The Big Ways AI Just Changed

17 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=WUEMQK_ABMY
https://www.youtube.com/watch?v=WUeMqk_ABmY

SUMMARY

June marked a pivotal month in AI, characterized by significant shifts in usage models, the introduction of powerful new technologies, and emerging regulatory challenges. The release of Fable 5 by Anthropic highlighted advancements in AI capabilities, while concurrent government interventions raised questions about access and licensing in the AI landscape.

- *June saw a transition from unlimited AI usage to token budgets, emphasizing token efficiency in enterprises.*
- *Anthropic's Fable 5 was released, showcasing substantial improvements in coding and technical tasks, prompting rapid adoption.*
- *The U.S. government intervened, leading to a suspension of Fable 5 access for foreign nationals, highlighting regulatory concerns.*
- *Companies began exploring alternative AI models and architectures due to cost and sovereignty issues.*
- *The concept of "bot sitting" emerged, revealing the hidden labor involved in managing AI outputs.*
- *CEOs increasingly recognized AI as a strategic priority, correlating with reported business value gains.*
- *The landscape is evolving towards a more diversified approach to AI, moving away from reliance on frontier closed-source models.*
- *Future developments will likely focus on navigating policy questions and optimizing AI architectures for broader enterprise adoption.*

Today on the AI Daily Brief,
why June was the most significant month
in AI in years. The AI Daily Brief is a
daily podcast and video about the most
important news and discussions in AI.
Well, friends, it is July 4th weekend.
Most of my American listeners, at least,
are washing summertime lakes, fireworks,
and patriotic feelings on the 250th
anniversary of this country. And yet,
over here in AI, we are shifting from
one month to another. It is a little
poetic that at the very, very end of the
month, Fable 5 got back just in time for
us to get back to building in July. And
yet, before we do so, it is worth
spending a moment, I believe, looking

back at the last month, which I would argue is one of the most significant in the post-ChatGPT history of AI.

By the way, for those of you wondering, this website companion experience was actually created not with Fable, but with Codex and GPT-5.

Before we get into June, let's actually go back to May. The historian in me thinks that these two months kind of make a matched pair, telling the same story but from different angles.

So, the story of May was all about the shift from the AI subsidy era to the token scarcity era.

Even before May, we had started to see providers shift away from their seat-based subscription models and move towards more usage-based models. This was, of course, the inevitable consequence of shifting from pre-agentic to agentic workloads, which consume just an absolutely massive amount more of intelligence than the type of queries that we were running back in '24 and '25. May was also when we started to see the chickens coming home to roost when it came to enterprises that had run out to start token maximizing.

Uber had been in the news for a couple months as it burned through its AI budget in the first 4 months of the year. We got more and more reports of companies turning off their token leaderboards, and all in all, May felt like the beginning of a shift to a new paradigm.

Now, at the beginning of June, that started to become real.

At the very beginning of the month, we had Walmart moving from unlimited usage of their internal tools to token

budgets. Uber made headlines when it set a $1,500 per month cap on AI spend. And these stories and the others like them reinforce the idea that token efficiency and token discipline were going to become important new aspects of the AI landscape, particularly in the enterprise. And starting then and throughout the month, new approaches, new architectures, efficiency became the name of the game.

Now, it's a little reductive to assume that before this every company was just applying the most advanced frontier model to every workload, but that's honestly not that far off from what I think the average situation was with most companies. Frankly, at most companies AI adoption hasn't proceeded to the point yet where they would really even need to be thinking about efficiency because most companies are just consuming such a vanishingly small portion of the total intelligence that they will ultimately consume. And yet for those companies on the vanguard, there was very clearly a new emphasis on new efficiencies, new model architectures, shifting to lower cost models, including Chinese open-weight models, which would become a little more fraud as we would see later in the month.

We also saw some indications of the infrastructure around AI adapting as well. Independent benchmarking company Artificial Analysis shifted around some of the metrics in its core intelligence index to better reflect agentic usage, and very quietly, in a story that I still think is wildly under discussed, is Microsoft pushing not only a new set

of proprietary models that they had trained from the ground up, but a new product where they would post-train models to the specific criterion requirements of a particular enterprise customer.

