Amazon's value in the Age of AI Agents

UNCOVERALPHA
MAR 05, 2026 · PAID

Hi everyone,

In this article, I'm breaking down my current thinking on Amazon. My goal here is to explain in detail the changes caused by AI on three pillars of the business: E-Commerce, AWS, and Advertising, and specifically how valuable each looks like in a world where AI agents increasingly sit between humans and the services they use. At the end, I'll do a sum-of-parts valuation that I think gives a useful anchor for where the stock sits today.

Let's start.

E-Commerce - the agentic threat and the logistics moat

Amazon captures roughly 40% of all U.S. e-commerce spending. It has 240+ million Prime subscribers globally (analyst estimates; Amazon last officially disclosed "over

200 million" in 2021), of which approximately 180–185 million are in the United States, representing penetration in about 80% of U.S. households. The Prime flywheel is well-documented: members spend on average $1,400/year, compared with $600 for non-Prime customers, and the retention rate after the first year is 99%, according to CIRP data. Amazon delivered over 8 billion items same or next day to U.S. Prime members in 2025, a 30%+ increase year-over-year.

This is the business everyone knows. But here's the question that matters for the next 3–5 years: what happens when AI agents start shopping for consumers?

The agentic shopping risk

I strongly believe that in the future, most e-commerce shopping will be done through AI agents acting as personal assistants to consumers, instead of direct consumers. I am not alone in those expectations. McKinsey projects agentic commerce could generate $1 trillion in U.S. retail revenue by 2030. Morgan Stanley expects nearly 50% of American shoppers will use AI agents by then, potentially adding $115 billion in e-commerce spending. Bain research shows that 30–45% of U.S. consumers already use GenAI for product research and comparison. During Cyber Week 2025, roughly 1 in 5 orders on Shopify involved an AI agent. AI-driven traffic to retailer sites has surged 7x since January 2025, according to Shopify data, with AI-driven orders up 11x.

What's happening is this: instead of opening the Amazon app, a consumer tells ChatGPT, Claude, or Gemini what they need. The agent searches across retailers, compares prices, checks reviews, and either completes the purchase or presents a shortlist. OpenAI has already embedded checkout directly into ChatGPT. Perplexity launched its Comet browser agent. Google is rolling out agentic AI shopping tools.

This is a major shift in consumer behaviour, and Amazon knows it. In November 2025, Amazon sued Perplexity for its AI browser agent making purchases on Amazon's marketplace. The company has blocked 47 AI bots from crawling its site. But at the same time, CEO Andy Jassy acknowledged on their recent earnings call that agentic commerce "has a chance to be really good for e-commerce." Amazon recently even posted a job for a principal corporate development officer specifically for "agentic commerce" partnerships.

Forrester retail analyst Sucharita Kodali captured the tension perfectly:

"With an agent on ChatGPT, retailers risk relinquishing transactions on their site to pay a toll on someone else's highway."

Amazon's shot at owning the application layer

That said, Amazon isn't conceding the front-end. They have several assets that give them a legitimate shot at being a surface where agentic shopping enters:

Rufus — Amazon's AI shopping assistant, used by more than 300 million customers in 2025. Customers using Rufus complete purchases at a 60% higher rate. It can now auto-purchase items when prices hit thresholds.

The most interesting recent project is the »Buy For Me« project. This is Amazon's experimental agent that can purchase from other retailers within the Amazon app. This is a smart flip from Amazon: instead of being the store that other agents shop, Amazon becomes the agent that shops everywhere else. Amazon does have some unique assets that make it valuable as the front-end touchpoint, and the key is around Prime Subscriptions.

Prime Video — 315 million ad-supported viewers globally, up from 200 million in early 2024. This is a massive surface for product discovery and agentic commerce integration, especially through interactive shoppable ads during live sports (Thursday Night Football averaged 15.3 million viewers, +16% YoY). Twitch — 105+ million monthly users, heavily Gen Z. An engaged, commerce-friendly audience. Alexa — still the most widely deployed voice assistant in smart home devices. If agentic commerce moves to a voice-first or ambient-first paradigm, Alexa has a head start.

