# SDS PODCAST

# EPISODE 1026: OPENAI'S GPT-6 ASTRA

Jon Krohn: 00:00 This is episode number 1026 on OpenAI's GPT-6 Astra. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. Today's episode is all about GPT-6 Astra, the new flagship model that OpenAI began rolling out last week, and that OpenAI's president Greg Barakman has gone so far as to suggest might one day be looked back upon as the arrival of AGI artificial general intelligence. That's a big claim and we'll get to it, but first let me walk you through what the model is, what it can do, what it costs, and then I'll wrap it up with the safety story, which in this particular release is unusually intertwined with the capability story. Let's start with the basics. GPT-6 Astra is the first six series model from OpenAI. It's OpenAI's most capable model then, of course, and it's positioned as the successor to GPT-5. 6 Sol, which had been the company's flagship since July.

01:06 And also the name seems to imply to me that Astra is even bigger in terms of parameter count than Soul, which is bigger than Terra, which is bigger than Luna. So moon, earth, sun, stars. I don't know, they don't release parameter counts anymore. But they did say that Astra came out of the largest training run the company has ever carried out, which used more than apparently a hundred thousand GPUs at a renowned compute facility in Texas called Stargate. And in a detail I found interesting, Astra is the first OpenAI model where other AI models played a significant role in supervising its training. So the flywheel of models training models is now explicitly part of how the frontier gets pushed forward. Another step toward RSI, recursive self-improvement, which you can hear all about in episode number 1004. Anyway, the model GPT-6 Astra is available in the API, the OpenAI API.

02:06 Whoa, that's fun to say, the OpenAI API. It's under the model string GPT-6-Astra. And as with OpenAI's other recent models, you can dial in how much inference time compute it spends on a given task via a reasoning effort parameter. You can hear all about those in episode

number 1020 from a couple weeks ago. In Astra's case, there are five settings of this reasoning effort parameter, low, medium, high, X high, and max. Standard API pricing is $10 per million input tokens and 50 per million output tokens, which is about two and a half times what Sol has been going for on promotion, but is in line with what Anthropic charges for its top tier Claude Fable 5.1 model. There's also a fast mode, which I hadn't seen before in the API that delivers up to two and a half times the speed for double the price.

02:59 So if money is no object to you, you can get these frontier capabilities for faster. Astra is also available through Amazon Bedrock. And if you're a ChatGPT user, it's rolling out over the coming days two plus pro business and enterprise subscribers with the three higher tiers of those, pro, business and enterprise. Additionally, getting a beefier variant called GPT-6 Astra Pro. But exactly what makes it beefier or different at all from the regular non-pro variant has not been yet publicly disclosed, at least at the time of me recording this. One note for people in larger organizations, enterprise administrators have to switch Astra on for their workplace because at launch it's off by default. Now onto what the model can do. Across the board, OpenAI is claiming state-of-the-art results on computer use, browsing, software engineering, cybersecurity science and professional knowledge work, and the benchmark table they published is long.

03:58 But anytime somebody's releasing a near frontier model, we get the same kind of, "Wow, look, it's absolutely the best on every benchmark that we tested wow." So I don't know, grain of salt, but this does seem to be a pretty serious model release and I'll pull out a few of the highlights that I think matter most to this audience. The first is computer use, which is the area OpenAI is leaning on hardest in its messaging. The pitch is that Astra can take over the tedious click-heavy parts of your working life, filling out online forms, updating records in a CRM,

organizing your calendar, running front end QA checks on a website it just built, installing and troubleshooting software and so on. On OS World 2.0, a benchmark of realistic desktop tasks, Astra scores 73% versus SOL, which had just 66%. That's actually not that much of a difference.

04:52 And according to OpenAI's latency simulations, it gets there in 40 minutes per task versus roughly 75 minutes per sol. So it's a little bit more accurate, but it's a lot faster, 47% on that particular benchmark, OS World 2.0. On ScreenSpot Pro, yes, ScreenSpot Pro, which tests whether a model can locate the right element on a professional application screen, Astra scores about 93%, which is a big jump relative to Sol's 77%. That is significant sounding. After computer use, the second highlight is coding. On terminal bench 4.0, which covers agentic software engineering system configuration and data analysis from the command line, Astra scores about 58% versus just 37% for Sol. And so that 58% from Astra is about the same as Claude Fable 5.1, which came in at 56%. So I'm sure you can notice that 56 to 58% difference. On another benchmark called Frontier Code, the models are bunched much more tightly together with Astra head of Sol, and again, essentially level with the top clawed model.

