

The world's biggest AI models aren't very transparent, Stanford study says

No prominent developer of AI foundation models – a list including companies like OpenAI and Meta – is releasing sufficient information about their potential impact on society, determines a new report from Stanford HAI (Human-Centered Artificial Intelligence).

Today, Stanford HAI released its Foundation Model Transparency Index, which tracked whether creators of the 10 most popular AI models disclose information about their work and how people use their systems. Among the models it tested, Meta's Llama 2 scored the highest, followed by BloomZ and then OpenAI's GPT-4. But none of them, it turned out, got particularly high marks.

Other models evaluated include Stability's Stable Diffusion, Anthropic's Claude, Google's PaLM 2, Command from Cohere, AI21 Labs' Jurassic 2, Inflection-1 from Inflection, and Amazon's Titan.

The researchers acknowledged to *The Verge* that transparency can be a broad concept. Their definition is based on 100 indicators for information about how the models are built, how they work, and how people use them. They parsed publicly available information on the model and gave each a score, noting if the companies disclosed partners and third-party developers, if they tell customers whether their model used private information, and a host of other questions.

Meta scored 54 percent, scoring highest on model basics, as

the company released its research into model creation. BloomZ, an open-source model, followed closely at 53 percent and GPT-4 at 48 percent – followed by Stable Diffusion at 47 percent despite OpenAI's relatively locked-down design approach.

OpenAI refuses to release much of its research and does not disclose data sources, but GPT-4 managed to rank high because there's a great deal of available information about its partners. OpenAI works with many different companies that integrate GPT-4 into their products, producing a lot of public details to look at.

The Verge reached out to Meta, OpenAI, Stability, Google, and Anthropic but has not received comments yet.

However, none of the models' creators disclosed any information about societal impact, Stanford researchers found – including where to direct privacy, copyright, or bias complaints.

Rishi Bommasani, society lead at the Stanford Center for Research on Foundation Models and one of the researchers in the index, says the goal of the index is to provide a benchmark for governments and companies. Some proposed regulations, like the EU's AI Act, could soon compel developers of large foundation models to provide transparency reports.

"What we're trying to achieve with the index is to make models more transparent and disaggregate that very amorphous concept into more concrete matters that can be measured," Bommasani says. The group focused on one model per company to make comparisons easier.

Generative AI has a large and vocal open-source community, but some of the biggest companies in the space do not publicly share research or their codes. OpenAI, despite having the word "open" in its name, no longer distributes its research, citing competitiveness and safety concerns.

Bommasani says the group is open to expanding the scope of the index but, in the meantime, will stick to the 10 foundation models it's already evaluated.

<https://www.theverge.com/2023/10/18/23922973/stanford-ai-foundation-model-transparency-index>