BY **MBI DEEP DIVES** IN **OPENAI** — MAR 31, 2026

The Pendulum Between Intelligence and Knowledge

Over the last week or so, I have started noticing an interesting development that is worth highlighting.

Let's start with Microsoft's announcement yesterday. From Microsoft's blog yesterday (emphasis mine):

> "Today, Researcher—Microsoft 365 Copilot's deep research agent for work—takes a significant step forward. Designed to tackle complex research in the flow of work, Researcher now goes further with two new multi-model capabilities that raise the bar for accuracy, depth, and confidence: Critique and Council.
>
> **Critique is a new multi model** deep research system designed for complex research tasks. It separates generation from evaluation and utilizes a **combination of models** from Frontier labs including Anthropic and OpenAI. One model leads the generation phase, planning the task, iterating through retrieval, and producing an initial draft, while a second model focuses on review and refinement, acting as an expert reviewer before the final report is produced. **Our evaluations show that this architecture exceeds traditional single model approaches** and delivers best in class deep research quality. This design provides clear optionality across generator and reviewer roles, with the ability to support and expand these roles over time as the system evolves.
>
> **Council brings multiple model responses side-by-side in the Researcher experience**. Additionally, a cover letter provides valuable insights on **where the models agree, where they diverge, and the unique insights each brings on the topic**."

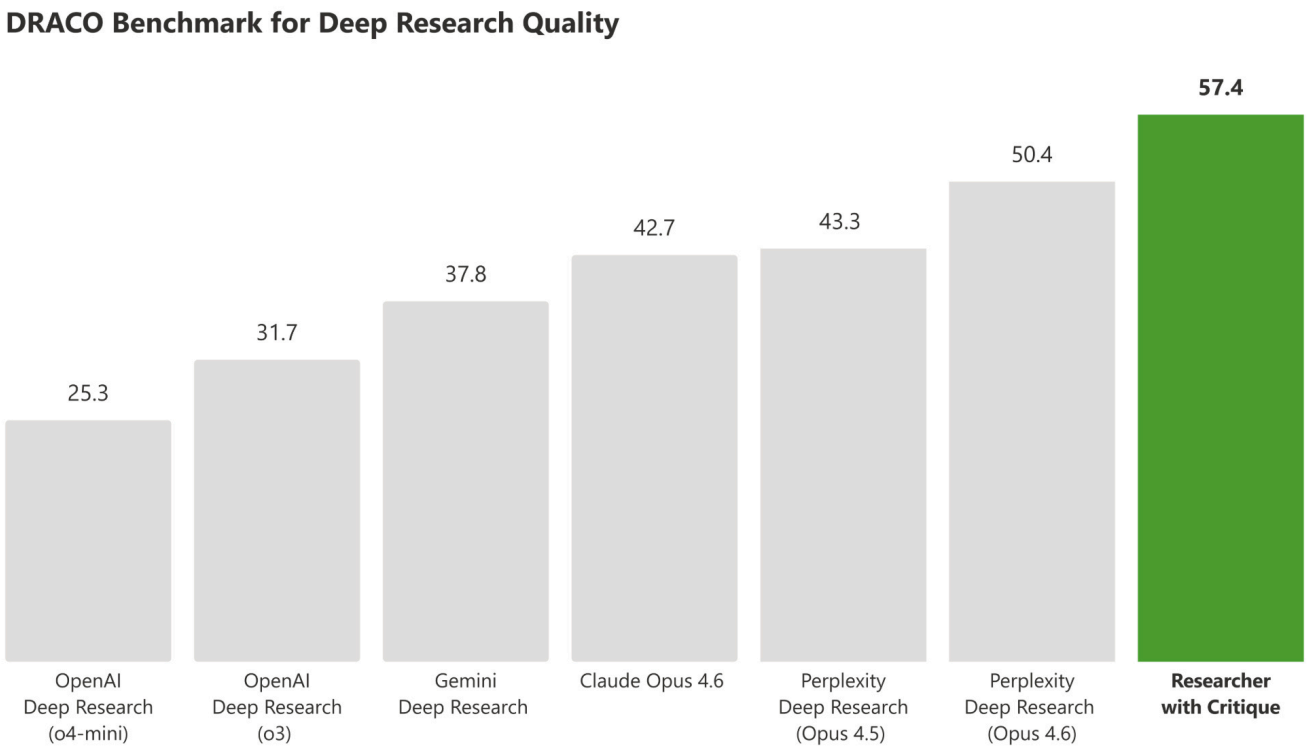
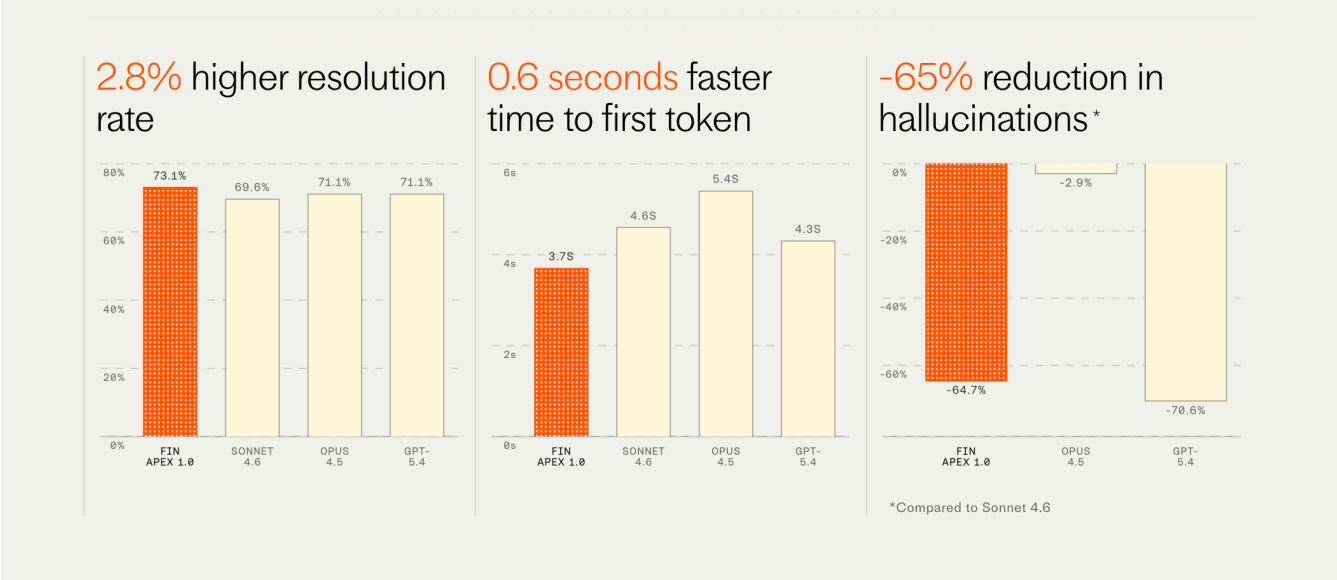


