SuperDataScience

SDS PODCAST

EPISODE 976: NVIDIA'S NEMOTRON 3 SUPER: THE PERFECT LLM FOR MULTI-AGENT SYSTEMS

Jon Krohn: 00:00 This is episode number 976 on NVIDIA's Nemotron three Super. Welcome back to the SuperDataScience podcast. I'm your host, Jon Krohn. Today's topic is NVIDIA's brand new Nemotron three super model, which is a mouthful, and it was announced to coincide with this week's big NVIDIA conference, GTC. Nmustron three Super is an openly available model that deserves your attention, not only because of its impressive technical specs, but because of what it signals about where the AI industry is headed, specifically toward agentic AI systems that can reason, use tools and operate autonomously over extended workflows. So let's start with the basics. Nimotron3Super is a 120 billion parameter model, but only 12 billion of those parameters, 10%, are active at any given time during inference. This is because the model uses a mixture of experts, architecture, or MOE for short, where different subsets of the model's parameters, the so-called experts, are selectively activated depending on the input.

01:05 So you get the knowledge capacity of a 120 billion parameter model, but with the computational cost closer to a 12 billion parameter one. That's a massive efficiency win. And if you'd like to hear more about mixture of experts, refer back to episode number 778 of this show. But now, Nimo Tron3Super isn't just any mixture of experts model. It's built on a hybrid architecture that combines two fundamentally different approaches to sequence processing. The common transformer-based attention layers that predominate LLMs today, as well as relatively exotic, though increasingly common Mamba layers. I did a whole episode on Mamba back in episode number 758, but quickly, Mamba is a so- called states-based model that processes sequences in linear time with respect to sequence length, which is way more efficient than the quadratic scaling you get with the traditional transformer self-attention. This is what makes Nemotron three supers one million token context window practical rather than theoretical.

02:04 But pure space-based models can struggle with precise retrieval tasks, finding one specific piece of information buried deep in a long context. So NVIDIA interleaves a small number of transformer-based attention layers at key depths to preserve high-fidelity information retrieval. It's a best of both worlds design. Mamba for efficiency, transformers for precision. On top of this hybrid backbone, NVIDIA introduced a novel technique called latent MOE, latent mixture of experts. In a standard MOE setup, tokens are routed to experts in their full hidden dimension, which gets expensive as models scale. With latent MOE, tokens are first compressed into a smaller latent space before routing, which dramatically cuts computational overhead. The savings are reinvested to activate four times as many expert specialists for the cost of one in a traditional setup. More experts consultant per token means specialized knowledge being brought to bear on each prediction, and the data show this translates directly into better accuracy.

03:07 There's one more architectural innovation worth highlighting, multi-token prediction or MTP. Standard language models predict one token at a time. Nemotron3super predicts multiple future tokens simultaneously using specialized prediction heads. At inference time, these heads function as a built-in draft model for speculative decoding. You generate several candidate tokens quickly and verify them in a single forward pass. The result is up to a three times wall clock speed up for structured generation tasks like code or tool calls, and you don't need a separate external draft model to get it. These architectural choices together deliver impressive throughput numbers. On an 8,000 input token and 64,000 output token benchmark, Nemotron 3Super achieves up to 2.2 times higher throughput than the comparably sized GPT OSS 120B and up to 7.5 times higher throughput than Quen 3.5 at 122 billion parameters, while in both cases, matching or exceeding on accuracy. The model was also pre-trained natively in

NVIDIA's four bit NVFP4 precision, which on Blackwell GPUs pushes inference up to four times faster than FP8 on the previous generation hopper GPUs, and again, with no loss inaccuracy.

04:26 Now, why does all of this matter? As companies move beyond simple chatbot interactions into multi-agent AI applications, they run into two major bottlenecks. The first is what's called context explosion. Multi-agent workflows can generate up to 15 times more tokens than a standard chat because each interaction requires resending full histories, tool outputs, and intermediate reasoning. This ballooning context increases cost and can cause goal drift where agents gradually lose alignment with their original objective. Nimotron3Super's million token context window lets agents retain the full state of a workflow in memory without truncation. Cool. The second bottleneck is the thinking tax. Complex agents need to reason at every step, but deploying a large expensive model for every sub task requires or makes multi-agent pipelines too slow and too costly. Nemotron3super's combination of sparse mixture of expert's computation and Momba-based efficiency is aimed squarely at making step-by-step reasoning affordable at scale.

05:28 That's its core value proposition, frontier class reasoning at a fraction of the typical compute cost. And did it all work? Yes, it did indeed. The benchmark data bear this out. Nimitron 3Super currently powers the NVIDIA AIQ research agent to the number one position on both the deep research bench and deep research bench two leaderboards, which measure multi-step research capability across large document sets. The model has also claimed the top spot on artificial analysis for efficiency and openness in its size class, outputting tokens at around 450 to 480 tokens per second depending on the provider. Speaking of openness, as I briefly mentioned at the top of this episode, NVIDIA is releasing the model with open weights under a permissive commercial license, but

they went further than just releasing weights. They're also publishing over 10 trillion tokens of pre and post-training data sets, 15 reinforcement learning training environments and their full evaluation recipes.

06:22 For researchers and practitioners who want to reproduce the training, fine-tune for a specific domain, or build their own hybrid architecture models, these data and recipes are invaluable. The model was post-trained using reinforcement learning across diverse agentech environments via NVIDIA's open source NemoGym library, which evaluates the model on sequences of real actions, tool calls, functional code generation, verifiable multi-step plans, rather than just optimizing for single turn responses. And to coincide with the launch, there were companies evidently working in the background to make sure that NVIDIA could announce that adoption is already picking up. Perplexity, for example, is offering Nemotron3super for search. Software development agent companies like CodeRabbit and Greptile are integrating it into their coding assistance. And on the enterprise side, companies like Siemens, Palantir, and Cadence are deploying it for manufacturing, cybersecurity, and semiconductor design workflows. In terms of where you can access the model, weights are on Hugging Face for self-hosting.

07:21 For cloud deployment, it's available through Google Clouds, Vertex AI and Oracle cloud infrastructure with Amazon Bedrock and Azure reportedly coming soon. On the inference side, it's available through providers, including Base10, Deep Infra, Fireworks AI, and Lightning AI, where full disclosure, I hold a fellowship. So despite not being an objective information source, I can nevertheless provide objective third party data from artificial analysis showing that Lightning AI at the time of me recording delivers the fastest Nematron three super output speed of any inference provider coming in at 480 tokens per second. I provided a link to this in the show

notes plus anything else I cited in today's episode. So what this means is if you don't want to go through the hassle or expense of setting up Nemotron3super on your own infrastructure, working with an inference provider like Lightning will make your life super easy and you can just get going with this innovative mixture of experts model today.

08:16 If you're building multi-agent systems, whether autonomous coding assistants, research agents, or enterprise automation workflows, a model like this that combines open weights, frontier class reasoning, and blazing fast throughput at a fraction of typical compute costs is exactly the kind of tool that can take your project from prototype to production. All right, that's it for today's episode. If you enjoyed today's episode or know someone who might, consider sharing this with them, leave a review of the show on your favorite podcasting platform or on YouTube. If you tag me in a LinkedIn post with your thoughts, I will respond to those. And if you aren't already, of course, subscribe to the show. Most importantly, however, we hope you'll just keep on listening. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.