I think this missed notice A because it was surrounded by a million other Microsoft announcements, but B, it was just before we really started talking about token efficiency as the important idea de jure.

But the month really kicked into high gear when on June 10th Anthropic released Fable 5.

And while historically it has often been the case that labs have underwhelmed when they've shifted to entire new numerical categories, such as when OpenAI went from the GPT-4 class to the first GPT-5 model, Fable 5 was not that. It was immediately and clearly much more powerful, particularly around technical and coding use cases. But, honestly, as I've said a couple of times, as much as the initial narrative in those first couple of days was that it made more difference for those coding tasks and the improve Opus and GPT wouldn't be as apparent in other areas, I have found that not to be the case. I think the improvement over the other models in every area is extraordinarily clear.

Still, for the first 48 hours or so after Fable 5 was released, the name of the game was finding your most complex and challenging problems and letting Fable 5 just absolutely rip on them.

The best way that I can describe what was different about it, given that I am non-technical and can't compare the elegance or proficiency of the code of

48, for example, to Fable 5. One of the ways that I can describe how it felt different was that if I look back over all the things that I've done throughout the course of 2026 with any of the various coding models, I very frequently get to 80 or 90% of a project and then just don't finish it. Now, in some of cases, that's because the initial tests or results weren't exactly what I wanted or just my priorities shifted elsewhere, but in a number of cases, it was because while the coding models had made the activation energy low enough to just get started, they hadn't obviated the completion energy to actually get the thing done. Fable 5 was the first model that made it feel fairly insignificant not only to start those big coding projects, but to just finish them as well.

And any of you who have enjoyed the new AI Daily Brief website that chunks every episode down into individual shareable components is living in the benefit of that. This is a project that I had had kicking around for weeks at that point, and because Fable 5 came around, I just decided to get it done and done it got in one fell swoop. And thank goodness because as we know now, Fable 5 wouldn't be around all that long.

And in those first couple days, there were basically infinite other versions of people really seeing much more complex and much more complete work getting done. We had Riley Brown one-shotting a Replit mobile style app building app. We had creators testing 3D worlds. And one customer call story had Fable 5 building a requested product feature while the conversation with the

customer was still happening.
Which is not to say that everything was hunky-dory when Fable came out. There were a bunch of big questions that surfaced almost immediately. Now, some of this was about what people thought were over-aggressive guardrails around topics like biology. But one of the other guardrail policies was that Anthropic was instituting a 30-day retention policy where Anthropic said that prompts and outputs from Mythos class models, including Fable, would be retained for trust and safety review. That immediately made many enterprises say, "Absolutely not. We can't use this if you're going to keep that sort of data." It was a preview in many ways of a broader power issue where it wasn't just that one policy, but companies realizing how much their access to one of the most important assets in the business world going forward was mediated by a single or small handful of companies.
And yet all of that seems quaint in retrospect because by Friday of that first week, the Fable story had transformed and the model instead became a precedent for direct government intervention in frontier AI access. The US government used an export control directive to demand that Anthropic suspend Fable 5 and Mythos 5 access for foreign nationals. Anthropic said that the only way that they could actually comply with that was to shut down access to the model for everyone.
Now, initially the story was that this was all extremely abrupt and that Anthropic had had almost no time to react. Although reporting made it clear

that it was a little bit more complex over the next few days.

I'm not going to recount all of it given that it's been so much of the substance of the last few weeks, but suffice it to say we would later learn that a narrow jailbreak report from Amazon triggered this flurry of activity in the US government. But that in many ways it feels like it was a catalyst for various parts of the US government to wake up and realize that this class of models was significantly more powerful than what we had had access to before.

While yes, the specific jailbreak did remain a point of contention throughout the negotiations, there was clearly a broader catch-up process happening at the same time.

Now, for the next couple of weeks, the industry waited while the government and Anthropic negotiated. Pretty quickly, the ban extended beyond Fable as GPT-5.6 got delayed, too.

OpenAI announced that GPT-5.6 would actually be a set of three different models, but that for the time being, the US government would be approving every wave of new companies and people to have access to the model.

To many, it felt like the beginning of a messy ad-hoc AI licensing regime, not based in any sort of determination or legal precedent, but instead a licensing regime that was very much just shooting from the hip. Now, of course, the industry wasn't just sitting around as this all happened. Although, there were some folks who argued that Fable 5 was so much more powerful that it made more sense to just take a vacation for 2 weeks and then come back to it, since

Fable 5 was going to be fixing everything that GPT-5.5 or Opus 4 rate had done in the meantime, anyway. For most others, this became the second major reason, after cost considerations, for why individuals and companies needed to take another look at alternative approaches to just frontier closed-source models. Basically, we now had both a cost and a sovereignty dimension for companies to think about diversifying their architecture away from just OpenAI or Anthropic.
Throughout the month, we saw a ton of experimentation with routing companies that would help create more complex AI architectures that were better adept at routing different types of tasks to the right level of model, but we also saw a lot of interest in new models, with no model capturing more attention than Z.ai's GLM-5.2.
Ever since January of 2025, when the deep seek moment happened, where people discovered deep seek R1 and experienced, in many cases, reasoning models for the first time, leading to, among other things, hundreds of billions of dollars being ripped off Nvidia's market cap, every few months after that, someone was proclaiming that some new China model was having a deep seek moment, but it really wasn't until GLM 5.2 that I think you could legitimately apply that label.
Now, it wasn't that GLM 5.2 was as good as Fable 5, or even necessarily Opus 4.8 or GPT 5.5. But what it was was a model that exceeded the Opus 4.6 GPT 5.2 sort of level, which initiated the agentic era at the end of 2025 and the beginning of 2026 that jump-started the period that we've been living through ever

since. For many, GLM 5.2 was the first open weight model that made the fallback strategy feel less like compromise and more like genuine competition for the frontier. And what's more, it wasn't just models like GLM 5.2 in their raw state that were getting attention, but also custom post-trained models that were built on top of those open weights, things like Cursor's Composer 2.5, which was built off of Kimmy, as well as integrated architectures that basically had multi-model systems built into them. We saw Harvey and Fireworks pair an open weight GLM worker with an Opus advisor for legal tasks, seeing improved performance over Opus alone for a fraction of the cost, and we also got OpenRouter's Fusion, which used a panel of model, a judge, and a synthesizer for hard tasks, once again promising state-of-the-art level capacity at a lower cost.

Now, it would be wildly overstating it to say that everyone switched en masse during this period of forced pause from Fable. But for honestly the first time since I've been doing this show, local AI became a serious question for a much wider set of actors than it ever had before. You had genuine enterprise boardroom conversations all around the world asking what their policy relative to local AI and open weight models was and whether that should be re-evaluated.

Now, the other thing that happened in the absence of Fable was that since we didn't have a new model to play with, there was a lot more emphasis on the harnesses and ecosystems that surrounded the models as an equally important part of how AI gets integrated into

real-world work systems. Now, obviously, one of the big themes of 2026 has been harness engineering, kicking off from Open Claw and running right on through. So, in some ways, that's nothing new. But, there were a slew of new features and announcements and experiments that really put a fine point on all of this in that fable pause period. Both Anthropic and OpenAI pushed some version of a more dedicated HTML or website artifact builder, getting knowledge workers to think differently about the traditional artifacts that they had used to do their work. I did a whole episode about all the different types of knowledge work where you should think about building websites instead of the spreadsheets or slide decks that you used to use.

There was also a growing sense that AI strategy needed to be an ecosystem strategy. In the immediate aftermath of Fable 5 being taken offline, Microsoft CEO Satya Nadella wrote a long post on X about how every company needed to build a learning loop and a learning system around its AI usage. Basically, firms didn't just need to choose the right model. They needed to own the compounding context, decisions, evaluations, and institutional memory that surrounded the usage of models.

