Benchmarking and evaluation 2

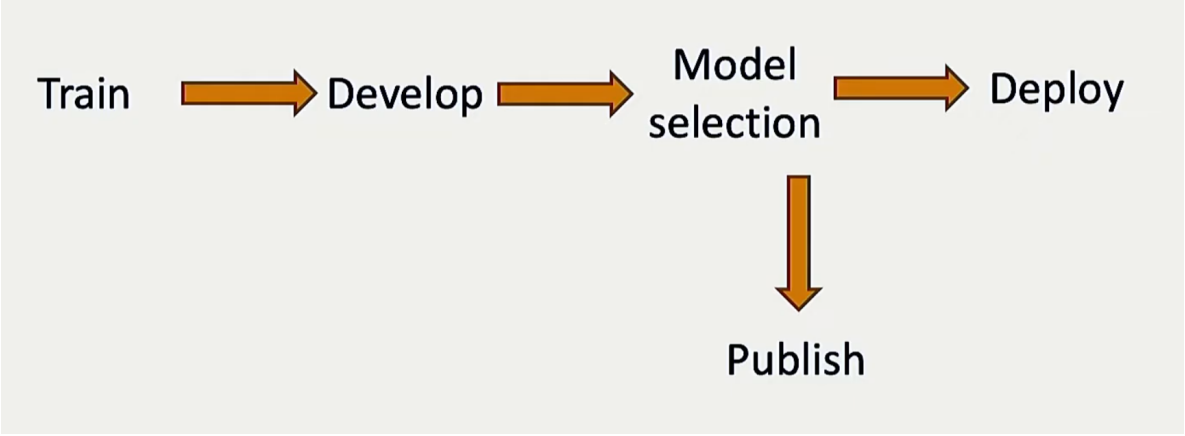
Apr 14, 2026

*Acknowledgment: Slides based on materials by CS224N @ Stanford University (Lecture by Yann Dubois).

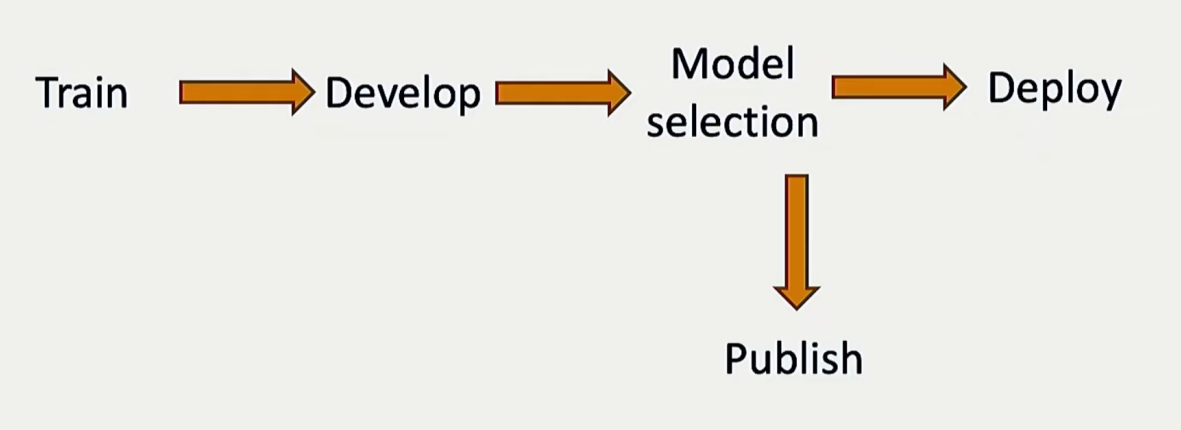
Outline

Outline

Different desiderata for measuring performance



Different desiderata for measuring performance



- Deployment / product-oriented research: task-specific, reliability-critical, requiring high trust
- Publication-oriented research: prioritizes reproducibility; simpler or approximate metrics may be acceptable

