AI Research

AI Debates and Emerging Trends

Equities

Americas

Semiconductors

Karl Keirstead
Analyst
karl.keirstead@ubs.com
+1-310-734 2455

Taylor McGinnis
Analyst
taylor.mcginnis@ubs.com
+1-212-713 1332

Roger Boyd
Analyst
roger.boyd@ubs.com
+1-212-713 1256

Radi Sultan, CFA
Analyst
radi.sultan@ubs.com
+1-212-882 0094

Summary

Over the last few weeks we've attended various AI conferences and spoken to a dozen+ F500 enterprise customers about their AI initiatives and tech stacks. In this report, we focus on the new/emerging AI trends that we're picking up, including custom model fine-tuning and the rising demand for AI security, in addition to addressing investor debates about AI model safety/alignment. Bottom line, we remain constructive on the (no-longer-consensus) long-AI trade.

What Feels "New"

We'd flag: 1. As enterprises roll-out AI agents, and especially post the OpenAI-Hugging Face incident, we're hearing a much bigger focus on AI security, sandboxing and real-time agent observability. While the frontier labs (especially OpenAI via Astra's security features) are emphasizing security use cases, in our view demand is ramping materially enough for the security pure-plays (as well as Microsoft and Datadog) to also win. 2. We're hearing a clear uptick in enterprise demand to fine-tune (mostly open) foundation models to build custom or private AI models, a trend that Palantir (via their "sovereign AI" message), Microsoft and more recently Nvidia have been espousing. 3. Nvidia, via Nemotron and OpenShell, has quickly become a real presence up the AI stack. 4. Token spend optimization (via model routing, per employee spending caps) has gone mainstream as AI bills have risen so quickly YTD but importantly, our checks suggest that AI spend is STILL ramping materially (key if the long-AI trade is to gain momentum).

Other Observations

These include: 1. The "pacing the frontier" debate has settled down, influenced by OpenAI's compute head saying last week at the AI Infra conference in Santa Clara (we were in the audience) that a greater focus on safety and alignment will require MORE compute (ensuring alignment is actually compute-intensive itself). 2. At the model layer, open/custom models are gaining share but the share going to frontier labs remains very high, with very positive early feedback on OpenAI Astra. 3. We remain impressed with how even non-tech F500 enterprises are custom-building so many parts of their AI tech stacks (cautious read to the 3P AI products of many software firms). 4. This has become consensus, but the revenue per GW of AI data center capacity is moving higher, a tailwind for all providers of AI compute infrastructure.

This report has been prepared by UBS Securities LLC. **ANALYST CERTIFICATION AND REQUIRED DISCLOSURES, INCLUDING INFORMATION ON THE QUANTITATIVE RESEARCH REVIEW PUBLISHED BY UBS, BEGIN ON PAGE 14.**

Impact of Safety/Alignment "Pacing"

The recent "pacing" commentary from the AI ecosystem (slowing down the pace of training runs for new next-gen or "frontier" AI models in order to ensure that they are safe/aligned) has received so much coverage that we won't add too much here, but we offer a few observations below. We conclude that the impact of self-imposed speed limits on compute demand could turn out to be modest. As a result, we remain constructive on the long-AI compute trade (long Microsoft, Oracle as well as CoreWeave shares having launched on the stock with a Buy rating just this week).

- **Impact on Training Compute Demand:** We were in the audience at the AI Infra conference in Santa Clara last week when the head of compute at OpenAI, responding to a question about the impact of pacing on compute needs, answered that perhaps counter-intuitively, OpenAI may need MORE compute because ensuring the safety of new models is actually compute-intensive itself. This view was echoed by the OpenAI CFO, who said last week that "we are still focused on acquiring more compute". Moreover, a) training runs only consume a portion of total compute needs (although this represents a high portion of the OpenAI spend mix for Oracle and CoreWeave), b) the duration and magnitude of any slowing could end up being modest, and c) the frontier labs have very recently signed large compute commitments with 3P providers or in the case of OpenAI moved forward with their enormous 8 GW data center campus in Ohio, moves that do not signal a need for LESS compute.
- **Impact on Inference Demand:** Inference compute demand could remain high as usage of existing (and still under-utilized) frontier models continues to ramp. For instance, we understand that OpenAI's Astra model became the most popular model for Palantir devs/engineers after just 8 days. Moreover, if the rationale to pace in the first place is rooted in a material inflection in the pace of model improvement (via RSI and other techniques), then this of course could lead to a ramp in enterprise AI ROI and therefore usage. The last time we went through an inflection in model performance (Dec/Jan), it served as a major catalyst for enterprise AI spend.

