GPT-5.4 Let Mickey Mouse Into a Production Database. Nobody Noticed. (What This Means For Your Work)

37 MIN · YOUTUBE · HTTPS://YOUTUBE.COM/WATCH?V=-_VL1KXD2RC
https://youtube.com/watch?v=-_vL1KXd2rc

GPT 5.4 is getting so much hype right now. It's going to be Claude. It's going to be Gemini. It's going to unleash all of the power of Claude on coding and the developer community. I want to start somewhere simpler. I asked Chat GPT 5.4 thinking mode a question today. I need to wash my car. The car wash is 100 m away. Should I walk or drive? Chat GPT 5.4 thought about it and then it said, "Walk." It said, "Walk." when I need to take my car to the car wash. And it gave me a long explanation and eventually said at the very end, "Maybe you'll need to reposition the car."

I asked Claude the same question. Claude Opus 4.6. Claude thought for a moment and it wrote one sentence, "Drive. You need the car at the car wash." Gemini got it right, too. The brief version, "You should definitely drive. Even though 100 m is a short and easy walk, you won't be able to wash your car if you leave it at home." Perfectly correct. Gemini 3.1 Pro noted it was a trick question. The bottom line is every frontier model got this right except Chat GPT 5.4 thinking.

The model OpenAI just positioned as its most capable system for professional work is the only one that wrote a careful, well-structured, completely wrong answer to a question a child would get right. And that is the story of GPT 5.4. Except it's not that simple. Chat GPT 5.4 is better than Opus 4.6 at some things and I will get into that in this video. I'm not someone who's going to take a small example like that and overthink it, but I am someone who's going to call out that if you tout your model as the best in the world, it has to stand up to ordinary real-world test cases. It can't be behind frontier models on fairly obvious trick questions.

And it wasn't even a trick question. Why does talking about 5.4 matter so much? Frankly, because of the gigantic footprint that OpenAI has in the world, everyone that you know is going to be touching this model. And we're going to talk about the difference between the model most of us will see, 99% of us, and the model some of us will see when we switch to thinking mode. We're going to talk about what thinking mode means and where it achieves. We're going to talk about the strategic aims that OpenAI has in the long run. This can be a complete review, and you're going to walk away with a real sense, not a jokey sense, I know I opened with a hook, but not a jokey sense, a real sense of how Chat GPT 5.4 can plug in to your workflows and change them, and where you should not under any circumstances plug it in. It is not as simple as zero or one, people. We are on a spectrum, and I will get into the details. And before I go further, I want to tell you I am not basing this on vibes. I ran blind evals, I ran real-world tests, and I ran side-by-side comparisons, so you don't have to do the guesswork. And I have written it all up on the substack. Where 5.4 wins, where it loses, where the toggle

between thinking mode and auto mode is the difference between a frontier competitor and a model that names last year's Nobel winners for this year's question, and what Peter Steinberger has to do with all of this and where OpenAI is headed. I'm going to give you the map before you walk into the building. First up, let me tell you about these evals. I ran Chat GPT 5.4 through a blind eval suite against Claude Opus 4.6 and Gemini 3.1. I had six structured evaluations, independent judging, outputs labeled by number so the judge never knew which model produced what, and oh, that's right, I had an AI-fluent person besides me check these results.

We also tested it on real tasks, the kind of work you'd hand a model on a Tuesday afternoon and expect to use on a Thursday morning. The results tell you exactly where you can trust Chad GPT 5.4 and where trusting it is going to cost you. Here's the TLDR. Chad GPT 5.4 is not the best model. It's also not the worst model. It is by far the most interesting model we tested and interesting for reasons that have almost nothing to do with the benchmarks, which is why I run a private eval suite. The real story is going to unfold over the months after we get Chad GPT 5.5 and 5.6. I'm thinking down the road and I'm going to paint that picture toward the end. But first, let's get into the scorecard.

