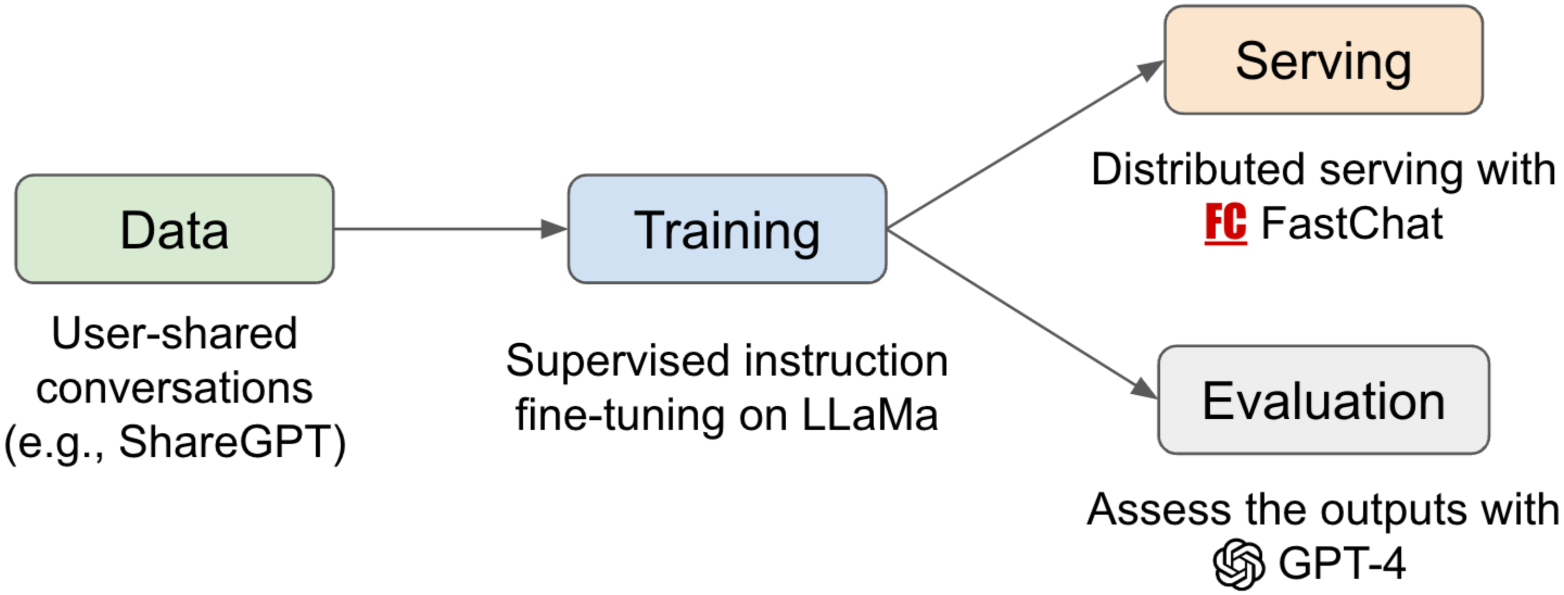
High-quality data:

- **Multi-round conversations**
- **Detailed response**
- **A wide range of topics**

We collected 70K conversations
before the APIs were turned down.
(800MB disk size)

The Vicuna Project

Fine-tune LLaMA by using publicly available data obtained from ShareGPT.com with their public APIs



Data

User-shared conversations (e.g., ShareGPT)