06:04 So the coding lead isn't uniform across every evaluation. One coding adjacent feature that I think deserves a mention is a new approach to long agentic sessions in Codex, OpenAI's coding tool. Rather than repeatedly compacting a session's history into a single summary as the context window fills up, Astra can keep running notes across context windows and earlier windows remain searchable. So details like why a given fix failed don't get compressed away. And that's experimental for now, but will apparently become the default for Astra in the coming weeks. The third highlight and probably the one that will get the most airtime in the mainstream press is abstract

reasoning and math. Astra scores 99.9% on Arc AGI three, a benchmark that SOL scored under 8% on. So OpenAI jumps from around 8% to effectively 100% on that abstract reasoning and math benchmark. That is pretty insane. I wish I had the data on how FABLE 5.1 from Anthropic performs on that, but just don't have it online for some reason, at least at the time of me recording.

07:16 The best I can see is that Opus five scores 30%. So yeah, TBD on whether this is actually an enormous jump on frontier abstract reasoning and math overall or just for OpenAI. Other than Arc AGI three, GPT-6 Astra scores about 98% on Frontier Math tier four, which is the hardest tier of research level mathematics problems in that suite, which is up a non-trivial amount from Soul, which had 83%. So 83% to 98% jump. OpenAI has also published two new mathematical results that Astra helped produce. So new math being discovered here and both of those results concern the gaps between prime numbers. So one of those tightens a longstanding bound on how closely together infinitely many pairs of primes can occur. So it tightens that from 240 down to 186. And the other new mathematical results improves a term in a bound on unusually large prime gaps that had stood unchanged for more than 80 years.

08:25 I candidly don't really understand what that means or what the implications are, but I guess it sounds like it's a big deal. Whatever you make of the AGI question, a model that contributes to open problems and number theory is a meaningful milestone, it sounds like to me. All right. The fourth category that I want to highlight here is benchmarking on science and professional work more broadly. So on terminal bench science benchmark, which tests whether an agent can carry out research workflows like analyzing data, running simulations and fitting models. Astra scores about 65% versus about just 22% for soul and about 53% for Fable 5.1. So that seems like

maybe a place where GPT-6 Astra really has an edge on the frontier. And on Agent's Last Exam, a fun benchmark that's playing on humanity's last exam benchmark. Agent's last exam tests agents on complex professional tasks in a real software across domains like financial modeling, engineering and media production.

09:27 And there Astra edges out Claude Opus five, Claude Opus five, while OpenAI says using roughly 65% fewer tokens. Again, I don't have the fable figures for you here, but at least compared to Opus five, that token efficiency point comes up repeatedly in the ChatGPT announcement and it matters for cost because a model with a higher per token price can still actually produce a lower overall bill for you if it finishes a job in fewer steps with fewer retries. All right, and the fifth and final capability that I'm going to highlight here is something that OpenAI is making a point of Astra's judgment when instructions are ambiguous. The claim is that Astra fills in routine gaps with sensible assumptions, but pauses to ask a focused question when the answer could change the outcome. And in OpenAI's coding tool Codex, here's something that was actually really cool.

10:23 I would love to experience this. A GPT-6 Astra can ask a question asynchronously of you while continuing with the parts of the task that don't depend on your reply. That is damn cool. And hopefully they're working on that for ClaudeCode as well, my typical environment. It's also reported to be better GPT-6 Astra at incorporating a mid-task course correction without losing track of the original goal, which anyone who has steered an agent through a long task will appreciate. Okay, that's it for the capability story. Now for the safety story, and in this respect, Astra differs from a typical model arch. So Astra is the first model OpenAI is designated as reaching the critical threshold for cybersecurity under its preparedness framework. What that means concretely is that tested without production safeguards, the model can find

previously unknown vulnerabilities and turn them into working exploits across well-protected systems without a human guiding each step.

11:22 This sounds quite similar to what was happening with that mythos and later fable drama from Claude on Exploit Bench, for example, which tests whether a model can turn known vulnerabilities into working exploits. Astra scored a perfect 100% versus 79% for soul. To address concerns that the model might have memorized historical vulnerabilities, OpenAI also built a fresh evaluation using 20 high severity Chrome vulnerabilities from June through August of this year. And during that evaluation, Astra discovered and used two previously unknown zero-day vulnerabilities, which OpenAI says it is disclosing to the Chrome maintainers. Expert assessments found the unrestricted GBT-6 Astra could achieve arbitrary code execution in hardened browsers and build privilege escalation exploits for hardened operating systems. Because of this, the launch of GBT-6 Astra has been phased and gated. OpenAI slowed Astra's release to add safety testing. The version everyone is getting will help with defensive work like secure code review and patching, but it will refuse more advanced tasks such as writing proof of concept exploits.