The risk here is that those surfaces might not be enough and that Amazon might not be aggressive enough in the early days of where we are today. From today's vantage point, the dominant surfaces if I had to choose would still be the smartphone assistant, or a standalone AI app (similar to ChatGPT, Gemini, Claude), and later on the AI glasses and personal assistant given by that provider. Prime Video and Twitch will still serve as important discovery platforms and could turn out to be much more valuable in terms of ads in a world where it will become increasingly hard to reach a human via digital channels, as internet usage will be dominated by AI agents instead of humans. Still, it doesn't solve the fact that the application layer, where most of the e-commerce

starts, moves to other providers. Even if Amazon were to launch an independent AI shopping assistant app, I don't think in the long-term that would be »moaty« enough. My view is that the dominant provider will be the one that can offer a full AI personal assistant, with shopping as one of its features, not the only or main one. For that to be Amazon, they would need to make an aggressive pivot from current levels and a possibly strong shift into consumer hardware, which I don't think is their plan.

With all that said, my base case is that Amazon will not be the application layer of agentic shopping and that its e-commerce business will move to the backend part of the shopping experience (still being important). Even in this scenario, Amazon still makes a decent margin given the logistics, payment, and fulfillment infrastructure that it offers at scale.

Advertising

Amazon's advertising revenue hit $68.6B in 2025, growing 22% YoY in Q4. This is now 9.6% of Amazon's total revenue, up from 5.9% in 2021. To put it in context, Amazon's ad business alone is larger than the total revenue of companies like Netflix, Uber, or Salesforce.

But here's the nuance that most analysts don't discuss: Amazon's ad business is really two very different businesses glued together.

Search ads

The vast majority of Amazon's advertising revenue comes from Sponsored Products: essentially search ads within Amazon's marketplace. When you search for "wireless headphones" on Amazon, the first several results are paid placements. Amazon doesn't break this out precisely, but based on WARC data, the retail media component (primarily search ads) accounts for roughly $60.6B of the estimated total, with Prime Video and other upper-funnel formats making up the incremental portion.

Here is my concern: search ads on Amazon are fundamentally tied to humans browsing Amazon's website and app. If an AI agent shops for you, it doesn't look at sponsored listings. It doesn't scroll past display ads. It skips right to the product that best matches your criteria and places the order. As Bain research noted, about 65% of retail media spending still occurs onsite, and that entire bucket is at risk if product discovery shifts to AI-driven search.

This is why I think the search ad portion of Amazon's advertising business is on a disruption clock. Not tomorrow, not next quarter, but over a 3–5 year horizon, the economics of Sponsored Products face a structural headwind as agentic interfaces capture more of the purchase journey and as we talked in the previous section I give it a low probabilitiy chance that Amazon is able to capture the AI agent assistant application layer so the eyeballs switch from amazon's site and apps towards the AI assistant owners.

Prime Video ads

The other side of Amazon's ad business is Prime Video advertising, and this is the piece I think is defensible. Amazon introduced ads on Prime Video in January 2024. S&P Global Market Intelligence Kagan estimated Prime Video's ad revenue at $433M in 2024 and forecast it to reach $806M in 2025. This is still a small fraction of total ad revenue, but it's growing fast and serves a different function: brand advertising through streaming video is not susceptible to agentic disintermediation the same way search ads are.

Prime Video reaches 315 million monthly ad-supported viewers globally. That's larger than Netflix's ad-supported tier at 190 million. Thursday Night Football alone averaged 15.3 million viewers with 16% growth YoY, and the Packers-Bears wild-card playoff game drew 31.6 million viewers, the most-streamed NFL game in history. Amazon has also integrated Netflix and Spotify inventory into its Amazon DSP, giving advertisers a broader programmatic buying platform.

My estimate is that by 2027–2028, Prime Video ads could reasonably be a $3–5B annual revenue stream, growing at 40%+ rates as ad loads increase and live sports inventory expands (NBA deal kicks in, international sports expansion). This business is much more structurally defensible because people watch content — AI agents don't.

But even that revenue doesn't materially change my thesis that the majority of Amazon's ad business is at risk of serious disruption.

For the sum-of-parts analysis in the last part of this article, I'm splitting the ad business into two buckets. For the search/retail media portion (~$60–63B), I'm assigning it a terminal value as if profits only last 4 more years with zero terminal value after that. That's deliberately punitive - I'm assuming this revenue stream is

structurally impaired. For Prime Video ads, I'll fold it into the e-commerce/subscription ecosystem, where it has long-term durability.

AWS - the cloud business

AWS is the most important reason why I own Amazon stock and why it has now become my biggest portfolio position.

The biggest fear around AWS has been that AI-related capital expenditures would permanently compress margins. And yes, there was a dip: AWS's operating margin fell to 32.9% in Q2 2025 as the company ramped up spending aggressively. But by Q4, it had recovered to 35.0%, and the full-year margin was 35.4%.

