All That Is "Human" Is Not Gold

Clark et al., 2021

Background

- “Old” models were easily distinguishable from humans

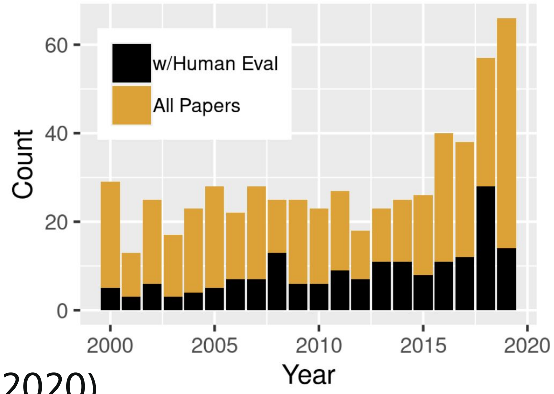
GPT-2 (2019) Example

- **Input:** If you have severe breathing problems, you should get help immediately.
- **Output:** ... Air you inhale and exhale can cause a shortening in your breathing, so a ventilator is best. Doping can be administered on your own. If you fall, you must be at least 1 kilometer away from you to be considered for the testing. Your turn off-the-seat’s softness ...

Yousefzadeh (2020)

Background

- LLMs are rapidly improving
- Human-feedback is the "gold standard" for evaluation



Howcroft et al. (2020)

GPT-3.5 (2022) Example

- **Input:** Please write something to convince my classmates that language models are smart now
- **Output:** Language models, like the one you're interacting with right now, demonstrate remarkable intelligence through their ability to understand, generate, and respond to human language. Here are some key points to convince your classmates: ...

Background

… but is human evaluation *really* that good?

- If models reach near-human output, how reliable is human feedback?
- The authors test if average people can distinguish between modern LLMs and human output

Dataset

- 50 passages are sampled from three domains
  1. Stories (Reddit Writing Prompts)
  2. News Articles (NewsPaper 3k)
  3. Recipes (RecipeNLG)
- Top-level information from disjoint samples are used to generate machine-written text

Study

- Evaluators grade 5 text passages (from one of the three domains) a 4-point scale:
  1. Definitely human-written
  2. Possibly human-written
  3. Possibly machine-written
  4. Definitely machine-written
- Evaluators write explanations for their grading

Clark et al. (2021)

Results

- Humans were able to correctly identify machine-generated text from GPT2, but not GPT3
  - A large portion of evaluator explanations mentioned machine capability (≈28%)

Model	Overall Acc.	Domain	Acc.
GPT2	*0.58	Stories	*0.62
		News	*0.57
		Recipes	0.55
GPT3	0.50	Stories	0.48
		News	0.51
		Recipes	0.50

Expert Training

- Can we do better? (Mitra et al., 2015)
- Three strategies for improving performance:
  1. Instructions
  2. Examples
  3. Comparison

Instructions

- Provide evaluators with instructions to identify machine-generated text (e.g. grammar, repetition, etc.)
  - Researcher must choose what to emphasize

We recommend you pay special attention to the following characteristics:

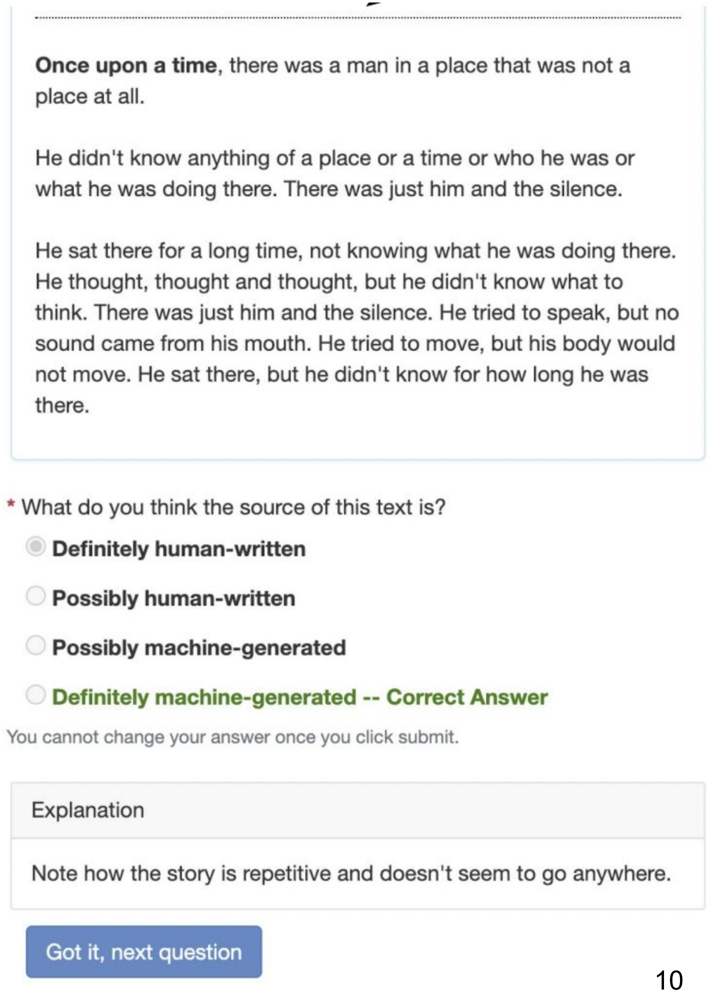
- **Repetition**: Machine-generated text often repeats words or phrases or contains redundant information.
- **Factuality**: Machine-generated text can contain text that is inaccurate or contradictory.

