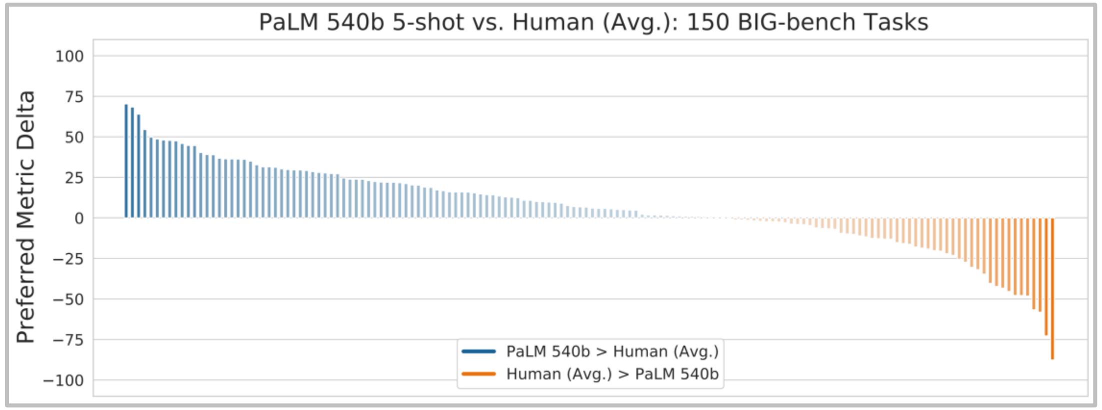
Developing and studying instruction-following models

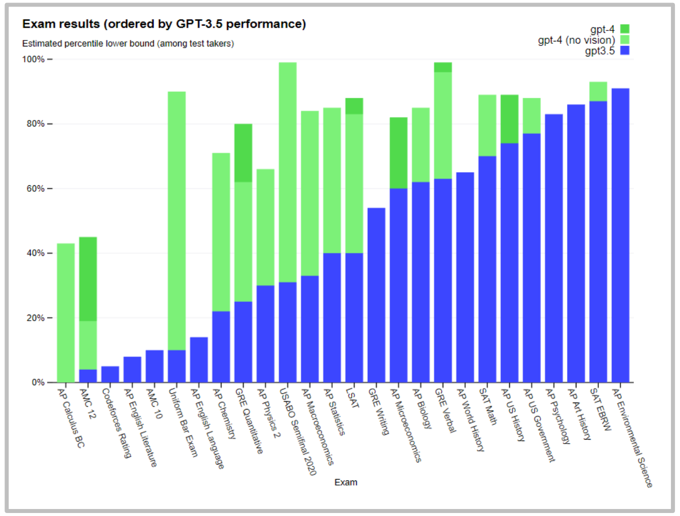
Tatsunori Hashimoto, Stanford CS

Future of Decentralization, AI, and Computing Summit

LLMs in the spotlight



Google PaLM on BigBench



GPT4 on a range of exams

Impressive, ongoing advances in NLP and AI from large language models!

These models are increasingly closed off

"On the competitive landscape front — it's competitive out there," said Sutskever. "GPT-4 is not easy to develop. It took pretty much all of OpenAI working together for a very long time to produce this thing. And there are many many companies who want to do the same thing, so from a competitive side, you can see this as a maturation of the field."

