The GPT-4 barrier has finally been broken

8th March 2024

Four weeks ago, GPT-4 remained the undisputed champion: consistently at the top of every key benchmark, but more importantly the clear winner in terms of "vibes". Almost everyone investing serious time exploring LLMs agreed that it was the most capable default model for the majority of tasks—and had been for more than a year.

Today that barrier has finally been smashed. We have four new models, all released to the public in the last four weeks, that are benchmarking near or even above GPT-4. And the all-important vibes are good, too!

Those models come from four different vendors.

- Google Gemini 1.5, February 15th. I wrote about this the other week: the signature feature is an incredible one million long token context, nearly 8 times the length of GPT-4 Turbo. It can also process video, which it does by breaking it up into one frame per second—but you can fit a LOT of frames (256 tokens each) in a million tokens.
- Mistral Large, February 26th. I have a big soft spot for a mistral given how exceptional their openly licensed models are—Mistral 7B runs on my iPhone, and Mixtral-8x7B is the best model I've successfully run on my laptop. Medium and Large are their two hosted but closed models, and while Large may not be quite outperform GPT-4 it's clearly in the same class. I can't wait to see what they put out next.
- Claude 3 Opus, March 4th. This is just a few days old and wow: the vibes on this one are *really* strong. People I know who evaluate LLMs closely are rating it as the first clear GPT-4 beater. I've switched to it as my default model for a bunch of things, most conclusively for code—I've had several experiences recently where a complex GPT-4 prompt that produced broken JavaScript gave me a perfect working answer when run through Opus instead (recent example). I also enjoyed Anthropic research engineer Amanda Askell's detailed breakdown of their system prompt.
- Inflection-2.5, March 7th. This one came out of left field for me: Inflection make Pi, a conversation-focused chat interface that felt a little gimmicky to me when I first tried it. Then just the other day they announced that their brand new 2.5 model benchmarks favorably against GPT-4, and Ethan Mollick—one of my favourite LLM sommeliers—noted that it deserves more attention.

Not every one of these models is a clear GPT-4 beater, but every one of them is a contender. And like I said, a month ago we had none at all.

There are a couple of disappointments here.

Firstly, none of those models are openly licensed or weights available. I imagine the resources they need to run would make them impractical for most people, but at after a year that has seen enormous leaps forward in the openly licensed model category it's sad to see the very best models remain strictly proprietary.

And unless I've missed something, none of these models are being transparent about their training data. This also isn't surprising: the lawsuits have started flying now over training on unlicensed copyrighted data, and negative public sentiment continues to grow over the murky ethical ground on which these models are built.

It's still disappointing to me. While I'd love to see a model trained entirely on public domain or licensed content—and it feels like we should start to see some strong examples of that pretty soon—it's not clear to me that it's possible to build something that competes with GPT-4 without dipping deep into unlicensed content for the training. I'd love to be proved wrong on that!

In the absence of such a vegan model I'll take training transparency over what we are seeing today. I use these models a lot, and knowing how a model was trained is a powerful factor in helping decide which questions and tasks a model is likely suited for. Without training transparency we are all left reading tea leaves, sharing conspiracy theories and desperately trying to figure out the vibes.

Posted 8th March 2024 at 6:02 pm · Follow me on Mastodon or Twitter or subscribe to my newsletter

More recent articles

- Prompt injection and jailbreaking are not the same thing - 5th March 2024
- Interesting ideas in Observable Framework - 3rd March 2024
- Weeknotes: Getting ready for NICAR - 27th February 2024
- The killer app of Gemini Pro 1.5 is video - 21st February 2024
- Weeknotes: a Datasette release, an LLM release and a bunch of new plugins - 9th February 2024
- Datasette 1.0a8: JavaScript plugins, new plugin hooks and plugin configuration in datasette.yaml - 7th February 2024
- LLM 0.13: The annotated release notes - 26th January 2024
- Weeknotes: datasette-test, datasette-build, PSF board retreat - 21st January 2024
- Talking about Open Source LLMs on Oxide and Friends - 17th January 2024

ai 467 openai 121 generativeai 404 gpt4 30 llms 378 anthropic 17 claude 17 mistral 7

Previous: Prompt injection and jailbreaking are not the same thing