Here is my core argument: we are severely compute-constrained for the foreseeable future. Amazon has invested $131.8B in capex for 2025 and has guided to approximately $200B for 2026, predominantly for AWS infrastructure. The company added more than 1 gigawatt of data center capacity in Q4 alone and 3.9 gigawatts in the trailing 12 months, which is double what AWS had in total in 2022. And Andy Jassy expects to double power capacity again by the end of 2027.

Despite this massive buildout, demand continues to outstrip supply. Jassy noted on the Q1 call that GPU and motherboard shortages were limiting the pace of AI workload onboarding. Bedrock (Amazon's managed AI service) reached a multi-billion-dollar annualized run rate with customer spend growing 60% quarter-over-quarter to a base of over 100,000 customers. Trainium2 is fully subscribed with 1.4 million chips deployed.

In this environment, there is no incentive for hyperscalers to engage in a pricing war. When every chip you install is immediately monetized, you don't cut prices — you add capacity. Until compute supply catches up with demand (which I don't expect before 2029 at the earliest), AWS can maintain mid-30%+ operating margins without sacrificing growth. The margin should hold around pre-AI era levels (AWS operated in the 28–35% range historically, with 2024 averaging 37%) because the scarcity dynamic supports pricing power.

Trainium and Custom Silicon are key things for long-term margins

This is a point I don't think gets enough attention. NVIDIA's gross margin sits at roughly 73–75%. Every cloud provider that is 100% dependent on NVIDIA for AI

compute is paying that tax on every GPU. That cost flows through to the cloud provider's cost of revenue and structurally limits the margin they can earn on AI workloads.

Amazon, through its Annapurna Labs subsidiary, has developed Trainium and Inferentia custom ASICs, as well as Graviton CPUs for general compute. Combined, these custom chips have surpassed a $10B annualized revenue run rate, growing at triple-digit percentages YoY. According to Amazon, Graviton provides 40% better price-performance than x86 processors and is adopted by 90% of AWS's top 1,000 customers.

Trainium2 powers Project Rainier, the world's largest operational AI compute cluster with 500,000+ Trainium2 chips, which Anthropic uses to train its Claude models. Trainium3 is in preview with broader volumes expected in early 2026, and Trainium4 is targeted for 2027.

I am sharing here the chart that we made some months ago in our detailed Amazon Trainium piece, where we calculated the manufacturing costs of Amazon Trainium, Google TPUs, and Nvidia's Blackwell B200:

Metric	AWS Trainium3	Google TPU v7 (Ironwood)	NVIDIA B200 (Blackwell)
Manufacturing Cost (Est. BOM + industry experts)	**~$3,000 – $3,500**	**~$4,500 – $5,200**	**~$6,400 – $7,000 (selling at ~$35,000-$40,000)**
HBM Memory	144GB HBM3e	192GB HBM3e	192GB HBM3e

You can see the most significant difference: if it costs Amazon $ 3,000-$ 3,500 to produce a Trainium3 chip, it costs them $35k-$40k to buy an Nvidia B200 chip. Even though B200 is much more performant from a cost-of-ownership perspective, Trainium3 gives B200 a run for its money.

The margin math is straightforward. When you design and manufacture your own silicon (using TSMC for fabrication and a design partner like Broadcom, Marvell, MediaTek), your cost per unit of compute is significantly lower than buying merchant silicon from NVIDIA at a 73–75% gross margin. This gives AWS a structural margin advantage for AI workloads vs. a competitor that sources 100% from NVIDIA. It doesn't mean AWS abandons NVIDIA (it still offers NVIDIA instances), but having an