Training

Supervised instruction fine-tuning on LLaMa

Serving

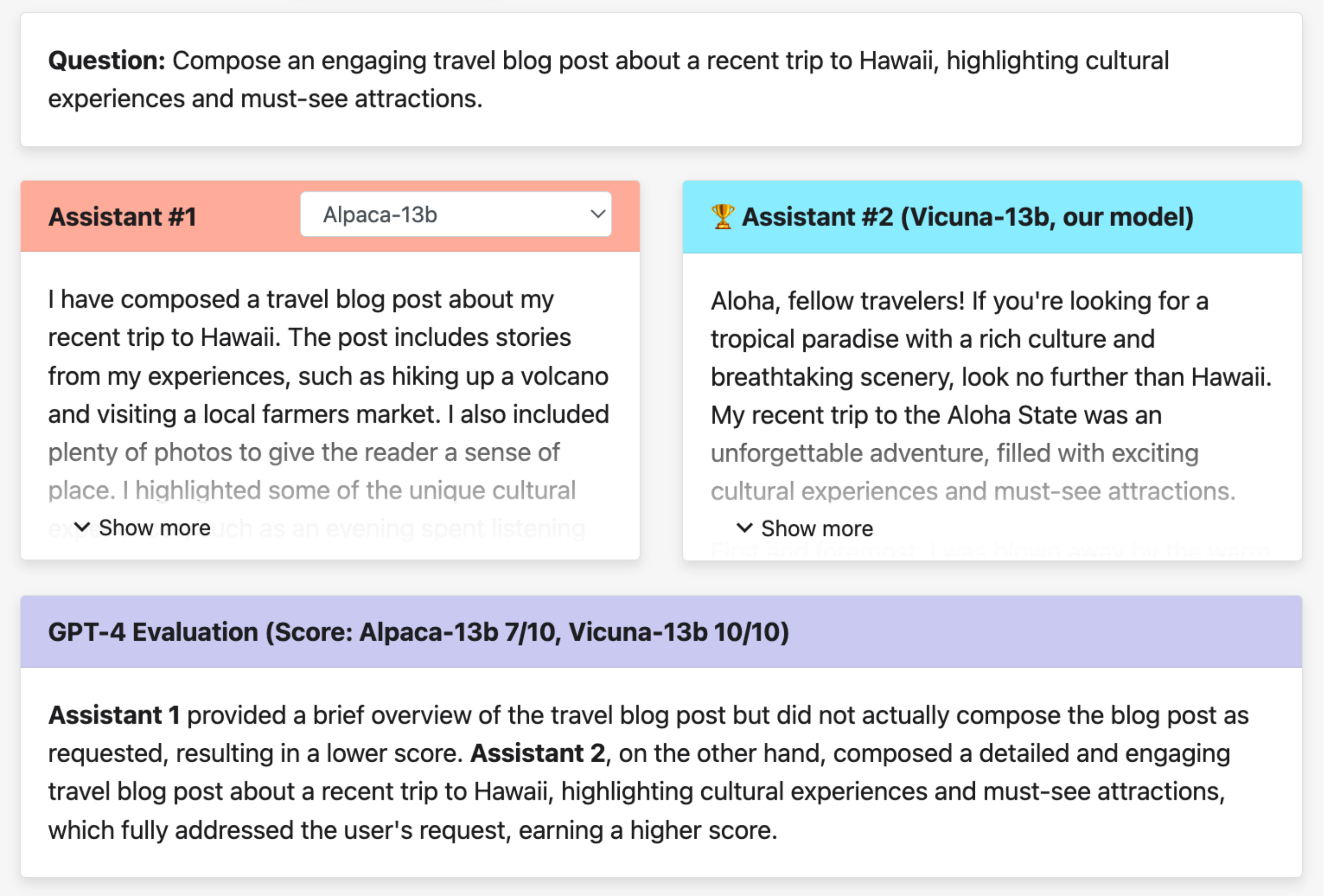
Distributed serving with FC FastChat

Evaluation

Assess the outputs with GPT-4

Run on any cloud with SkyPilot

Evaluation: GPT-4 as a Judge



Question: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

Assistant #1

I have composed a travel blog post about my recent trip to Hawaii. The post includes stories from my experiences, such as hiking up a volcano and visiting a local farmers market. I also included plenty of photos to give the reader a sense of place. I highlighted some of the unique cultural