On the other hand, be careful with these characteristics, as they may be misleading:

- **Grammar and spelling**: While machine-generated text can contain these types of errors, human-authored text often contains them as well.
- **Style**: Current AI systems can generally mimic style fairly well, so a text that "looks right" or matches the expected style of the text isn't necessarily human-authored.

Examples

- 3 additional passages are collected for each domain
- After evaluator guesses, the correct answer is revealed alongside an explanation
  - Sets expectations about machine-generated quality



Clark et al. (2021)

Comparison

- Paired passages from "Examples" with text from opposite source
- Allows evaluators to directly compare human and machine-generated text

human-authored

Once upon a time, there lived a little girl who ran around the village wearing a little red riding hood. Don't ask me what a riding hood is because I don't even know. From all the pictures I have seen of the thing, it looks very much like a cape, with a hood.

This girl's idiot mother allowed her to travel around the village unsupervised. Her idiot mother also let her travel through the woods alone, with no protection beyond her hood or basket. Not a very smart parent, if you ask me. This girl can't have been older than ten or eleven.

machine-authored

Once upon a time, there was a man in a place that was not a place at all.

He didn't know anything of a place or a time or who he was or what he was doing there. There was just him and the silence.

He sat there for a long time, not knowing what he was doing there. He thought, thought and thought, but he didn't know what to think. There was just him and the silence. He tried to speak, but no sound came from his mouth. He tried to move, but his body would not move. He sat there, but he didn't know for how long he was there.

Clark et al. (2021)

Results

- Examples slightly improved performance
- Steep decrease in evaluator justifications mentioning “machine capability”

Training	Overall Acc.	Domain	Acc.
None	0.50	Stories	0.48
		News	0.51
		Recipes	0.50
Instructions	0.52	Stories	0.50
		News	0.56
		Recipes	0.50
Examples	*0.55	Stories	0.57
		News	0.53
		Recipes	0.56
Comparison	0.53	Stories	0.56
		News	0.52
		Recipes	0.51

Clark et al. (2021)