Enterprise AI Security in the Spotlight

While most investors are evaluating the pacing discussion through the lens of how it might impact compute demand and the AI semis/infra trade, we wonder if it might impact the pacing of enterprise AI agent deployment. The risk of a Hugging Face-like swarm of autonomous AI agents hacking internal systems is surely being absorbed by enterprise CIOs, CISOs and Heads of AI. Even if it doesn't weigh on the pace of AI agent adoption, it may result in a heightened focus on AI security and governance.

What We've Heard

Over the last few weeks we've heard the following from enterprise checks. While we have yet to hear any enterprises reacting by slowing down their AI roll-outs, it is clear that the Hugging Face incident as well as the recent debate about safety and alignment have focused the attention of enterprise IT/security execs on the cyber-security layer, especially for multi-agent environments.

- **Customer 3:** *The security software incumbents are focusing on the smaller elements of AI security, for instance Crowdstrike's efforts to secure coding agents. What we need is a more comprehensive enterprise AI platform, the industry needs an enterprise AI control plane for governance and security, for risk classification, identity access and run-time security (protection of IT assets while AI agents are actively operating). The start-ups too are too focused on individual elements.*
- **Customer 5:** *We are now deploying multi-agent architectures in production but it was a long journey to get there because we were worried about the security risks. We finally addressed this issue by custom-building a secure multi-agent sandbox on Nvidia's open source OpenShell platform.*

- **Customer 10:** *The age of what I would call "useful AI" has arrived, AI that is running in secure and sandboxed environments, in which agents can still run autonomously but where you bound their autonomy via virtual sandboxes or Kubernetes-based isolation. You want to bound or control the ability of AI agents to manipulate systems of record and resolve Jira tickets. Agent observability is also key. We've increased our focus on AI security just in the last few months, we're trying to create a clean environment.*
- **Customer 11:** *In the OpenAI-Hugging Face incident, the agents simply didn't have good enough guardrails. You need strong visibility, to observe and audit every move and decision they make. Most enterprises don't have adequate agentic guardrails and traceability in place, that's why AI tech like OpenClaw, Instinct and Grok Bot aren't deployed in enterprises. Security is a blocker to enterprise adoption, you need real-time observability (not logs that you analyze days later) and governance, more and better sandbox technology so that model behavior and agents can be controlled. At the time we moved forward with AI agents in production, Nvidia didn't have OpenShell and there were no good AI sandboxing solutions, so we built our own AI sandbox.*
- **Customer 13:** *Two things are important in the next 12 months. First is improving the trust and governance of AI agents. Second, data is the main instrumentation of building high-ROI agents. Making data AI-ready, helping agents narrow-down context, is key. You need some type of data ontology, a way to organize your corporate IP or knowledge, you need knowledge ontologies.*
- **Customer 14:** *We have not stopped in any shape or form the AI advancements or focus on AI, but we are keeping a close watch on this frontier pacing discussion and how it develops further. The use case of AI in the manufacturing industry is to save cost and improve time to market. That depends perhaps partially, but not directly, on how much Dario decides to slow down R&D. The cyber risk is growing. We are trying to create not only role-based access control but also attribute-based access control. Apart from the standard identity policy setup, what we're trying to do is create a kind of architecture where we are working on data protection, tool execution, authorization, and agent identity.*
- **Customer 15:** *We believe we utilize a prudent and well practiced governance approach to our AI. We believe that our observability, human-in-the-loop approval flows, model selection, or security and governance of autonomous agents protects against security risks. Our CISO ecosystem represents the largest % capex budget increase on a YoY basis to ensure we can continue to deploy automation with safety and soundness. We are also evaluating / deploying AI orchestration and observability services at scale. The timeline for these vary from a quarter for discreet tools to 24 months for end-to-end governance, observability, audibility, training & tuning inference services. Could security risks slow us down? Of course. We are a financial services firm and so we are risk adverse by nature however we are realizing tremendous value / ROI from our AI implementations and continue to have an aggressive roadmap for scaling these services across front, middle, and back office use cases. The ethical question of AI is one we tackle on a regular basis but we frame this relative to model sophistication (i.e. Sonnet vs. Fable). We also remain convicted to a distribution of our AI utilization enabling multiple models including frontier, open-weight, and proprietary - this may shift in the coming months pending risk concerns. We do believe there is role to play from AWS and Azure in our security posture as well as SaaS providers like ServiceNow.*
- **Partner 1:** *Customers in general have become a lot more edgy with AI, but that hasn't dimmed their interest or their plans to expand or essentially onboard Anthropic or OpenAI or Google or Mistral or other frontiers into their workflows. I don't see that the current trajectory will slow down, potentially just reordering or resequencing of a particular workflow based on sensitivity or overall criticality. This is going to further enhance the opportunities for the cybersecurity sector, which are already strong. The number one focus area is identity, and second is governance.*
- **Partner 2:** *It was never the case where we were able to put AI solutions in production without proper guardrails. When we talk about enterprise-grade scale, it's security, responsible AI, control plane, observability, monitoring, tracking, these are at the heart of things right now. The idea of let's build the use case and then security, governance and everything else comes as an afterthought has no value anymore.*