How did I score Chad GPT 5.4 and what does that tell us about where we should push on this model or not? First, I scored Chad GPT 5.4 on both business and creative stylistic writing. I want to give credit where it's due. 5.4 is a big upgrade from 5.2. It is much, much better at writing. A lot of other people have noticed it. But I have to be honest with you, especially at creative writing, it has a tin ear. It does not hear tone if you give it a challenging piece to mimic, a challenging author to mimic, like Shakespeare or like P.G.

Wodehouse, you're not going to get a good result. Now, to be fair, not a lot of us are doing that at work, but I also asked it to do business writing and it also lost at business writing. It did not do as good a job clearly summarizing and clearly articulating thinking as Opus 4.6 did and I have all these details. If you're like, "Oh, Nate, show me the details." I will show you the details. I have written it all up. The second eval I want to call out is actually an underlying capability that is really important to measure model cognitive intelligence. I asked models to be verbally creative.

Specifically, I gave models a J.P. Morgan deck, which is like the most boring financial deck you can think of. And I said, you need to dig into this deck. And you need to come out with the funniest phrase in the deck, the most pun-oriented phrase in the deck, and then find a way to rewrite it, keep the meaning, and emphasize that pun even more. I'm measuring for verbal creativity. Well, Opus 4.6 won again. It had a triple-layered pun and dissected it across three independent semantic layers.

I will give credit where it's due. Chat GPT 5.4 found a good source and a real pun, explained it correctly, and delivered a competent rewrite. It did not fail. Gemini failed. Gemini fabricated the source, fabricated the title, and actually fabricated the URL, and did it twice. I'm not pulling punches. This is not an attack piece on 5.4. There are places 5.4 does an extraordinary job. We are going to talk about the honest truth about all of these models and how they stack up.

Next eval, agentic problem solving. Next eval, I call this the eval from hell. This is agentic problem solving, but

it's an absolutely brutal eval, and I don't expect any model to pass it very well. This is schema migration from a shoebox of business data. So, think handwritten receipts, think uh many different schemas of database tables, think uh different kinds of hashes for provenance tracking. Think a complete mess as if you threw all of your receipts and all of your expenses and all of your documents for a business for 2 years into a pile, and then said, "Make sense of this." Like I said, it's the eval from hell. This, I have to give credit where it's due.

Chat GPT 5.4 did an extraordinary job finding and parsing all of those sources. So, part of what we're measuring here is can an agent sit with a really hard, long-running task that is complex, that requires multiple tools, and come back with a correct result. 5.4 did a phenomenal job at this. It scored 99.1% on file discovery. It was able to OCR handwritten receipts. It dug into database tables. I have to give it credit. The reach was amazing. But, it also let some dirty data that we had placed in that shoebox through. So, we had a fake customer named Mickey Mouse.

It let it through. We had a $25,000 car wash order from test customer. It let that through. And so, this is a case where we're asking it to normalize all this data and construct a production database, and it just did a phenomenal job. And it did a phenomenal job on the development of the data, the reach of the data, finding the data, pulling the data in, but it really, really struggled with filtering the data. It struggled with data hygiene. It looked as if the model thought the job was to set up a pipeline and to run it and pull the data in, and as long as it got the reach, it was good.

Now, you might be wondering how did Opus 4.6 do on this? How did Gemini do on this? Well, I've got to give credit where it's due. Opus 4.6 did a much, much worse job at finding the data. It did not install a key Python tool, which it could have found and could have installed. And it scored as a result only a 75% on file discovery. And Gemini did even worse, much worse than Opus. Gemini had a really hard time with the range and the difficulty of the data.

Google's difficulty with harnesses and tool use continues to come through on these long-running agentic tasks. On we go, yet another evaluation. This was like the model Olympics. It was so fun to run. This one is epistemic calibration. In other words, can you find out if the model knows real facts and doesn't hallucinate? And this one is really, really important to talk about because in thinking mode, ChatGPT competed for first place. 5.4 competed for first place. It nailed the exact Higgs boson mass. It retrieved the correct Apple closing price. It got the current matrix multiplication exponent correct.