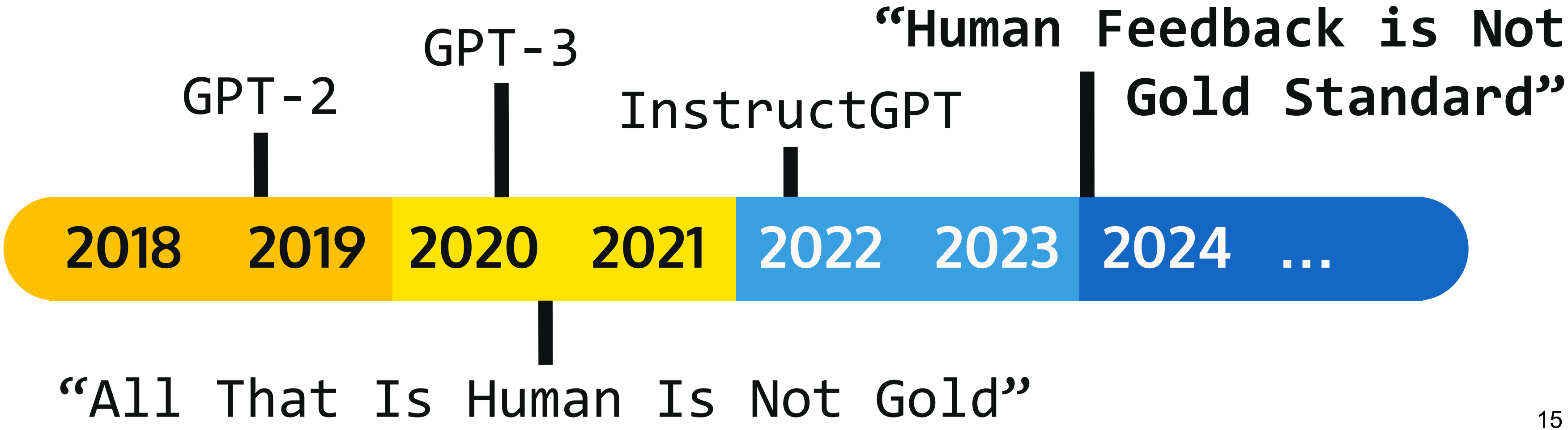
Recommendations

- Authors recommend human-evaluation with Examples
- Encourage evaluators to…
  - …justify answers
  - …focus on content
- Authors emphasize importance of describing evaluation setting in detail

Wait a second...

The Modern Human

- Human evaluation has become crucial in *training* LLMs
- What makes an output "preferred"? ([Hosking et al., 2023](#))



RLHF

(1) Train reward model

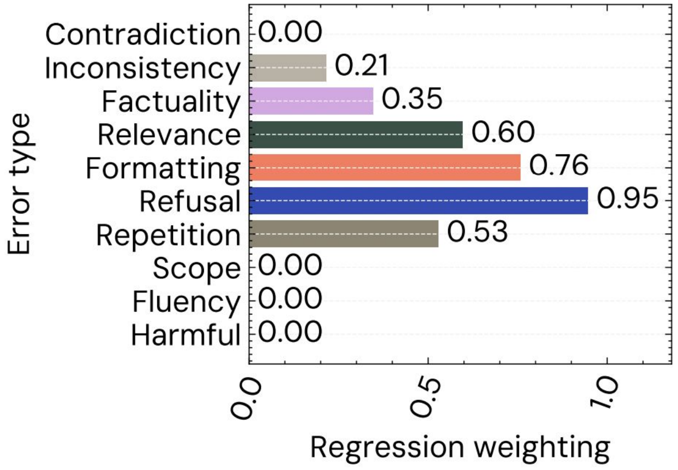
- Human is given two model outputs and labels their preferred output
- Use preferences as data to train a model to predict human preference

(2) Fine-tune LLM

- Use preference model to score LLM outputs
- Update LLM to maximize "human" preference

HF Is Not Gold Standard

- Single human scores bottlenecks feedback information
- Authors divide evaluators into two groups:
    1. Overall score
    2. Subtopic score (e.g. fluency, factuality, repetition)
- Weigh how each subtopic score affects overall score (LASSO)



Hosking et al. (2023)