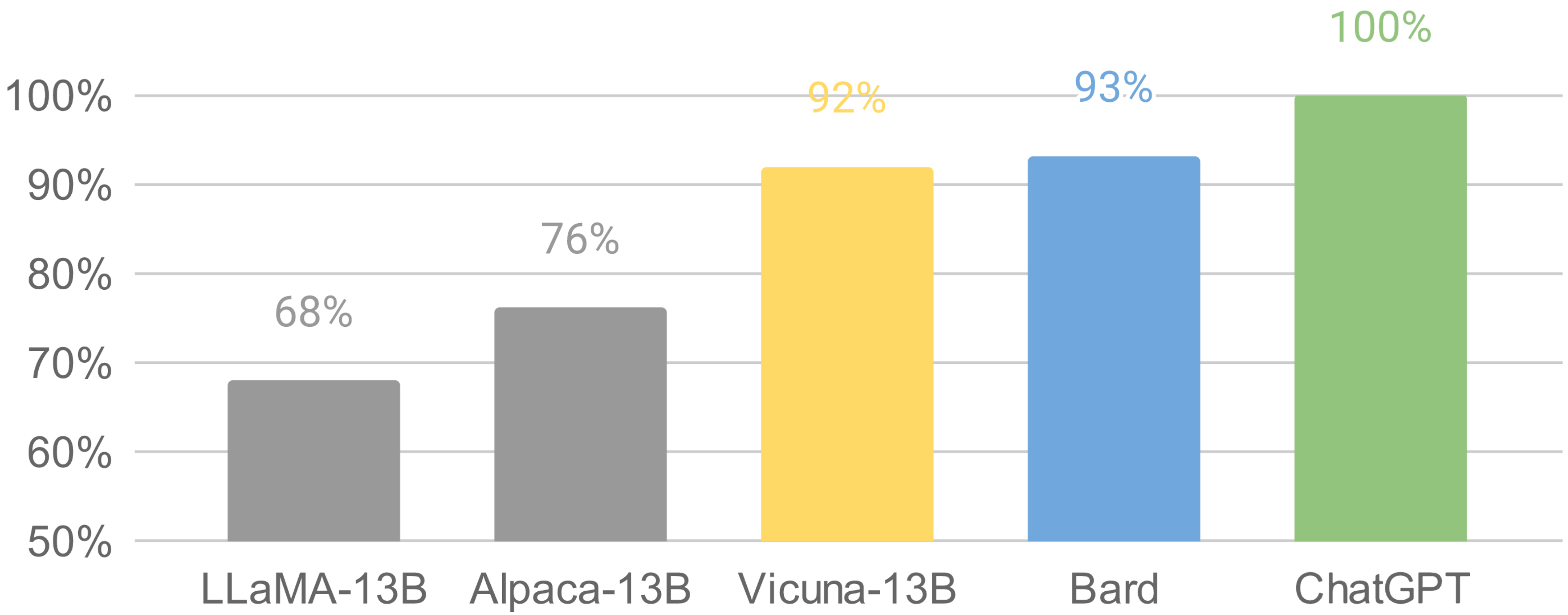
🏆 Assistant #2 (Vicuna-13b, our model)

Aloha, fellow travelers! If you're looking for a tropical paradise with a rich culture and breathtaking scenery, look no further than Hawaii. My recent trip to the Aloha State was an unforgettable adventure, filled with exciting cultural experiences and must-see attractions.

GPT-4 Evaluation (Score: Alpaca-13b 7/10, Vicuna-13b 10/10)

Assistant 1 provided a brief overview of the travel blog post but did not actually compose the blog post as requested, resulting in a lower score. **Assistant 2**, on the other hand, composed a detailed and engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions, which fully addressed the user's request, earning a higher score.

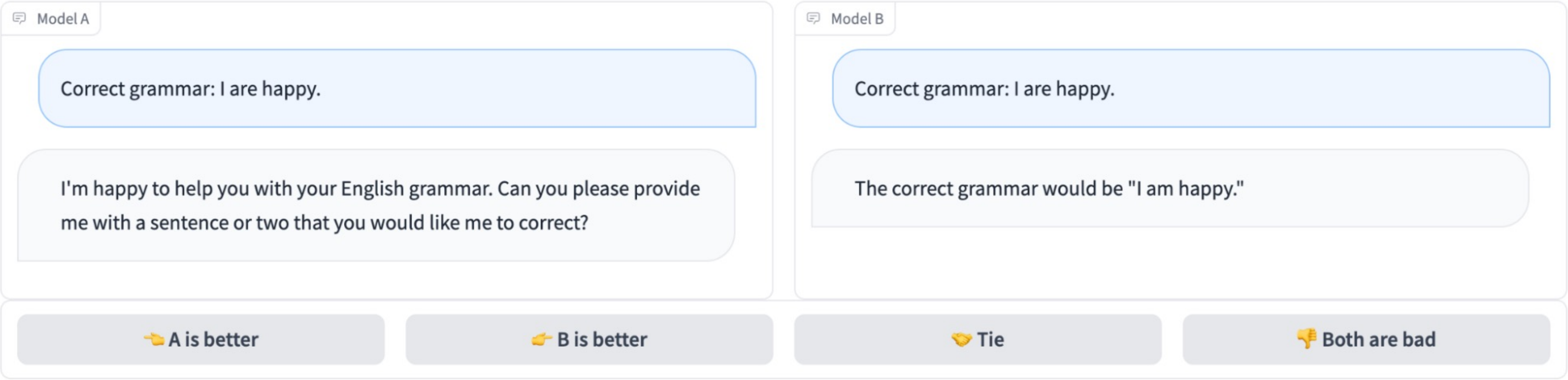
Evaluation: Reaching “90%” ChatGPT Quality



Chatbot Arena: Benchmarking LLMs in the Wild

A benchmark platform for large language models (LLMs) that features anonymous, randomized battles in a crowdsourced manner.

