SUPER DATASCIENCE
MAKING THE COMPLEX SIMPLE

SDS PODCAST EPISODE 778 FIVE-MINUTE FRIDAY: MIXTRAL 8X22B: SOTA OPEN-SOURCE LLM CAPABILITIES AT A FRACTION OF THE COMPUTE

(00:05):

This is Five-Minute Friday on Mixtral 8x22B.

(00:27):

Welcome back to The Super Data Science Podcast. I'm your host, Jon Krohn. It's been a long long time since a Five-Minute Friday episode actually was only five minutes long. I'm going to do my best today to give you an actually-only-about-five-minute update on a groundbreaking new open-source Large Language Model called Mixtral 8x22B out of an extremely hot French startup called Mistral.

(00:51):

Given the Mistral name, you are correct to assume that their Mixtral model is related to their company name. The "mix" part of "Mixtral" comes from the fact that the Mixtral 8x22B model is a mixture-of-experts model consisting of eight 22-billion-parameter "expert" submodels. I've discussed this mixture-of-experts approach before in the context of OpenAI's GPT-4, which is rumored to also consist of eight expert submodels, each of which specializes in a different kind of natural-language-generation task. It's probably not as clear cut as this in reality, but for the sake of providing an illustrative example, you could think of one of the submodel experts being called upon to handle code generation tasks while another one of the submodels is called upon to specifically handle say math-related tasks.

(01:44):

The big advantage of this mixture-of-experts approach is that, at inference time, when you actually use the model in production for real-world tasks, only a fraction of the full model needs to be used. Building on my caricature example and applying it to Mixtral 8x22B, if you ask a math question of this mixture-of-experts model a small part of the model would be used to triage your prompt to the particular 22-billion-parameter math-expert submodel while the other seven 22-billion-parameter submodels can remain unused. This means that, to provide a given response to your prompt, the Mixtral 8x22B LLM uses only 39 billion of its 141 billion total model parameters, saving about 75% of the cost and 75% of the time relative to if you were to use all of the model's 141 billion neurons.

(02:42):

Previously, Mistral released a 7-billion parameter model not a mixture of experts, just a standalone dense, what they call dense 7-billion parameter model, meaning that you need to use all that 7-billion parameters on every call, and that 7-billion parameter model outperformed other leading open-source LLMs, even larger ones, such as Meta's Llama 13B and 33B LLMs and it did this on the most popular LLM benchmark called MMLU, which stands for multi-task language understanding. More recently, Mistral released their first mixture-of-experts model, so this was a couple of months ago, and this consists of eight 7-billion-parameter submodels. So first they had a single 7-billion parameter model dense, you had to call all of it, and then they had their first mixture of experts model which consists of 8x7-billion parameter submodels. And that one, the 8x7B eclipsed the much more expensive-to-run Llama 70B to make the Mixtral 8x7B model the most capable open-source LLM yet according to the MMLU benchmark, and that was a couple of months ago.

(03:56):

With the Mixtral 8x22B released just last week, Mistral outdoes their 8x7B mixture-of-experts model to set new high watermarks across the gamut of all major natural-language common sense, reasoning and knowledge benchmarks. It also outperforms all other open-source models on French, German, Spanish and Italian tasks. So it also does very well on major non-English languages. And it beats all the other open-source models on coding and math benchmarks as well.

(04:41):

Unlike other so-called “open-source” models like Meta’s Llama 2 which have restrictions on their use, Mistral is actually releasing the Mixtral 8x22B under an Apache 2.0 license, which is the most permissive open-source license — it allows anyone anywhere to use Mixtral 8x22B without any limitations whatsoever. It also has a pretty darn solid context window of 64k tokens and it is natively capable of calling functions so Mixtral 8x22B can convert natural-language into API calls inside of a software application, allowing developers to dramatically modernize their software by allowing the software to respond to natural-language requests from users.

(05:32):

All in all, Mixtral 8x22B is an important update from Mistral for data scientists and software developers all over the world, allowing for state-of-the-art open-source LLM performance across many natural human languages, coding and math-related text-generation tasks, all while being less expensive to run in production than the models like Llama 70B that it overtook on all these benchmarks. While LLM benchmarks should not be trusted on their own, the anecdotal response of users of Mixtral 8x22B online suggests that it lives up to its quantitative hype. You can download the model today to adapt to your own personal or professional uses; as always, I’ve got a link for you in the show notes.

(06:16):

All right, that’s it for today’s episode. If you enjoyed today’s episode or know someone who might, consider sharing this episode with them, leave a review of the show on your favorite podcasting platform or on YouTube, tag me in a LinkedIn or Twitter post with your thoughts, I’ll respond to those, or if you aren’t already, of course, subscribe to the show. Most importantly, however, we just hope you’ll just keep on listening. Until next time, keep on rockin' it out there and I'm looking forward to enjoying another round of the Super Data Science podcast with you very soon.