Sora the App, Sonnet 4.5 and the Question of Models as Processors

Wednesday, October 1, 2025 Listen to Podcast

Listen to this post:

0:00 / 13:30

Good morning,

On **this morning's Sharp China**, Bill and Andrew break down the final formulation of the TikTok deal, and China's focus on the status of Taiwan in ongoing trade talks.

On to the Update:

Sora the App

From **The Verge**:

> OpenAI has a new version of the Sora AI video generator that it launched at the end of last year, and it's arriving today alongside a new social video app, also called Sora, for iPhones. The currently invite-only app resembles TikTok with a feed of videos you can shuffle through. But instead of encouraging people to stitch together duets, it asks you to record short videos that anyone can spin into new AI-generated deepfakes — with your consent.
>
> In a briefing with reporters on Monday, employees called it the potential "ChatGPT moment for video generation." The Sora app is currently only available to US and Canada users, with other countries set to follow, and when someone receives access, they also get four additional invites to share with friends. There's no word on when an Android version might be released.

Some things are old, and some things are new. File iPhone dominance of anything new and interesting into the former; the coolest new apps came first to iPhone in the smartphone era, and if that is still the case in the AI era, can you really say that Apple is behind?

What is new is the reality that compelling AI video generation is very much here, and it's widespread: over the last three weeks we have actual AI video products from Google, Meta, and now OpenAI. What is so fascinating, however, is just how different these products are.

Google is building features for YouTube; Meta is building the aptly named Vibes to take you away into fantastical worlds; OpenAI is letting you make as many variation of Sam Altman as you can handle, if my day one Sora feed is any indication.

Indeed, it feels like each company has an entirely different target audience: YouTube is making tools for creators, Meta is building the ultimate lean back dream-like experience, and OpenAI is making an app that is, in my estimation, the easiest for normal people to use.

In this new competition, I prefer the Meta experience, by a significant margin, and the reason why goes back to one of the oldest axioms in technology: the 90/9/1 rule.

- 90% of users consume
- 9% of users edit/distribute
- 1% of users create

If you were to categorize the target market of these three AI video entrants, you might say that YouTube is focused on the 1% of creators; OpenAI is focused on the 9% of editors/distributors; Meta is focused on the 90% of users who consume. Speaking as someone who is, at least for now, more interested in consuming AI content than in distributing or creating it, I find Meta's Vibes app genuinely compelling; the Sora app feels like a parlor trick, if I'm being honest, and I tired of my feed pretty quickly. I'm going to refrain on passing judgment on YouTube, given that my current primary YouTube use case is watching vocal coaches breakdown songs from KPop Demon Hunters.

I honestly have no idea if my evaluation of these apps is broadly applicable; as I've noted repeatedly, I'm hesitant to make any pronouncements about what resonates with society broadly given that I am the weirdo in the room. Still, I do think it's striking how this target market evaluation tracks with the companies themselves: YouTube has always prioritized creators, while OpenAI's business model is predicated on people actively using AI; it's Meta that has stayed focused on the silent majority that simply consumes, and as a silent consumer, I still like Vibes!

To that end, I thought that bit in the excerpt above about OpenAI employees calling Sora 2 the "ChatGPT moment for video generation" was interesting; OpenAI itself said as much in **its post about the launch**:

> The original Sora model from February 2024 was in many ways the GPT-1 moment for video—the first time video generation started to seem like it was working, and simple behaviors like object permanence emerged from scaling up pre-training compute. Since then, the Sora team has been focused on training models with more advanced world simulation capabilities. We believe such systems will be critical for training AI models that

> deeply understand the physical world. A major milestone for this is mastering pre-training and post-training on large-scale video data, which are in their infancy compared to language.
>
> With Sora 2, we are jumping straight to what we think may be the GPT-3.5 moment for video. Sora 2 can do things that are exceptionally difficult—and in some instances outright impossible—for prior video generation models: Olympic gymnastics routines, backflips on a paddleboard that accurately model the dynamics of buoyancy and rigidity, and triple axels while a cat holds on for dear life.

