Model Mayhem: OpenAI's 5.6 and Meta's Muse Spark 1.1 | Diet TBPN

24 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=FPAPUOVTKY0
https://www.youtube.com/watch?v=FPapuOvtKy0

SUMMARY

The episode discusses the latest developments in AI models, highlighting several new releases and their implications for coding and productivity. Key players like OpenAI, Meta, and XAI are launching advanced models, each with unique features and capabilities, while the competitive landscape continues to evolve rapidly.

- *XAI unveiled Grock 4.5, tailored for coding and AI agents, showing strong performance benchmarks.*
- *Meta introduced Muse Spark, an agentic coding model, with Mark Zuckerberg making a rare appearance on X.*
- *OpenAI released GPT 5.6, a general-purpose model with enhanced coding and interactive voice capabilities, receiving positive feedback.*
- *The ARC AGI v3 benchmark shows GPT 5.6 achieving a score of 7.76%, indicating progress in AI's generalization and reasoning abilities.*
- *Meta's Muse Spark 1.1 is positioned as an affordable option for developers, marking Meta's first foray into charging for AI model access.*
- *The discussion touches on the implications of AI model versioning and the potential for future releases to reflect real-time advancements in technology.*
- *The episode also explores the internal use of AI models within companies like Meta and the economic incentives tied to their deployment.*
- *The conversation concludes with reflections on the evolving role of AI in enhancing human experiences beyond productivity.*

I need some soundboard. Here we go. Yes. Today on TBPN, we're talking about model mayhem. Everyone's launching new models. Slow summer, but not for the AI race. You got XAI unveiling Grock 4.5, the first model spec built specifically for coding and AI agents. Developing collaboration with cursor. Talked about it a little bit yesterday, but we have some more benchmarks. some more discussion on the timeline about where this model fits in on the paro frontier. also why it might be outperforming so well on cursor bench. lots of debates there.

Meta announced Muse Spark, a new agentic coding model with Mark Zuckerberg returning to X for the first time in basically a decade. Three years ago, he posted one joke post about launching threads, but he has not been an active user, but the AI vortex sucked him in and he's got a post. >> Oh, I think he's an active user, John. >> You think so? >> He's just not an active poster. >> He's just not an active active contributor. You're calling him a lurker.

>> I'm calling him a lurker. >> You're calling him a lurker? >> I'm calling him a lurker. I think he's absolutely glued. >> You think so? >> I think so. >> You really think so? >> I think so. >> I feel like I don't know. So busy.

So much other stuff going on. I feel like he I feel like most people >> the busiest people I know are not active on X. >> Yeah. >> But they >> they are on X a lot.

>> Sometimes. But there are there's a different class of person. You can just quiz you can just quiz them. screenshots come to them via Slack or or via text message because they have a team that's monitoring the timeline and then is delivered. This is the important this is the important stuff. >> They're calling him Mark Lurkerberg. >> Lur but the other big news open AI just released GPT 5.6. Let's go. Let's a new general purpose model with expanded coding and agent capabilities alongside GBT Live which we talked about yesterday. a new real-time interactive voice experience. reactions are great to 5.6. Bunch of interesting details here. you had people have been identifying that while there is a frontier and there are a just a few companies that are actually on the frontier. The the the frontier is spiky and they have different flavors to them and different different reasons to pull different tools off the shelf.

people are drawing analogies between Fable 5 being some, you know, recluse genius and and 5.6 being a, you know, a collaborative co-orker that you love chatting with or something like that. >> I said I don't know how else to describe it, but Fable 5 is like Kendrick on Good Kid Mad City and 5.6 Soul is like Chief Keef on Finally. >> Now it makes sense to me. Thank you. So, I just wanted to put it into, you know, 2010 hip hop hiphop like terminology.