Frontier Labs Pivot to Security

The primary expressed rationale for frontier lab pacing is the security risks, made clear via the OpenAI-Hugging Face incident. On a recent pod, OpenAI's CEO said that we could be sitting in front of a new wave of cyber-attacks, while OpenAI's CFO said last week that we're in a "cyber-defender moment". With OpenAI's DevDay coming up on Sept 29th, it seems very likely that OpenAI will focus on the security features of its models. The company is already laying the groundwork for the security use case, with OpenAI's CEO saying that they trained their new Astra model to combat these threats (OpenAI Daybreak is the company's broader security defense initiative) and the CFO saying that Astra is scoring very well on security benchmarks such that users can "point Astra at software, find vulnerabilities and then patch at machine speed with Codex".

This of course opens up a debate for investors – will the frontier labs a) position their frontier models for security use cases in a way that crowds-out pure-play security software firms (a negative) or b) partner with the pure-plays, with enough AI security TAM to lift all boats (a positive)? At a high level, this is the debate that investors have had around the apps and data software segments all year. We side with the more constructive view and conclude that several pure-play cyber-security firms can get pulled along, as we see AI security similar to cloud, where enterprises will primarily want to work with security specialists. This appears to be the investor consensus, as the stocks of several larger-cap security firms are up materially over the last few weeks as investors already anticipate that cyber-security could be a major new AI use case.

How to Play the AI Security Theme

How exactly are we recommending that investors play this theme? Among the pure-play security software stocks, we prefer CrowdStrike and Okta. CrowdStrike has what we see as the broadest portfolio for addressing AI security concerns, and additionally has led the frontier amongst cyber firms when it comes to leveraging and building their own AI models to discover and remediate risk. As AI security threats have evolved, Agent Identity has continued to attract more enterprise attention, with access management and governance for agents representing a focus point. Okta's tone around pipeline for and potential uplift from its Okta for AI Agents and Auth0 for AI Agents products continues to uptick in our view.

In our view, there are two other plays on this theme, with potential that AI security becomes a more meaningful narrative in the quarters ahead:

- **Microsoft:** In our view, if there is an enterprise-focused software provider that is well-positioned to play the role of trusted AI security partner, it may be Microsoft. Microsoft has a $30 billion security software revenue stream, much of which is captured in the Azure and M365 segments. On his most recent podcast, Microsoft's CEO said that enterprises "need aggressive containment and monitoring, more robust security" in response to the OpenAI-Hugging Face incident, and Microsoft has flagged its new MAI-Cyber-1-Flash model, saying that "it achieves better performance than a much larger Mythos model but at half the cost when combined with our multi-agent security harness". We expect Microsoft's AI security messaging to get louder in the coming quarters, with the mid-November Ignite conference a likely occasion to showcase its AI security models and features. While the Azure and M365 revs growth rates are already poised to accelerate in the coming quarters, a more material AI security pull-through might add to the Street's confidence in the acceleration story unfolding.
- **Datadog:** As noted in the customer conversations above, several of these customers were expressing a need for real-time AI agent observability. Datadog is early-stage with Bits AI and other agent observability features, but if enterprise demand for real-time agent monitoring takes off, it seems very likely that Datadog (and ClickHouse) get pulled along, potentially Dynatrace as well. This is an opportunity that is getting some investor attention but that might result in a narrative shift if/as AI observability takes off.

Token Optimization Now Mainstream

When we first spotted this trend back in May, we wrote that we believed that perhaps 60% of all large enterprises were attempting to control their AI spend (which from Jan-May 2026 had ramped significantly as adoption spiked) via a number of "token spend optimization" techniques, including a) more deliberate routing to cheaper closed or in a growing number of cases, open Chinese models (using non-frontier models for non-frontier tasks), b) usage or token spend caps on employees, and c) prompt caching.

Fresh Checks on Token Optimization

With the passage of a few months, what are we hearing now? Our last published note on this subject (see here) was in mid-Aug post a meeting with about a dozen Bay Area enterprise (mostly tech sector) CFOs, with the key topics being AI usage, ROI and token efficiency. We flagged a widespread focus on token efficiency via several tactics, namely model routing, per-employee usage caps and answer caching to bring AI spending under control. Over the last month we've heard the following. Our key take is that token efficiency or optimization efforts are now even more widespread, mainstream in fact.