Image Source: Microsoft Blog

As someone who often copies and pastes the same queries to ChatGPT, Gemini, and Claude and then **manually** reads their responses and evaluates the quality afterwards, I can certainly see the appeal for a product such as "Council". "Critique" is perhaps more useful if you want to automate something (hence you're not in the loop) and want to ensure AI's work goes through multiple phases of refinement before it presents the final work to you.

Let's move onto Intercom now which is a customer service suite software company. Intercom shipped their own customer service AI model called "Fin Apex 1.0". From Intercom's blog this week (emphasis mine):

> "As of last week, ~100% of all (English language, chat and email) customer conversations are now running on Apex. Since day 1, **the Fin engine has comprised a system of models, and last year we started replacing the off-the-shelf models with our own, custom trained on our proprietary data**. But the core answering model was always a frontier labs offering—initially versions of GPT and recently Sonnet 4.0. But now that core answering model is Apex 1.0.
>
> This model resolves customer issues at a materially higher rate than any other model available. One of our largest customers in the gaming space saw their resolution rate improve overnight from 68% to 75% (i.e. a reduction in unresolved conversations of 22%). We've never seen a jump this large from a single improvement since we started Fin.
>
> But **importantly it's also dramatically faster, has fewer hallucinations, and is far cheaper than all other available models—all factors that weigh significantly in the consideration of companies deploying these agents to their service operations**."