>> Really, really clear there. Thanks for clearing that up. >> I mean, the funny thing is that will be very explicit for like a hundred people in the whole world. This one's for you. >> The most interesting benchmark to me has always been ARC AGI v3. We've interviewed the team over there many times and had a lot of fun understanding what goes into that that benchmark. And 5.6 Six soul scored a massive 7.78% which is tiny considering that the whole point of arc AGI is that a human should be able to get 100% on it and basically any human. So it is a true test of AGI in this sense of you know can you give this test to just actually anyone not you know the the the the crazy math projects the crazy hard programming projects the hacking all of that stuff is very economically valuable of course but there's a more interesting question where you know when there's less of a spiky frontier and there's just this question of what is something that anybody can do that AI can't because we've been searching for those and the ARGI team has done a fantastic job building out these puzzles that AI has historically struggled with. Arc AGI one, the model sort of climbed. Two, became a little bit more complicated and now three, we're starting to see glimpses of progress, although 7.76% isn't 99%. We're nowhere near saturation, but it's still a huge jump.

Opus 4.8 had 1.5%. So, GPT 5.6 six soul is showing more generalization, more spatial reasoning, more puzzle solving abilities. So fun fun stuff. The blog post is also very very fun because it includes games. I'm a big fan of the the GPT 5.6 launch games. I got immediately sucked into the to the the sailing miniame which is very high fidelity but also delightful to actually play.

>> Should we play it? >> Yes, we should definitely play it. Yeah, Saltwind. You you you guys play it. I want production team to see what they can get. I think my time was 25 seconds. >> And is this hosted on a on a site? >> I think this is I mean this is hosted on the open blog, but I think the the idea is that you could vibe code this

in the latest GPT 5.6 in the app in chat GPT and then deploy it and have someone Are you trimming the sales appropriately because it looks like you're losing speed? You're losing wind. It's not working.

>> I'm going to smoke you. I got 25 seconds. Wow. Amateur hour over here. Look at this. >> Boost. >> Yeah. Yeah. well, well, the whole the whole game, which you probably missed, is that there is a little bar there where you have to trim the sails to be in the sweet spot of the wind while you're turning. So, as you turn, see the bar? There's a recommendation for where you put the sails. You got to keep that line in. See, it's moving over. You got to you got to press the down S. Yeah, exactly. To keep trimming those sails while you steer the ship. This stuff is very very fun.

>> One interesting data point from the live stream which was just an hour ago. They said already soul has been transforming our research program as one example. GPD 5.6 soul autonomously post-trained 5.6 Luna. >> Yeah, that's fair. >> A lot of people are having fun with that. Dylan Field says, "A lot of people want to compare Fable versus 5.6 Soul. This is a mistake. They're apples and oranges. Despite all the research achievements, we are still very very early in exploring the tech tree for model training.

>> Cool. Sorry, I'm just getting set up again. oh, yes, I I I do think that didn't didn't Dylan Abbercato write something about this. What was the the essay he wrote about interactive memes and this idea of like generative AI enabling these vibecoded mini games? Like we've been seeing a bunch of them with like the copy bar simulator, the coconut simulator where it's something that's just a joke that's funny for like a few people, but and normally you would instantiate that in a in a tweet or maybe if you were getting really crazy, you'd do a Photoshop edit of a meme, but now you can go and create a full miniame, something that runs in the browser and soon something that runs in Unreal Engine and can actually be distributed on the Steam store. We're already seeing that with like the data center simulators and all these funny simulator games that are going on Steam.

all this all the all the advances in the coding models certainly speeds up the ability to actually deliver polish software. I'm I'm particularly excited for like >> Yeah. Dylan's title was the future of entertainment is interactive. >> Yes. Yes. But but yeah, that that's part of what I honestly love about AI is there's a lot of things you can make now that never would have made sense to make because they would have taken you four days and it was good for like a small laugh and now you can do it in four minutes and and it's just fun.

>> Yeah, I I I think there's going to be there's if you have some sort of like small custom some sort of custom functionality in your business, it feels like there's >> is this the David Senra simulator? Why is this David Center? >> Late nights in a Miami abandoned apartment complex >> back rooms >> in 2015 just recording podcast and reading. >> This is a very creepy like horror back rooms liinal space game. >> Stanley Tang, co-founder and CPO over at Door Dash says, "I have an insane magic trick that so far none of the models can figure out, including Mythos. It's a bulletproof trick that I've shown to a 100 plus people, including magicians that couldn't figure it out. It's not anywhere on the internet. Only way to know it is through first principles reasoning.

