U.S. Internet

A Primer on AI Lab & AI Hyperscaler Unit Economics

We expect some confusion as AI labs start reporting GAAP financials, as business mix and revenue recognition can vary. Similar to how UBER and LYFT are in the same core business yet gross bookings and net revenue have different definitions, comparing AI labs may prove difficult.

U.S. Internet
POSITIVE

U.S. Internet
Ross Sandler
+1 415 263 4470
ross.sandler@barclays.com
BCI, US

Alex Hughes
+1 212 526 3069
alexander.hughes@barclays.com
BCI, US

Michael DiSanto
+1 212 526 1054
michael.disanto@barclays.com
BCI, US

Alexander Kessinger
+1 212 526 1324
alexander.kessinger@barclays.com
BCI, US

The Key Take-Away: The unit economics of AI labs are similar across like-for-like product lines, but business mix, revenue recognition, and partnerships can lead to fairly significant differences in how financials are presented. This report attempts to break down these nuances and how revenue and costs ultimately flow through the P&L at the AI lab. We also look at how the unit economics of AI labs flow into the unit economics of the supporting hyperscalers.

Business Mix Differences Explain a Lot

AI labs have revenue lines across a number of different products and services. API was the original AI lab business model, but in recent years applications and subscription products have entered the mix. Shown below in Figure 1, hypothetical frontier lab "A" has upwards of 70% of revenue coming from API and 30% from subscription. Hypothetical frontier lab "B" has the inverse, closer to 80% of revenue from subscription. As we will illustrate further down in this report, API tends to have higher inference margin than subscription, which in addition to training costs being completely different, explains most of the difference in gross margins (after training) at the aggregate AI lab level. It's possible that some of the training costs may be recorded in R&D for various AI labs, so for the purposes of this report, we are showing ~10% of training costs in COGS, which represents the final training run for an AI model. It's unclear at this point how training costs are likely to be allocated across P&L lines by AI labs.

As we will discuss below, other major factors explaining revenue and margin differences are: 1) partnership dynamics (with hyperscalers), and 2) the indirect API gross revenue recognition – discussed in more detail below.

Barclays Capital Inc. and/or one of its affiliates does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision.

Please see analyst certifications and important disclosures beginning on page 9.
Completed: 27-Aug-26, 22:18 GMT Released: 28-Aug-26, 04:10 GMT Restricted - External

FIGURE 1. **Business Mix Explains Most of the Differences in AI Lab Unit Economics in '26**

AI Lab Economics	Lab A	Lab B
Customer Spend	**$100**	**$102**
Revenue	**100.0**	**100.0**
Subscription	30.0	80.0
Direct API	45.0	10.0
Indirect API	25.0	10.0
Costs	**35.0**	**52.0**
Indirect Cloud Fees	5.0	2.0
Strategic Partner Revenue Share	0.0	20.0
Inference Compute Cost	30.0	30.0
AI Lab Paid Inference Profit	**65.0**	**48.0**
% Margin	*65%*	*48%*
Model Production Training Cost	10.0	10.0
AI Lab Gross Profit (After Training)	**55.0**	**38.0**
% Margin	*55%*	*38%*
R&D Training Cost	70.0	96.0



2026 Mix Estimates; Inference Costs Exclude Free Users
Source: Barclays Research Estimates

AI lab unit economics have improved greatly in '26 (vs. '25) driven by agentic workflows and enterprise becoming the "must buy" products in the market today. API and enterprise accounts likely carry higher inference margins; hence, as mix shifts in that direction, and as AI models are more token efficient per task, inference margins have surged higher. Based on our estimates shown in Figure 2, we think frontier labs "A" and "B" may have increased paid inference margins to upwards of 50%-65%+ in '26 (up from mid-teens % in '25), and possibly much higher on select products and services. After factoring in final training costs (10% of overall training costs), labs have seen the adjusted gross margin increase ~30-50 points vs. '25. AI lab "B" has strategic partner revenue sharing which impacts profitability until certain thresholds are achieved, likely eliminated to zero sometime after '28.

FIGURE 2. **The Shift to Enterprise Has Markedly Improved AI Lab Unit Economics in '26**

	2025			2026	
AI Lab Economics:	Lab A	Lab B		Lab A	Lab B
Revenue	**$100**	**$100**		**$100**	**$100**
Subscription	15.0	90.0		30.0	80.0
Direct API	50.0	5.0		45.0	10.0
Indirect API	35.0	5.0		25.0	10.0
Costs	**86.1**	**82.1**		**35.0**	**52.0**
Indirect Cloud Fees	7.0	1.0		5.0	2.0
Strategic Partner Revenue Share	0.0	20.0		0.0	20.0
Inference Compute Cost	79.1	61.1		30.0	30.0
% of Revenue	79%	61%		30%	30%
AI Lab Paid Inference Profit	**13.9**	**17.9**	Model Efficiencies, Pricing Power, and Shift Towards API/Enterprise Driving Higher Margins	**65.0**	**48.0**
% Margin	*14%*	*18%*		*65%*	*48%*
Model Production Training Cost	10.0	10.0		10.0	10.0
AI Lab Gross Profit (After Training)	**3.9**	**7.9**		**55.0**	**38.0**
% Margin	*4%*	*8%*		*55%*	*38%*
R&D Training Cost	81.0	53.0		70.0	96.0