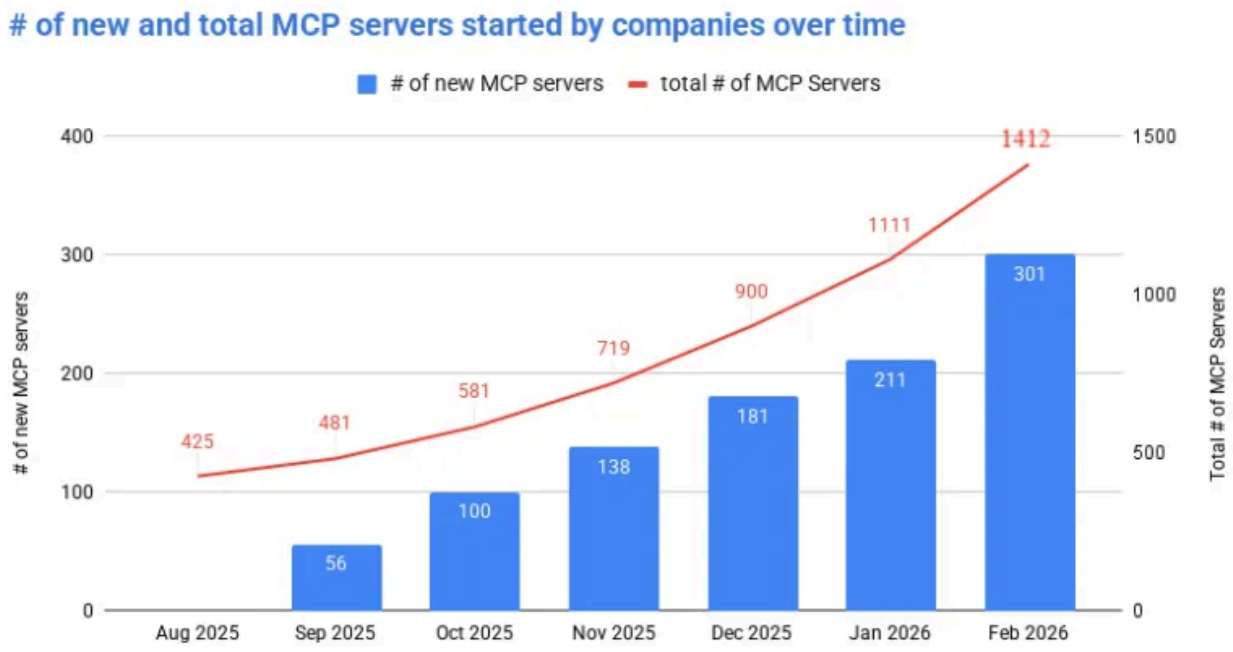
alternative lets AWS capture more of the AI value chain and maintain margin in ways that someone who is entirely dependent on NVIDIA simply cannot.

This key difference will prove even more important in the coming years, especially once demand/supply for compute is more in balance and the hyperscalers' focus shifts from capturing revenue growth to profitability and customer optimization.

Traditional cloud demand is actually accelerating because of AI agents

There's a narrative that AI is all that matters for AWS growth. That misses something important: AI agents themselves create enormous demand for traditional cloud services, as we already discussed in part in our The Forgotten Chip: CPU the New Bottleneck of the Agentic AI era article. Every AI agent needs storage (S3), compute (EC2, powered increasingly by Graviton), databases, networking, and monitoring. The more AI agents there are in production, the more traditional cloud infrastructure gets consumed.

The number of AI agents and their deployment is rapidly surging right now. Here is an alt provider that tracks the Model Context Protocol (MCP), an open-source standard for connecting AI applications to external systems.



source: Bloomberry

The number of MCP servers being set up every month is growing exponentially, and the MoM pace is accelerating. The market is still in very early stages, as the current

number of MCP servers is probably less than 1% of the API market. But the interesting thing was comparing where these MCP servers were being deployed with where API deployments are. Both Azure and GCP % of these MCP server deployments were lower compared to their API deployments, while AWS MCP deployments actually rose compared to API deployments:

Provider	api.*	mcp.*
AWS	53%	60%
Google Cloud	14%	12%
Azure	19%	7%
Cloudflare Workers	< 1%	4.6%*
Vercel	2%	5%
DigitalOcean	6%	4%
Railway	< 1%	2.4%
Render	< 1%	1.8%
Heroku	0.9%	0.6%

source: Bloomberry

While this data is not large enough yet, it could indicate that more smaller companies are on AWS than on the other two hyperscalers and that they are early adopters. The data in some form also shows the importance of AWS's »legacy« cloud infrastructure, which is very much needed in the Agentic AI phase.

The demand for traditional AI infrastructure is skyrocketing, and you can see it from comments from CEOs of AMD and Intel, where they are basically sold out of CPUs, and you can see it when talking to other industry experts.

This is a former Amazon employee talking about the usage of the AWS S3 service (storage):

»Right now, S3, nobody is thinking about how S3 is exploding. It's quite an explosion because what are AI systems doing? They're generating embeddings, they're storing the prompts and the responses. Where are they storing this? S3. They're logging every interaction for auditing, tuning, safety. All of this is going into S3. Remember when we used to think of S3 as just their cheap storage? The storage is still cheap, but you're using more of it«.

source: AlphaSense

The key takeaway from this comment is the need to store this data for audit and safety purposes. Companies running these AI agents need clear, auditable trails of what an AI agent has done, so they can track and monitor as problems arise and fix them. Nobody wants to give an AI agent full permission to run freely across the company's stack and make changes and tasks that nobody has visibility into. This human visibility means services like Storage grow even more in usage.