Source: Intercom Blog

Last week, I also highlighted a paper by Meta in my piece "Meta's Agentic AI Ambitions". From my piece:

> The most interesting takeaway from the paper is that **a great setup can compensate for a less powerful AI**. The researchers proved that a weaker model (Claude 4.5 Sonnet) using the Confucius scaffolding successfully fixed more bugs (52.7%) than a stronger, more expensive model (Claude 4.5 Opus) using Anthropic's standard setup (52.0%). When powered by the GPT-5.2 model, Confucius Code Agent successfully resolved **59%** of the real-world bugs on the SWE-Bench-Pro test, beating both prior academic research and the official corporate systems built by OpenAI and Anthropic under identical conditions.

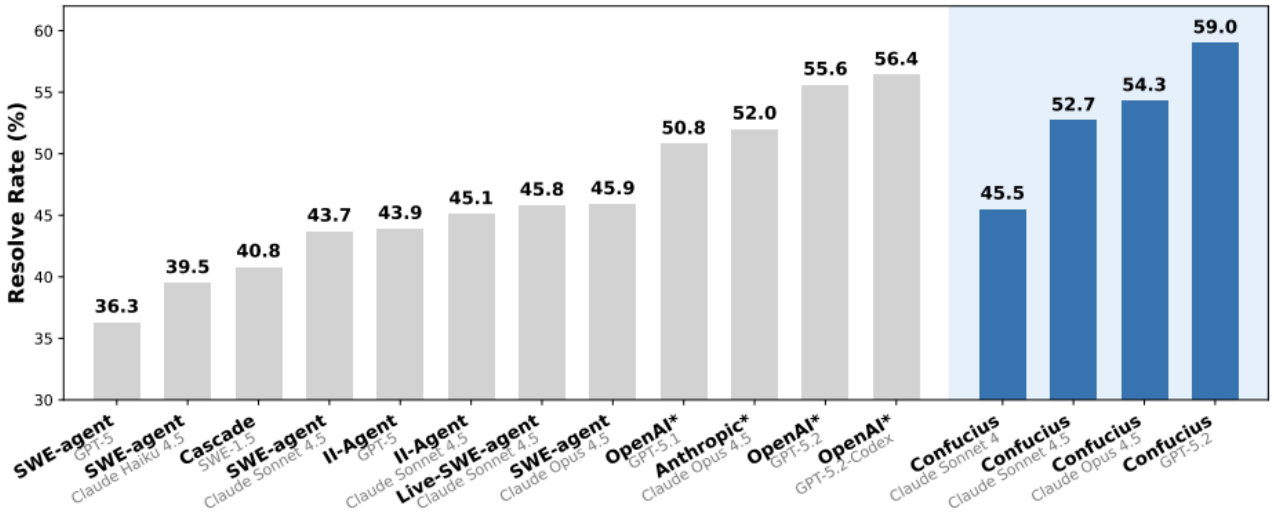


Figure 1 Performance comparison on SWE-Bench-Pro benchmark. (* reported from Anthropic's Claude Opus 4.5 system card / OpenAI's GPT-5.2-Codex system card.)

Image Source: From paper "Confucius Code Agent: Scalable Agent Scaffolding for Real-World Codebases"

Behind the paywall, I will share some thoughts what these different announcements hint at the future of AI and potential implications for frontier model developers.

*In addition to "Daily Dose" (yes, **DAILY**) like this, MBI Deep Dives publishes one Deep Dive on a publicly listed company every month. You can find all the 67 Deep Dives here.*

What do all the aforementioned different announcements/research have in common? They collectively demonstrate that raw, general-purpose intelligence from frontier models is increasingly being outperformed by specialized system architectures and domain-specific adaptations. If vertical companies can use cheaper, open-weights models as a base and outperform proprietary giants like GPT-5.4 or Claude 4.6 in specific use cases, frontier AI labs may increasingly struggle to justify premium API pricing for those **specific** enterprise tasks.

Meta's Confucius Code Agent and Microsoft's Critique/Council systems prove that the **system surrounding the model** can potentially be more important than the model itself. By using advanced scaffolding such as hierarchical working memory, persistent note-taking, multi-agent peer review etc. developers are achieving state-of-the-art results without needing the largest, most expensive models. If a mid-sized model wrapped in superior software engineering beats a massive frontier model, API spend will inevitably shift toward cheaper, faster models for those tasks.

In fact, by using pre-trained models as a starting point, vertical companies can bypass **the most expensive and compute-intensive phase** of AI development. However, the improvements they are making are not merely "incremental". Thanks to their own proprietary data, they represent the complex "last mile" required to make AI useful in the real world.