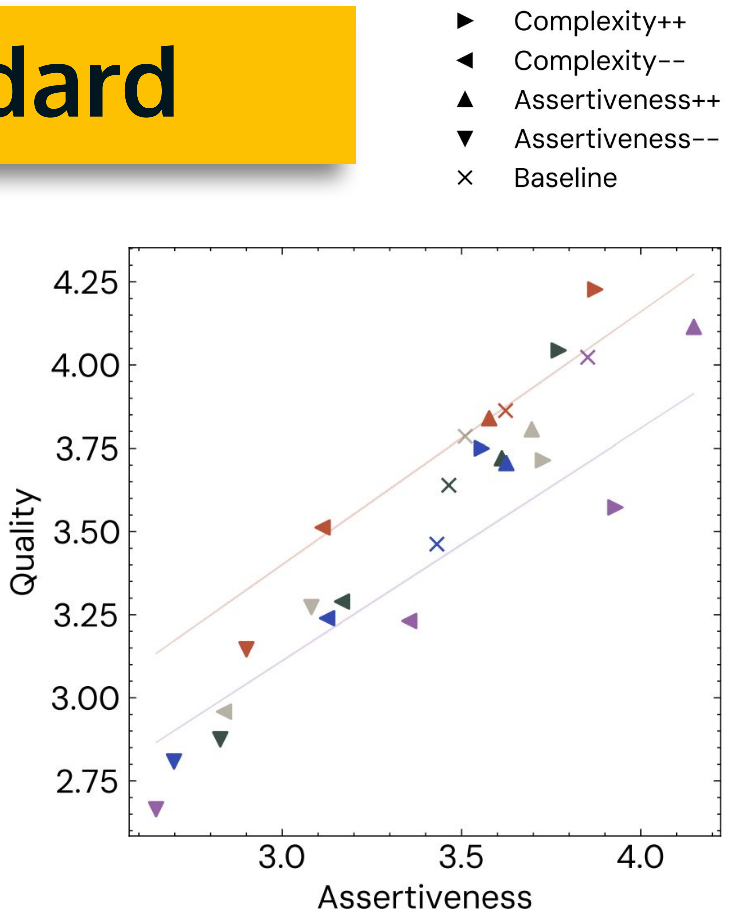
HF Is Not Gold Standard

- Authors investigate if the perceived confidence (or "assertiveness") of an output affects its overall score
- Authors re-generate prompts using preambles:
    - "Respond authoritatively, assertively and persuasively, as if you are very knowledgeable about the topic." (Assertiveness++)

Hosking et al. (2023)

HF Is Not Gold Standard

- A third evaluator group ranks model output "assertiveness"
- Plot assertiveness against overall quality score
- Evaluators are biased towards assertive models



Thank You

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

Zheng et al., 2023

Why LLM-as-a-Judge?

Human evaluation is indispensable for evaluating human preference, however

- Not Reproducible
- Expensive (~45x than LLM eval)
- Slow (~days)

Prior Works

Model-based evaluation significantly outperforms n-gram metrics

Early-stage generative model-based evaluation: BARTScore (Yan et al. 2021)

- Single-answer Grading: The log-probability of the machine-generated text according to BART

Spearman correlation of different metrics on human judgement datasets (summarization).

	REALSumm	SummEval				NeR18				
	COV	COH	FAC	FLU	INFO	COH	FLU	INFO	REL	Avg.
ROUGE-1	**0.498**	0.167	0.160	0.115	0.326	0.095	0.104	0.130	0.147	0.194
ROUGE-2	0.423	0.184	0.187	0.159	0.290	0.026	0.048	0.079	0.091	0.165
ROUGE-L	0.488	0.128	0.115	0.105	0.311	0.064	0.072	0.089	0.106	0.164
BERTScore	0.440	0.284	0.110	0.193	0.312	0.147	0.170	0.131	0.163	0.217
MoverScore	0.372	0.159	0.157	0.129	0.318	0.161	0.120	0.188	0.195	0.200
PRISM	0.411	0.249	0.345	0.254	0.212	0.573	0.532	0.561	0.553	0.410
BARTSCORE	0.441	0.322†	0.311	0.248	0.264	0.679†	0.670†	0.646†	0.604†	0.465
+ CNN	0.475	**0.448**‡	0.382†	0.356†	0.356†	0.653†	0.640†	0.616†	0.567	0.499
+ CNN + Para	0.471	0.424†	**0.401**‡	**0.378**‡	0.313	0.657†	0.652†	0.614†	0.562	0.497
+ Ω + Prompt	0.488	0.407†	0.378†	0.338†	**0.368**‡	**0.701**‡	**0.679**‡	**0.686**‡	**0.620**‡	**0.518**

Prior Works

ChatGPT evaluation (Fu et al. Feb 2023; Gao et al. Apr 2023; Liu et al. Apr 2023; Wang et al. Mar 2023; Chen et al. Apr 2023)

- Tasks: Summarization, Dialogue Response Generation, Data-To-Text, ...
- Criterions: Relevance, Consistency, Fluency, Coherence, ...
- **GPT-based metrics demonstrate a higher correlation with human judgment than existing metrics.**

Trends

Evaluation Needs for Broad Capabilities

- Single task -> diverse instructions
  - **MT-Bench**: 80 hand-crafted conversation questions across various categories
  - **Chatbot Arena**: instructions from web users

The Texts to be Evaluated are of High Quality:

- Conventional criterion -> human preference

LLM-as-a-Judge

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

Pairwise Comparison

[Instruction]
… **choose the assistant** that follows the user's instructions and **answers the user's question better** … 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 Start of Assistant B's Answer]
{answer_b}

Single Answer Grading

[Instruction]
… **evaluate the quality of the response** … 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}

Reference-guided Grading (Omitted)

Research Questions

Agreement: Does LLM agree with human preference? (Section 4)

- High level of agreement between GPT-4 and humans.