Two major types of evaluations

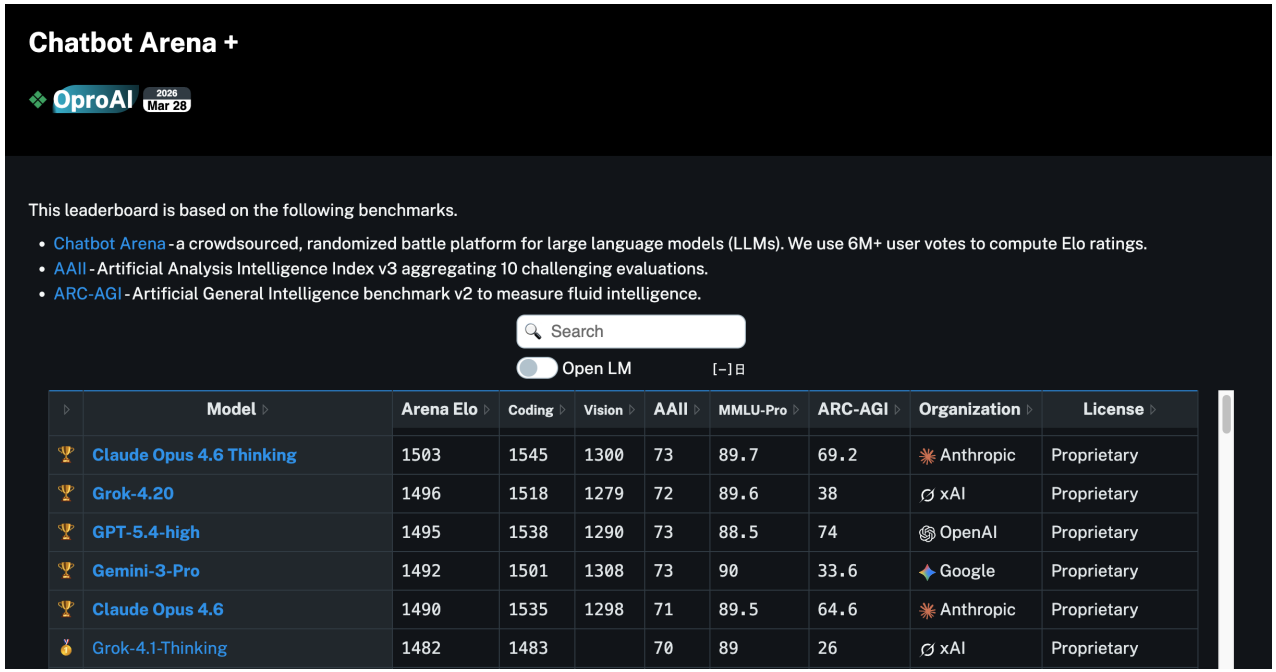
- Close-ended evaluations
- Open-ended evaluations

Outline

Human evaluations

- Automatic metrics fall short of matching human decisions
- Human evaluation is most important form of evaluation for text generation
- Gold standard in developing new automatic metrics
  - New automated metrics must correlate well with human evaluations
- Issues with human evaluations

Chatbot Arena+



Chatbot Arena +

OproAI 2026 Mar 28

This leaderboard is based on the following benchmarks.

- Chatbot Arena - a crowdsourced, randomized battle platform for large language models (LLMs). We use 6M+ user votes to compute Elo ratings.
- AAII - Artificial Analysis Intelligence Index v3 aggregating 10 challenging evaluations.
- ARC-AGI - Artificial General Intelligence benchmark v2 to measure fluid intelligence.

Search

Open LM

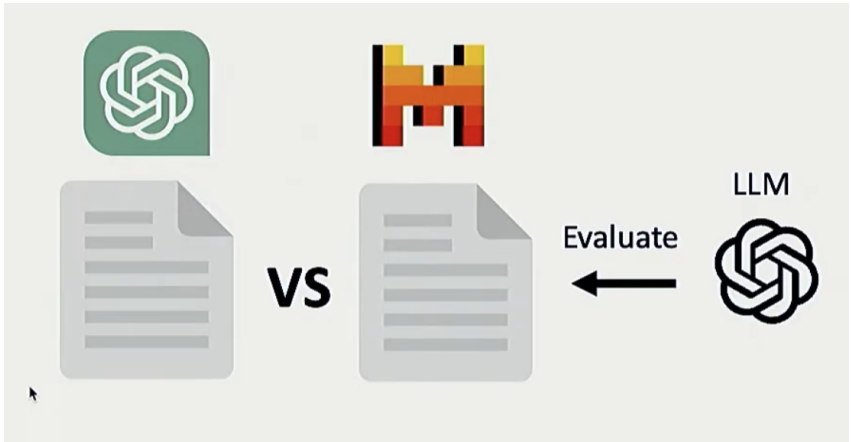
	Model	Arena Elo	Coding	Vision	AAII	MMLU-Pro	ARC-AGI	Organization	License
🏆	Claude Opus 4.6 Thinking	1503	1545	1300	73	89.7	69.2	Anthropic	Proprietary
🏆	Grok-4.20	1496	1518	1279	72	89.6	38	xAI	Proprietary
🏆	GPT-5.4-high	1495	1538	1290	73	88.5	74	OpenAI	Proprietary
🏆	Gemini-3-Pro	1492	1501	1308	73	90	33.6	Google	Proprietary
🏆	Claude Opus 4.6	1490	1535	1298	71	89.5	64.6	Anthropic	Proprietary
🥇	Grok-4.1-Thinking	1482	1483		70	89	26	xAI	Proprietary

https://openlm.ai/chatbot-arena/

What's missing with side-by-side human eval?