So the company who resolutely claims that they had no idea ChatGPT would be popular now feels confident declaring that they're about to repeat one of the most seminal moments in the entire history of technology; I'm not so sure.

Sora 2 is, to be clear, amazing, and the app is very easy to use. What matters in terms of creating moments that matter, however, is consumption, and while creation is obviously a prerequisite to consumption, it's not the main thing. What made ChatGPT unique was that LLM's immediately delivered perfectly customized content for each individual user to consume based on the most basic of prompts; in this the fact that text is cheap and easy was critical. People create video, however, for others, and I'm just not sold that the AI video the Sora app enables, easy though it may be to make, is that interesting to anyone other than the creator.

OpenAI does seem aware of this limitation; again from the company's post:

> This app is **made to be used with your friends**. Overwhelming feedback from testers is that cameos [the ability to use other people's likenesses to make AI videos] are what make this feel different and fun to use — you have to try it to really get it, but it is a new and unique way to communicate with people. We're rolling this out as an invite-based app to make sure you come in with your friends. At a time when all major platforms are moving away from the social graph, we think cameos will reinforce community.

This is an interesting inversion of my observation about how **Facebook needed to move beyond social networking**; focusing on something smaller than being an entertainment app is a luxury afforded to a new entrant in the social media space. AI generated videos of my friends, however, still depends on my friends being interesting enough to make me want to consume more, and I'm not sure that's going to be enough to make Sora the app more than a passing fancy.

Sonnet 4.5 and the Question of Models as Processors

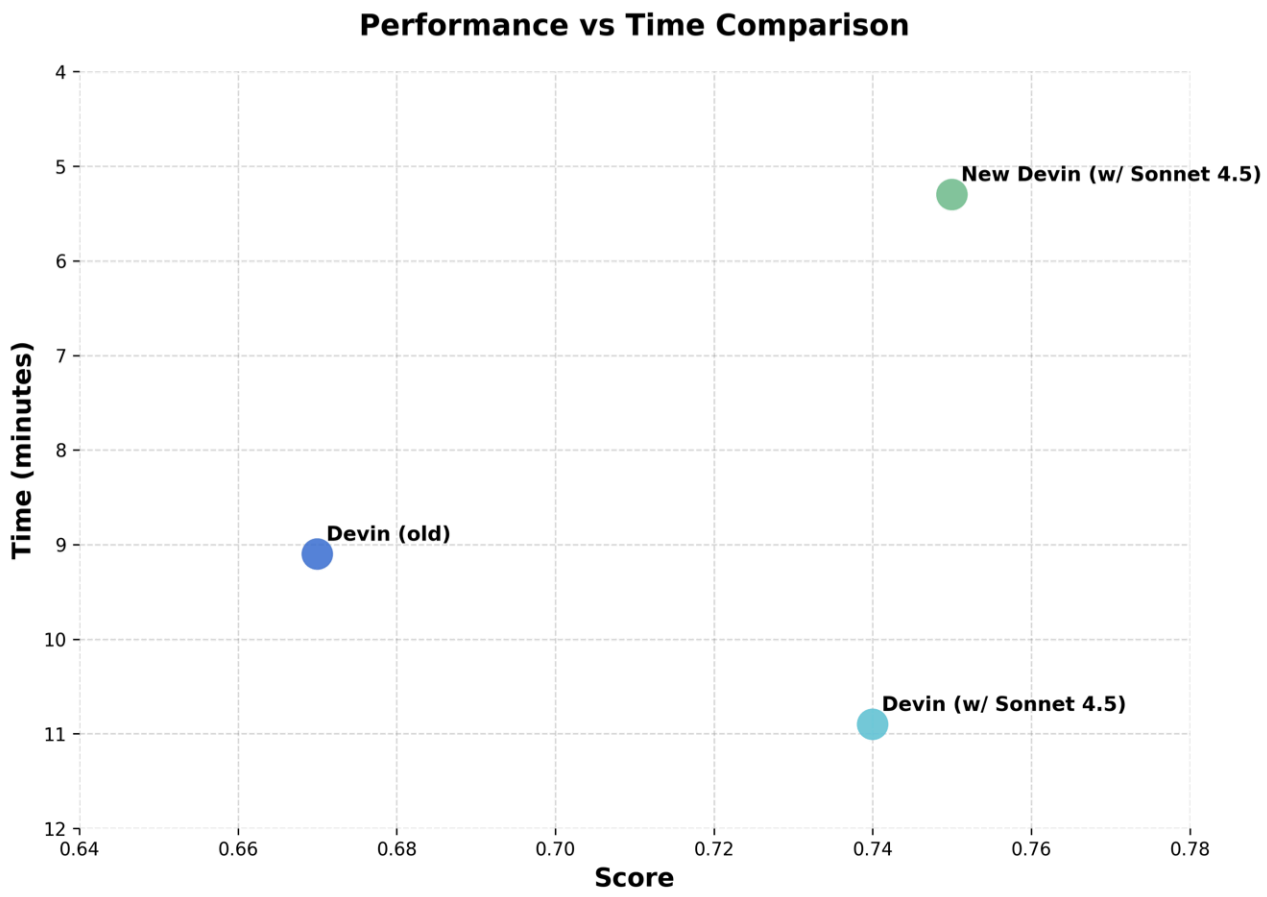
From **Bloomberg**:

> Anthropic is releasing a new artificial intelligence model that is designed to code longer and more effectively than prior versions, its latest attempt to stay ahead of rivals like OpenAI in offering tools for software developers.
>
> The new model, called Claude Sonnet 4.5, is better at following instructions and can code on its own for up to 30 hours straight, the company said on Monday. By comparison, a previous model called Claude Opus 4 is said to be able to field coding tasks for up to seven hours by itself. The updated version of Sonnet is also intended to excel at using a person's computer to take actions for them, improving on a feature Anthropic introduced a year ago.
>
> Anthropic has been an early leader in building so-called AI agents that field complex tasks on a user's behalf, particularly for streamlining the process of writing and debugging code. The company, now valued at $183 billion, reached $5 billion in run-rate revenue in August, fueled in part by traction for its coding software. But other companies, including OpenAI and Alphabet Inc.'s Google, are also vying to win over programmers with similar capabilities. Anthropic's latest release comes a week before OpenAI is set to hold its annual developer event.

It's challenging to cover new model releases; I'm not a model evaluator, and at this point the various evaluations seem both too targeted by the model makers and too narrow to give a proper estimate of how good a model is. What generally seems to happen is that a consensus about model quality forms over the following weeks, and given both Anthropic's history and success with coding in particular, my assumption is that Bloomberg was right to focus on the coding angles in their write-up.

What I did find very interesting about this release was **this writeup from Cognition about rebuilding Devin for Claude Sonnet 4.5**:

> We rebuilt Devin for Claude Sonnet 4.5. The new version is 2x faster, 12% better on our Junior Developer Evals, and it's available now in Agent Preview. For users who prefer the old Devin, that remains available.
>
> Why rebuild instead of just dropping the new Sonnet in place and calling it a day? Because this model works differently — in ways that broke our assumptions about how agents should be architected. Here's what we learned:



Because Devin is an agent that plans, executes, and iterates rather than just autocompleting code (or acting as a copilot), we get an unusual window into model capabilities. Each improvement compounds across our feedback loops, giving us a perspective on what's genuinely changed. With Sonnet 4.5, we're seeing the biggest leap since Sonnet 3.6 (the model that was used with Devin's GA): planning performance is up 18%, end-to-end eval scores up 12%, and multi-hour sessions are dramatically faster and more reliable.

In order to get these improvements, we had to rework Devin not just around some of the model's new capabilities, but also a few new behaviors we never noticed in previous generations of models.

