How AI Changes if Open Source Gets Banned

23 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=KPGZ83XQDQA
https://www.youtube.com/watch?v=kpGZ83XqDqA

SUMMARY

The AI Daily Brief discusses the rapid advancements in AI models and the implications of potential restrictions on open-source models from China. As new models like OpenAI's GPT 5.6 and SpaceX's Grok 4.5 are set to launch, the conversation shifts to how the AI landscape might change if China limits the distribution of its leading models, potentially reshaping the competitive dynamics in the global AI race.

- *OpenAI's GPT 5.6 family of models is receiving positive early feedback, with significant improvements in execution and usability.*
- *SpaceX's Grok 4.5 is expected to be released soon, emphasizing efficiency and lower costs.*
- *China may consider restricting the overseas distribution of its advanced AI models, viewing them as national security assets.*
- *The potential restriction could lead to increased focus on open weights and alternative models from Western labs.*
- *Companies like Nvidia and Google are enhancing their model offerings to fill any gaps left by restricted Chinese models.*
- *Microsoft is exploring a new approach called Frontier Tuning, allowing businesses to customize AI models for specific tasks, potentially reducing costs.*
- *The rise of model routers could facilitate better model selection based on task requirements and risk management.*
- *Overall, the AI market is evolving rapidly, with new opportunities emerging regardless of geopolitical changes.*

Today on the AI Daily Brief, how does AI change if access to open weight models starts to get cut off? Before that in the headlines, all the new models you have access to right now and all the ones that are coming. The AI Daily Brief is a daily podcast and video about the most important news and discussions in AI. Welcome back to the AI Daily Brief headlines edition, all the daily AI news you need in around 5 minutes. And my friends, if you thought that this was going to be a slow summer, think again.

We are just absolutely drowning in model announcements or announcements of announcements in some cases today. So, let's get into everything that is here and everything that is coming. Now, the first one is not a surprise as this was announced during the period that Fable was offline, but the GPT 5.6 family of models including Soul, Terra, and Luna, OpenAI announced in the middle of the night for some reason that they would be officially coming on Thursday. And in addition to that announcement of the announcement, they also unlocked early testers to begin sharing their impressions.

We're going to go much deeper into this when the model actually comes out, but a lot of those first impressions

are pretty positive. Ali K. Miller calls the model an execution beast. So much she says that I think 5.6 is the absolute wrong name considering how big of a leap this felt to me. Her conclusion, when Sonnet 3.7 came out, I think we no longer tolerated bad writing. Having GPT 5.6 and Fable 5 out in the world, I think we will no longer tolerate bad execution or slow bug fixes or unhelpful customer support or at least tolerate a whole lot less.

Magic Path CEO Pietro Schirano wrote, "I can finally talk about 5.6. I've been testing it for months and without exaggeration, it's the best model I've ever used. Fast, smart, genuinely creative, and you guessed it, they finally fixed front-end design. I haven't needed to check the code I've written in 2 months." YouTuber and AI entrepreneur Theo wrote, "It's a damn good model. Not quite as {quote} smart as Fable, but it's incredibly capable. Fixed all the problems I had with GPT 5.5. It's incredibly determined, will run for a day without even using a slash goal. It understands sub agents incredibly well and is great at orchestrating. It's super pleasant in use cases like open claw and Hermes agent. It knows iOS dev incredibly well.

It has rough edges too, but far fewer than 5.5 did. For many things, GPT 5.6 Soul will become my obvious default. Now, of course, the question that many will have is how does it compare to Fable? And not everyone was convinced that 5.6 beats it. Nat Schumer writes, "5.6 Soul is an amazing model, but for almost every task I tested Fable was quite a bit better and more agentic to boot. I.e., one Fable turn does the same things many 5.6 turns do."

Interestingly, however, according to Ethan Mollick, that mode of interaction, where Fable goes off and does more things on its own, and 5.6 Soul sticks closer to the user, might be more intentional than it at first seems. Ethan wrote, "5.6 Soul is of similar ability, but quite different feel than Fable. Fable wants to go off and do work on its own pace. Soul is faster, but works with you in steps more." Now, for Ethan, this wasn't an either or. He continued, "I found myself switching between Fable and Soul depending on task. Soul for back-and-forth tasks, especially when I had not yet figured out what I needed exactly. Fable for very long tasks, where I could define what I wanted. And Soul Pro for really hard problems." Still, for some, the biggest and most interesting hint from this commentary was around just how long some folks said that they had been testing this. Remember, Pietro Schirano wrote, "I've been testing it for months." Chubby Communist Miss writes, "Wait, he'd already been testing 5.6 for months? That means 5.6 had already finished training when Mythos and Fable 5 had their reveal. And of course, the implication is that these are not the most state-of-the-art models that these labs have access to."