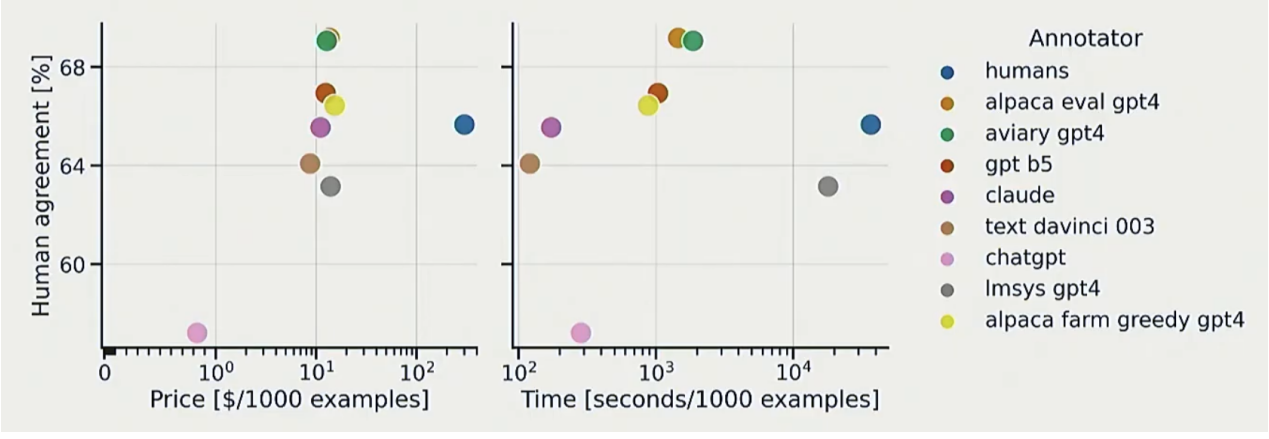
- Current gold standard for evaluation for chat LLM
- External validity
  - Typing random questions into a head-to-head website may not be representative
- Cost
  - Human annotation takes large, community effort
  - New models take a long time to benchmark
  - Only notable models get benchmarked

Lowering the costs - use a LM evaluator



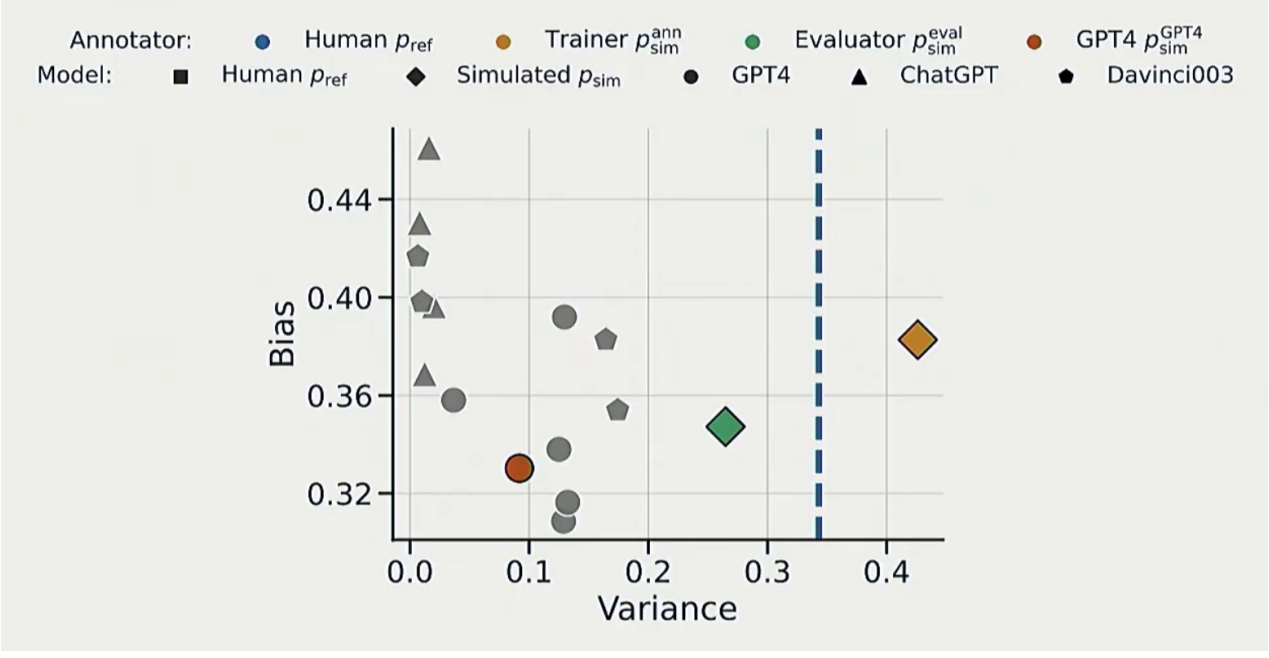
- Use a LM as a reference free evaluator
- Surprisingly high correlations with human
- Common versions: AlpacaEval, MT-bench