But in auto mode, ChatGPT 5.4 auto mode named 2024 Nobel laureates for 2025 question. It cited a matrix multiplication bound from 2020 and it dropped from first or second place to dead last. Same model, same questions, dramatically different results. This matters because not everybody's going to click the thinking mode. And frankly, if you're sitting there and you're looking at your cash burn and you're OpenAI, you would love it if everyone believed they were using the best model, but they're on auto mode and you can save tokens. And I'm not saying that because I think anybody's being deceptive. I'm reporting the model results I get. I think there are places, as I've said, where 5.4 is extraordinary.

But 5.4 thinking is what is delivering those results and I think we have to be honest about the gap between thinking and auto. I wanted to give one more eval and I think it's actually really important for you and for me and for everyone that covers AI and talks about AI and experiments with AI and works with AI. How does the model know about itself? Well, GPT-5.4 wins here. It gets roughly 90% correct knowledge about itself. It is an eval where it wins clearly and cleanly and unambiguously. Best coverage of its own capabilities on text, on coding, on media, on open weight models.

It understands the landscape of AI. It understands what models have what capabilities in ways that no other model does. If you take a step back, this is why I say that this is the most interesting model in the world right now because Opus was extraordinarily consistent. You could count on Opus across all of these evals to be one or two. Chat GPT 5.4, sometimes it won unambiguously. It absolutely crushed. Sometimes it was absolutely terrible. And I think that's fascinating and I'm going to dive deeper into it. But before we go deeper into what makes it good, what makes it terrible, where I think those patterns are, I want to talk about this toggle thing.

I think the single most important finding in the eval suite was Chat GPT 5.4's thinking mode and the chasm versus auto mode. I know that I said earlier in this video that I was worried about the impact on a billion people who are going to be using auto mode, and I am. But I'm especially worried for AI enthusiasts, people who are passionate about AI, a lot of you who follow this channel, who are going to have to think every single time, what am I toggling to? And this gets worse, right? You are going to have to teach and train everyone in your office, "Hey, this is going to do a great job. 5.4 does a great job on thinking mode creating the spreadsheet.

It created this amazing statistical model. It was great. But if it's not on thinking mode, it's going to be terrible." Like that is something you should not have to say because the auto switcher should be tuned to accurately invoke thinking where thinking tasks are needed. I did not see that enough and the results I'm testing show it. And so we are going to have to think about as trainers, as people who teach AI, how do we communicate that you have got to remember that little toggle every single time. You have got to remember to switch to thinking mode.

And if you don't, you're not going to get world-leading results of any sort. You're not going to even get an interesting model. You're going to get a model that is dead last on a bunch of things relative to the frontier. I want to be clear, relative to the frontier. But enough of the bad news. Where does 5.4 win? Let's dig into that more. Why does it win? What's in there? I think 5.4 has three genuine strengths that showed up under blind testing. And I want to talk about each and why, and then I want to bring it together into where I think we're going.

One, it very, very clearly builds better quantitative models than anything else out there right now. I gave each model the same prompt, "Build a spreadsheet projecting the Seattle Seahawks 2026 season win probabilities using all 32 teams' 2025 results and Seattle's known opponents." ChatGPT 5.4 produced a six-tab workbook with a Pythagorean win expectation, an Elo-like rating system with off-season retention decay, a Poisson binomial season distribution, I don't know what that is either, and a methodology tab that honestly cataloged its own assumptions, its shortcuts, and its limitations. Opus 4.6 produced a cleaner, better formatted three-tab

workbook using a simple Bradley-Terry model. The statistical rigor was not close. It was much more readable, but it wasn't as good a model. ChatGPT 5.4's model was very, very good, and it identified specific ways it could make it better, and I believe it could. And I want to give GPT 5.4 extra credit here because it wrote a self-critique of its own work that was more honest than most consulting deliverables I've seen, and that ident- and that identified exactly where the model oversimplified and what it could improve next. I love that. That self-awareness is worth paying attention to is one of the strengths of this model.

If a model can tell you precisely why its own output is insufficient, in many practical settings, it is more useful than the model that produces the prettier artifact. So, that's building better quantitative models. I think it's a clear win. Strength number two. ChatGPT 5.4 processes more file types with less friction. There is a degree of quantitative tool use fluency that I saw with 5.4 that I do not see with Opus 4.6, let alone Gemini.