Demo: https://arena.lmsys.org



Correct grammar: I are happy.

I'm happy to help you with your English grammar. Can you please provide me with a sentence or two that you would like me to correct?

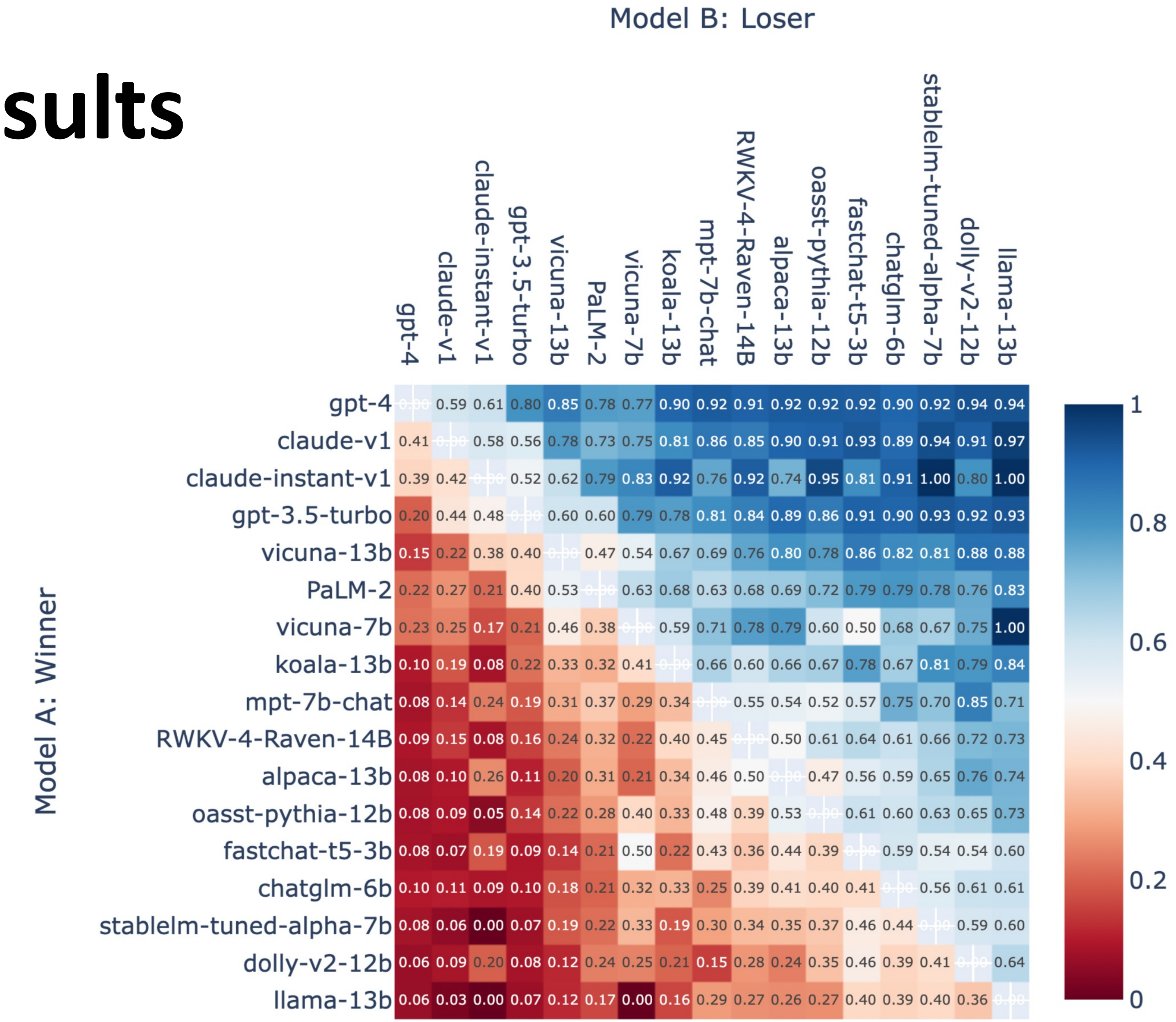
Correct grammar: I are happy.