Now, while any new state-of-the-art model captures more attention than anything else, it is increasingly the case that people are thinking not just about raw model performance, but also model efficiency. And you can feel increasingly people getting excited not just about the frontier model releases, but models which offer something discrete and specific as part of an overall robust and complex model architecture. And that is potentially where SpaceX and Cursors new model comes in. Yesterday afternoon, The Information reported that the model release was imminent and could be coming as soon as Wednesday. The memo stated the release was pushed back from earlier this week to allow for efficiency tweaks.

Now, when it comes to this particular model, we've had a few breadcrumbs over recent months. Last month, for

example, Cursor CEO Michael Truel announced that they had finished pre-training their first model from scratch using SpaceX AI infrastructure. He said the model had 1.5 trillion parameters and also hinted that the model would be intelligent beyond coding, suggesting that this could end up a more general purpose model as opposed to the composer series, which has been very specifically designed for coding tasks. Elon Musk has also hinted at multiple large training runs taking place at Colossus 2 and a little over a week ago said that Grok 4.5 had entered private beta at SpaceX and Tesla late last month. Grok 4.5, he said, is based on what he called their 1.5 trillion parameter V9 foundation models with Cursor data added in post-training. At the end of June, Musk wrote, "Early eval show performance close to, perhaps exceeding, Opus." And that was reinforced when late last night Elon Musk confirmed the rumors and said that, yes, indeed, Grok 4.5 would be coming today. In fact, by the time that you are listening to this, it is highly likely that Grok 4.5 is out. Elon tweeted, "Based on strong positive feedback from customers in our beta test program, SpaceX AI will make Grok 4.5 available to the public tomorrow. It is an open class model," he wrote, "but faster, more token efficient, and lower cost."

Now, I want you to hold in mind that token efficiency and cost positioning, especially as we get to the main part of our episode in a few minutes. And by the way, you might have noticed that I keep referring to SpaceX as SpaceX AI. That's because that is the new official name of the company. The full integration of Elon's empire continues unabated. SpaceX AI lives. Now, one more small note on SpaceX/SpaceX AI. The post-IPO quiet period ended on Tuesday, meaning we have the first bank analyst ratings for the stock. Everything that I've seen so far is pretty wildly bullish. Morgan Stanley gave a $300 target, Bernstein at 239, and JP Morgan was also wildly bullish, expecting 5,000 Starship launches, i.e.

14 per day by 2031. Now, keep in mind the IPO price was $135, and SpaceXAI stock is currently trading at a buck 60, meaning that these price targets represent a significant increase. Now, one model that you don't have to wait for, but you do get more time with, is Fable 5. Tuesday was expected to be the last day to use Fable as part of cloud subscriptions, with Anthropic switching their flagship model over to usage-based pricing after that. However, you now have until Sunday to make the most of your bundled Fable usage, as they have extended access to Fable 5 on all paid plans through July 12th. That is, of course, unless you've already maxed out your usage, which Andrew Curran thinks is all part of the plan, commenting, "With the number of people distraught that they have used up all their Fable usage for the week already, under the assumption that today was the last day, it's almost certain that Anthropic will announce a surprise reset. This is how you feed a heroic aura, set the stage, and then save the day."

Now, one thing that some folks have been asking me is whether the renewed Fable 5 has lived up to not only the hype, but the experience we had a couple of weeks ago before it got turned off. And so far for me, it absolutely has, although I'll come back and talk in more detail about that at some episode in the near future. Certainly, there are a lot of impressive things that people have done in this very short period that we've had it back. Omar Rashid, for example, the product lead at Google AI Studio, showed off an iPad port of the 2003 game Command & Conquer: claiming Fable had ported all the code across, rewriting it to run natively on the ARM 64 with touchpad controls.

Not to let the big labs all the fun, Business Insider reports that Perplexity has quietly cooked up a coding agent

to take on Claude Code and Codex. The tool is named Teammate, and has been deployed internally since May. An internal announcement viewed by Business Insider said the team eight is designed to oversee software projects from start to finish with the announcement saying it's built for long horizon engineering work owning projects investigating issues and monitoring services. We also have some Google rumors with some vague chatter about Gemini 4. And even meta has rejoined the party in the model game.