GPT-4

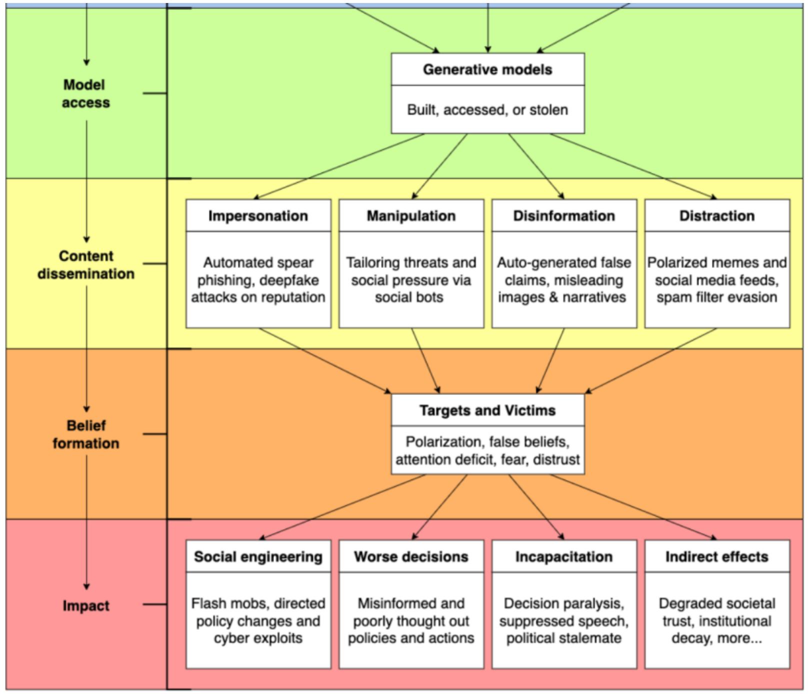
Jan Leike @janleike · Oct 24, 2022

I agree. While OpenAI doesn't like talking about exact model sizes / parameter counts anymore, documentation should definitely be better.

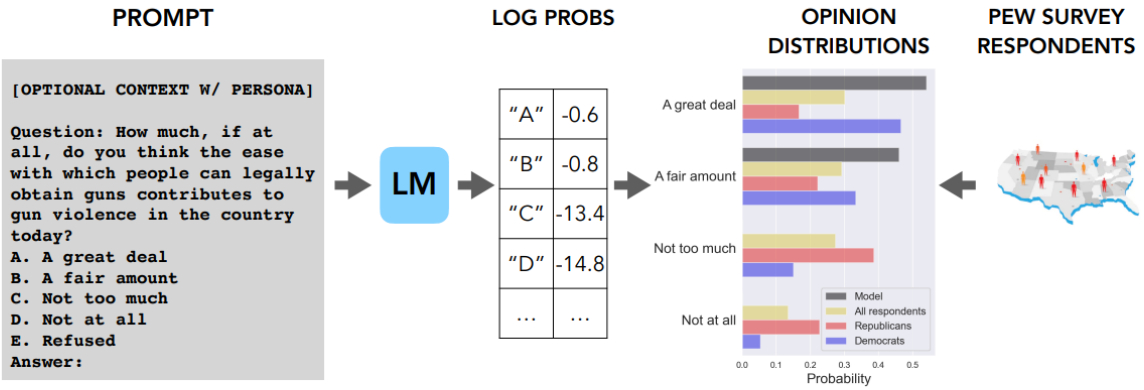
text-davinci-002 isn't the model from the InstructGPT paper. The closest to the paper is text-davinciplus-002.

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	BigScience	Meta	AI21 labs	ALEPH ALPHA	EleutherAI	
Draft AI Act Requirements	**GPT-4**	**Cohere Command**	**Stable Diffusion v2**	**Claude**	**PaLM 2**	**BLOOM**	**LLaMA**	**Jurassic-2**	**Luminous**	**GPT-NeoX**	**Totals**
Data sources	● ○ ○ ○	● ● ● ○	● ● ● ●	○ ○ ○ ○	● ● ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	22
Data governance	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	19
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	7
Compute	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ○ ○ ○	● ● ● ●	17
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	16

Closed models are hard to study and improve



Dual-use / misuse [Kang 2023]



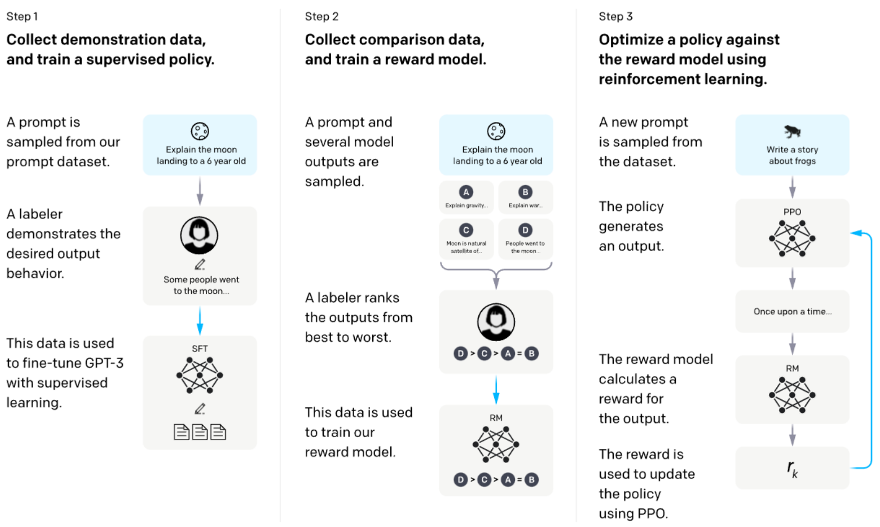
Political values / biases [Santurkar 2023]

API-only access makes it difficult to do deep analysis or propose improvements

Reproducible low-cost environments for LLM experiments

Reproducing instruction-following models

- **Cost :** high cost of human annotation
- **Replicability :** crowdsourcing doesn’t replicate
- **Reference :** no known working PPO implementation

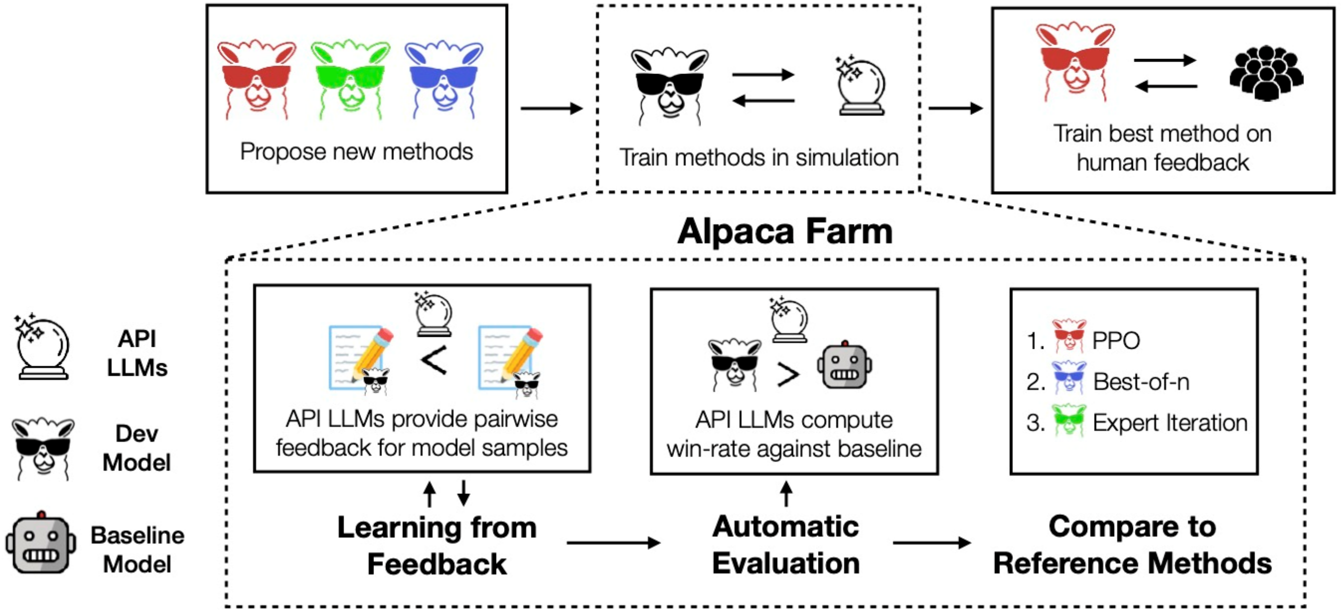


[Ouyang 2020]

- What’s the impact of instruction tuning?
- Does reinforcement learning actually help?
- What changes does RL actually make?

Why is this hard? Figuring this out (in full) requires replicating instructGPT/chatGPT

Alpaca trio: low-cost experiments for instruction-following



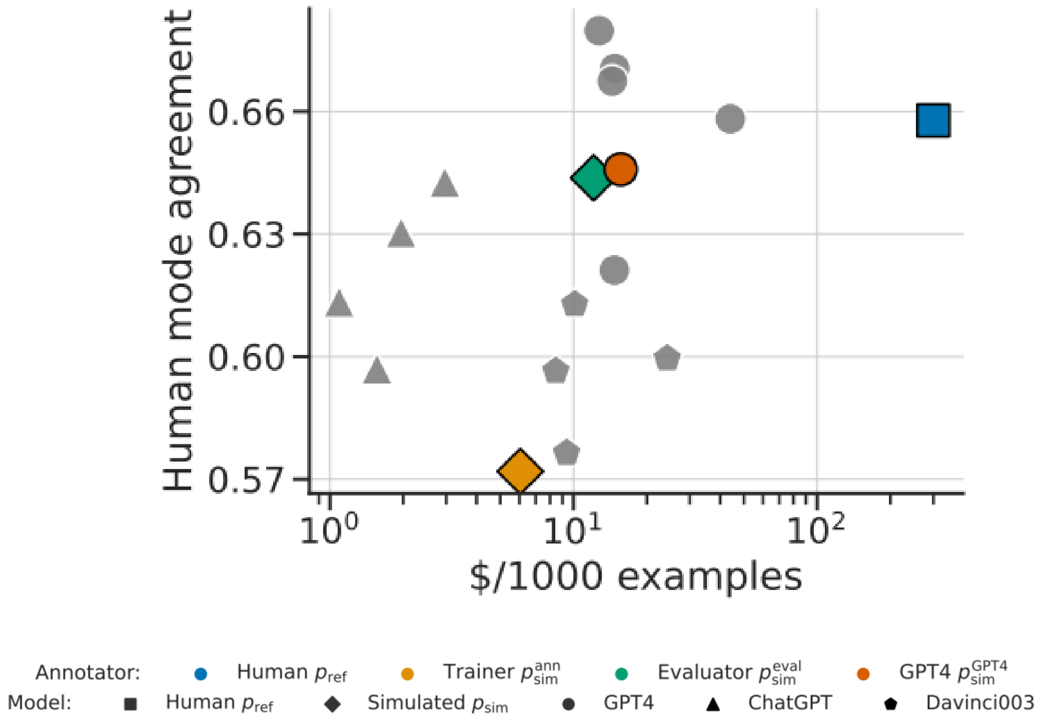
	Feedback Source	Experiment Cost	Iteration time	Reproducible
Alpaca Farm		$70	Hours	✅
Human Feedback		$3,150	Days	❌

Step 1 (SFT) – Alpaca **Step 2 (RLHF)** – AlpacaFarm **Step 3 (Evals)** - AlpacaEval

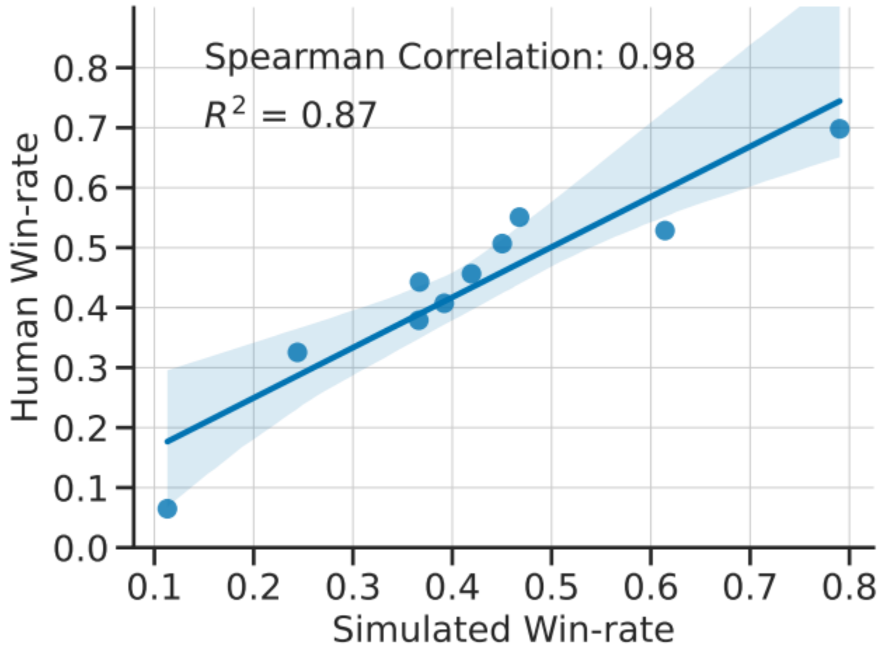
Simulating annotators (via GPT4) enables fast, low-cost prototyping and R&D of LLMs

[Dubois, Li, Taori, Zhang et al 2023]

Validating the accuracy of simulated annotations



Agreement near human inter-annotator levels



Near-perfect rank correlation at the system level

[Dubois, Li, Taori, Zhang et al 2023]

High-performance, reference methods for RLHF

Method	Simulated win-rate (%)
GPT-4	79.0 ± 1.4
ChatGPT	61.4 ± 1.7
PPO	46.8 ± 1.8
Best-of-n	45.0 ± 1.7
Expert Iteration	41.9 ± 1.7
SFT 52k (Alpaca 7B)	39.2 ± 1.7
SFT 10k	36.7 ± 1.7
Binary FeedME	36.6 ± 1.7
Quark	35.6 ± 1.7
Binary Reward Conditioning	32.4 ± 1.6
Davinci001	24.4 ± 1.5
LLaMA 7B	11.3 ± 1.1

Our findings replicate RLHF's effectiveness, and these results hold outside the simulator

[Dubois, Li, Taori, Zhang et al 2023]

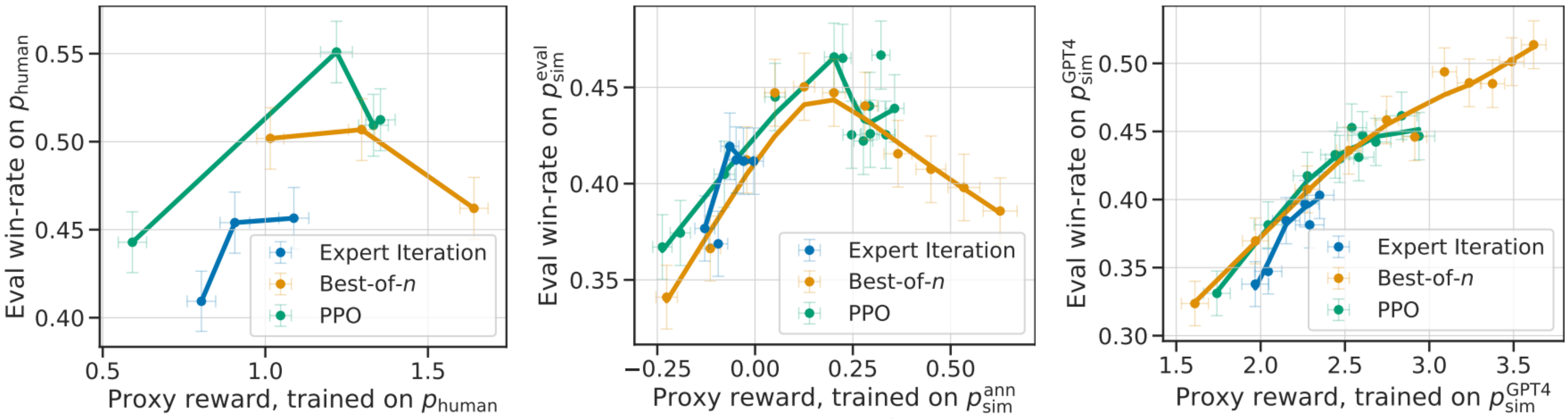
High-performance, reference methods for RLHF

Method	Simulated win-rate (%)	Human win-rate (%)
GPT-4	79.0 ± 1.4	69.8 ± 1.6
ChatGPT	61.4 ± 1.7	52.9 ± 1.7
PPO	46.8 ± 1.8	55.1 ± 1.7
Best-of-n	45.0 ± 1.7	50.7 ± 1.8
Expert Iteration	41.9 ± 1.7	45.7 ± 1.7
SFT 52k (Alpaca 7B)	39.2 ± 1.7	40.7 ± 1.7
SFT 10k	36.7 ± 1.7	44.3 ± 1.7
Binary FeedME	36.6 ± 1.7	37.9 ± 1.7
Quark	35.6 ± 1.7	-
Binary Reward Conditioning	32.4 ± 1.6	-
Davinci001	24.4 ± 1.5	32.5 ± 1.6
LLaMA 7B	11.3 ± 1.1	6.5 ± 0.9