- **Customer 1:** *Our 1H26 was marked by a material spending burn and so we turned to token spending controls such as token compression, prompt caching and model routing to become more efficient.*
- **Customer 3:** *The whole industry recently began to realize that AI is not cheap and yet everyone is in a race to deploy. The AI bill arrived and we all realized that we need to better control token consumption. We are doing this too, and have put governance channels in place, gating control.*
- **Customer 4:** *We are trying to lower our cost per token. We're doing this via smarter model routing (using Cursor) and by capping the token allocation per employee. Each user is capped at X number of tokens/day, devs have a higher allocation, and if they need more they have to ask for approval.*
- **Customer 6:** *We have seen 100% y/y growth in token spend and we're now looking to be more efficient, to token optimize.*
- **Customer 12:** *Our AI token costs were running at $1 million per month to start the year and are now $10 million per month, a 10x to $120 million annualized just 8-9 months later. I cannot push for even greater AI adoption and see another 10x spending increase, so I have to bring down my cost curve. I'm doing this primarily by putting per employee spending caps in place as well as model routing.*
- **Customer 13:** *Caching of often-repeated user queries and model output, basically saving the LLM calls, is a key part of driving token efficiency. If a query output can be shown in a pre-populated drop-down based on a certain prompt, that saves LLM calls, that's basically key value caching and re-use. The cache itself consumes GPUs, for sure, but the savings can be material.*
- **Customer 14:** *Right now our volume is not so high that we have to do FinOps for tokens. Once we reach a certain level of token economics, then we would have to think about it.*
- **Partner 1:** *Token-maxxing became a prominent buzzword because you had industry leaders and industry-leading companies talking about it, like Microsoft saying they were using too much OpenAI or the Uber example. But the average customer was adopting AI capabilities in a generally gradual manner, then they also work with companies like mine to optimize it.*

Net Impact on AI Spend

Beginning in June, these enterprise token optimization efforts sparked concern among tech investors that the net impact might be a slower trajectory in AI spend in 2H26/2027 that could weigh on frontier lab growth and hence the long-AI trade. There is now no doubt that token efficiency efforts have become mainstream, but have they indeed weighed on enterprise AI/token spend? OpenAI is not signaling any slowdown, with the CEO recently saying that "every time we find efficiency gains in our models, the world token demand just goes up and up". We've had more enterprise IT exec conversations over the last few weeks on this subject and conclude that in fact these token optimization efforts have NOT materially slowed enterprise AI spending, an important conclusion against a backdrop of investors weighing a potential rotation back into "AI winner" stocks. We heard:

- **Customer 1:** *We're still seeing an explosion of AI use cases across the company.*
- **Customer 2:** *AI is being deployed everywhere, for both internal and external use cases.*
- **Customer 4:** *We are seeing our AI spend ramp materially, it is still ballooning despite our token efficiency efforts. These are helpful little strategies for sure, but our AI spend is still moving higher, up and to the right for the next 5 years. Adoption is not even close to being 100%, it takes a long time to drive usage. But it's possible that our AI spending growth curve could moderate over time compared to 2023-2025 when we had bad or inefficient practices, we're now a lot more ROI-sensitive.*
- **Partner 2:** *The spend on AI is increasing disproportionately. I think there will be 30-40% more AI budget dollars in 2027. It was harder to track in 2026 as budgets were disjointed, but I think that was a 20-25% increase overall.*

Share Shifts at the Model Layer

How are efforts to drive higher ROI and be more token-efficient impacting the choice of AI models? From our mid-Aug Bay Area dinner, we picked up some clear hesitation in using open (Chinese or US) models. Instead, CFOs were pushing for greater use of cheaper closed models where possible (for "non-frontier" tasks). In this latest round of checks, it was clear that open models were gaining share but importantly, it was also clear that frontier lab models were still grabbing the lion's share of tokens and spend. Among the open models, several were using (US-hosted) Chinese models but we were struck with the momentum behind Nvidia's Nemotron models (more on this below).