A big tell was also the November OpenAI- AWS deal. The press release stated that OpenAI would access "hundreds of thousands of state-of-the-art NVIDIA GPUs, **with the ability to expand to tens of millions of CPUs** to rapidly scale agentic workloads."

The GPU part is known, but the CPU part is the most interesting one. We need »legacy« cloud workloads and CPUs to enable the AI agent economy; that is just the way it is, and this is a big uplift for AWS, which has the largest fleet of optimized cloud services out there.

Amazon noted that more of the top 500 U.S. startups use AWS as their primary cloud provider than the next two providers combined. That startup and scale-up cohort is building AI-native applications that are heavily cloud-intensive.

The on-prem fallacy and SMB lock-in

Some investors argue that AI inference will eventually move to the edge or on-prem, killing the cloud growth story. Let me push back on this.

First, even if 90% of personal AI assistant use cases eventually run on edge devices (phones, laptops, local hardware), the remaining 10% that stays on cloud or on-prem infrastructure is still an enormous market. These are the "god-like AI" use cases:

complex enterprise reasoning, multi-step agentic workflows, financial modeling, drug discovery, code generation at scale. These require the kind of compute density and model size that doesn't fit on a phone. And these use-cases are the most profitable, as their outputs are the most valuable.

Second, on-prem AI infrastructure is radically more complex than anything businesses have managed before. Running an AI inference cluster on-prem means managing GPU and CPU servers, networking fabric, cooling systems, model deployment pipelines, and monitoring at a level of sophistication that most IT departments have never dealt with. For any small or medium-sized business, the cost and complexity of running your own AI infrastructure to have your "AI accountant" or "AI customer service agent" simply doesn't make sense when you can rent it from AWS for a fraction of the upfront cost with zero operational hassle.

The cloud is the natural home for AI workloads for the vast majority of companies, and that reality isn't changing anytime soon. If anything, as AI becomes more central to knowledge work, more companies will move to the cloud specifically to access AI capabilities they can't build or run themselves.

The revenue trajectory of AWS

With all that in mind, I believe AWS, with its power capacity availability, which I already discussed in my previous articles, is well-positioned for multiple quarters of accelerating growth. I believe, despite AWS's size, we will soon see the segment grow by +30% YoY. AWS also exited Q4 2025 with a backlog of $244B with a weighted average remaining life of 4.1 years. Capacity is being installed and monetized as fast as it comes online.

If AI agents truly absorb a meaningful portion of knowledge work over the next 5–10 years — and companies like Anthropic ($19B ARR rate up from $9B just two months ago) and OpenAI are building the models to do exactly that — then the total demand for cloud inference is going to be multiples of what it is today. Every AI-powered accountant, lawyer, engineer, customer service agent, and analyst running in the cloud creates recurring compute demand.

The Anthropic and OpenAI stakes are hedges

Besides the already mentioned segments, Amazon also has other important aspects such as Project Kupier, Subscription business, and stakes in Anthropic and OpenAI,

which are now becoming increasingly important.

Amazon has invested approximately $8B in Anthropic (capped below 33% ownership) and recently announced a strategic partnership with OpenAI that includes an investment of up to $50B (starting with an initial $15B commitment, with the remainder tied to milestones and a potential OpenAI IPO).

Anthropic just closed a $30B funding round at a $380B post-money valuation in February 2026. If Amazon holds roughly 20% of Anthropic (estimates vary given the cap structure), that stake is worth $76B on paper. But in the last few weeks, Anthropic has accelerated its adoption and revenue growth so much that a $500B valuation for a company that will probably exit 2026 at a $50B ARR growing 5x YoY and disrupting the whole knowledge work economy is nothing extraordinary, which would add $100B of value or almost 5% of Amazon's current market cap.

For OpenAI, the proposed $100B funding round would value the company at approximately $830B. Amazon's $50B investment at those terms would represent roughly a 6% stake.

Combined, these stakes could be worth +$145B. And here's the real value: in a world where Anthropic, OpenAI, and Gemini become the application layer, having significant stakes in two of those companies isn't just financial investments. They are Amazon's guarantee that the biggest AI consumers remain AWS customers. OpenAI has committed to spending $100B on AWS over the next eight years. Anthropic is using Project Rainier (500,000+ Trainium2 chips) for training. Both are locked in as massive cloud customers.

Valuation

Now let's put the numbers together. I'm deliberately being conservative in places and factoring in serious disruption risk. Here are my numbers: