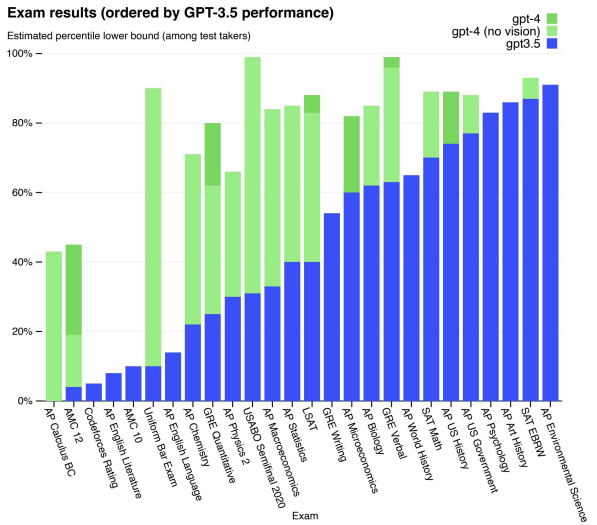
Section 1: Research

GPT-4 is out and it crushes every other LLM, and many humans

GPT-4 is OpenAI's latest Large Language Model. In contrast with text-only GPT-3 and follow-ups, GPT-4 is multimodal: it was trained on both text and images; it can among other capabilities generate text based on images. At 8,192 tokens when it was released, it had already exceeded the previous-best GPT-3.5 in possible input size. It is, of course, trained using RLHF. Equipped with these advances, GPT-4 is, as of the release of this report, the uncontested most generally capable AI model.

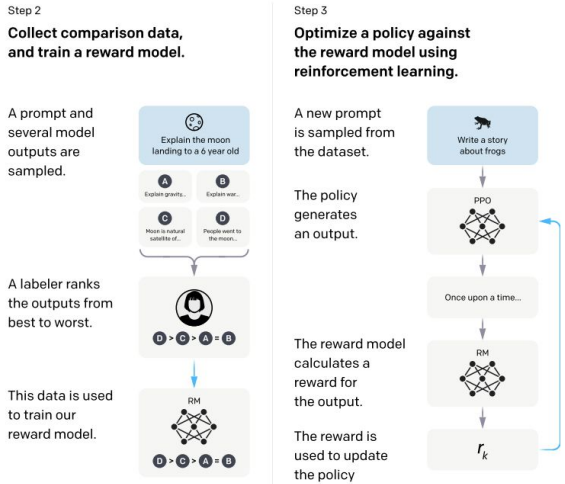
- OpenAI did a comprehensive evaluation of GPT-4 not only on classical NLP benchmarks, but also on exams designed to evaluate humans (e.g. Bar exam, GRE, Leetcode).
- GPT-4 is the best model across the board. It solves some tasks that GPT-3.5 was unable to, like the Uniform Bar Exam where GPT-4 scores 90% compared to 10% for GPT-3.5. On most tasks, the added vision component had only a minor impact, but it helped tremendously on others.
- OpenAI reports that although GPT-4 still suffers from hallucinations, it is factually correct 40% more often than the previous-best ChatGPT model on an adversarial truthfulness dataset (generated to fool AI models).



Fueled by ChatGPT's success, RLHF becomes MVP

In last year's Safety section (Slide 100), we highlighted how Reinforcement Learning from Human Feedback (RLHF) – used in InstructGPT – helped make OpenAI's models safer and more helpful for users. Despite a few hiccups, ChatGPT's success proved the technique's viability at a massive scale.

- *"RLHF involves humans ranking language model outputs sampled for a given input, using these rankings to learn a reward model of human preferences, and then using this as a reward signal to finetune the language model with using RL."* In its modern form, it dates back to 2017, when OpenAI and DeepMind researchers applied it to incorporate human feedback in training agents on Atari games and to other RL applications.
- RLHF is now central to the success of state of the art LLMs, especially those designed for chat applications. These include Anthropic's Claude, Google's Bard, Meta's LLaMa-2-chat, and, of course, OpenAI's ChatGPT.
- RLHF requires hiring humans to evaluate and rank model outputs, and then models their preferences. This makes this technique hard, expensive, and biased[1]. This motivated researchers to look for alternatives.



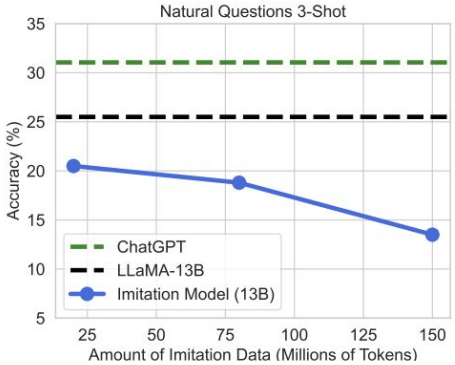
Typical steps of RLHF, which follow an initial step of supervised fine-tuning of a pre-trained language model, e.g. GPT-3.

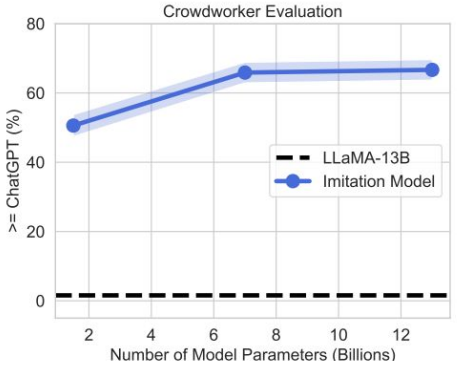
[1] *We will cover other issues of RLHF in the Safety section.*

The false promise of imitating proprietary LLMs, or how RLHF is still king

Berkeley researchers show that fine-tuning small LLMs on the outputs of larger, more capable LLMs results in models which are stylistically impressive but which often produce inaccurate text.

- The researchers examine a range of pretrained LLMs of different sizes and pre-trained on a varying amount of data. They show that at a fixed model size, using more imitation data actually hurts the quality of the output. In turn, larger models benefit from using imitation data.
- By using model size as a proxy for quality, the authors argue that more attention should be paid to better pre-training rather than fine-tuning on more imitation data.
- In the near future, RLHF seems here to stay. After careful ablation studies, Meta researchers concluded in their LLaMa-2 paper: *"We posit that the superior writing abilities of LLMs, as manifested in surpassing human annotators in certain tasks, are fundamentally driven by RLHF".*





Berkeley UNIVERSITY OF CALIFORNIA

Even so, researchers rush to find scalable alternatives to RLHF

In the wake of ChatGPT, many labs set out to answer the question: Can we create models as capable and safe as OpenAI's LLMs, but that drastically reduce human supervision?

- Anthropic proposed RL from AI feedback, which we cover in the safety section.
- Other approaches entirely do a way with reinforcement learning. In *Less is More for Alignment* (LIMA), Meta argues for using a few (1,000 in their paper) very carefully curated prompts and responses. According to human evaluations of model outputs, LIMA is competitive with GPT-4 in 43% of cases.
- In *LLMs can self-improve*, Google researchers showed that LLMs can improve by training on their own outputs. In a similar vein, Self-Instruct is a framework in which a model generates its own instructions, input and output samples, and curates them to finetune its parameters. Yet another work in this direction is Meta's Self-Alignment with Instruction Backtranslation.
- Stanford researchers used this last approach to generate instructions and outputs using GPT-3.5 and fine-tune Meta's LLaMa-7B.

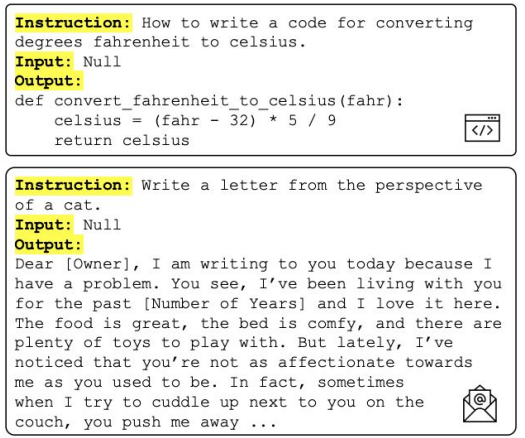
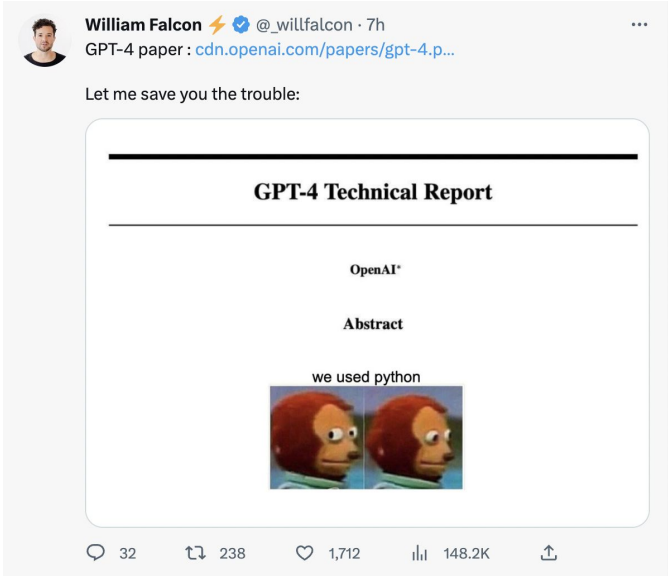


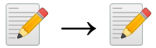
Figure 1: Selected tasks from the generated instruction data using vanilla GPT3. Some texts are reformatted for presentation. See Table 10 for more examples.

The GPT-4 technical report puts the nail in the coffin of SOTA LLM research...

OpenAI published a technical report on GPT-4 where it didn't disclose any useful information for AI researchers, signalling the definitive industrialization of AI research. Google's PaLM-2 technical report suffered the same fate, while (OpenAI spinoff) Anthropic didn't bother releasing a technical report for its Claude models.