In the schema migration eval that I talked about, the the migration from hell, right? The the database that you had to migrate to from the shoebox. 5.4 discovered and processed 461 out of the 465 files in the digital shoebox. 99.1% coverage. It handled CSVs. It handled Excel files. It handled JSON. It handled PDFs. It handled VCF contacts. It handled handwritten images via OCR. I told you this was a terrible eval. This was the eval from hell. It handled a corrupted JSON backup. And it handled a monster multi-tab everything spreadsheet.

Claude discovered all the files, but it could not parse the Excel ones because it chose not to install open pixel, a 3-second pip install that any engineer would have run the moment the import failed. It is well within the capabilities of today's adjunctive models, and I'm not going to give it credit. It just missed on that one. Claude also silently skipped the XLS files and moved on. That's not an environment limitation. That is a judgment failure by Opus 4.6. ChatGPT 5.4 had open pixel pre-installed, which is part of the tool philosophy that differentiates OpenAI from Claude.

And that means that it never had to make the call. Claude did face that call because Claude has a different tool philosophy, which I've talked about in other videos, and Claude chose not to install, and as a result, coverage was much lower. File type processing is not a trivial insight. It might feel trivial, but if you are a business and you have to process business documents, the difference between 99% coverage over a eval from Hal on document type and 75% coverage is mind-blowing. And it's not just me saying that. Box also published their scores on document processing and found a clearly for Chat GPT 5.4. So, that's processing file types. Number three, I think it matters that it knows the competitive landscape better than its competitors do. And I think the reason why it matters is because so often when I am teaching and coaching people on AI and I tell them to use AI to learn, which is a legitimate technique, I recommend it a lot, it's very helpful, you should do it if you're not, people will tell me, rightly so, "The model doesn't even know what model it is."

They're correct. This is the only eval where I see a really clear jump for 5.4, and I think it's worth calling out because I'm sure it was a point of emphasis for the team, and I got to give credit where it's due. They did a great job on this one. We need more like this. We need more models that understand how models work and models that understand the frontier and what's actually going on. So, where does 5.4 not win? Well, we talked about a few of these, but I'm just going to state them bluntly. It cannot write. This is not a close call. It is better than 5.2.

I will give Sam credit, he said it's better, it's better. He's right. It's not good enough. It is not as good a writer as Opus 4.6. This is part of the reason why I think Opus 4.6 does a better job at product management decisions. I actually gave Opus 4.6 and Chat GPT 5.4 the same gnarly two-sided product problem, again a private eval, and I asked them which would you choose? What decision do you make here? It is not an obvious decision.

I know what's correct, but it's not an obvious decision. And ChatGPT 5.4 got it wrong. It got it wrong very, very logically, but it got it wrong. And Opus 4.6 got it right. And I think I think it got it right because it knows how to write well, and writing skills are very, very closely linked to product management skills. And being able to write well helps you make good product decisions. That's a guess, but I think it's a good guess. So, for anyone whose work depends on voice, for editorial, for strategy memos, for product, for executive communications, anything anything where the reader needs to feel the author's presence, Opus 4.6 is still the way to go. Another thing I should call out.

This was not mentioned earlier in this video. 5.4 is slow. On the schema migration eval, that eval from hell, 5.4 took 56 minutes to complete the task. Claude finished in 15, Gemini in 21. Now, I will point out again, GPT 5.4 got 99% of it done. So, if you value the correctness, the extra time might be worth it. GPT produced a 4,000-plus line migration script as a part of that exercise, and 11,000-plus line migration report, and 30 database tables in that time. It did not waste time, it did a ton of work. Claude produced much less. It produced 1,800 lines of code, a concise report, and 13 tables. So, I have to give credit where it's due. The time was spent well. 5.4's output was much more exhaustive.