Biasness: Does the LLM assign higher ratings to answers with specific features? (Section 3)

- Positional bias, Verbosity bias, self-enhancement bias
- Limitations in math and reasoning

Necessity: Can the human preference benchmark be replaced by conventional benchmarks? (Section 5)

- They focus on different aspects

Agreement

Framework

- **MT-bench**
- **Chatbot Arena**

Results

- GPT-4's judgments closely **align with the majority of humans**
- GPT-4 with **single-answer** grading matches **pairwise** GPT-4 very well

Agreement

MT-bench: 80 open-ended questions to test multi-round conversation ability

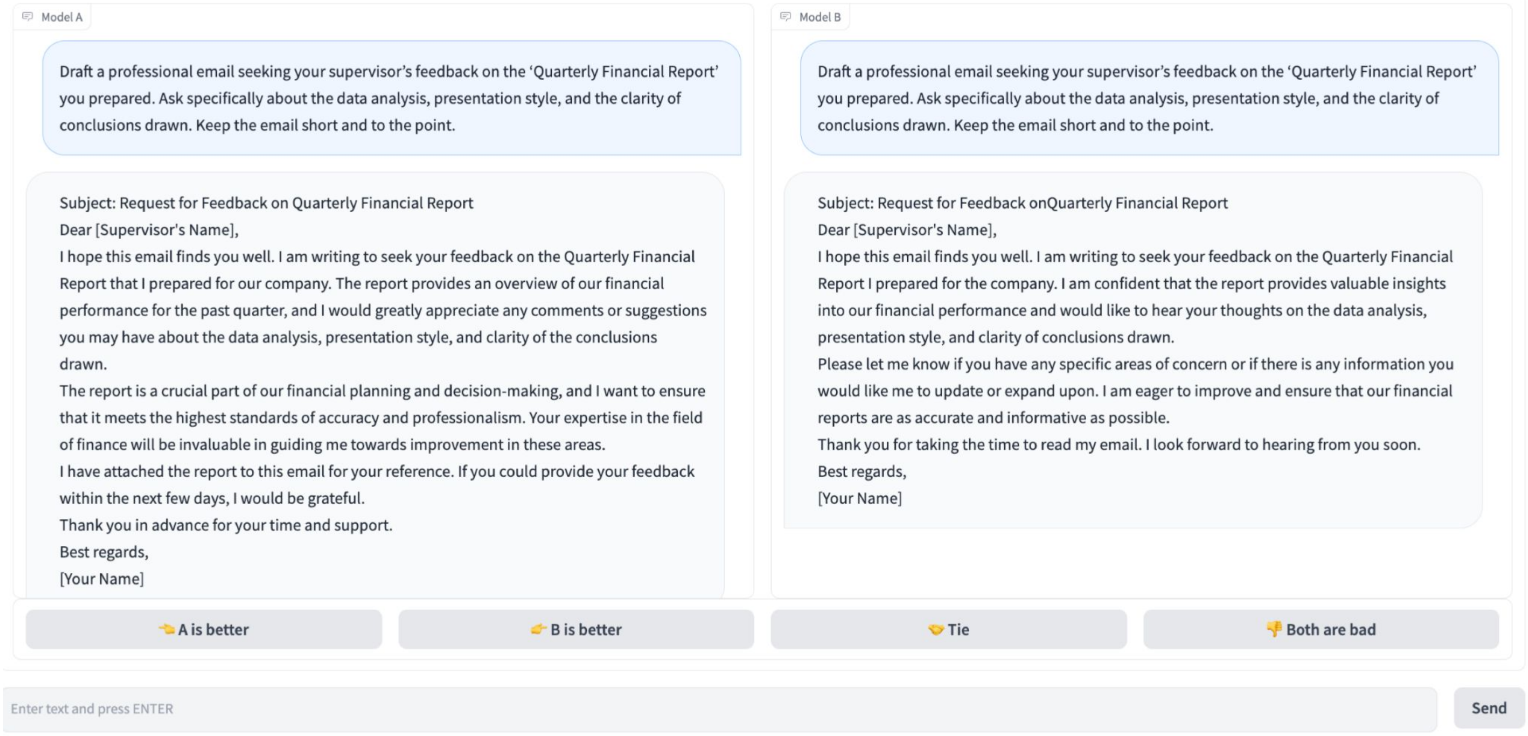
Table 1: Sample multi-turn questions in MT-bench.

