TechScape: Will Meta’s open-source LLM make AI safer – or put it into the wrong hands?

Publication Date: 2023-07-25

Author: Alex Hern

Section: Technology

Tags: Apple, TechScape newsletter, Computing, Meta, Microsoft, OpenAI, X, Elon Musk, newsletters

Article URL: https://www.theguardian.com/technology/2023/jul/25/techscape-meta-open-source-large-language-models-llm-ai-twitter-x-apple

The AI summer is well and truly upon us. (This gag may not play as well for readers in the southern hemisphere.) Whether we call this period the peak of the “hype cycle” or simply the moment the curve goes vertical will only be obvious in hindsight, but the cadence of big news in the field has gone from weekly to almost daily. Let’s catch up with what the biggest players in AI – Meta, Microsoft, Apple and OpenAI – are doing. Apple Always one to keep its cards close to its chest, don’t expect to hear of many R&D breakthroughs from Cupertino. Even the AI work that has made it into shipping products is hidden rather than shouted from the rooftops, with the company talking about “machine learning” and “transformers” at its annual worldwide developer conference (WWDC) last month, but conspicuously steering clear of saying “AI”. But that doesn’t mean it’s not playing the same game as everyone else. Per Bloomberg (£): The iPhone maker has built its own framework to create large language models — the AI-based systems at the heart of new offerings like ChatGPT and Google’s Bard – according to people with knowledge of the efforts. With that foundation, known as “Ajax”, Apple also has created a chatbot service that some engineers call “Apple GPT”. In recent months, the AI push has become a major effort for Apple, with several teams collaborating on the project, said the people, who asked not to be identified because the matter is private. The work includes trying to address potential privacy concerns related to the technology. On one hand: of course they are. It’s hard to remember, because it’s been on the back foot for so long, but Apple led the industry with voice assistants when it launched Siri in 2011. But within a few years – certainly by the launch of the Echo smart speaker in 2014 – it had fallen behind, and has now been relegated almost to the status of a joke. Fixing Siri is a hard job, but it’s one that the cutting edge of LLM work is perfect for. So it’s no surprise that the company is working at it. On the other: building a foundation model is hard, expensive – and perhaps unnecessary. Apple has built on top of open-source roots before (every single one of its operating systems, for instance, ultimately sits on top of the open-source Darwin kernel), and has licensed technology from third parties (most notably, these days, Arm, which still provides the core designs for its chips). And there’s plenty of opportunities for either of those approaches … Meta and Microsoft Meta’s Llama foundation model has become the accidental, er, foundation of an entire research community. The GPT competitor was released for download to a select group of researchers, who had signed NDAs and promised not to share it more broadly … when it promptly leaked. Samizdat copies have been shared across the net, as has a whole system for collaborating without ever openly publishing the stolen LLM. The whole thing was against Meta’s terms, but the company didn’t seem too unhappy about being central to a computing revolution. And now, it’s official. Meta’s released Llama 2 with terms of service that legitimise that ecosystem. From its announcement: We’re now ready to open source the next version of Llama 2 and are making it available free of charge for research and commercial use. We're including model weights and starting code for the pretrained model and conversational fine-tuned versions too. And the company has partnered with Microsoft to expand access: Starting today, Llama 2 is available in the Azure AI model catalog, enabling developers using Microsoft Azure to build with it and leverage their cloud-native tools for content filtering and safety features. It is also optimized to run locally on Windows, giving developers a seamless workflow as they bring generative AI experiences to customers across different platforms The model is free-as-in-beer, rather than free-as-in-speech, though. Meta's commercial terms require a license from any company with more than 700 million monthly active users – essentially, every other company discussed in today's newsletter and very few others. On top of that, it prevents anyone from using Llama 2 to improve other LLMs. It might be free, in other words, but it isn't open source. OpenAI But it's still more open than the competition. Allowing users, researchers, and (smaller) competitors to download the full model and poke around to see how it ticks obviously helps anyone who wants to build on top of what you've made, but it also helps build trust with potential partners. For a sign of the pitfalls that come with the opposite approach, take a look at OpenAI. From Ars Technica: In a study titled "How is ChatGPT's behavior changing over time?" published on arXiv, Lingjiao Chen, Matei Zaharia, and James Zou, cast doubt on the consistent performance of OpenAI's large language models (LLMs), specifically GPT-3.5 and GPT-4. Using API access, they tested the March and June 2023 versions of these models on tasks like math problem-solving, answering sensitive questions, code generation, and visual reasoning. Most notably, GPT-4's ability to identify prime numbers reportedly plunged dramatically from an accuracy of 97.6 percent in March to just 2.4 percent in June. Strangely, GPT-3.5 showed improved performance in the same period. The results play into a widely held fear that efforts to improve the safety of GPT are making it dumber. OpenAI certainly releases tweaks to GPT on a regular basis, and given the regularity with which chief executive Sam Altman talks about AI safety, it's perfectly plausible that those tweaks are largely safety-focused. And so if the system is getting worse, not better, it's perhaps because of that trade-off. But the paper itself doesn't hold up. Ars Technica, again: AI researcher Simon Willison also challenges the paper's conclusions. "I don't find it very convincing," he told Ars. "A decent portion of their criticism involves whether or not code output is wrapped in Markdown backticks or not"… So far, Willison thinks that any perceived change in GPT-4's capabilities comes from the novelty of LLMs wearing off. After all, GPT-4 sparked a wave of AGI panic shortly after launch and was once tested to see if it could take over the world. Now that the technology has become more mundane, its faults seem glaring. But the accusations strike at the heart of OpenAI's (ironically) closed model. The company rolls out changes to GPT on a regular basis, with little explanation, and no ability for users to understand why or how each new model differs. Inspecting any LLM is a "black box" problem, with little ability to peek inside and see how it thinks – but those problems are far worse when your only way of interacting at all is through an API to a version hosted by a third party. Ars Tehcnica, one last time: Willison agrees. "Honestly, the lack of release notes and transparency may be the biggest story here," he told Ars. "How are we meant to build dependable software on top of a platform that changes in completely undocumented and mysterious ways every few months?" X marks the spot So Twitter has a new name: here's everything we know so far. Elon Musk reveals Twitter's new logo X, part of a risky WeChat-inspired rebrand that is to be "centred in audio, video, messaging, payments/banking". Dan Milmo is very good on whether the rebrand can turn around Twitter and make it an "everything app". Hiccup No 1: police mistakenly stop workers from changing Twitter's sign at its San Francisco HQ. Hiccup No 2: Meta already appears to hold the rights to 'X'. It could make Twitter's rebrand complicated, reports Business Insider. Why is Musk so obsessed with X? Andrew Lawrence reports. Every TikTok is worth 1,000 words: in a response to Twitter and Threads, the video-sharing platform now offers the option to create lengthy text-only posts. If you want to read the complete version of the newsletter please subscribe to receive TechScape in your inbox every Tuesday.