AlpacaFarm: Human agreement



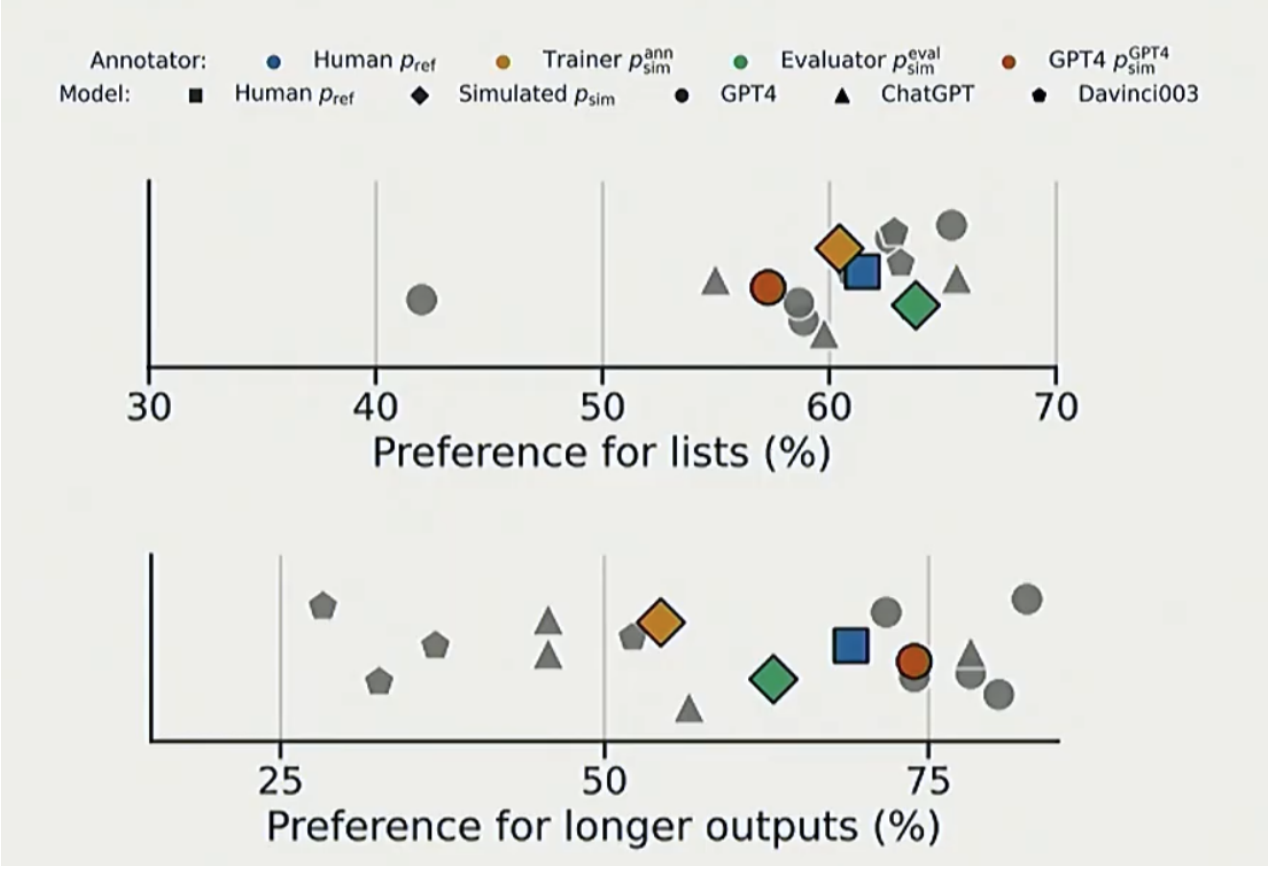
- 100x cheaper, 100x faster, and higher agreement than humans
- Note: can also use for RLAIF!

AlpacaFarm: Human agreement



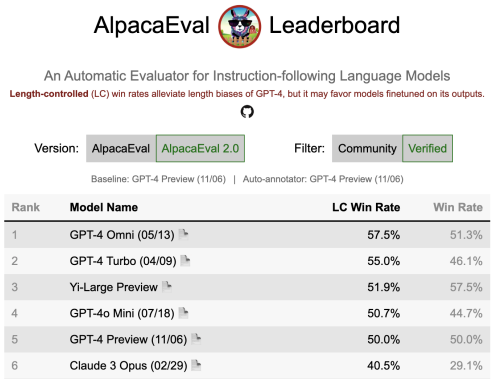
- Humans have low agreement because of variance!

Things to be careful with



- Same issues as before: Spurious correlations
  - length
  - position (left? right?)
  - GPT-4 self bias

AlpacaEval



AlpacaEval Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Length-controlled (LC) win rates alleviate length biases of GPT-4, but it may favor models finetuned on its outputs.