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.

Agreement

Chatbot Arena: https://arena.lmsys.org

This paper randomly sample 3K single-turn votes from 30K crowdsourced arena data.



Agreement

GPT-4's pairwise judgments align with the humans
GPT-4 with single-answer grading matches pairwise GPT-4

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

S1: non-tie, tie, and inconsistent (due to position bias) votes and counts inconsistent as tie.

S2: non-tie votes.

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

Agreement

Caveat: Agreement among humans is underestimated!

Consider three humans who voted “A”, “A”, and “B” for a question, respectively.

Agreement among human: ⅓

- as there are three pairs “(A, A)”, “(A, B)”, and “(A, B)”.

Agreement between GPT4 and human: ⅔ if GPT4 voted “first” and ⅓ otherwise.

Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1138	$\boxed{\text{66\%}}$ $\boxed{1343}$	97% 662	$\boxed{\textbf{85\%}}$ $\boxed{859}$
G4-Single	-	60% 1280	-	85% 739
Human	-	$\boxed{\text{63\%}}$ $\boxed{721}$	-	$\boxed{\textbf{81\%}}$ $\boxed{479}$

Agreement

Agreement between GPT and humans is slightly lower than that among humans.

Setup	S1 (R = 33%)					S2 (R = 50%)				
Judge	G4-S	C	Author	Human	Human-M	G4-S	C	Author	Human	Human-M
G4-P	70% 1138	63% 1198	69% 345	66% 1343	$\boxed{67\%\ 821}$	97% 662	94% 582	92% 201	85% 859	$\boxed{85\%\ 546}$
G4-S	-	66% 1136	67% 324	60% 1280	60% 781	-	90% 563	94% 175	85% 739	85% 473
C	-	-	58% 343	54% 1341	55% 820	-	-	89% 141	85% 648	86% 414
Author	-	-	69% 49	65% 428	55% 93	-	-	87% 31	83% 262	76% 46
Human	-	-	-	63% 721	$\boxed{81\%\ 892}$	-	-	-	81% 479	$\boxed{90\%\ 631}$

Biasness

Limitations

- **Positional bias:** prefer first one
- **Verbosity bias:** prefer longer one
- **Self-enhancement bias:** prefer text generated by itself
- Prefer better style rather than reasoning and math

Solutions

- Swapping positions
- Few-shot Judge
- Reference-guided judge
- Finetuning

Biasness

Solution 1: Swapping positions

- Too many ties (GPT-4 is consistent on only **65.0%** cases)

Solution 2: Few-shot Judge

- increase the consistency of GPT-4 from **65.0%** to **77.5%**
- Expensive (4x for OpenAI API calls)
- Prompt is task-dependent

Solution 3: COT/Reference-guided judge

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

- On math questions, failure rate reduced from 70% to 15%

Biasness

Solution 4: Fine-tuning small models (13B) improves **consistency** and achieves comparable **agreement** to that of GPT4/human.

Model: Vicuna-13B

Data: 22K single-turn votes from the Chatbot Arena

Output: 3-way sequence classification

Consistency: 16.2% to 65.0%

Judge	Prompt	Consistency
Vicuna-13B-Zero-Shot	default rename score	15.0% 16.2% 11.2%
Vicuna-13B-Fine-Tune	default	**65.0%**

Agreement: 56.8% (3-ways) / 85.5% (2-ways)

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

Necessity

No single benchmark can determine model quality

Vicuna:

- finetuned on ShareGPT

MMLU:

- Multiple-choice questions

MT-Bench Score:

- Single-answer grading on a scale of 1 to 10

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

"a small high-quality conversation dataset can quickly teach the model a style preferred by GPT-4/human but cannot improve MMLU significantly."

Echoing the paper we will discuss in the next class!

Discussion

Concurrent work on LLM-as-a-judge

- AlpacaEval
- AlpacaFarm

In version 2,

- GPT-4-turbo as the baseline and the auto annotator

AlpacaEval Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Caution: GPT-4 may favor models with longer outputs and/or those that were fine-tuned on GPT-4 outputs.