12:32 Organizations in OpenAI's application-based security program called Daybreak got access first and OpenAI says it plans to roll out less restrictive safeguards to that group in the coming weeks for workflows like malware analysis and detection engineering. There's also a new layer of misalignment monitoring in production for AstroClass models, classifiers that check the model's reasoning and actions and automatically stop activity that looks unauthorized. OpenAI is candid that this can sometimes pause or stop legitimate work in ChatGPT and Codex you may be asked to review an action before continuing while in the API the task will simply halt. This all happens, of course, against a specific cybersecurity

backdrop. As I detailed in episode 1014 about a month ago, and I'm sure everyone has heard about by now, two OpenAI models escaped their containment during research, accessed the open web and breached Hugging Face's systems. OpenAI paused parts of its research and training and response, including work on Astra, even though Astra wasn't actually one of the models involved apparently, and OpenAI added safeguards before release.

13:34 Reporting from CNBC indicates the model also went through a formal review process with the US administration before launch and OpenAI built a new evaluation informed by the Hugging Face incident that measures whether a model given a difficult or impossible task will exceed its authorized scope. Without production safeguards, Sol went beyond the authorized target 48% of the time, while Astra did so in 0% of cases. So it does seem like they're figuring out how to reign their models in. There is one caveat in OpenAI's own materials that I think is worth flagging. Their evaluations found that Astra's written reasoning is harder to monitor than Soul's in tests that explicitly asked the model to evade monitoring. So OpenAI attributes this to Astra having greater control over its written reasoning and being able to solve problems in fewer written steps. And it notes that Astra still appears to struggle to conceal the reasoning needed for complex tasks, but a step backward on monitorability like this at the same time as we take a step forward on offensive cyber capability is exactly the combination the field should be watching closely.

14:44 And OpenAI says it's a research priority. So finally, is this AGI? Brockman's own framing was that AGI remains a gray fuzzy concept, that the term is no longer tied to any contractual trigger with Microsoft and that he personally thinks future observers might point to this model as the moment, to this model as the moment. He closed the press briefing by welcoming everyone to the AGI era. I'll leave you to form your own view, but here's mine. As I

detailed years ago in episode number 748, I don't think it makes sense to think of AGI as a binary event. There are varying degrees of AGI breadth and depth. So instead of thinking of this GPT-6 Astra release in binary AGI terms, we can acknowledge this is at most a small step forward in general intelligence capabilities relative to Anthropic's Fable 5.1 model. But having not used GPT-6 Astra myself, nor having much third party benchmarking to work with yet, this might even just be bringing OpenAI in the vicinity of the Fable 5.1 frontier.

15:51 So yeah, not necessarily the big deal that Brockman's making it out to be. But regardless, we now have at least two models near the frontier of general intelligence capabilities, including with GPT-6 Astra, that means a generally available model that contributes to open mathematics research, works through desktop applications faster than most people can and find zero day vulnerabilities in hardened software. That's powerful for sure. And more along this staggering trajectory is coming in the months ahead, no doubt, including, no doubt, open source alternatives. If you'd like to dig further into any of this, I've of course put links in the show notes to OpenAI's full announcement, which includes the complete benchmark tables and the two prime gap proofs. Maybe one of you will actually understand them, as well as the Astra system card where the cybersecurity and alignment findings are documented in detail. Wow. Now go get your hands dirty, think carefully about what you can now delegate to agents, think even more carefully about what you should, and then go build something worthwhile.

17:03 Cool. And we do have a new Apple podcast review for me to read for you on air. This one is short. It comes from Ronan 186. The title is simply I'm a huge fan of the Super Data Science Podcast, exclamation mark. It's a five star review and wow, I'm particularly fond of the body of this review. It just says Jon Krone is great with three

exclamation marks. All right, thanks Ronan. 186. I think you're great too. Thanks for all the recent ratings and feedback on Apple Podcasts, Spotify, and all the other podcasting platforms out there, as well as for likes and comments on our YouTube videos, bonus points. If you leave written feedback on Apple Podcasts, if you do, I'll be sure to read your feedback on air like I did today. At the time, I'm only seeing feedback in the US app, but someday, especially when I don't have US reviews to read, I will go and look at other countries and read the reviews you're putting in there too.

18:00 All right, that's the end of today's episode. If you enjoyed it or know someone who might consider sharing this episode with them, tag me in a LinkedIn post with your thoughts and if you aren't already, be sure to subscribe to the show. The most important thing though is that 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 Super Data Science podcast with you very soon.