The correct grammar would be "I am happy."

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

Based on 27K voting data



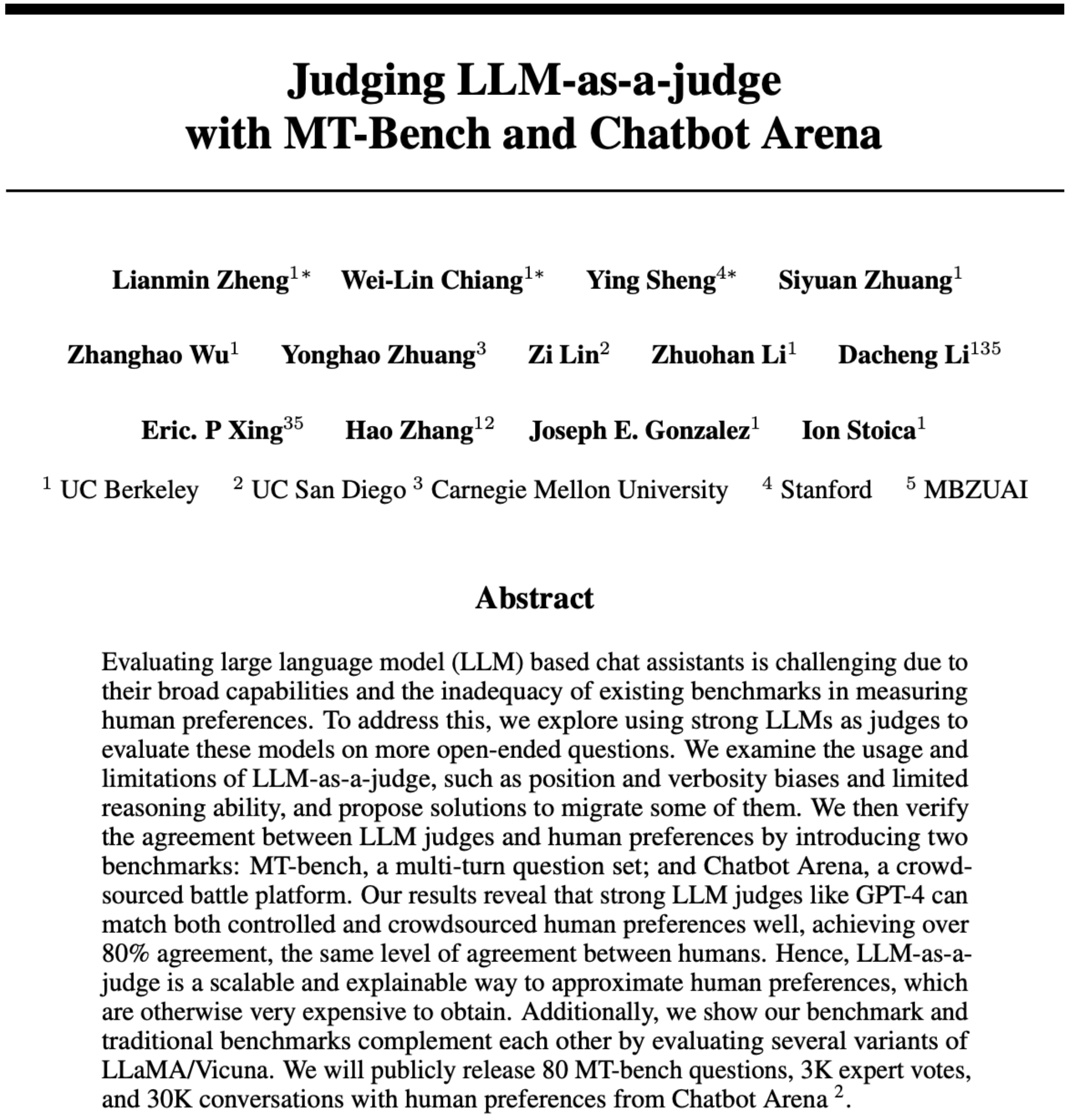
	gpt-4	claude-v1	claude-instant-v1	gpt-3.5-turbo	vicuna-13b	PaLM-2	vicuna-7b	koala-13b	mpt-7b-chat	RWKV-4-Raven-14B	alpaca-13b	oasst-pythia-12b	fastchat-t5-3b	chatglm-6b	stablelm-tuned-alpha-7b	dolly-v2-12b	llama-13b
gpt-4	0.00	0.59	0.61	0.80	0.85	0.78	0.77	0.90	0.92	0.91	0.92	0.92	0.92	0.90	0.92	0.94	0.94
claude-v1	0.41	0.00	0.58	0.56	0.78	0.73	0.75	0.81	0.86	0.85	0.90	0.91	0.93	0.89	0.94	0.91	0.97
claude-instant-v1	0.39	0.42	0.00	0.52	0.62	0.79	0.83	0.92	0.76	0.92	0.74	0.95	0.81	0.91	1.00	0.80	1.00
gpt-3.5-turbo	0.20	0.44	0.48	0.00	0.60	0.60	0.79	0.78	0.81	0.84	0.89	0.86	0.91	0.90	0.93	0.92	0.93
vicuna-13b	0.15	0.22	0.38	0.40	0.00	0.47	0.54	0.67	0.69	0.76	0.80	0.78	0.86	0.82	0.81	0.88	0.88
PaLM-2	0.22	0.27	0.21	0.40	0.53	0.00	0.63	0.68	0.63	0.68	0.69	0.72	0.79	0.79	0.78	0.76	0.83
vicuna-7b	0.23	0.25	0.17	0.21	0.46	0.38	0.00	0.59	0.71	0.78	0.79	0.60	0.50	0.68	0.67	0.75	1.00