Version: AlpacaEval | AlpacaEval 2.0 Filter: Community | Verified

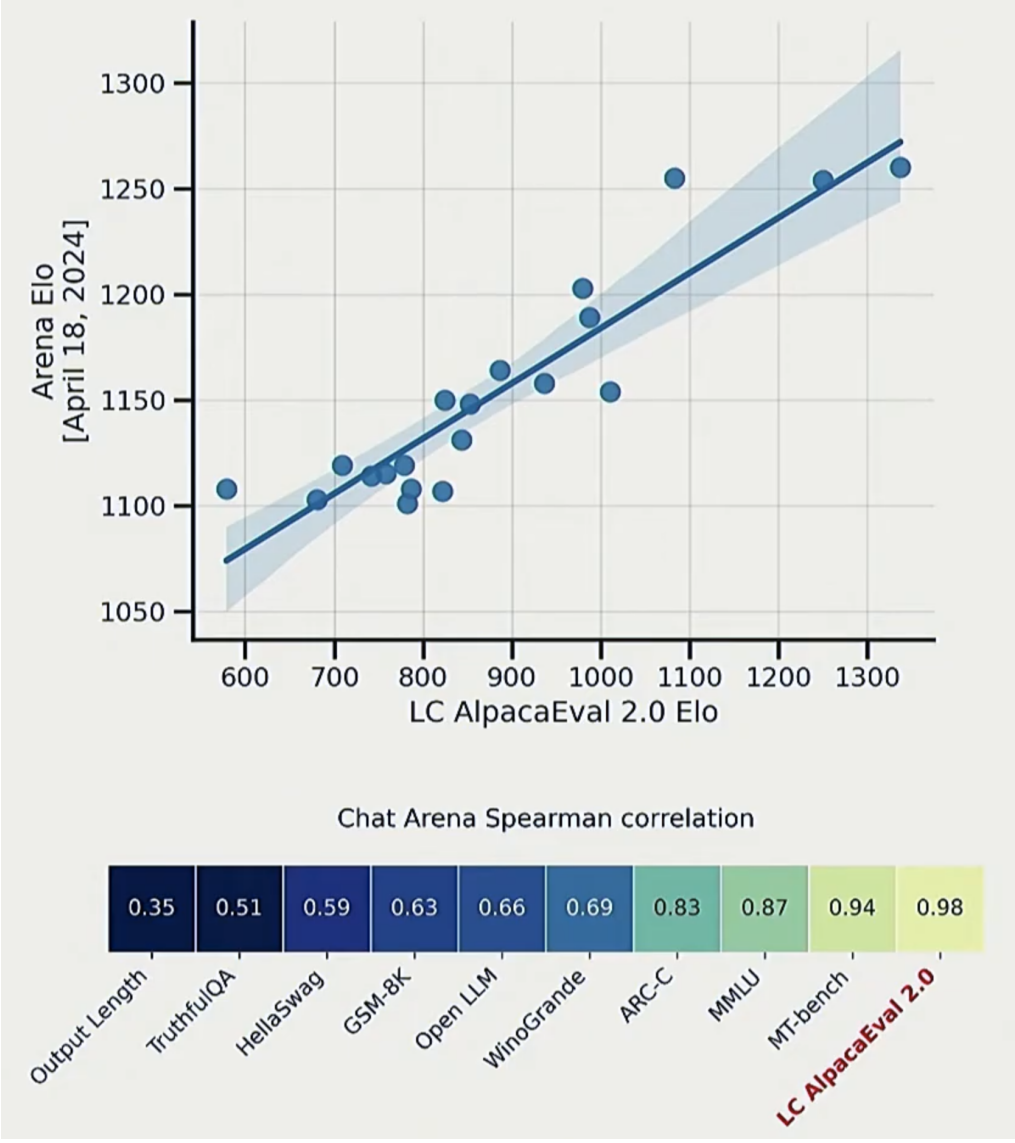
Baseline: GPT-4 Preview (11/06) | Auto-annotator: GPT-4 Preview (11/06)

Rank	Model Name	LC Win Rate	Win Rate
1	GPT-4 Omni (05/13)	57.5%	51.3%
2	GPT-4 Turbo (04/09)	55.0%	46.1%
3	Yi-Large Preview	51.9%	57.5%
4	GPT-4o Mini (07/18)	50.7%	44.7%
5	GPT-4 Preview (11/06)	50.0%	50.0%
6	Claude 3 Opus (02/29)	40.5%	29.1%

- Internal benchmark for developing Alpaca
- 98% correlation with Chatbot Arena
- Less than three mins; 10 dollars
- How it works:
  - For each instruction, generate an output by baseline and model to evaluate
  - Ask GPT-4 the probability that the model's output is better
  - Reweight win-probability based on length of outputs
  - Average win-probability $\rightarrow$ win rate

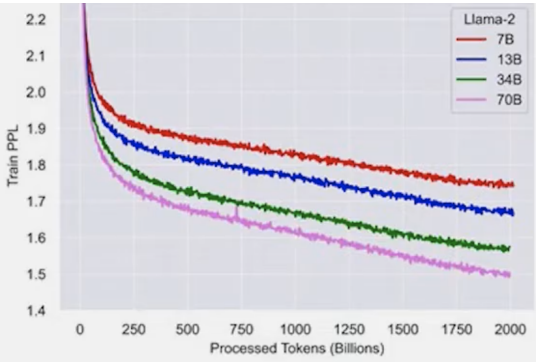
RIT Rochester Institute of Technology

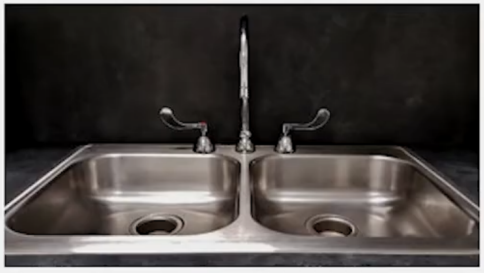
AlpacaEval


Arena Elo
[April 18, 2024]
1300
1250
1200
1150
1100
1050
600
700
800
900
1000
1100
1200
1300
LC AlpacaEval 2.0 Elo
Chat Arena Spearman correlation
0.35
0.51
0.59
0.63
0.66
0.69
0.83
0.87
0.94
0.98
Output Length
TruthfulQA
HellaSwag
GSM-8K
Open LLM
WinoGrande
ARC-C
MMLU
MT-bench
LC AlpacaEval 2.0