One of the big questions I have been pondering for the last several years is the extent to which model evolution would be similar to processor evolution in the 1980s and 1990s: back then the best way to approach software development was to write to the edge of what was possible on the processors you had; by the time you were finished, or shortly thereafter, your customers would have new processors that made all of your slow code fast, while otherwise working pretty much the same. This is what Microsoft CTO Kevin Scott predicted in **a 2024 Stratechery Interview**:

> I think x86 is probably a pretty apt comparison here, because the thing that made x86 interesting is it was a general purpose piece of infrastructure that allowed lots and lots

> and lots of software to be written, and the power of the system, the platform, just increase over time because it was getting cheaper every 18 months or so and more powerful simultaneously. And so you just had this rapid progression of capability flowing into the hands of lots of people that were building things on top of it.
>
> There was a clear separation between the x86 and the operating system and the PC manufacturers and the people who are building applications on top of it. And sometimes, Microsoft built both applications and operating systems so there's a little bit of the both, but there was a whole universe of possibility there for people to do things on top of the Wintel platform that had nothing to do with Microsoft predicting what all of the useful things were and people could trust that it was an interesting platform because you had this exponential called Moore's Law that was just going to ultimately result in the thing being completely ubiquitous.

If AI models — the new processors, in many respects — follow a similar path, then the best way to develop AI applications would be to write to the edge of the capabilities of the current models, confident that new models would come along who were just like the old models but better in basically every respect, allowing your application to harvest the gains. This is a rich world for app developers, who can focus on building for specific niches and use cases and get the AI model improvements with little more than a change in their API targeting.

Cognition, however, is giving one data point that this might not be the case: notice how in their graph the new model actually decreased performance in some vectors; to truly take advantage of the new model Cognition had to completely rework their application.

Now an app like Devin is, to be fair, very close to the metal in terms of model behavior; it's possible that more bog-standard apps won't notice the difference. At the same time, this is a point of evidence that being an AI app developer might be more difficult than expected: prior investments might not be worth as much, which translates to much higher ongoing R&D costs, and meanwhile the model itself will have more runway to simply eat your business directly.

This Update will be available as a podcast later today. To receive it in your podcast player, **visit Stratechery**.

The Stratechery Update is intended for a single recipient, but occasional forwarding is totally fine! If you would like to order multiple subscriptions for your team with a group discount (minimum 5), please contact me directly.

Thanks for being a subscriber, and have a great day!

Related

Sora, Groq, and Virtual Reality

Tuesday, February 20, 2024

Checking In on AI and the Big Five

Monday, June 23, 2025

Tech Philosophy and AI Opportunity

Tuesday, July 8, 2025

← **OpenAI Instant Checkout, AI and Long Tail E-Commerce, Is AI Different?**

An Interview with Ben Bajarin About AI Infrastructure →

Stratechery *Plus* UPDATES

An Interview with Ben Bajarin About AI Infrastructure
Thursday, October 2, 2025

Sora the App, Sonnet 4.5 and the Question of Models as Processors
Wednesday, October 1, 2025

OpenAI Instant Checkout, AI and Long Tail E-Commerce, Is AI Different?
Tuesday, September 30, 2025

View All

Stratechery *Plus* PODCASTS

The Magic and the Top 100, Surveying the Southeast Division, Kuminga's Deal and a LeBron Injury
Greatest Of All Talk | Oct 2

TikTok in the Clear; Taiwan Questions and Soybean Angst; The K Visa Goes Live; Jensen Huang on China Hawks
Sharpchina | Oct 1

An Afternoon with the Clippers, 600 Episodes of the GOAT, Notes on the Lakers, Nuggets, Warriors, and More
Greatest Of All Talk | Sep 30

Stratechery *Plus* INTERVIEWS

An Interview with Ben Bajarin About AI Infrastructure
Thursday, October 2, 2025

An Interview with Booking CEO Glenn Fogel About Travel and Aggregation
Thursday, September 25, 2025

An Interview with YouTube CEO Neal Mohan About Building a Stage for Creators
Wednesday, September 17, 2025

View All