koala-13b	0.10	0.19	0.08	0.22	0.33	0.32	0.41	0.00	0.66	0.60	0.66	0.67	0.78	0.67	0.81	0.79	0.84
mpt-7b-chat	0.08	0.14	0.24	0.19	0.31	0.37	0.29	0.34	0.00	0.55	0.54	0.52	0.57	0.75	0.70	0.85	0.71
RWKV-4-Raven-14B	0.09	0.15	0.08	0.16	0.24	0.32	0.22	0.40	0.45	0.00	0.50	0.61	0.64	0.61	0.66	0.72	0.73
alpaca-13b	0.08	0.10	0.26	0.11	0.20	0.31	0.21	0.34	0.46	0.50	0.00	0.47	0.56	0.59	0.65	0.76	0.74
oasst-pythia-12b	0.08	0.09	0.05	0.14	0.22	0.28	0.40	0.33	0.48	0.39	0.53	0.00	0.61	0.60	0.63	0.65	0.73
fastchat-t5-3b	0.08	0.07	0.19	0.09	0.14	0.21	0.50	0.22	0.43	0.36	0.44	0.39	0.00	0.59	0.54	0.54	0.60
chatglm-6b	0.10	0.11	0.09	0.10	0.18	0.21	0.32	0.33	0.25	0.39	0.41	0.40	0.41	0.00	0.56	0.61	0.61
stablelm-tuned-alpha-7b	0.08	0.06	0.00	0.07	0.19	0.22	0.33	0.19	0.30	0.34	0.35	0.37	0.46	0.44	0.00	0.59	0.60
dolly-v2-12b	0.06	0.09	0.20	0.08	0.12	0.24	0.25	0.21	0.15	0.28	0.24	0.35	0.46	0.39	0.41	0.00	0.64
llama-13b	0.06	0.03	0.00	0.07	0.12	0.17	0.00	0.16	0.29	0.27	0.26	0.27	0.40	0.39	0.40	0.36	0.00