- **Customer 1:** *At the model layer, we're mostly using frontier lab models. Anthropic is about 60% of our model spend (Haiku, Sonnet and Opus, but not Fable yet) and OpenAI is about 30% (especially the cheaper GPT Sol, Terra and Luna models). We also use open Chinese models, namely Alibaba/Qwen and Moonshot/Kimi. We do so out of cost control and run these on-premise. I'd say 10% of our spend is now related to open models (including Llama), up from zero at the beginning of the year. We use LiteLLM for model routing.*
- **Customer 3:** *We use a combination of custom-built as well as open-weight models. We prefer to use Chinese open models, deployed on-premise to be safe, as they are cheap and very easy to use.*
- **Customer 4:** *We spend a lot with the frontier labs and have AI agents built mostly on top of Anthropic (both Opus and Sonnet). We've started using Cursor more for their model routing software, as they have a great model router, Cursor's software decides what tasks should go to which model, it auto-routes. In terms of our direct use of open models, we evaluated Moonshot/Kimi as well as DeepSeek but bank regulations make this tough, given data security concerns. It's too risky for us so we don't directly deploy them, we don't even use Nvidia Nemotron models. We'll continue to use frontier models or custom-build our own. Anthropic is likely 60-70% of our token spend, up from 35% before. OpenAI is likely 20-25% of our spend with the rest from Amazon models, as well as Meta Llama for very specific developer tasks. Google used to have a much larger share but Gemini models are now less than 5% of our spend.*

- **Customer 9:** *At the beginning of the year, closed models were 99% of our token consumption. Since then, open models have closed a lot of the performance gaps and use is increasing for us outside of developer use cases. Our internal devs/ engineers still prefer frontier lab models, and I think the frontier labs will still succeed because the demand for intelligence is unbounded. We rolled out OpenAI Astra to our developer teams just 8 days ago and it is already the preferred model for one-third of them, it's the real deal. OpenAI Luna is also cheap and very good. We do now use the open models, including Chinese models (but only if US-hosted) as well as Nemotron, and are very interested in the inference platforms such as Fireworks to host these open models as AWS Bedrock actually does a poor job for open model hosting. Nemotron is good at some tasks (especially Ultra, better than Super) but not yet great overall. Whatever the frontier lab models could do a year ago, Nemotron Ultra 3 can do today, that's the timing gap.*

- **Customer 11:** *We use a mix of frontier and non-frontier models, you need optimized model routing. On open models, we were a material user to Meta Llama but have now joined the Nvidia Nemotron Coalition and have switched to Nemotron. We are a very regulated bank and won't use any Chinese open models, but we're also fine-tuning to create a lot of custom models.*

- **Customer 12:** *As we try to token optimize, our use of open Chinese models is up for sure, especially Moonshot/Kimi and Alibaba/Qwen. To be secure, we're hosting these on-premise but this is getting expensive so I have two options. First, we can (and are) using Cursor more, we can just access the Chinese models through Cursor and Cursor just hosts the models itself. Or, I can look at Fireworks and Baseten.*

- **Customer 14:** *Not everything needs ChatGPT or Anthropic in the future. I believe the future belongs in large enterprises to a governance layer which differentiates between what kind of workload will get small controlled models, enterprise models, or frontier models. This is an ongoing discussion. There's no clear decision, but there is a discussion to be independent eventually if the Anthropic's and OpenAI's of this world decide to change their strategy.*

- **Partner 1:** *Open source usage is also sort of helping companies normalize token usage because it's not just that you can avail AI's benefits by using the Claude's and the GPT's in the world. You can also do it using Kimi or GLM or other open source models. I don't know if many people are tracking this, but Nvidia is on its way to become the first fully vertically integrated AI company. The reason I say this is because open source is definitely here to stay, not just because China is promoting it, but there are enough within the U.S. and other regions as well who are trying to catch up. An average corporate will eventually access 6 to 7 models in my view for all their requirements. But the value will still be centered on to the frontiers. In fact, that is why these Fables and Astras will command a super premium.*

Nvidia – Now a Presence Up the AI Stack

Nvidia has obviously been the dominant provider of AI hardware since the AI boom began in early-2023 but as we discussed in the sections above, Nvidia is suddenly (we use this word carefully because it's truly been in the last few months that Nemotron and OpenShell have hit our radars) a material presence up the stack in the model and AI sandboxing markets. Nvidia is open-sourcing these products, clearly with a strategy of removing friction to enterprise AI adoption (a strategy of Nvidia's for years at the AI infrastructure layer) and ultimately driving demand for Nvidia GPUs.

- **Nemotron:** The anecdotes above make clear that Nemotron may now be the most widely-used open-weight model by enterprises (displacing Meta Llama), and we expect this to increasingly show up in 3P model share data. We spoke with one customer and partner of Nvidia, who relayed that Nvidia is "very serious" about improving Nemotron, that this is not some "check the open model box" effort. This tech partner surmised that Nvidia realizes that many enterprises plan to fine-tune larger models to create smaller specialized models (discussed below) and that Nvidia wants to ensure that they are a key part of this new AI stack.

- **OpenShell:** While we've been hearing enterprises flag their use of Nemotron for several months now, it has only been in the last several weeks that enterprises have started calling out the need for a secure "sandbox" for AI agents and that they were customizing Nvidia's open source OpenShell framework for this. We spoke with one check who said that several of Nvidia's tech partners, primarily IBM Red Hat, were scaling their OpenShell support and distribution efforts.