One more feature that was announced that I do think is worth specific note was Claude Tag. But, Claude Tag wasn't just another way to interact with Claude via Slack. It was, instead, a way for people in any part of Slack to call upon the power of Claude Code. This democratizes access to the advanced technical capabilities of Claude Code. It gives

Claude Code access to more persistent context. And, it started to shift AI from an individual experience to a group experience in ways that apparently have had fairly dramatic impacts. One of the reason that people took notice of this was the reverence, almost, with which Anthropic's team was talking about it. One of the biggest headline-grabbing claims was Anthropic saying that 65% of its product team code was now being produced not in the Claude app or in the Claude Code terminal experience, but by initiating Claude Code from Slack.

Now, going back to May and the shift from the token subsidy to the token scarcity era, the approximate causes of that shift were not just the increase in workloads that came along with agentic usage. It was also the fact that those shortages are going to be amplified as we run up against the limits not only of our existing computer infrastructure, but the surrounding physical infrastructure that's needed to expand that compute.

One of the big themes in markets this month was the outperformance of memory companies as the memory shortage came into focus. Compute itself is becoming a market of its own. This has certainly been led by SpaceX who expanded their entropic deal to other similar deals with Google and reflection AI. And now it's being reported that Meta and Zuckerberg are following Elon and SpaceX into that sort of accidental neo cloud space.

And with every month that goes on, AI specifically via data centers become more and more of a hot button when it comes to political discourse.

June in some ways was actually a fairly low ebb, but you can feel things brewing from the left and from the right. Anytime you've got Erin Brockovich on the one hand and former Tea Party conservatives on the other mobilizing against the same thing, it's going to be part of the political discourse.

Now for enterprises, some of this discourse around the frontier is going to seem so far outside of their lived experience just based on where they are in the adoption cycle. Indeed, while a tiny sliver of early adopter companies are dealing with things like token efficiency, companies that fall more in the average band when it comes to AI adoption are uncovering new challenges around agentic work like this new phenomenon of bot sitting that was identified in a Glean report. Bot sitting is basically all the work that goes around in making agents work. And Glean found workers in their survey spending an average of 6.4 hours per week making AI usable through things like feeding it context, checking its outputs, and rerunning underwhelming results.

In many ways, June reinforced that the capability overhang is not just going to be solved by new models. In fact, new models are going to make it worse and that it's only going to be solved by real change management. One big jump we saw in the most recent KPMG quarterly pulse survey was the growth of CEOs actively owning AI as a strategic priority. They also uncovered some value around that, finding that organizations where CEOs were accountable for AI versus CEOs were not accountable for AI

were more than twice as likely to report meaningful business value being gained from using AI.

So, as we head into July, what are the big questions?

Despite the fact that we have Fable back, the situation remains very unresolved. We don't, for example, know right now how whatever agreement the government reached with Anthropic impacts the release of GPT-5.6.

And given that we've got reports that there are now even more advanced GPT and Anthropic models waiting in the wings, it isn't clear how this ad hoc informal licensing regime is going to deal with those new models, either. So, one strand of what happens next is going to be inevitably more and more questions on the policy side. Meanwhile, for companies, I do think that there will be a lasting legacy of this period of a pretty significant shifted Overton window when it comes to not just being locked into whatever the state of the art closed frontier model is. And yet you'll note that these types of questions, policy and AI licensing regimes, companies redesigning their token architectures or thinking about customer open models, these are not short-term changes. Instead, these are setting them up for the rest of the year and beyond.

Now, in the immediate term, we do have Fable back. And I think there is actually a fairly unique opportunity in July and August for people to take advantage of that to race out ahead.

No matter how cool the new technology is, there are big chunks of the corporate world that really turn off for

this part of the summer. And honestly, especially if you are operating in that world, if you do not turn off and instead use this time to really see what this new class of models can do, you have a chance to significantly increase your value to whoever you need to be valuable for.

As always, I am thinking about ways we we be able to help with that, but that I see as the real opportunity for the rest of the summer.

Anyways, guys, it'll come as no surprise just how significant I think June will go down in history. Appropriately, if the beginning of 2026 was all about the explosion of real agentic use cases, the middle of '26 was about recognizing the consequences and challenge of that increasing capacity, and the rest of 2026 is going to be all about figuring it out from here.

Anyways, friends, I hope that wherever you are, you are heading into a wonderful weekend. Enjoy friends and family. Happy birthday, America. Appreciate you listening or watching, as always, and until next time, peace.