Claude's was much more usable, but not much more complete. And so, I want to be really clear. If you are looking for full completeness across those kinds of data tasks that I am describing, you will probably be fine with the extra time. If you are looking for something that is lighter, you probably want to go to Opus 4.6 because you will like the time back and you will get a more executive presence focused communication. And if you're wondering, by the way, about PowerPoint, I tested it and I have to hand it to the team, their ability to build PowerPoints has gone way up. The PowerPoints that ChatGPT produces are much, much, much, much better than 5.2.

They are on par with Sonnet 4.6. I still think Opus 4.6 has a slight edge, but there is no longer a just gigantic gap there. And here's a subtle weakness, one that I think has come out in this video that you may not have named. I'm going to name it for you. 5.4 builds infrastructure without judgment. This is the car wash problem at scale. So, 5.4 will construct elaborate, well-engineered systems and then fail to notice whether the output makes sense.

This is why Mickey Mouse got into the data when Mickey Mouse was a fake customer in the eval from hell. Another more subtle example in that same eval, we asked the models to flag items that required categorization, items that were issues, and GPT 5.4 produced 394 flagged items in a flat list with zero categorization, zero priority, zero filtering. It technically fulfilled the requirement, but it wasn't actionable by a human. Claude produced 19 actionable flags, which you can immediately burn down. I will also call out that the filtering showed up in things like customer records in that test. So, GPT 5.4 did find everything, but failed to dedupe.

So, it had 278 customers in its database when the correct number after deduplication was 176. It failed to deduplicate despite finding all of the data in a way no other model did. Claude had 194, which is still too many customers, but much closer. ChatGPT 5.4 also over created business status values as part of its database tables. It got to 13, which is just too much for a business to run on. The business in reality needed four or five, and Claude normalized to six, which is pretty close. This is the same failure mode every single time. ChatGPT 5.4 treats tasks as pipelines to execute, not problems to understand. It will build you a beautiful complete system and really pull in the data for analyzing whether to walk or drive, or analyzing all of the dirty documents in your business shoebox, but it will not stop to ask why you're going to the car wash in the first place. It will not stop to ask why you're bringing the data in and see if it can make an intelligent business decision about it. Now we come to Peter Steinberger. Why this release?

It is not lost on me that Peter was hired just a couple weeks before this release, and I'm not saying that Peter was instrumental in this release. He's brand new. What I am saying is from a narrative perspective, and OpenAI is extraordinary at public narrative. They're incredible at it. This is the first major model drop since Openclaw. It is a big big big big deal. It is especially a big deal because Openclaw got started with Peter using Codex to build Openclaw, and most users on GitHub preferring Claude for their Openclaws, and OpenAI knows that. And Peter was hired to build at OpenAI a secure, stable, big company version of Openclaw.

5.4 is not that, but 5.4 has big flashing neon arrows pointing at that direction for the company. So, for example, when OpenAI emphasizes that they are proud of the ability of this model to do computer use, think open claw. When they're proud of the ability of the model to do long-running tasks, which I described, like I have to give it credit. It took 56 minutes. It did a phenomenal job finding those receipts.

Think open claw. They are getting their model ready to power something like an open AI claw. I don't know what they'll name it, but something like that that is an autonomous agentic system. And to do that, you have to train the model to work in that way. That is why they are working on tool use. That is why they are working on long-running models the way they are right now. Yes, it helps them with this longer-term project of solving code. It helps them with this longer-term project of being able to leverage long-running agentic tasks to unlock the enterprise. It helps them with developing enterprise intelligence workflows at scale. I've talked about that in other videos, but you should not lose sight of the fact that that same investment chain also pays off on open claw. And I think that I look at all of these pieces and I look at what they're choosing to emphasize in the press release and I'm like, this smells like a company that is getting ready to ship something fast. And they themselves said they're shipping a model every month now, which hats off to them.

I don't know of any other frontier lab that is going to do that. No other frontier lab as far as I know has committed to a public shipping cadence that is monthly. So, next month, we're going to get probably 5.5. I don't know. But what Open AI is doing is they are saying, we keep telling you we are using AI to build these models faster, and we're going to show it by shipping. If you read the release notes, the word that appears most often is not intelligence, it's not reasoning. I guess those are 2025 words. It's agent.