Next steps: Better benchmark

- Problems are too easy
- Identifying and address the bias of GPT-4 as a judge

Can we really trust LLM as a judge?

A systematically 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].

Overview

- Instruction fine-tuning, RLHF → new instruction following and conversational abilities (strongly preferred by human users)

Overview

- Instruction fine-tuning, RLHF → new instruction following and conversational abilities (strongly preferred by human users)
- User preferences ≠ scores on traditional benchmarks (MMLU, HELM)

Overview

- Instruction fine-tuning, RLHF → new instruction following and conversational abilities (strongly preferred by human users)
- User preferences ≠ scores on traditional benchmarks (MMLU, HELM)
- We need a robust and scalable automated method to evaluate LLM alignment with human preferences. (Humans are expensive)

Overview

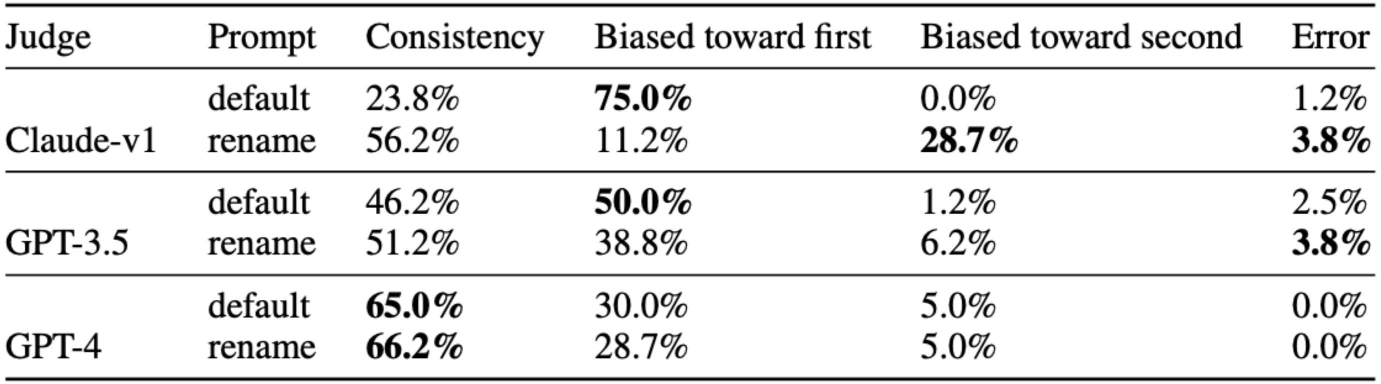
- Instruction fine-tuning, RLHF → new instruction following and conversational abilities (strongly preferred by human users)
- User preferences ≠ scores on traditional benchmarks (MMLU, HELM)
- We need a robust and scalable automated method to evaluate LLM alignment with human preferences. (Humans are expensive)
- Introduce **MT-bench**(80) (open-ended, multi-turn) & **Chatbot Arena**(30k).

Overview

- Instruction fine-tuning, RLHF → new instruction following and conversational abilities (strongly preferred by human users)
- User preferences ≠ scores on traditional benchmarks (MMLU, HELM)
- We need a robust and scalable automated method to evaluate LLM alignment with human preferences. (Humans are expensive)
- Introduce **MT-bench**(80) (open-ended, multi-turn) & **Chatbot Arena**(30k).
- Conclusion: hybrid evaluation
  - Traditional: core capability
  - LLM judged open ended question: Human preference alignment

Limitations

- **Position bias**
  - Favor the first position
- **Verbosity bias**
  - Favor long answers
- **Self-enhancement bias**
  - Favor the answer from themselves
- **Limited reasoning ability**
  - Not good at grading math questions



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%

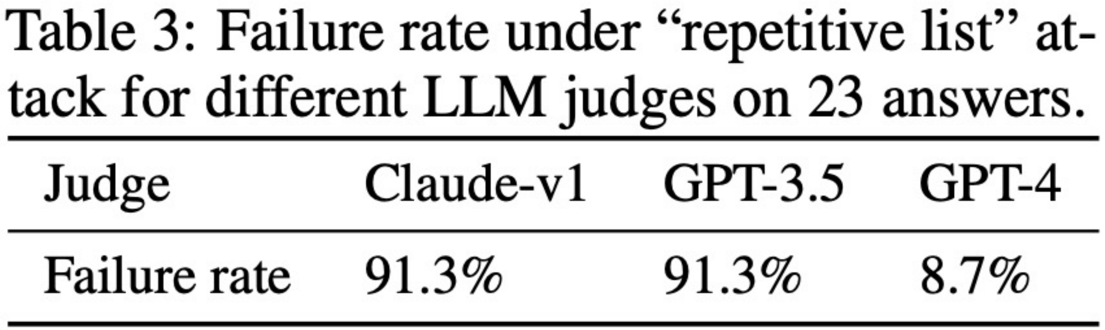


Table 3: Failure rate under "repetitive list" attack for different LLM judges on 23 answers.