Foundation labs train on public internet data and increasingly also on synthetic data. On the other hand, vertical companies like Intercom have access to proprietary data i.e. millions of historical, resolved actual customer service interactions. By applying post-training and reinforcement learning against real-world business outcomes, they can create highly specialized intelligence that foundation labs cannot replicate without that specific data. Frontier models can be much more intelligent in general sense than these models, but still fall short on executing specific tasks at hand compared to these vertical models.

But aren't these frontier models just going to be more and more intelligent over time? Dean Ball recently wrote a thoughtful piece in which he eloquently made the argument that intelligence and knowledge can often diverge, and knowledge can be far more distributed even if AI models keep getting better every year. From Dean Ball (emphasis mine):

> "There is one further observation that follows from the disentanglement of knowledge and intelligence. This is that knowledge itself is distributed throughout the world in highly uneven and imperfect ways. **Anyone who thinks that "all the**

> **world's knowledge" is on the internet is deeply mistaken. There is information that exists within a firm like Taiwan Semiconductor Manufacturing Corporation that is, first of all, not only unavailable on the internet but literally against Taiwanese law to make public**. Even more importantly, though, there is knowledge within that firm that cannot be written down *and* is only held collectively. No single employee knows it all; it is the network—the meta-organism of TSMC itself—that holds this knowledge. It cannot be replicated so easily. This is all merely a restatement of the knowledge problem most memorably elucidated by the economist Friedrich Hayek.
>
> The implicit, and sometimes even explicit, argument of "the doomers" is that intelligence is the sole bottleneck on capability (because any other bottlenecks can be resolved with more intelligence), and that everything else follows instantly once that bottleneck is removed. I believe this is just flatly untrue, and thus I doubt many "AI doom" scenarios. **Intelligence is neither omniscience nor omnipotence**."

The more successful vertical companies become at post-training, the less differentiation will come from the raw base model. If Intercom can beat Opus 4.5 on customer service, and Meta's scaffold can make Sonnet outperform Opus, then the willingness-to-pay premium for the very best frontier model can shrink in production use cases. Of course, the labs' strongest remaining moat can still be on **novel, general-purpose tasks where no domain-specific data exists** but will such novel, potentially low volume tasks be enough to justify the skyrocketing compute spent on frontier model training?

Indeed, Andrej Karpathy recently mentioned in the "No Priors" podcast that while he does expect more domain specific, smaller models to thrive, frontier models will still have a large enough pie to keep their momentum. From Karpathy (emphasis mine):

> "I think currently my impression is the labs are trying to have a single sort of like monoculture of a model that is arbitrarily intelligent in all these different domains, and they just stuff into the parameters. I do think that we will, I do think **we should expect more speciation in the intelligences**...there's lots of different niches of nature. And some animals have overdeveloped visual cortex or other kind of parts. And I think we should be able to see more speciation. And you don't need like this oracle that knows everything. You kind of speciate it. And then you put it on a specific task. And we should be seeing some of that because **you should be able to have like much smaller models that still have the cognitive core**. **Like they're still competent. But then they specialize. And then they can become more efficient in terms of latency or throughput on specific that you really care about**"
>
> ...there's going to be always like some demand for like frontier intelligence. And **that can actually be extremely large piece of the pie. But it could be that the frontier, the need for frontier intelligence is going to be like, you know, Nobel Prize kind of work**.
>
> ...open source is kind of like going to eat through a lot of the more basic use cases or something like that. At some point, what is Frontier Today is going to be, you know, probably later this year, what's Frontier Today in terms of what I'm

> using right now from the closed labs might be open source and that's going to be doing a lot of work. So I kind of expect that this dynamic will actually basically continue."

I am not sure if I find Karpathy's argument convincing on the case of frontier model being able to keep large piece of the pie by focusing on "Nobel Prize" worthy of work. My skepticism stems from the general observation that even in the status quo world, our society doesn't seem to pay a hefty premium for "Nobel Prize" worthy intelligence. These data are not public, but I do suspect the life-time earnings of most Nobel laureates over the last 20 years likely fell short of majority of the people who spent their entire career in investment banking, hedge funds, law, or different big tech companies. I'm sure people in those industries are plenty intelligent but hopefully none of us wants to claim that most of them are doing anything close to "Nobel prize" worthy work. Intelligence can certainly be loosely correlated with income, but it may be too simplistic to assume higher intelligence will always lead to commensurate ability to capture value from your output. Far more people likely became much wealthier by commercializing Nobel laureates' ideas than the other way around. Perhaps the future world will be different than the world in the past, but I did want to point out this dichotomy in the past.