The Pivot to Private/Custom Models

In previous AI reports (see "Open Models and the Impact on the AI Trade" here) we discussed share shifts between the three broad AI model types – closed, open and custom (using a foundation model, usually an open model, to fine-tune against the customer's own corporate IP/data to create smaller and very specialized domain-specific models). While our customer conversations point to continued strong demand for all three model types (see above), over the last month we've noticed a marked uptick in anecdotes from enterprises and AI vendors about a pivot to custom-built models, typically via "supervised fine-tuning" on a foundation model, either closed or (in more cases) open-weight models (usually Nvidia Nemotron, some Chinese). The idea here is that in order to deliver on the ROI promise inside enterprises, AI needs context, context from being exposed to corporate IP. This is why we and investors are suddenly hearing the term "context engineering", "data ontology", "data semantic layer" and "knowledge graph" everywhere we turn. Along with the uptick in focus on multi-agent security and sandboxing (spurred-on by the OpenAI-Hugging Face incident), this is the most significant emerging AI trend that we're hearing.

The custom or fine-tuning option can be cheaper (helps to contain token costs), smaller models can be faster and just as performant as large general-purpose models for certain tasks, it allows enterprises to "own their own weights/intelligence" and protects against the perceived risk that any data leaks to the labs (in our view this is an overstated risk, one that Palantir has been very vocal about). Some have argued that custom or specialized models make sense when the alternative (using a frontier lab) is simply too expensive, when speed matters and when performance really comes from exposure to very proprietary data sets. Palantir, Microsoft, Fireworks and Baseten are all vocal about this shift and the Nvidia CEO talked about it during his Dreamforce discussion, referring to this trend as a "private models". We're told by several checks that Nvidia is investing very seriously into its Nemotron model in part to secure a position as part of this "custom model stack".

What Customers Are Saying

Importantly though, customers are also telling us that this is a high-priority route, in addition to continuing to use frontier lab models. We heard:

- **Customer 2:** *The emerging inference players such as Fireworks and Basten believe that the next big thing in the enterprise, the next frontier, is around fine-tuning open models on the proprietary data of the customer. This creates a new wave of inference demand for these new inference clouds, as they offer customization features for fine-tuning efforts.*
- **Customer 3:** *For our big AI projects, we are fine-tuning on open models or actually building internal homegrown models, both using our own data. The path depends on the use case.*
- **Customer 4:** *We are planning to utilize frontier lab models to build our own tailored or specialized models. This may be the future, custom models trained on our data sets or context. We'll likely end up fine-tuning a closed model, but we are very early-stage here, it's more in the "thinking through it" stage. We'll move in this direction as our devs get better and given a need for speed, ease of use and to reduce complexity. The frontier labs have actually pitched the idea of building custom models for us.*
- **Customer 14:** *There's no decision, but there's a school of thought that large enterprises could have their own managed models for sensitive or intellectual property-related workflows, and broad enterprise large language models or frontier models for advanced reasoning and research.*

Palantir as a Fine-Tuning Platform

Palantir might be pushing this concept the hardest. Palantir refers to this as "sovereign AI", encouraging enterprises to do this to better control their AI stack and protect their IP ("tribal knowledge") and because LLMs lack precision without the context from corporate data. We recently attended Palantir's AIPCon event in the Bay Area, met with customers and management and published our key takes (see note here). We heard the following from customers (given the event, Palantir's customers likely skew more to custom-building models/apps on Palantir than typical enterprises):

- **Customer 5:** *We use Palantir to help optimize our global supply chain. We are fine-tuning Nvidia Nemotron models on their ontology layer to create smaller and more specialized post-trained models that we have found actually outperform larger models with the added benefit that they end up using fewer GPUs. AI bills pile up quickly with the number of AI agents in production and with so many LLM calls. Everyone measures the cost per million tokens but you need to measure the cost per achieved outcome to reflect factors beyond cost, such as accuracy and latency. The cost/performance trade-off skews in favor of fine-tuned open models. We're tuning domain-specific models, tuning our own harness and model routing, all to lower costs so that the ROI of AI can begin to make sense.*
- **Customer 6:** *We use Palantir as our cyber-security data provider and combine this with Nvidia Nemotron models to fine-tune custom models on this data set.*
- **Customer 7:** *We are building our own AI agents on top of the Palantir data layer, allowing us to use AI on our terms, more control.*
- **Customer 8:** *We care a lot about AI sovereignty and don't want to hand over our IP, so we're fine-tuning Google Gemma open models on our data and hosting them on-premise. We've found that they outperform frontier lab models at 95% of the cost.*
- **Partner 1:** *The concern is there that based on my data, which I am letting you access and then you give me a beautiful output, but then that's still my data and you're learning from it. Palantir's pitch is that the model is portable, your data is portable. The ontology on which the entire business is based, that is going to be the asset. The premise is definitely compelling. But Palantir doesn't have the same functionality in terms of capability of the models. So tomorrow, if Palantir comes up with a model, which is more or less in the same capability - even if it is on par with 5.6, not Astra - then that would be a very interesting development.*

How to Play the Custom Model Trend?