Specifically, meta has launched a new image model that actually looks pretty impressive. The model is called Muse image and it's the first image model released by meta since restructuring the AI division to launch super intelligence labs. The model looks pretty close to state of the art. It can handle photo realistic images as well as various stylized effects. Now, benchmarks are inherently a little tricky for image models, but on the image edit version of Arena AI, image managed to rank in second place behind only GPT image 2.

Meta's AI CEO Alexander Wang gave us a look under the hood to see what meta was doing with this model. The model is paired with Muse spark meta's LLM to apply reasoning to a prompt before producing an output. Now, this approach was of course pioneered with nano banana and as we've seen create some fairly significant upgrades to the capability set. Wang said that he was particularly impressed by three things. Self-refinement, i.e. the model improves its own output within its chain of thought, which he said emerged during reinforcement learning not by design.

Second, multi-reference composition, i.e. many images blended into one coherent generation. And third, multi-turn editing, iterating without losing coherence or starting over. Wang also previewed the upcoming Muse video model, which he suggests will be competitive on prompt adherence, visual fidelity, and temporal consistency. Now, a lot of the coverage is focused on how this model is being introduced. Like their previous image model, this one will be available in the standalone meta AI app, but it's also being dropped straight into Instagram and WhatsApp with a bunch of social features like being able to generate an image according to what's trending. However, the feature that's causing controversy is the ability to tag someone else in a prompt and use their public photos to insert them into a generation. For most, this will just be harmless fun, but people are also worried about it being a one-click deep fake machine.

Users can of course opt out, but considering the current state of consumer AI sentiment, I would not be surprised to see it cause some controversy in the coming weeks. Now, one interesting note from a very functional perspective is that Meta is planning an advertiser-specific version of Muse image designed to allow brands to quickly generate product images. There are a ton of AI startups out there who have been focused on exactly that type of use case, but given how deeply integrated advertisers are already into the Meta ecosystem, this could be one that drives business value from AI for Meta very, very quickly.

Lastly today, we're also getting rumors out of China of MiniMax working on a very new LLM with 2.7 trillion parameters, which is larger than any other Chinese AI model that is currently on the market. According to the information sources, the model could be released as soon as the third quarter and is known as M3 Pro internally. As of now, MiniMax is planning to open source the model, but as we will see, depending on how things proceed with the Chinese government, that might not be the way it plays out.

For more on that, we will close the headlines and move over now to the main episode. Welcome back to the AI Daily Brief. Today, we're doing something a little bit different. While our episode starts in a specific news report from Reuters, the main substance of the episode is actually an exploration of the potential implications of how a particular type of action, which I think is not at all implausible, would change the nature of the AI race and a lot of the key questions that we've been asking for the past several months. So, the specific query that we're going to be exploring is how AI changes and how the answers to all these challenges of token efficiency and token cost change if China were to decide to stop letting their companies' premier models be open sourced overseas.

Now, the reason we're having this conversation is a report from Reuters that Beijing is potentially exploring ways to block the overseas distribution of leading models. Reportedly, representatives from Alibaba, ByteDance, and Z dot AI have attended meetings with Chinese authorities over the past month. The meetings were led by the Ministry of Commerce, itself an indication that this goes beyond the tech industry regulator and involves the much more powerful economic planning officials. Officials discussed placing limits on the distribution of the most advanced AI models being developed in China, including both open and proprietary models. Sources said the measures under consideration include limits on who can invest in Chinese AI companies, all the way up to making the leaking of AI technology a criminal offense under stringent national security laws. Now, no decisions have yet been made, but the reporting certainly suggests that China, just like the US, now views frontier AI technology as a national security asset, not just a consumer product. Some of the discussion centers around a Mythos-style approach where lesser models are able to be distributed, but the rollout of next-generation models are under government control. And it is worth noting that officials at this point seem to be mostly thinking about this policy applying to future models, not trying to scrub already released model weights from the internet. Now, I think that most discussions of the AI China US race start from the position that the model of open-source distribution that the Chinese companies have used so far is always going to be their strategy, and if nothing else, this reporting calls that assumption into question.

Now, some folks are fairly incredulous. CNBC's Deirdre Bosa wrote, "Anthropic shutting down access to Fable and Mythos gave Chinese open-source models a huge opening. Unless this is about control and leverage over distribution, why would Beijing restrict access now?" Now, a number of Chinese language accounts on Twitter popped up to basically say that Reuters got it wrong. As evidence, they pointed to a sourced document that they were able to find of a public dialogue in the Chinese courts about this set of issues. The process sounds a bit like the EU trilogs, where in advance of big decisions being made, there's a broader discussion period. And effectively, the Chinese language X accounts that were reading those documents were arguing that Reuters exaggerated its conclusions. TechBuzz China's Rui Ma wrote, "It's important to note that this is not an official policy document, but rather a discussion featuring a Supreme People's Court IP judge alongside leading legal and AI scholars. They say that the 10 biggest themes that repeatedly emerged across the experts were one, open source is no longer presumed to be pro-competition.