Outline

Current evaluations of LLMs





Perplexity

Everything

Arena-like

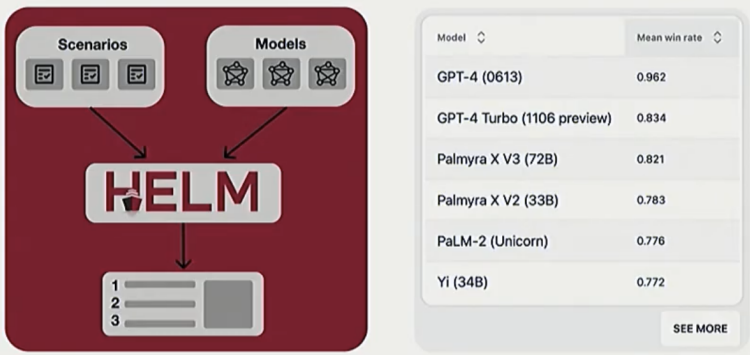
pretraining

finetuned

Everything: HELM and open-llm leaderboard

Collect many automatically evalutable benchmarks, evaluate across them

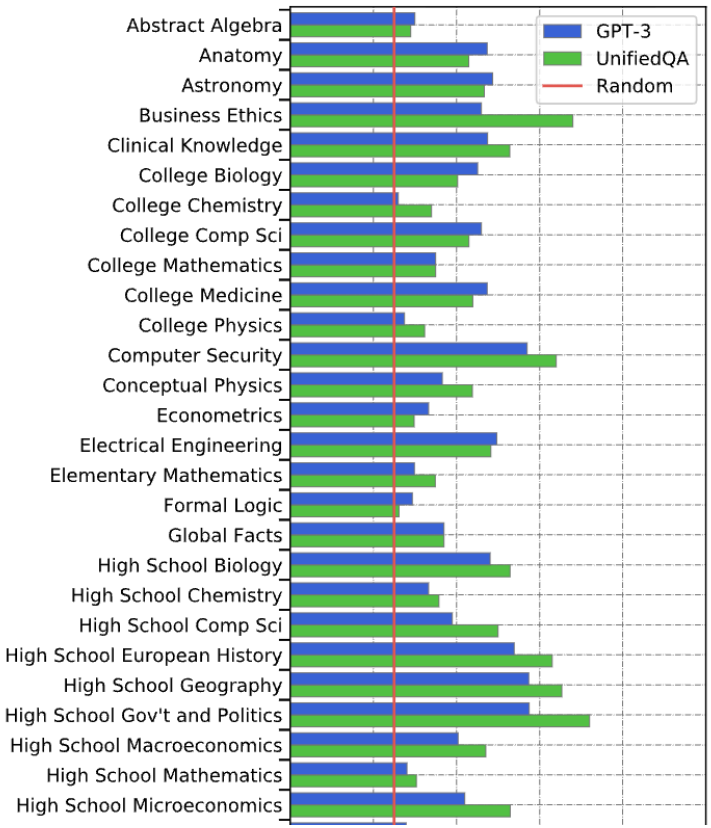
Holistic evaluation of language models (HELM)