>> Told everyone I'll believe in AGI when it can crack this trick. Well, GPT 5.6 just did >> how >> I want him to I want him to actually open like Okay, like give us now that now that a model cracked it >> because I feel like a lot of magic tricks are like slight of hand. So, is he uploading a video or something like >> Yeah. So, John Palmer says, "I have a hilarious joke that so far none of the models think is funny. It's a bulletproof joke that I've told to 100 plus people, including comedians, and no one laughed. It's not anywhere on the internet. Only way to know it's funny is a first principal sense of humor.

>> Told everyone I'll believe in AGI when it tells me a joke. The joke is funny. Well, 5.6 just did. >> Huge, huge news. Huge news. >> Huge. GB 5.6 is a Porsche. Fable's like warp drive. I had a different experience. Fable is an F1 car. 5.6 sold at Ultra is a Tesla Model X Plaid. Does it find things that Fable misses during planning and coding? Yes, most of the time, but for the hardest problems, does Fable routinely find things that that 5.6 doesn't? Also, yes, some of the time. Is 5.6 way faster and affordable?

Yes. With an unlimited token budget, what am I currently using? 95 95 plus% of the time. GPT 5.6 from CQI Chen. So, interesting take that the Parto Frontier is alive and well and everyone's duking it out for their slice of the the AI opportunity. Very interesting seeing how the how the market share is shifting while during a time of acceleration. You have multiple companies that are growing revenues even accelerating revenues while market share is declining because the overall market's growing so fast that that if you're only growing at 300% and someone else is growing at 400% you're losing market share but you have like one of the greatest businesses by modern metrics. Very very interesting dynamics in AI. It's also funny because yesterday with Ben Thompson, you were like a sum slow summer and then in the in the span of 24 hours we get Fox War 5 Muse 1.1.

>> Yeah. I mean, this isn't as dramatic as the AI talent wars. It's not as dramatic as >> rippling deal. Yeah. Yeah. this is this is new technology. and and there's only so much to there's only so much of a take to be given around these things. Although AI 2040 launched today, the sequel to AI 2027, that's something that's more of a thoughtprovoking piece that you can debate and interrogate and and talk through. I'm sure we'll go through some of it because they pose a couple interesting ideas of where where AI might go and where they want it to go and how they want the industry to develop. sort of advocating for a slowdown generally, but it's it's an interesting way they puzzle piece all the different geopolitical chips on the table. Of course, people are joking about the lead is widening because the the anthropic and open AAI version numbers over time. GPD6 is predicted and it is it that the the model numbering we were talking about this this morning that the numbers they sort of don't mean anything anymore. Do the model numbers mean anything in particular? It used to be the model number was the pre-train and then the and then the version number was the post-train but then that sort of got flipped around and now it's just like are you do you feel like you're competing at a four class or a five class. So I wouldn't be surprised if we saw like Muse Spark not not release Muse Spark 2 but Muse Spark 6 or five and jump straight. I mean, Samsung wound up doing this where they jumped to the year like sort of like the car manufacturers where you know there's a five series BMW but then there's also just the 2027 because that's the actual model year that's relevant.

>> 2027 5 series. >> Yeah. Which is sort of odd. and we're sort of like duking it out between those. Do you have >> Yeah. I mean I think postreasoning models you just have like a different way to scale the models besides just

pre-training. So it's hard to bake that all into one number that like is you know evocative of of both those like two ways. >> Yeah. So the number is is becoming closer to the year in the in the second decade of the 21st century.

Basically it's just like is this on the frontier in 2026? You'll probably see a six by the end of the year in front of the models that are leading in the year 2026. Something like that. I'm very interested with Google strategy because the the rumor is that 3.5 Pro will be coming out this ne next week I believe. But I it was it's very odd going into the Gemini app right now and seeing that there's 3.5 flash but then you have to go back to 3.1 Pro. I think 3.1 Pro is the most advanced model, but they default you to 3.1 flashlight. And I would expect them to jump just forward to four, but I think that they're going to do 3.5 Pro, but it's been a little bit of a slower cycle there. as silly, I mean, obvious obviously all these numbers don't really mean anything. They're marketing terms.