Source: Barclays Research Estimates

Based on the mix noted above and some of the revenue sharing agreements, training costs, and inference costs, we conclude that around $35-$40 of every $100 in AI Lab revenue flows to hyperscaler revenue in '26, on which the hyperscalers generate nearly $10-$20 in OI (at high ~35%-45% operating margin). Partner revenue share results in higher hyperscaler OI margin, but the actual profit on a per token basis is likely the same excluding these fees. We think there is a decent possibility that AI lab margins and hyperscalers' OI margins are higher than what we illustrate here currently in '26, but we would assume these margins come down a bit over time as competition increases at the frontier and as compute scarcity eases.

FIGURE 3. **For Every $100 of AI Lab Revenue, $35-$40 Flows to the Hyperscalers in the Form of Inference Revenue, or $10-$20 in Hyperscaler OI (in '26)**

Per $100 of AI Lab Revenue	Subscription	Direct API	Indirect API	Total
AI Lab A				
AI Lab Revenue Mix	30%	45%	25%	**100%**
Hyperscaler Revenue	9.0	13.5	12.5	**35.0**
Less: Infrastructure Costs	6.3	9.5	7.5	**23.3**
Hyperscaler Profit	2.7	4.1	5.0	**11.8**
% Margin	*30%*	*30%*	*40%*	***34%***
AI Lab B				
AI Lab Revenue Mix	80%	10%	10%	**100%**
Strategic Partners	40%	5%	0%	
Other Partners	40%	5%	10%	
Hyperscaler Revenue	32.0	4.0	5.0	**41.0**
Less: Infrastructure Costs	16.8	2.1	3.0	**21.9**
Hyperscaler Profit	15.2	1.9	2.0	**19.1**
% Margin	*48%*	*48%*	*40%*	***47%***

Source: Barclays Research Estimates

AI Products Have a Wide Range of Unit Economics, and Are Often Recorded Differently

Noted above, the product mix at frontier labs can be dramatically different. Figures 4-6 below illustrate that even within the same products (like for like) there can be nuances that drive different unit economics including partnerships (revenue shares) and how actual revenues are recorded (gross vs. net). Like UBER/LYFT, despite being the same core business, the revenue and margin can be completely different.

Subscription products (like Claude Code or Codex) generally have lower inference margins than API as some users benefit from AI labs' willingness to subsidize token cost to remain competitive and retain users. Subscription products are both monthly-fee based and usage based, often with caps which can be reset from time to time (we have seen an increase in cap resets of late, likely in response to churn reduction efforts and model efficiency gains). Subscription products tend to default to a specific model, but the user can toggle between models and latency modes for various jobs, which can burn through token credits faster or slower depending on model size. Enterprise subscriptions are often dollar-based by license (often at higher limits than Plus and Pro plans), with daily caps that can be allocated across the licenses. We estimate inference margins on subscriptions are in the 70% range. Noted above, we would guess that the '26 inference margin may be higher than this level, but over time the industry may drift back down towards this level to the degree that competitive alternatives show up and if/when compute moves into surplus (vs. today's shortages).

Hyperscalers generate solid unit economics on agentic subscription products as the compute being provisioned for these stateful runtime products often involves some "up the stack" software products like databases, etc., and in some cases have revenue sharing between the AI lab and the hyperscalers.

FIGURE 4. **Subscription Apps Tend to Have Slightly Lower Inference Margins (vs. API)**

Subscription Apps	AI Lab A	AI Lab B	
	All Partners	Strategic Partner	Other Partners
Customer Spend	**$100**	**$100**	**$100**
AI Lab Economics:			
AI Lab Revenue	100.0	100.0	100.0
% of Customer Spend	*100%*	*100%*	*100%*
Strategic Partner Revenue Share	0.0	20.0	20.0
% of Revenue	*0%*	*20%*	*20%*
Inference Compute Cost	30.0	30.0	30.0
% of Customer Spend	*30%*	*30%*	*30%*
AI Lab Paid Inference Profit	70.0	50.0	50.0
% Margin	*70%*	*50%*	*50%*
Hyperscaler Economics:			
Hyperscaler Revenue	30.0	50.0	30.0
% of Customer Spend	*30%*	*50%*	*30%*
Revenue Share	--	*20.0*	--
Inference Compute Revenue	*30.0*	*30.0*	*30.0*
Hyperscaler Infrastructure Costs	21.0	21.0	21.0
% of Compute Revenue	*70%*	*70%*	*70%*
Hyperscaler Profit	9.0	29.0	9.0
% Margin	*30%*	*58%*	*30%*