Huggingface open LLM leaderboard

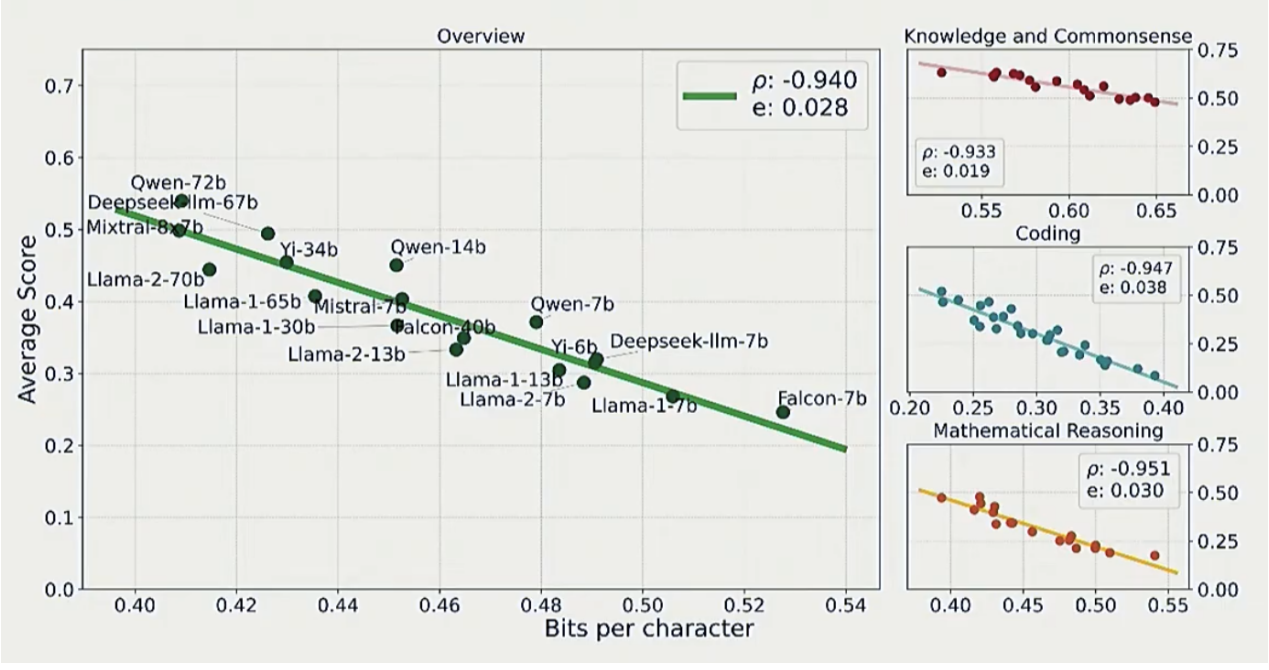
MMLU

- Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021)
- Some examples in the paper



Perplexity

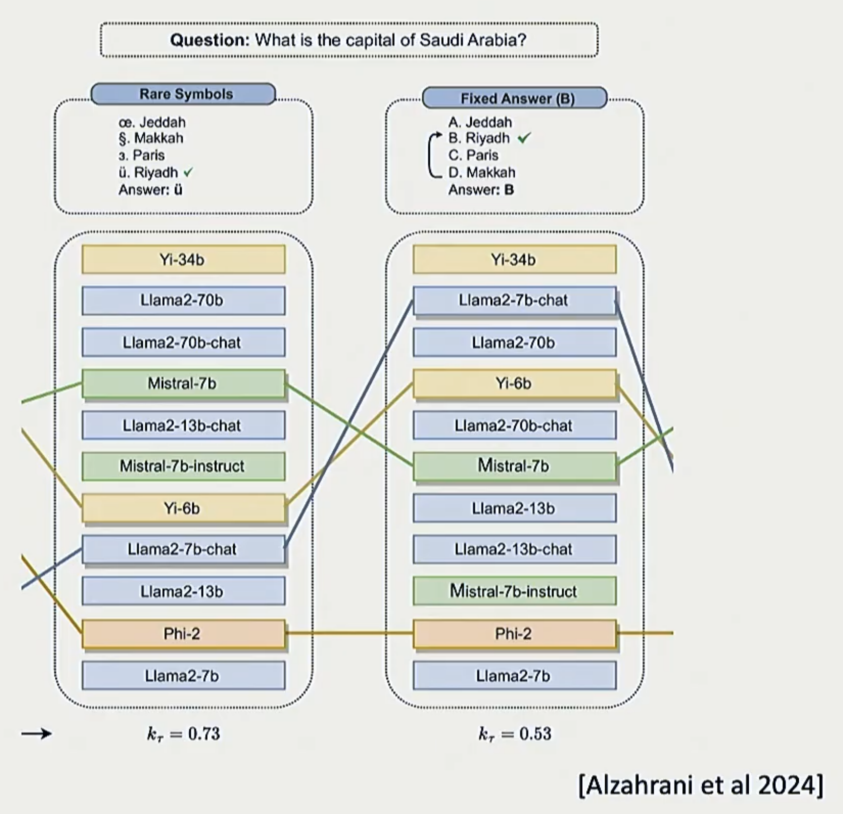
Perplexity is highly correlated with downstream performance (but depends on data & tokenizer)



Outline

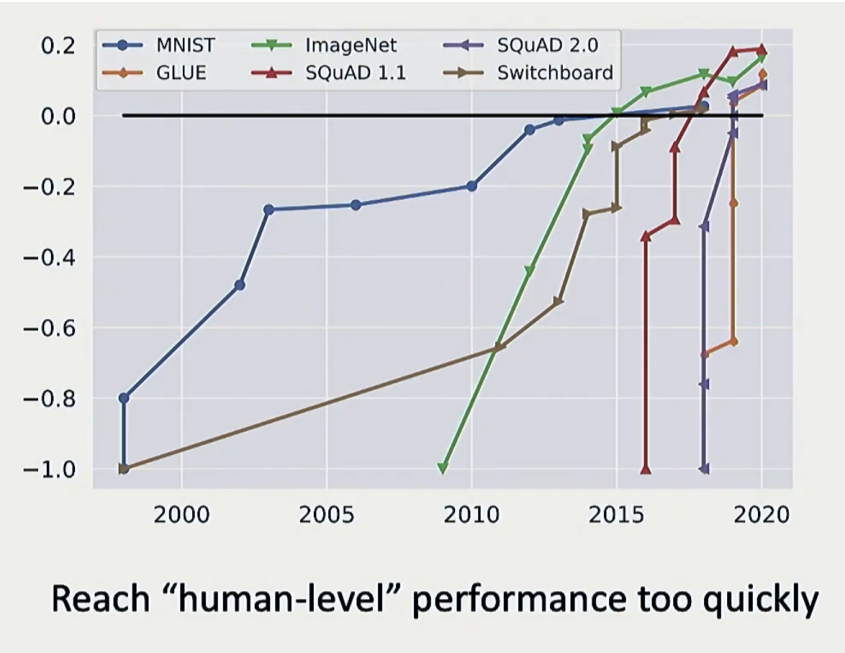
Consistency issues

For multiple choice questions, if you change ABCD to random symbols, generation will become different; ranking will also become different.