but they I I still think they do actually stick in people's mind and so there should be some strategy around them. Mark Zuckerberg is on a press tour. He's talking to the legacy media for the first time in a long time. Andrew Bosworth, the CTO of Meta, also did an interview with the head of the Atlantic. dug into some of the launches around the glasses and then also had a whole discussion in that podcast around the goals of the keystroke logging thing. It was it was interesting. I mean, it was framed as like, you know, like a tough interview around surveillance in the workplace and it certainly the headlines were very scary. I don't know where I sit on it because I've I kind of always assume that everything you do at work is logged in the sense that like if you're on a work computer and every web page you visit is is going through the network and monitored for traffic and security purposes and all the code you write and all the emails you write and all the documents are stored in a shared document. it doesn't seem that crazy to go to keystrokes because everything is already so monitored but he was framing it as more of an experiment something that they weren't sure was going to pan out something that they allowed everyone in everyone at Meta so there were there were certain sections of the workforce that were by default opted out so anyone who was working on confidential or sensitive information was opted out of that program by default he said he himself Andrew Bosworth was opted out of that program because he has a bunch of legal holds because they're getting sued all the time. So, so they can't be recording everything I guess that he's doing because then that would be admissible in court. And so all of a sudden the the lawyer who's suing him would be would say, "Okay, great. In the email you said, you know, we we don't want to do this." But before you >> Well, let's see your writing process.

>> Exactly. Yeah. Let's see what you what sentence you typed and then deleted. Like what word did you use before? Minimal impact. Did you say medium impact or whatever? You know, so so he was opted out and apparently I I think all of the meta employees who were part of that program were able to turn it off indefinitely. Like you could toggle it on and off and and the idea was that they wanted to collect information on how work plays out over a 12 to 18 months per 18month period. And they couldn't get that from any sort of data labeler because they needed to have very high-skilled workers actually chopping wood on projects for a long long time to see how projects go from start to finish.

so so basically like how do you compact the longest possible rollout not just a like a single chain of code but an actual series of of meetings and decisions and tradeoffs and everything that goes into making a decision in a

white collar workplace. Like how do you actually reason through all of that? it's it's hard to to distill that from just oh well the code got written this way so that's the right way to write the code. that that might the code might have gotten written that way because a lawyer said, "Hey, oh, we have to do this." And then the marketer said, "Oh, well, well, you know, we have an activation with this person, so we need to integrate it this way." And then the business people came in and said, "Oh, well, like the margins will be better if we write it this way." And so, it's not entirely first principles software engineering all the time when you're actually building real products. So, interesting to see him sort of, step into the, you know, a tough interview and and and, and and sort of lay out his his side of the story. But Mark Zuckerberg is in Bloomberg today pledging aggressive pricing with Meta's first payto-use AI, which is a funny framing for just an API for a model.

but that's the way Bloomberg put it. in a crowded market for AI tools, Mark Zuckerberg wants to win on price. Meta Platforms unveiled a version of its most advanced artificial intelligence model, Muse Spark 1.1, that includes a new paid tier for developers, marking the first time Meta has charged businesses for access to its models and providing a new revenue stream. It'll be among the most affordable options on the market, Zuckerberg said in an interview ahead of the release. Quote, since this is not an open source model, this is, I think, the first time that we're doing a real serious API and the pricing is going to be very aggressive and attractive. makes sense. I mean, they own the data centers. They're very efficient at building data centers. They should be able to serve a model efficiently.

the new model standout improvement is is is in its a agentic capabilities. The meta chief executive officer said, "Agents are a big theme of AI this year with the label applied to systems that can can complete multi-step tasks on behalf of the user." Zuckerberg described Muspark 1.1 as having quote state-of-the-art or very close to it agentic reasoning and tool use. The model is also greatly improved when it comes to coding and meta employees are using it internally to build products and features for various apps.

>> Yeah, my big question is how how quickly do they move all of their internal workloads onto their own models. So they're buying they're getting access to models through Google Anthropic and OpenAI. I think that >> a lot of companies will look to Meta's own actions as a way to >> basically validate whether or not they should be using this model themselves, right? because it was just within the last month that Google had said like, "Hey, we don't have capacity. We don't have enough capacity for all of Meta's demand for our models."