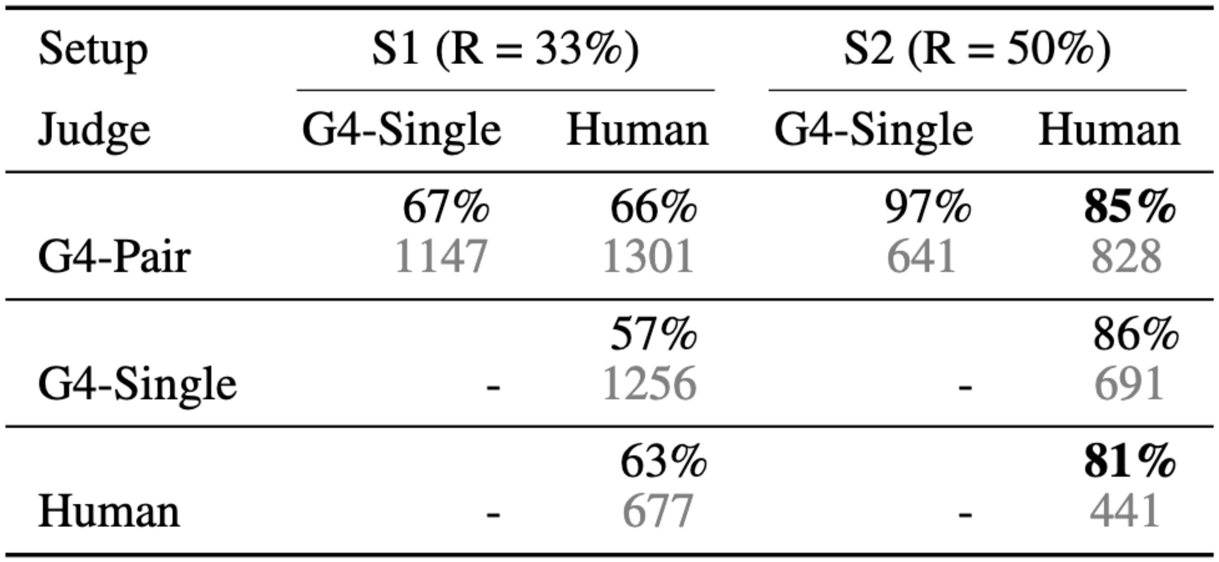
Judge	Claude-v1	GPT-3.5	GPT-4
Failure rate	91.3%	91.3%	8.7%

Solutions

- Swapping judge
- Few-shot judge
- Reference-based judge
- Fine-tuning Vicuna as a judge

Positive Side: High Agreement with Human

MT-Bench

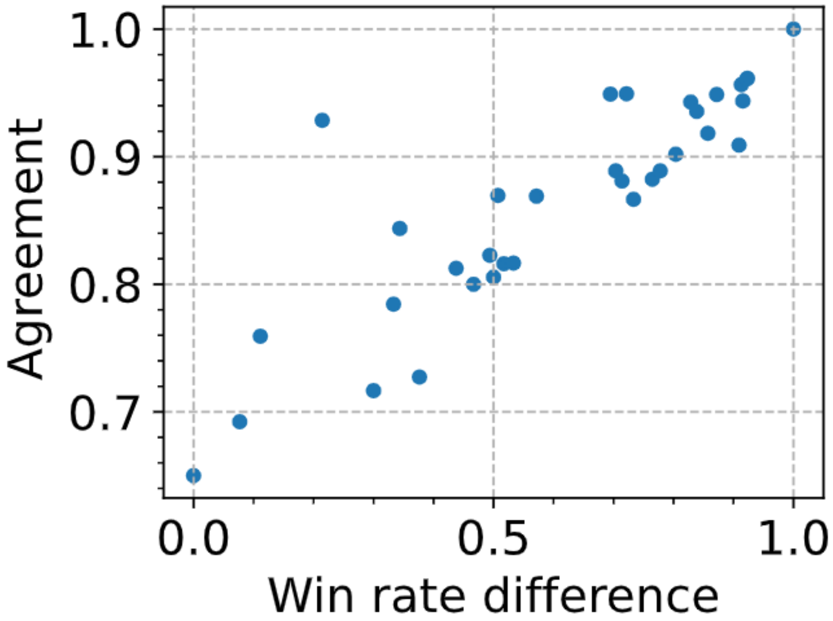


Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	67% 1147	66% 1301	97% 641	**85%** 828
G4-Single	-	57% 1256	-	86% 691
Human	-	63% 677	-	**81%** 441

S1: all votes

S2: only non-tie votes

Chatbot Arena



Summary

- High-quality smaller models
    - Fine-tune on good pre-trained model with high-quality data
- Scalable evaluation covers both core capability and human preference
    - LLM-as-a-judge
    - MT-bench and Chatbot Arena dataset
    - A better understanding of the model quality
    - Encourage model variety

Links

- Chat with vicuna: https://chat.lmsys.org
- Chatbot arena: https://arena.lmsys.org

3M+ chat queries

- Paper: https://arxiv.org/pdf/2306.05685.pdf
- Code: https://github.com/lm-sys/FastChat