SDS PODCAST EPISODE 944: GEMINI 3 PRO: GOOGLE'S BACK ON TOP

Jon Krohn: 00:00 This is episode number 944 on Gemini 3 Pro. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. There's been a ton of excitement around Google's new Gemini 3 Pro model. It appears that Google, after trailing behind OpenAI since the release of ChatGBT 3 years ago and then later lagging behind Anthropic as well. It appears Google has regained the top spot at the frontier of AI capability. In today's episode, I'll fill you in on everything you need to know about Gemini 3 Pro's performance and why it matters first, Gemini 3 Pro is topping major evaluation leaderboards on the popular LM Arena leaderboard, for example, which we cover in detail back in episode number 707 with professor Joey Gonzalez whose lab devised the unique arena. But in a nutshell, what makes LM Arena unique and hard to game relative to most benchmarks is it uses human evaluators.

00:54 And Gemini 3 Pro debuted in first place across all of the key tracks, text reasoning, vision, coding, and web development. In fact, the Ella Marina team noted that Gemini 3 surpassed even brand new rivals like X's GR 4.1 and open AI's latest GBT five class models on a wide range of tasks from math and long form Q&A to creative writing. This is huge news because in recent years, open ais and anthropics best models like GPT 5.1 and Claude Sonnet 4.5 were seen as the ones to beat Gemini 3 Pro upended. That and is currently at the time that I'm recording this topping the LM Arena leaderboard overall beyond LM Arena. Gemini 3 pro's performance on a wide array of benchmark evaluations is also impressive. For example, on humanity's last exam, Gemini 3 Pro scores a 38% with no tools like web search or code execution while its next closest competitor, GBT 5.1 scores only 27%.

01:56 That's a big jump on what is still a tough benchmark for LMS in 2025. On another also tough benchmark, a math contest benchmark called Math Arena Apex Gemini 3 Pro solved about 23% of the problems, which might sound

low until I tell you that the next best performing models GPT 5.1 and Claude 4.5 managed only about 1% on that test. So this is a jump from 1% to 23% with his Gemini 3 Pro release. That's pretty crazy. Gemini 3 Pro also set new high watermarks, including the AIM Math benchmark, the GPQA Diamond Benchmark of Scientific Knowledge and the Triple Pro multimodal Understanding and reasoning benchmark. It also tackled tricky visual reasoning puzzles in the Arc agi I two challenge quite a lot better than Rivals scoring 31% versus just 18% and 14% for GPT 5.1 and clawed 4.5 respectively. That is a remarkable jump in a notoriously hard test of abstract problem solving.

03:02 For those of us using LLMs to help us with coding or for agentic AI tasks, Gemini 3 Pro made some cool leaps there too. While it performed comparably to Claude and GBT models on the popular suite be verified Benchmark, it crushed all other models on Live Code Bench Pro. It also shines in tool use and long horizon planning. For instance, on Vending Bench 2.0 where a commercial vending machine simulation is carried out, Gemini 3 Pro made a simulated profit of over $5,500. While its closest competitors clots on at 4.5 and GPT 5.1 net profits of only 4,000 and $1,500 respectively. I guess you'd actually, if you added together the performance of both clots on at 4.5 and GBT 5.1, that would still be a little bit less than Gemini 3 Pro made on its shot at Vending bench. And finally, for those of us interested in developing agents that can understand what's happening on a computer screen to be able to act autonomously using that information, Gemini 3 Pro might be the first LLM.

04:06 We'd actually trust with that because it scored 73% on the screen Spot Pro Benchmark, while its closest rival sonnet 4.5 scored less than half that only 36%. And what about the model itself? Well, Google hasn't revealed much about Gemini 3's architecture. No surprise. The details

are proprietary, but one technical spec worth sharing or worth mentioning is that it boasts a big million token context window that corresponds to about eight novels worth of natural language. Not such a big deal because million token context windows are becoming more and more common with frontier models, but still worth mentioning. And so how is it being used? Well, Google is wasting no time deploying Gemini 3. The launch is one of Google's most extensive ever with Gemini 3 Pro being rolled out across many of their products simultaneously. It's powering features in Google search, the new Gemini chatbot app, vertex AI cloud services and developer tools beyond Google's own ecosystem.

05:01 Some early partners are already test driving it in real workflows. For example, Wayfair's CTO shared that they've been piloting Gemini 3 Pro to convert complex support documents into clear data, accurate infographics for their fuel teams. That's a great illustration of Gemini's multimodal power in an industry setting. It's taking long text heavy manuals and automatically producing visual digestible summaries. We're also hearing about Gemini 3 being evaluated for coding copilots and enterprise knowledge assistance. GitHub, for example, noted a 35% boost in code accuracy when they switched to Gemini 3 Pro in an early test compared to its predecessor, Gemini 2.5. The broader significance is that this release signals that Google is back on top in the AI race, at least for now. Over the past year, open AI's GBT models and anthropics Claude models often grabbed the most advanced LLM headlines. But Gemini 3 Pro has emphatically Google's flag in the ground as the lab to now beat it, leapfrog not only OpenAI and Anthropics latest, but also up and comers like Xai.

06:06 In fact, the team behind Ella Marina noted Gemini 3 Pro was the first model to ever break an ELO score of 1500 on their platform. A testament to just how far Google pushed

the envelope here for Google, this is a strategic win. It validates their investment in combining deep minds research with other Google resources for us, leveraging frontier models and bringing 'em to the real world within applications. As I suspect many listeners are, this shows us that the competition at the frontier of AI is alive and well. As each new model raises the stakes, we all benefit from the rapid improvements in what these systems can do. On that note, I'm not going to cover it in any detail in today's episode, but you might also want to check out Google's brand new Nano Banana Pro. I've got a link to that for you in the show notes to see what's possible for the state-of-the-art in text to image generation, as well as image editing, perhaps unsurprisingly, Gemini 3 Pro is the LLM running in behind on Nano Banana Pros backend well with new capabilities at your fingertips.

07:12 Now for image generation, image editing, text generation coding, and especially Ag agentic AI relevant capabilities like screen understanding and long-term reasoning, I hope your human brain is buzzing with ideas about what you could do with Gemini 3 pro. If not, chat with whatever your favorite LLM is to get ideas on what you could do given your particular industry network background, et cetera, while being able to predict what company will be leading the AI race in any given month may be tough. One thing's for sure, folks like us leveraging and deploying these rapidly improving LMS are all benefiting. Alright, that's it for today's episode. I'm Jon Krohn and you've been listening to the SuperDataScience Podcast. 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, please, and if you aren't already, be sure to subscribe to the show. Most importantly, we hope you'll just keep on listening. Until next time, keep on. We're rocking it out there, and I'm looking forward to enjoying another round of the SuperDataScience podcast with you very soon.