Our findings replicate RLHF's effectiveness, and these results hold outside the simulator

[Dubois, Li, Taori, Zhang et al 2023]

AlpacaFarm highlights the complexity of instruction RLHF



(a) Human preferences (b) AlpacaFarm (c) Single-prompt GPT-4

AlpacaFarm replicates important, complex phenomena like overoptimization

[Dubois, Li, Taori, Zhang et al 2023]

Beyond this work: LLM driven prototyping lowers the cost of R&D

Textbooks Are All You Need

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, Yuanzhi Li

Microsoft Research

AlpacaEval : An Automatic Evaluator for Instruction-following Language Models

Code License Apache 2.0 | Data License CC By NC 4.0 | python 3.10+ | discord server

How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources

Yizhong Wang*♣♠, Hamish Ivison*♣, Pradeep Dasigi♣, Jack Hessel♣, Tushar Khot♣, Khyathi Raghavi Chandu♣, David Wadden♣, Kelsey MacMillan♣, Noah A. Smith♣♠, Iz Beltagy♣, Hannaneh Hajishirzi♣♠

Self-Alignment with Instruction Backtranslation

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, Mike Lewis

Meta AI

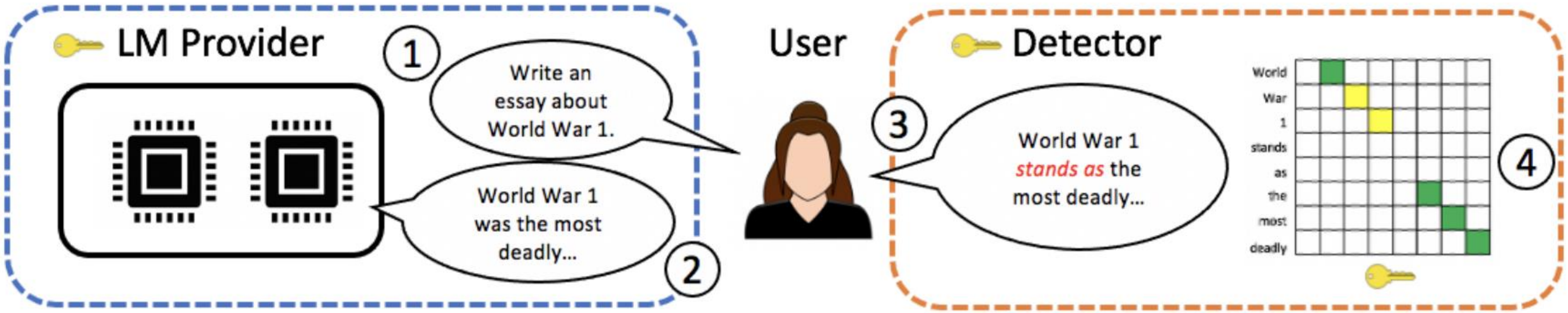
Studying fine-tuning data

Development metrics

Caveat: development and deployment needs more than automated data/evals

Development metrics, synthetic data → Crowdsourced data + evaluation → Live evaluation

Case study: watermarking LLMs



Watermarking enables tracking of LLM-generated text (see Kirchenbauer et al)

Challenges:

- Watermarks induce distortion (hard sell for LLM vendors)
- Many watermarks highly non-robust (to deletion of a few words, or cropping)

Development of a distortion-free, robust watermark.

In recent work [Kuditipudi et al 2023], we derive a distortion free and robust watermark.