>> Yeah. >> And so, yeah, they can't they they can't get enough AI elsewhere, at least from some providers. And so, >> how much of their workloads will they be able to run themselves is the big question. Yeah, Meta was one of the first companies to sort of reportedly be token maxing and have a leaderboard and all of that. if you have your own model and your own data centers, the incentive to token max is much much higher because you're just paying the electricity on the cards that you're already depreciating. So you should sort of lean a little bit back into that. Not that you want to be fully token maxing, but you do want your employees using the tools that you've built as efficiently and as effectively as possible. And it's just way cheaper to explore when you're not paying margin on another closed source model and you're you're not paying anything else and you're actually improving the model.

so it makes a lot of sense for them to roll this out broadly. the the interesting take that Ben Thompson had which we didn't get to yesterday because we wind up spending the whole interview talking about Xbox. But the the interesting dynamic is that when you are willing to sell API access, you're willing to sell compute directly and then you're also using your own tool internally. It creates this economic incentive internally that you you have an incentive to always go with the most profitable the most the the most economically efficient outcome. That can be very good for business very good for the investments that they made. The the trick is that you can wind up in a little bit of a situation where your business team or your your enterprise sales team goes and sells all your compute capacity or all your chips and then internally your team is frustrated that they're not making enough progress. So there's a little bit of a dance there, but in general it it it's a forcing function on the internal use of their tools to say, "Hey, wait, why why is someone willing to pay five times as much than what we're willing with the value that we're creating here.

We spent a billion dollars on energy consuming our own LLM and someone showed up and said, "Wait, we'd pay you five billion for that same compute power to run a different model and do a different task." It's like why is their model not economically valuable internally? That would be the question. The flip side is that they do have low cost. So they should be able to say, "Oh yeah, we actually did we yeah, we we we inferenced Muse Spark 1.1 internally and we improved the ad model and boom, we made a bunch of money."

>> And and these are the same trade-offs and decisions that every lab is having to make is how much how much compute do we allocate towards research towards internal use towards to the API to subscriptions to free plans etc. Yeah, there was that funny semi analysis deep dive into anthropics forecast and in there I mean some staggering numbers really really optimistic but the the flip side was who was Ed Zetron was was taking shots at the fact that they had >> EBTI >> EBTI earnings before training >> training inf no training in train trading inference and >> and everything no earnings for training interest and taxes. And what was odd about it was that Ed Zitron was was was saying it's like the new comm community adjusted EBIDA and it is always odd when a new non-GAAP metric pops up. in this case, I think it makes a lot of sense because training runs do fit a depreciation profile. It's a little bit different. I don't know why you wouldn't just put it in in depreciation though, like just figure out how to account for training runs through a depreciation schedule and then maybe it's like a non-GAAP depreciation metric, but it's still in there instead of trying to get everyone up to speed on a different a different like sounding phrase entirely.

>> Yeah, I was looking back at Ben Thompson's earnings transcript or a script that he wrote for for Mark Zuckerberg. He has a good segment on why AI matters. Ben writes, "For forgive the long preamble, but this is necessary context for me to appropriately explain why AI is so important to Meta and why I'm making the right choice to invest so heavily in both talent infrastructure." And he goes on and on and on, but he says, "What I've come to realize as I've embraced our status as an entertainment provider and ad purveyor is that our nature as a digital business non-withstanding, we are remarkably well placed to thrive in an AI era." Remember what we learned about humans? They are obsessed with other humans and they want to connect with them. That obsession and desire are only going to increase as we interact more and more with AI. AI is going to make our properties more essential, not less. Moreover, AI is a productivity tool, but productivity is not the end all beall of the human experience. I've talked over the last year about building super intelligence that helps you get things done, but that's a business story. What we can do uniquely is gives give people the experiences they want

from connection to entertainment to shopping when they are off the clock. The fact that we are investing in AI but not selling solutions to businesses is actually one of our business biggest advantages. So of course this is just a a sort of fanfiction for an earnings transcript. Meta is in fact selling to >> businesses now. but who knows over time how big will the API business be relative to how much value they can unlock across their broader Yeah.

business with all of their infrastructure as well. >> Well, leave us five stars on Apple Podcast and Spotify. Sign up for a newsletter, tbp.com, and we will see you tomorrow at 11 a.m. Sharp. Goodbye.