Overfitting issue

Reach "human-level" performance too quickly



Y-axis = (Model performance − Human performance)
There are different ways to alleviate/detect contamination or overfitting.

Monoculture of NLP benchmarking

Most papers only evaluate on English and performance (accuracy).

Area	# papers	English	Accuracy / F1	Multilinguality	Fairness and bias	Efficiency	Interpretability	>1 dimension
ACL 2021 oral papers	461	69.4%	38.8%	13.9%	6.3%	17.8%	11.7%	6.1%
MT and Multilinguality	58	0.0%	15.5%	56.9%	5.2%	19.0%	6.9%	13.8%
Interpretability and Analysis	18	88.9%	27.8%	5.6%	0.0%	5.6%	66.7%	5.6%
Ethics in NLP	6	83.3%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Dialog and Interactive Systems	42	90.5%	21.4%	0.0%	9.5%	23.8%	2.4%	2.4%
Machine Learning for NLP	42	66.7%	40.5%	19.0%	4.8%	50.0%	4.8%	9.5%
Information Extraction	36	80.6%	91.7%	8.3%	0.0%	25.0%	5.6%	8.3%
Resources and Evaluation	35	77.1%	42.9%	5.7%	8.6%	5.7%	14.3%	5.7%
NLP Applications	30	73.3%	43.3%	0.0%	10.0%	20.0%	10.0%	0.0%

Multi-lingual benchmarking

- Benchmarks exist, we should use them!
- MEGA: Multilingual Evaluation of Generative AI
- GlobalGench
- XTREME
- Multilingual Large Language Models Evaluation Benchmark (MMLU, ARC, HellaSwag translated in 26 languages)

Reductive single metric issue

- Performance is not all we care about:
  - Computational efficiency
  - Biases ...
- Taking averages for aggregation is unfair for minoritized groups
- Different preferences for different people

Other biases

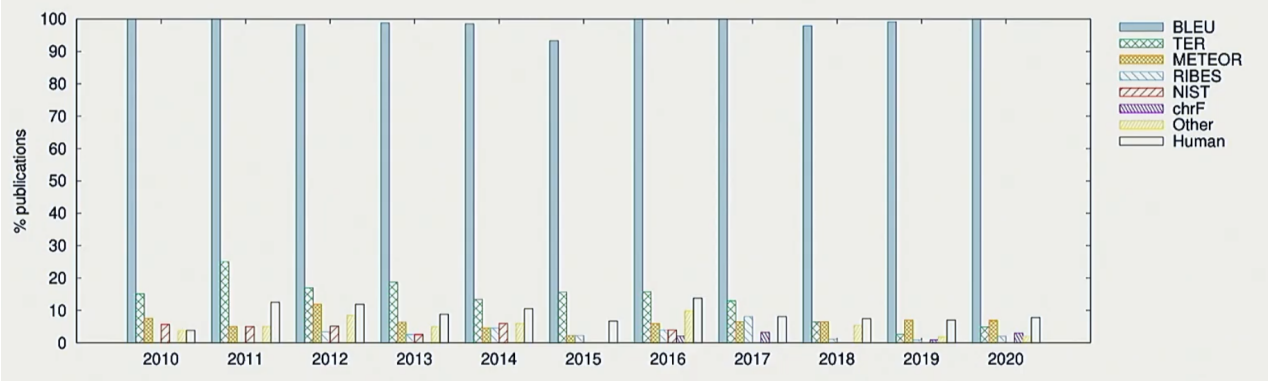
Whose Opinions Do Language Models Reflect?

Shibani Santurkar [1] Esin Durmus [1] Faisal Ladhak [2] Cinoo Lee [1] Percy Liang [1] Tatsunori Hashimoto [1]

Abstract

Language models (LMs) are increasingly being used in open-ended contexts, where the opinions they reflect in response to subjective queries can have a profound impact, both on user satisfaction, and shaping the views of society at large. We put forth a quantitative framework to investigate the opinions reflected by LMs – by leveraging high-quality public opinion polls. Using this framework, we create OpinionQA, a dataset for evaluating the alignment of LM opinions with those of 60 US demographic groups over topics ranging from abortion to automation. Across topics, we find substantial misalignment between the views reflected by current LMs and those of US demographic groups: on par with the Democrat-Republican divide on climate change. Notably, this misalignment persists even after explicitly steering the LMs towards particular groups. Our analysis not only confirms prior observations about the left-leaning tendencies of some human feedback-tuned LMs, but also surfaces groups whose opinions are poorly reflected by current LMs (e.g., 65+ and widowed individuals).

The challenges of challenges: status quo issue



- Academic researchers are incentivized to keep using the same benchmark to compare to previous work
- 82% papers of machine translation between 2019-2020 only evaluate on BLEU despite many metrics that correlate better with human judgment

Outline

Presentations

- Saeed: Wei et al. (2023). Chain-of-Thought Prompting.
- Karthik: Wang et al. (2023). Self-Consistency for Chain-of-Thought Reasoning.