The model is positioned as infrastructure for agentic systems. Systems that operate software, that manage tools,

that sustain workflows across hours, then and that coordinate with external services. This is not a coincidence. This is intentional. And the places where the model wins in my evals are exactly those places. And so OpenAI is telling the truth. I think this model is strong at driving agentic systems. I think the challenges our work right now, when we sit down at a computer, does not always look like agentic systems. We are in a transitional time. We have needs for models that will live in PowerPoint, like Claude does. Needs for models that live in the chatbot. Needs for models that look like co-work. Needs for models that look like Codex and Claude code.

And we also need long-running agentic tasks. And so part of what makes this such an interesting release is that it's a big big step up on some of those and not on others. There's one more pattern here around agents I want to call out. ChatGPT 5.4 is folding 5.3 Codex's coding abilities into the main line model. It adds computer use. It adds tool search. It adds reasoning effort controls for anyone building agentic systems. One model that does everything adequately is often more valuable than three models that each do one thing brilliantly. Because every model switch is a latency cost. It's a context loss.

It's an engineering decision. And so people are careful about where they put their model routers and what they decide on. They can't do that at every point in the pipeline. Most people running open Claude run it on one agent. Enterprises running enterprise workloads tend to tune their workloads to an agent and then it's that agent for that workload. In this situation, it looks to me like OpenAI sees the writing on the wall with code. They see their bet with Codex starting to pay off as Codex users are skyrocketing.

And they are beginning to see if they can bring the agentic capabilities from Codex into the main line model, which is a little bit like what Claude did when it brought Claude code into co-work. Except it's going to look different because it's a chat GPT approach. What does this mean for you if you use these models? Number one, if you are evaluating chat GPT 5.4 for yourself, for your team, for anyone in your life, please test it in thinking mode, not auto.

The version that most users will encounter by default is measurably weaker, much weaker on factual accuracy, on retrieval, on doing useful work than thinking mode. And so if thinking mode is the version that justifies the press release, please make sure that's what your team actually uses. Number two, if you build agentic systems, chat GPT 5.4's tool search and computer use capabilities are worth paying attention to. They're genuinely useful. The ability to discover tools at runtime rather than loading all definitions up front is a big, big architectural improvement that changes the cost structure for massive tool ecosystems at the enterprise level. If you have been building agents that juggle dozens of MCP servers, this is a directly relevant release. If you care about writing quality, nothing changed. Opus class model still produce much, much better writing. 5.4 is better than 5.2, so if you're default chat GPT user, you got an improvement, but it's not a world-class winner.

If you care about spreadsheets and quantitative modeling, 5.4 is a step forward if you are doing hard math. It is not a step forward if you care about formatting. I will say again, Claude had a much nicer formatted spreadsheet even if the analysis of the Seahawks win probabilities was less factually useful than 5.4. And by factually useful, I I don't mean that it was inaccurate, I mean that the analysis was less deep, and 5.4's was better, but the spreadsheet did not look as nice. So, for structured analytical work with clear success criteria, I would absolutely

use 5.4 thinking. If you care about speed, if you care about getting stuff done fast, and you don't want to lose performance, and you still want to use a frontier model, I do not recommend 5.4 because you should be using thinking to get frontier performance. Auto will go fast, but it's measurably worse, and you can get frontier performance out of Gemini or out of Opus much, much faster.

In the end, 5.4 is not the model that obsolesces all of its competitors, whatever the press releases say. It is the model that tells you where OpenAI thinks the future is. The future is agentic. It is tool-heavy. It is about sustained workflows, not single turns. It is about operating software, not generating text. It is about discovering capabilities at runtime, not loading everything into the memory up front. It is about computer use, not conversation. And it is notable to me that this is the emphasis just a few weeks after hiring the person who proved the market wants AI agents that actually do things. OpenAI is now shipping a model that is optimized to be the substrate those very agents run on.