- *"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar"*, OpenAI writes in the GPT-4 technical report published on arXiv.
- When Google released PaLM 2, its most capable LLM, the company wrote in the technical report: *"Further details of model size and architecture are withheld from external publication."*
- As the economic stakes and the safety concerns are getting higher (you can choose what to believe), traditionally open companies have embraced a culture of opacity about their most cutting edge research.





...unless LLaMas reverse the trend

In February'23, Meta released a series of models called LLaMa. At their release, they stood out as being the most capable models trained exclusively on publicly available datasets. Meta initially granted access to the LLaMa model weights on demand only to researchers, but the weights were quickly leaked and published online.

- The LLaMa-1 models use regular transformers, with slight changes to the architecture. The authors also made a few changes to the optimizer and to the implementation of attention. As a result, *"when training a 65B-parameter model, [their] code processes around 380 tokens/sec/GPU on 2048 A100 GPU with 80GB of RAM. This means that training over [their] dataset containing 1.4T tokens takes approximately 21 days."*
- The LLaMa-1 models outperform GPT-3 (the original one, not the InstructGPT variants) and are competitive with DeepMind's Chinchilla and Google's PaLM.
- LLaMa-1 didn't allow commercial use, prompting heavy criticism around the term "open-source" that Meta used to describe the model release. But a second LLaMa iteration appeased most of the open source community.

		RACE-middle	RACE-high
GPT-3	175B	58.4	45.5
PaLM	8B	57.9	42.3
	62B	64.3	47.5
	540B	**68.1**	49.1
LLaMA	7B	61.1	46.9
	13B	61.6	47.2
	33B	64.1	48.3
	65B	67.9	**51.6**

Table 6: **Reading Comprehension.** Zero-shot accuracy.

Meta

LLaMa sets off a race of open(ish) competitive Large Language Models

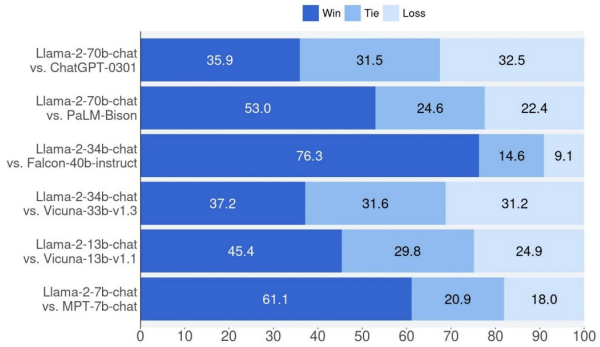
After Meta released LLaMa-1, other institutions joined the movement to release the weights of relatively large language models. A few of them stand out, like MosaicML's MPT-30B, TII UAE's Falcon-40B, Together's RedPajama, or Eleuther's Pythia. Meanwhile another dynamic was taking place, where the open-source community fine-tuned the smallest versions of LLaMa on specialized datasets and applied them to dozens of downstream applications. Mistral AI's 7B model also recently emerged as the strongest small model.

- Notably, RedPajama aimed to exactly replicate LLaMa-1 to make it fully open-source. Falcon 40B came from a new entrant in the LLM sweepstakes, TII UAE, and was quickly made open-source. Falcon-180B was later released, but was notably trained on very little code, and not tested on coding.
- Helped with parameter-efficient fine-tuning methods like LoRa (Low-rank adaptation of LLMs – initially by Microsoft), LM practitioners started fine-tuning these pre-trained LLMs for specific applications like (of course) chat. One example is LMSys's Vicuna which is LLaMa fine-tuned on user-shared conversations with ChatGPT.

LLaMa-2: the most generally capable and publicly accessible LLM?

In July'23, the LLaMa-2 series of models was released, giving (almost) everyone the right for commercial use. The base LLaMa-2 model is almost identical to LLaMa-1 but further fine-tuned using instruction tuning and RLHF and optimized for dialogue applications. In September 2023, Llama-2 as had almost 32M downloads.

- The pre-training corpus for LLaMa-2 has 2 trillion tokens (40% increase).
- For supervised fine-tuning, the researchers tried publicly available data, but what was most helpful was using a few (24,540) high-quality vendor-based annotations. For RLHF, they use binary comparison and split the RLHF process into prompts and answers designed to be helpful to the user and others designed to be safe.
- LLaMa-2 70B is competitive with ChatGPT on most tasks except for coding, where it significantly lags behind it. But CodeLLaMa, a fine-tuned version for code beats all non-GPT4 models (more on this later).
- Per Meta terms, anyone (with enough hardware to run the models) can use the LLaMa-2 models, as long as their commercial application didn't have more than 700M users at the time of LLaMa-2's release.



Human evaluation of LLaMa-2 helpfulness vs. other open source models

GPT and LLaMAs win the popularity contest

- **ChatGPT has the highest number of mentions on X (5430 times), followed by GPT-4 and LLaMA. While proprietary, closed-source models get the most attention, there's an increase in interest in LLMs that are open-source and allow commercial use.**