The easy (and perhaps accurate) answer is to be long the stocks of the tech firms that are most vocally supporting this path, namely Palantir (our preferred choice) and Microsoft. In a recent podcast, Microsoft's CEO reaffirmed that Microsoft's customers are "asking to own the weights" and to "use all of the models but be independent of all". Microsoft will surely make this a key talking point to customers and investors at the upcoming Nov 17-20 Ignite conference. The other clear way is to be long one of the data software firms (Palantir, Snowflake) that are benefiting from enterprise efforts to aggregate and sync their corporate IP/data to enable fine-tuning. As the CEO of Fireworks says, "data is the essence of tuning". Salesforce's Data Cloud also benefits from this theme and broadly speaking, some argue that SaaS/apps firms are likely to emerge as de facto "neo-labs" by building their own models.

This custom path requires inference compute, resulting in a new inference compute workload type for the hyperscalers and neo-clouds but perhaps more so the new inference platforms like Fireworks. Fine-tuned open models are often hosted on new inference platforms from the likes of Fireworks, Baseten, Together and Modal, such that these hyper-growth privately-held infra providers and their VC backers are singing the same "custom model" tune. The fact that Fireworks has exceeded $1 billion in ARR and that almost every neo-cloud is now more aggressively vertically integrating up into the cloud inference layer is a strong signal that demand at this layer is spiking.



Custom-Building an AI Stack

The desire to build/DIY and exert control over the AI stack goes beyond the model layer, the trend to build custom or specialized models. Large non-tech enterprises (see anecdotes below) are sharing with us that they want to custom-build other parts of their AI stack, even the harness as well as the security and sandboxing environment. It is frankly amazing to us that a shoe retailer (Customer 1) would be building their own AI harness or a financial services firm (Customer 4) would be custom-building an AI security sandbox. Why? In some cases, it's because enterprises are moving fast, faster than the tech community has been able to build enterprise-grade AI solutions. It may stem from a desire for control. This trend also exists in part because of the strength and popularity of open-source AI software (including Nvidia's Nemotron and OpenShell as well as LangChain's LangGraph/Deep Agents open source harnesses and Harbor for an open source evaluation framework), at least giving enterprises something to build and customize on.

Broadly speaking, this "build versus buy" trend is a negative for incumbent software firms in the business of selling off-the-shelf 3P AI products. Why bother to harness a model directly and custom-build (on AWS Bedrock) an alternative to an incumbent piece of software? We've been asking this question to customers for 2+ years now on the back of many claiming that they prefer to custom-build their own AI apps rather than leverage incumbent software. What are we hearing today? The short answer is that the desire to custom-build, for reasons associated with cost, performance or control, remains very healthy.

- **Customer 1:** *We evaluated the use of products such as Salesforce/Agentforce but decided to use our own proprietary AI stack, an internally-built architecture. At the model layer, this internal AI stack includes mostly frontier labs, but 10% of our spend is now related to open models (including Llama), up from zero at the beginning of the year. We've custom-built our own model harness based on open source (LangChain), we use Databricks as our data layer, AWS for infra/storage, LiteLLM for model routing, Datadog and Dynatrace for observability and LangSmith and DeepEval for internal evaluation. Context and memory management are key. If we use a 3P harness, we can't control what goes into the memory of the model, we don't want to dilute the model performance with data that is bad and impacts accuracy. We want fine-grained control, so we control the harness connectivity to 3P data from Salesforce, SAP, Atlassian and others. We don't want vendor lock-in, we want full integration control. If these 3P software firms start charging for MCP-based data access, we'll have to bite the bullet and pay, the ROI is that high.*
- **Customer 3:** *Everyone is saying they have an enterprise AI control plane for security and governance, but the reality is that they are only optimized for their own platforms. So, we decided to build our own AI control tower. It works, but admittedly it isn't great.*
- **Customer 4:** *We are one of the largest banks in the world so our choices may not be matched by others, but we are making a big shift to our own private on-premise data centers, including for our AI workloads. Even at the general-purpose chatbot level, we created our own custom sandboxed chatbot so that data never leaked to the frontier labs. We want control over our AI stack, so that we can better manage security controls, in addition to being highly regulated. Our non-AI workloads have mostly run on-premise anyway, so the shift here is that we're beginning to move more of our AI workloads on-premise as well.*