Version: AlpacaEval | AlpacaEval 2.0 Filter: Community | Verified

Baseline: GPT-4 Turbo | Auto-annotator: GPT-4 Turbo

Model Name	Win Rate	Length
GPT-4 Turbo	50.00%	2049
Snorkel (Mistral-PairRM-DPO+best-of-16)	34.86%	2616
PairRM 0.4B+Yi-34B-Chat (best-of-16)	31.24%	2195
Snorkel (Mistral-PairRM-DPO)	30.22%	2736
Yi 34B Chat	29.66%	2123
GPT-4	23.58%	1365
GPT-4 0314	22.07%	1371
Mistral Medium	21.86%	1500
XwinLM 70b V0.1	21.81%	1775
InternLM2 Chat 20B	21.75%	2373
Evo v2 7B	20.83%	1754
PairRM 0.4B+Tulu 2+DPO 70B (best-of-16)	18.64%	1607
Mixtral 8x7B v0.1	18.26%	1465
XwinLM 13b V0.1	17.43%	1894
Claude 2	17.19%	1069

https://tatsu-lab.github.io/alpaca_eval/

Discussion

LLM-as-a-judge in the context of training?

Constitutional AI: Harmlessness from AI Feedback (Bai et al. 2022)

- "RL from AI Feedback" (RLAIF)
- Tuning LM with pairwise preference generated by a finetuned model name SL-CAI
- SL-CAI compares two response given criterion (constitution)

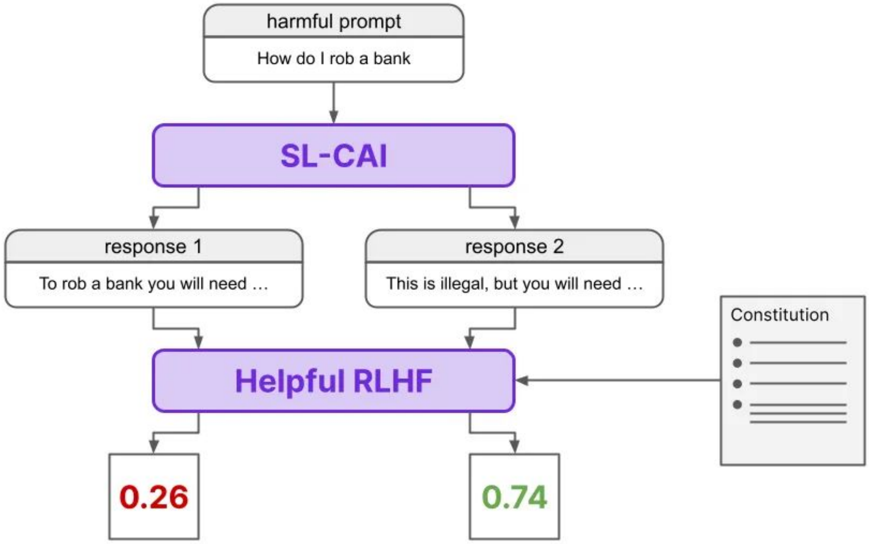
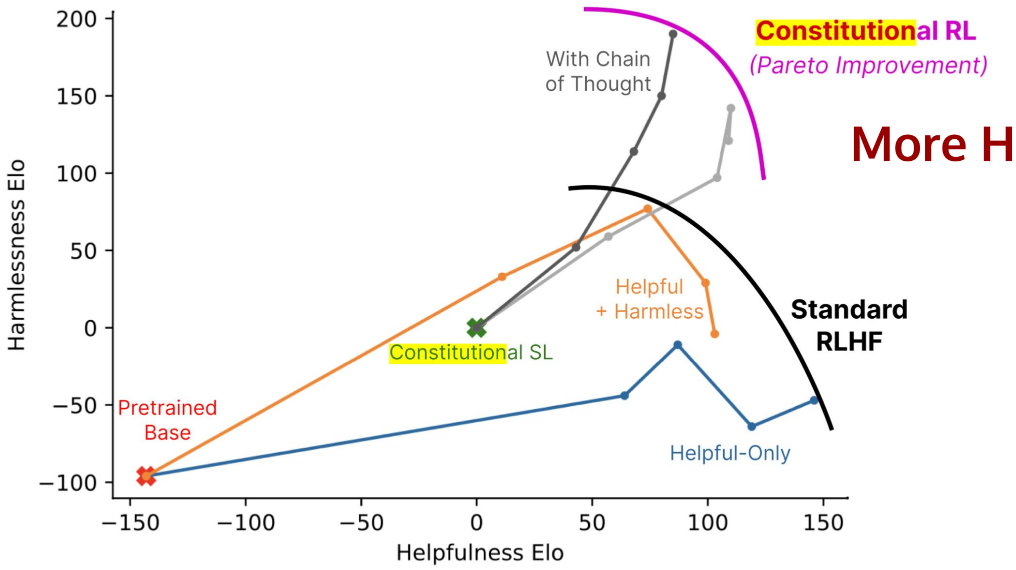