Two, the real source of market power is the ecosystem, not the model. Three, open source washing, where companies open source enough to attract developers while keeping the most valuable layers proprietary, is a major concern. Four, open weights and open AI should be regulated differently. Five, traditional antitrust tools in China are seen as insufficient. Six, cross-border governance is increasingly recognized to be fundamentally different for AI. And finally, China wants to become a global rule maker for AI open source. Rather than simply

adopting US or EU models, they wrote, China should develop its own legal framework, strengthen domestic open source infrastructure, and play a larger role in shaping international AI open source governance.

Now, I think that this is a good summary of the document that people went out and sourced. But it seems to me that these accounts are almost willfully misreading the Reuters piece. Reuters is not just referring to the public dialogue in the court. They are explicitly focused on meetings between Chinese authorities and companies including Alibaba, ByteDance, and Z. AI about these issues. Now, I suspect that what's happening here is that in conjunction and in the lead up to that public dialogue in the court, there have been a series of closed-door meetings to take input from those companies. And I do think it's fair to caveat all of this as saying that it feels to be pretty clearly in an exploratory phase. But as you can probably tell from the beginning of this episode, I don't think it's at all guaranteed that China doesn't decide to make a very different decision about their approach to open source than the norms are today. Ethan Malik agrees, posting the article and writing, "This is a key reason I don't expect the flow of frontier open weights models to continue indefinitely or even for very much longer. Sovereign AI strategies of all types are built on the assumption of continuous releases of open weight models that keep pace with the frontier giving cost privacy control gains at the expense of only a little worse performance, but that may no longer hold soon.

Now, what I don't think is particularly useful is to try to get into the minds of the Chinese government. For the sake of this episode, let's assume that there are reasons, ones we might agree with or disagree with, why the Chinese government, who clearly likes having control over the distribution of technology in their country, would want to have more control over the distribution of the technology that their country is making than releasing the open weights of models allows for.

So, let's say that China does start restricting access. Well, what happens then? The entire theme that we've been exploring for the past several months is the growing realization that in a world of agentic workloads, AI costs look radically different than the SAS style budgets that people were planning for before. Now, certainly those issues don't go away if China starts restricting access to open weight models because those issues don't have anything to do with China's open weight models, they have to do with Anthropic and OpenAI's cost of provisioning the frontier, the shortages in compute and all the infrastructure challenges that surround that. Now, so far, the first set of answers to the emerging token cost question have been fairly blunt force type of answers. They've been things like token spending caps, which we got another one from Tesla, applied evenly across the entire company, or on the other hand, there's simplified brute force approach of simply switching to a cheaper model.

So, what are the implications if that second option of switching to a cheaper Chinese model gets taken off the table? First of all, it obviously creates a lot more opportunity for emphasis on open weights and alternative model approaches from the US and Western labs. And already it's pretty clear that some companies are sensing this is an opportunity. Nvidia has recently been putting more and more emphasis on its NeMoTron model family, just announcing this week that it had reached 100 million downloads. About a month ago, the company introduced Nemo-Tron 3 Ultra, really pushing its differentiation from the other Western models and even the other Chinese models around things like output speed. In this world of China blocking access to the frontier of

open weights models, Google's emphasis around Gemma also gets a lot more interesting. In fact, DeepMind's series of Gemma models are quietly pretty popular. A couple of weeks ago at the end of June, Google announced that Gemma 4 had hit 200 million downloads in just its first 2 and 1/2 months. For context, they said the total downloads across the entire Gemma family of models when Gemma 3 was launched was at 100 million. Gemma are what Google calls its lightweight state-of-the-art open models and are clearly meant to serve a different part of the market than just the raw frontier. Now, we don't know how good Gemini 3.5 Pro or Gemini 4 are going to be. And it could be that once released, those models rocket Gemini right back into the conversation alongside the Fables and GPT 5.6's of the world. But even if they don't, Gemma represents this entirely different bet that at least at this stage, OpenAI and Anthropic aren't really making.