Generate (for each token y_i)

- Draw a random sequence $\xi_i \in [0,1]$, call this the key
- Sample according to $\min_i -\log \xi_i / p_i$ (From Aaronson)

This is distortion free (i.e. the marginal distribution over ξ is p)

Detect

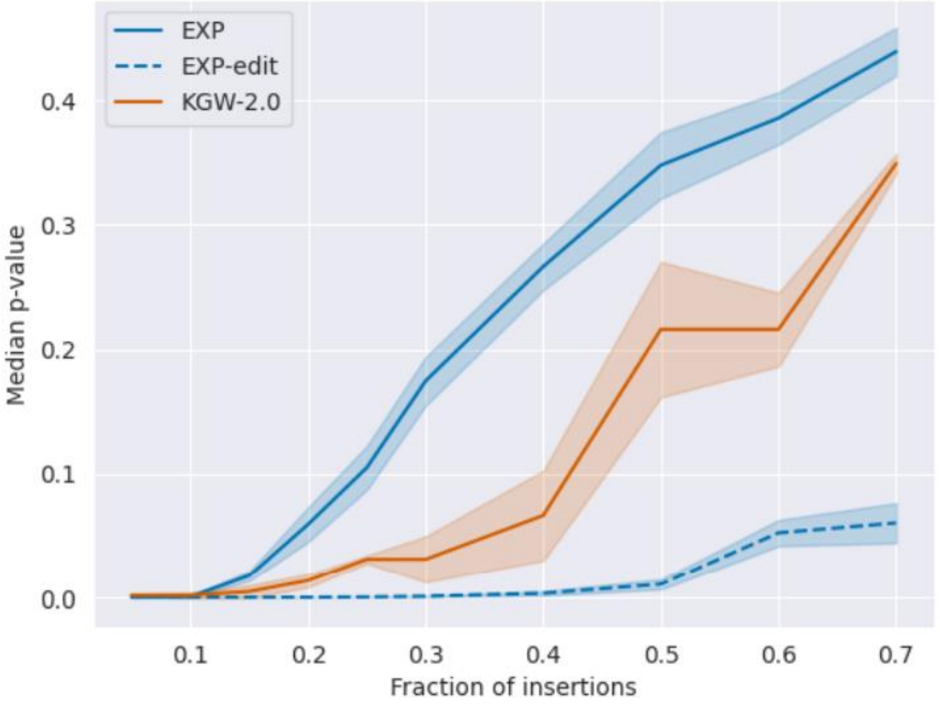
- Find the min-Levenshtein cost with $d(y, \xi) = \sum_i \log(1 - \xi_{i, y_i})$
- Compare vs the min-Levenshtein cost w/ random ξ

This is robust (i.e. can detect under small Levenshtein edits)

[Kuditipudi et al 2023]

Watermarks and open models

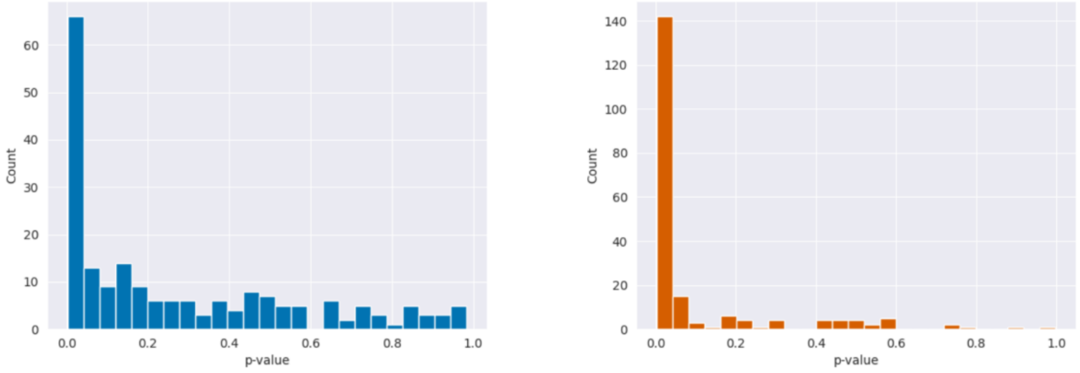
Open models (and access to logprobs) **enable watermarking research**



[Kuditipudi et al 2023]

Open instruction-tuning models and evals lead to new open problems

Distortion-free watermarks are weaker on instruction-tuned models



(a) EXP-edit

(b) KGW-2.0

Distortion-inducing watermarks lead to major drops in performance

70.86% → 51.2%

AlpacaEval Win Rate

Unwatermarked

KGW watermark

[Freeman and Hashimoto, unpublished]

Takeaways

Open models and trustworthiness

Open source provides important accountability and transparency

Research on LLMs

LLMs enable new research into instruction-following models

Enabling safer and more robust LLMs

New innovations and interventions based on open LLMs