Diminished Threat to Apps Software

Post reported Jun/Jul results, we've seen some sharp moves higher in several apps/SaaS stocks (Atlassian, Salesforce, Veeva) as well as broader re-rating higher across the SaaS/apps space since late-June (although the rally appears to have leveled-off more recently). In our view the move in these stocks has not been driven by material upside to numbers, but rather by a narrative shift regarding the perceived threat from the AI model providers. In particular:

- **Reduced Threat from 1P AI Labs Products:** One lingering Street concern has been that the frontier labs would be assertive over time in launching 1P products beyond coding agents into other software segments. But YTD we really haven't seen this 1P threat play out. The OpenAI CEO recently said that enterprise adoption and the displacement of incumbents has taken longer than he thought and that OpenAI is not keen to launch 1P products into every product category. Rather, the CEO conveyed that he'd prefer that OpenAI focus on offering models and tooling to enterprise customers to build what they want, rather than launch 1P products that end up competing with customers.

- **Value of Apps:** The signals of a tighter relationship between the AI model providers and apps firms like Salesforce also highlights the value of the data, incumbency, and workflow knowledge of the application software providers. Put another way, sentiment began to shift to a view that the performance and hence ROI of AI models – at least for non-coding use cases and in an enterprise context - is materially enhanced via a tighter integration with the application software firms than is possible via a simple MCP connection. The more nuanced and tougher question is HOW enterprises are likely to access this valuable data over time, via a simple MCP connector or via fully-integrated products.

- **The Move to Headless:** While many apps firms such as Salesforce have been "headless" (accessed via a user interface or surface outside of the Salesforce systems) for some time, Salesforce's "AIforce" could be a harbinger of a bigger shift to accessing apps through a 3P model interface. This has resulted in a healthy debate as investors judge the downside of apps firms "losing" the UI layer (and thereby becoming "dumb back-end databases") or the upside as headless access is inevitable and very likely to result in greater customer satisfaction and therefore usage. Salesforce has cited a 6x increase in usage of its platform when accessed via AI models given the productivity gains (but can Salesforce monetize this usage ramp?).

The AI Infra Layer and the Revs/GW Ramp

In our CoreWeave launch report earlier this week (see link to report here), we articulated our rationale for our non-consensus Buy rating on the stock (sentiment towards highly-levered neo-clouds, even Oracle, has materially deteriorated of late) and discussed key AI infrastructure market debates about credit risk/leverage, AI data center capacity build-outs, ultimate ROIC and profitability and competition in the AI infra market as the likes of SpaceX and Meta step into this space. We won't repeat all that content here but did want to flag one widely-discussed AI trend – the ramp in monetization per GW of capacity due to inflation in per-hour GPU hosting prices, the mix shift to higher-ASP chips (Nvidia Blackwell, now Vera Rubin's) and the advent of short-duration (3-6 month) AI compute contracts.

Last fall, Oracle guided to a revenue outlook that was rooted in the assumption that it could generate $10 billion of revenue per GW of capacity. That figure served as the Street's anchor metric to driving GW-based revenue estimates for AWS, Microsoft Azure, Google Cloud, Oracle and the neo-clouds. Not anymore. Particularly since May 2026, when SpaceX announced that it signed AI infra deals with revs/GW levels of closer to $30-50 billion, the Street has been forced to re-think the potential revenue pull-through per GW of AI data center capacity. Now, almost every AI infra provider is signaling that Oracle's $10 billion figure is too low and moving higher based on strong compute demand, GPU price inflation, a mix shift to next-gen Nvidia GPUs, early progress in moving up the stack as well as the mix of leased versus self-built data centers. We illustrate these latest data points in Figure 1 below. Microsoft is not revealing any specifics about their per-GW monetization, only that a) they are not interested in allocating scarce capacity to short-dated contracts in an effort to boost Azure growth "for short-term gain" but b) will absolutely mark-to-market both new capacity and contract roll-offs to take advantage of the stronger pricing backdrop. Bottom line, this could be a strong catalyst for hyperscaler growth upside.

Figure 1: Range of Revenue per GW Data Points



Source. Company disclosures

Valuation Method and Risk Statement

We use various valuation techniques such as P/E, EV/FCF for valuing the companies in this report. Risk factors include but are not limited to macroeconomic factors such as a downturn in the economy, a disruption of international trade, technological disruption due to new inventions, or business model innovation whereby structural changes in the industry alter the future course of revenues.