Does that become more interesting in this world where Beijing starts to restrict model access? One would think so. And then there's Microsoft. Earlier this year, Microsoft released a series of new models that they had trained in-house. Now, this announcement didn't get a ton of attention because I think the broad perception outside of Microsoft was just that this was them slowly working to catch up with this set of models mostly just functioning as a way to prove that they still had some chops and especially over time should not be considered out of the game. But I actually think that that misses a fairly important strategic change that Microsoft seems to be exploring. Alongside the models, Microsoft introduced something that they called Microsoft Frontier Tuning.

Microsoft AI CEO Mustafa Suleyman wrote, "It's time to move from renting intelligence to truly controlling your AI. Microsoft Frontier Tuning lets you take our models and make them uniquely your own, turning them from capable generalists to complete custom partners." He then goes on to explain how the process works and said that their early results have been really promising. He wrote, "Within Microsoft, we use our reinforcement learning environments combined with our MAI models to climb towards the best agentic use cases for Excel. Our MAI-tuned model is on par with GPT-5.4 on public and private benchmarks while being up to 10x more efficient." In the overall model announcement post, he wrote, "When we tuned our models for McKinsey's tasks, MAI delivered the highest win rate, outperforming GPT-5.5 on quality while being 10x lower on cost. Now, obviously the MAI models are not open, but what Frontier Tuning represents is effectively a commercialized and productized version of what a lot of other folks are exploring with these post-training approaches, but using Microsoft's models instead of these Chinese models that theoretically we might not always have access to." By the way, this doesn't just seem theoretical.

Bloomberg is reporting that Microsoft is actively considering using their MAI models for a number of different functions within their apps. What's interesting is that previous reporting had suggested that they were going to use DeepSeek, but now they're finding that their MAI models, when optimized for specific tasks like generating a chart from Excel data, are actually up to the task in a way that would both save money as well as avoid any weird sovereignty issues with China. Now, the announcement of Microsoft Frontier Tuning happened on June 3rd. The Fable 5 banning happened a week and a half later on June 12th. I believe that if this Frontier Tuning announcement had come out after Fable 5 got taken offline, during that period where we were all waiting for what was next, it would have had a lot more buzz. And by the way, Microsoft is far from the only company playing in this space. Back in October of last year, Thinking Machines Lab launched something called Tinker, which they labeled a flexible API for fine-tuning models. Just about a week ago, Thinking Machines' Mira Murati shared a case study of Bridgewater using the Tinker API to fine-tune a model using what she called their

unique financial knowledge. Thinking Machines' co-founder John Schulman wrote, "People sometimes ask why fine-tune when general-purpose models keep getting better. Bridgewater's work is a good reminder that with the right data, here expert judgments, you can beat prompting-only approaches by a lot."

Now, in the paper, they showed that whereas models between GPT-5.2 and Claude Opus 4.8 all had an average accuracy of between 74% and 78% for a cost of between $20 and about $90, their model got up near 85% at a cost of single-digit dollars. Google DeepMind's director of AGI economics, Alexis Imus, wrote, "The longer I've spent with this paper, the bigger of a deal it seems. The economic implications are quite significant. And we keep hearing about examples of exactly this sort of thing.

Composer's Cursor 2.5 is another example, built on a base of Moonshot's Kimmy, and then modified and post-trained by them to produce Opus and GPT-level performance at a tiny fraction of the cost." Now, in addition to China getting out of the open-source game, creating new model opportunities for Western companies, these sort of changes also put a fine point on an additional value proposition of model routers. Now, model routers are an increasingly popular approach, where instead of just interacting with a single model, you can plug in a model router, which can theoretically figure out the right model for any particular task, leading to efficiencies and cost savings.

Well, especially in a period of increasing regulatory grayness, all of a sudden model routers could start to play an important governance role as well, selecting models not only on the basis of capability, but on the basis of risk. Already, there is so much evidence of things changing. Forsee CEO Guillermo Rauch recently discussed the extent to which they've observed a shift from companies choosing a single AI lab to partner with to actually building complex model architectures. Now, I think what's interesting about this is that in many ways, these trend lines are already set. The need for lower-cost alternative models and better model architectures to get the right task to those models is going to be there whether it's Chinese models plugging in or not. And personally, I think that even the possibility or a growing recognition of the possibility of China cutting off access to the frontier is going to create incredible market opportunities for new model approaches from over here as well.

Certainly, the reality is if you are an AI buyer at an enterprise, your life is getting more not less complicated. At the same time, I can promise you'll never be bored. These are trends we will continue to watch, but for now that is going to do it for today's AI daily brief. Appreciate you listening or watching as always, and until next time, peace.