However, I also wanted to point out that while the research or blog posts indicate vertical companies are closing the gaps and even exceeding the frontier models, there can be divergence between theory and execution. "Every Consulting" explained a couple of days ago why customers may be better off buying the tools directly from the model companies:

> "When you evaluate AI-powered tools, you're also—whether you realize it or not—evaluating the tool vendor's choices and constraints, rather than what the underlying model provider (like Anthropic, Google, or OpenAI) is capable of. It's often faster to build your own Claude/Gemini/Codex skill with your own rules and preferences already built in.
>
> Companies are increasingly building, not buying, AI software on top of models, because it gives you flexibility. **I don't know how it's possible for companies that aren't the core model providers to keep up when the big labs know what models are coming, build their internal tools to align with those releases, and train them on how to operate within their own environments**."

Indeed, unless you want to participate in this "red queen race" that requires you to keep pace with the frontier model developers, the vertical models can lose their appeal over time. On that note, I would like to highlight what Brett Taylor mentioned in this podcast:

> "You have to be the best at every stage of your company's existence… **you're obviously having to be the best at something that you know is going to get commoditized**. That means you have teams who are putting a lot of their life force for two years into something that everybody knows is just for two years, but it still matters nonetheless."
>
> "**I'm building this and I'm 100% certain we'll throw it away in the next four**

> **months. But I have to build it, because if I don't, I can't serve the bank that has a big business in Hong Kong... that is the reality right now**."

Such R&D intensity doesn't auger too well for software's long-term margin. The core point I am trying to highlight today is the economics in the AI value chain is far from settled. My mental model for AI's eventual economics still gravitates towards Soren Larson's framework. A recent leak suggests Anthropic may launch a substantially more powerful model called "Mythos" soon and there are rumors that Anthropic may charge $15-25 for 1 million input tokens (vs $5 for Opus 4.6), and $75-150 for 1 million output tokens (vs $25 for Opus 4.6). Perhaps unlike the Nobel Laureates of the past, the companies building the frontier intelligence will be order of magnitude more efficient in capturing value for their associated intelligence.

While AI labs bulls must be optimistic about the size of the eventual market for frontier intelligence (and they may be right), I hope today's piece elucidated to you that the range of outcome remains fairly wide.

Current Portfolio:

Please note that these are **NOT** my recommendation to buy/sell these securities, but just disclosure from my end so that you can assess potential biases that I may have because of my own personal portfolio holdings. Always consider my write-up my personal investing journal and never forget my objectives, risk tolerance, and constraints may have no resemblance to yours.

My current portfolio is disclosed below:

Ticker	Avg. Cost	Current Weight*	Unrealized Gain (loss) %	First Bought	First Buy Price	Last Bought	Last Buy Price	Last Sold	Last Sell Price
ABNB	119.6	31.7%	2.9%	Apr'25	100.0	Feb'26	127.6	N/A	N/A
META	446.9	31.6%	20.0%	Aug'18	172.8	Mar'26	600.0	Sep'25	767.6
GOOGL	192.9	13.3%	41.8%	Apr'25	149.8	Mar'26	295.3	N/A	N/A
FND	51.9	8.3%	-5.0%	Mar'26	50.3	Mar'26	52.6	N/A	N/A
CNSWF	2,257	5.9%	-24.6%	Mar'22	1,733.8	Mar'26	1,800.0	Apr'25	3,350
RYAN	35.3	3.9%	-6.6%	Mar'26	35.3	Mar'26	35.3	N/A	N/A
AMZN	137.4	3.8%	46.2%	Feb'20	91.0	May'25	186.9	Mar'26	211.9
VEEV	216.8	3.2%	-18.8%	Jan'26	220.0	Jan'26	210.0	N/A	N/A
Cash		-1.8%							
Total		100.0%							

**Based on closing prices as of March 30, 2026 (time-weighted YTD: -17.0%); Since inception (August 24, 2018) time-weighted annualized return CAGR +13.1%*

***Disclaimer:** All posts on "MBI Deep Dives" are for informational purposes only. This is NOT a recommendation to buy or sell securities discussed. Please do your own work before investing your money.*