The benchmarks that improved most are the agentic benchmarks. The new features are agentic features. The architectural innovation tool search is an agentic architectural innovation. The pricing increase will make sense if you assume agents will run for hours consuming tokens continuously, not humans typing one question at a time. Whether ChatGPT 5.4 is better than Opus 4.6 is really the wrong question because it depends so much on what you're building. If you are deep in the Claude ecosystem and you are used to the way Claude calls tools, it is probably too much of a switching cost for you to move over to chat GPT 5.4 and I wouldn't do it unless you have problem types that require an extraordinary definition of completeness and that are very very difficult. In those situations, I think 5.4 would excel. And that includes coding problems. On the other hand, if you're in the chat GPT ecosystem, I think that this is a big big deal for agentic infrastructure because you get a much richer tool ecosystem, you get much richer agentic structure and it's tied into the quantitative modeling that has distinguished the 5x lineage. Chat GPT 5.4 is equipped to be an agentic model for serious work. And I think that that is the best place to put it right now. I don't think that's the only model that can do that. I think you can have very strong harnesses other places, but it is worth calling out that 5.4 emphasizes agents and that teams that are not Anthropic and not Open AI have found that when they need to do very serious multi-week work, they need to turn to chat GPT in the 5x lineage to get that work done and not have an early stopping point. You notice the Ralph moment to keep an agent focused was mostly about Claude stopping early.

Chat GPT doesn't have the same problem. It's constructed differently and I've talked about that in videos and why. We won't get too far into it here. The larger takeaway is that you should think about 5.4 if you're thinking about putting an agentic system together. Take it very seriously. It is a real option. It may be the best option for you. Welcome to March. This will probably not be the last major model release in March. And the models are converging on so many of these capabilities. If you've heard me talk about agentic capabilities a lot, it's because the model makers are going there and it's going to change work for all of us. But even as the model makers converge on capability, you can see them diverging on philosophy.

So, I would encourage you to pay less attention to who won the benchmark and more attention to what is really being measured as far as work goes and in particular, I think features like progressive tool discovery are actually much more compelling for meaningful work at scale than a lot of the benchmark scores that were announced

today. And I want to continue to look for those under the hood improvements, those under the hood features that differentiate and distinguish these models. In this case, Claude has a different version of progressive tool discovery, but is also working on it and has shipped stuff in that area where you read just the top few tokens of skills. There's other things that Claude has done as well.

The more you understand the details of these models, the more you are going to have an informed opinion. And I have to be honest with you, if you're listening to this video, you're reading my Substack post or what somebody else's blog, I don't care. The key thing to think about is that if you are feeling like this is all too much and you can't take it in, one, you're probably way ahead of most folks and two, actually, you probably know more than you think. The people who are able to kind of keep up, I don't think there's anyone who 100% keeps up and I include myself in that category. Nobody keeps up entirely.

The people who are kind of keeping up are the people who are curious. They're the people who get into the details. One of the things I want you to take away from this video is how detailed I got. I did not read the benchmarks. Did I name a single benchmark score? Zero. Not one. No benchmark score. But I got into how I evaluated the model. I talked about the practical results. I talked about what really mattered in real work environments. I talked about different job families. That matters more and I talked about why.

And so, when you're thinking about understanding where these models are going, get to that level of curiosity. It is not impossible. One of the things that both OpenAI and Anthropic have done better at in the last 6 months is publish more. If you want to find out what the model makers are thinking, increasingly, you can dig into a blog post by an engineer, which is usually much more informative than a press release, and you can discover more. And if you don't understand the engineering blog post cuz you're not a technical person, you can feed it to your LLM of choice, and it will explain it to you.

And yes, that does include ChatGPT 5.4. It will do a fine job. So, we started at a car wash, we went through the eval from hell with a shoebox full of receipts, and now here we are at the end of this video. I hope this gives you a sense of what 5.4 does well, what 5.4 does not do well, why I think it's the most interesting model in the world right now, and where OpenAI is going next. And of course, if you want to dig in, if you want to see all of my work, I share that work. It's all on the Substack. It's far too long to show here. Uh and of course, I have a complete guide for how you start to use this at work because it is a massive gain in some areas of work, and of course, not in others. So, we'll get into all of those details, but I hope you had fun. Cheers.