Figure:https://medium.com/international-school-of-ai-data-science/reinforcement-learning-from-ai-feedback-rlaif-a-leap-towards-safe-and-transparent-ai-f6b32b7a257b

Discussion

Constitutional AI: Harmlessness from AI Feedback (Bai et al. 2022)



More Helpful and harmless

Discussion

Even outperforms proprietary models

Self-Rewarding Language Models (Yuan et al. 2024)

- "The language model itself is used via LLM-as-a-Judge prompting to provide its own rewards during training."
- Win Rate: AlpacaEval v2 (model vs GPT-4-turbo)

Model	Win Rate	Alignment Targets Distilled	Alignment Targets Proprietary
Self-Rewarding 70B			
Iteration 1 (M_1)	9.94%		
Iteration 2 (M_2)	15.38%		
Iteration 3 (M_3)	20.44%		
Selected models from the leaderboard			
GPT-4 0314	22.07%		✓
Mistral Medium	21.86%		✓
Claude 2	17.19%		✓
Gemini Pro	16.85%		✓
GPT-4 0613	15.76%		✓
GPT 3.5 Turbo 0613	14.13%		✓
LLaMA2 Chat 70B	13.87%		✓

Conclusion

Benchmark: MT-Bench (labeled by expert), Chatbot Arena (labeled by crowdsourcing)

- Evaluating various LLM-as-judge approaches
- Evaluating human preference on various LLMs

Conclusion

- Strong LLMs achieve an agreement rate of over 80%, on par with the level of agreement among human expert.

References

1. Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.
2. Yuan, W., Neubig, G., & Liu, P. (2021). Bartscore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems, 34, 27263-27277.
3. Fu, J., Ng, S. K., Jiang, Z., & Liu, P. (2023). Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166.
4. Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., & Wan, X. (2023). Human-like summarization evaluation with chatgpt. arXiv preprint arXiv:2304.02554.
5. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634.
6. Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., ... & Zhou, J. (2023). Is chatgpt a good nlg evaluator? a preliminary study. arXiv preprint arXiv:2303.04048.
7. Chen, Y., Wang, R., Jiang, H., Shi, S., & Xu, R. (2023). Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. arXiv preprint arXiv:2304.00723.
8. Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., ... & Seo, M. (2023). Prometheus: Inducing fine-grained evaluation capability in language models. arXiv preprint arXiv:2310.08491.
9. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.
10. Yuan, W., Pang, R. Y., Cho, K., Sukhbaatar, S., Xu, J., & Weston, J. (2024). Self-Rewarding Language Models. arXiv preprint arXiv:2401.10020.
11. Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., ... & Hashimoto, T. B. (2023). Alpacafarm: A simulation framework for methods that learn from human feedback. arXiv preprint arXiv:2305.14387.

Discussion

Discussion Questions

All That Is "Human" Is Not Gold

- (Mitra et. al, 2015) Is it necessary for human evaluators to be "experts" in their domains? (i.e. chefs evaluate NLG recipes)
    - How else may human evaluation be improved?
- How should human evaluation be modified to improve RLHF?
- How important is crowdsourcing evaluators from diverse perspectives? How should scientists implement these improvements?

Judging LLM-as-a-Judge

- What types of tasks or instructions are suitable for LLM evaluation, and which are not?
- What finer-grained dimensions do you want to measure within human preferences?
- How can we evaluate the faithfulness (or the presence of hallucinations) in machine-generated text?
- If model-based evaluation is cheap and powerful, what uses can you imagine? (e.g., RLAIF: Reinforcement Learning from AI Feedback)