Lab B 20% Strategic Partner Revenue Share, Up To A Cumulative Cap

Assuming AI Lab Subscriptions ~70% Margin*

IaaS Estimated To Run At ~30%-ish OI Margin

* Paid User Subscription Margins
Source: Barclays Research Estimates

Direct API is the original business model for most frontier labs (before even products like ChatGPT were introduced) and carries very attractive unit economics for both the AI lab and the hyperscaler. Direct API is where a developer like Cursor or Figma embeds features inside their products that call on the AI lab's API to generate tokens. The business model for the AI lab is usage based on token consumption (i.e., pay-as-you-go). Inference margins for API have been trending up as the models have improved token efficiency (i.e., the volume of tokens to complete a task) while headline API token prices have increased. Inference providers have similarly improved the efficiency of serving models at the infrastructure layer (via quantization, speculators, and next-gen compute), supporting higher hyperscaler margins. Lastly, revenue sharing between the AI lab and the hyperscaler further boosts the latter's margins (at the expense of the former).

We have heard that AI lab inference margins on API are running well north of the levels noted in Fig 5 in 2Q26, but we would guess this may normalize lower at some point (but it's not entirely clear when).

FIGURE 5. **Direct API Has Very Solid Unit Economics**

Direct API	AI Lab A	AI Lab B	
	All Partners	Strategic Partner	Other Partners
Customer Spend	**$100**	**$100**	**$100**
AI Lab Economics:			
AI Lab Revenue	100.0	100.0	100.0
% of Customer Spend	*100%*	*100%*	*100%*
Strategic Partner Revenue Share	0.0	20.0	20.0
% of Revenue	*0%*	*20%*	*20%*
Inference Compute Cost	30.0	30.0	30.0
% of Customer Spend	*30%*	*30%*	*30%*
AI Lab Paid Inference Profit	70.0	50.0	50.0
% Margin	*70%*	*50%*	*50%*
Hyperscaler Economics:			
Hyperscaler Revenue	30.0	50.0	30.0
% of Customer Spend	*30%*	*50%*	*30%*
Revenue Share	--	*20.0*	--
Inference Compute Revenue	*30.0*	*30.0*	*30.0*
Hyperscaler Infrastructure Costs	21.0	21.0	21.0
% of Compute Revenue	*70%*	*70%*	*70%*
Hyperscaler Profit	9.0	29.0	9.0
% Margin	*30%*	*58%*	*30%*

Current Inference Margins Estimated North of 80%, But We Assume Comes Down Over Time

Source: Barclays Research Estimates

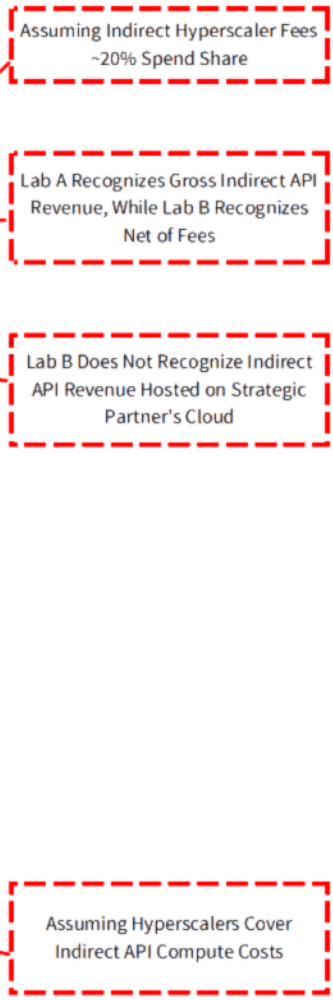
The Indirect API business model is more or less the same as Direct in terms of what the end user sees, but billing and the go-to-market relationship is between the user and the hyperscaler (without direct engagement with the AI lab). AI lab "A" records indirect API revenue on a gross basis, and it is becoming a larger part of the story. AI lab "B" doesn't have significant revenue here and records revenue on a net basis, and may exclude revenue recognition entirely when its strategic partner runs indirect API.

So similar to UBER and LYFT, as indirect grows as a percent of the overall AI lab story, it likely creates distortion between reported revenues, and hence can distort the narrative around relative AI lab progress.

人在草木间

FIGURE 6. **Indirect API Has Solid Unit Economics, But Different Revenue Recognition**

Indirect API	AI Lab A	AI Lab B	
	All Partners	Strategic Partner	Other Partners
Customer Spend	**$100**	**$100**	**$100**
Indirect Hyperscaler Fees	20.0	--	20.0
% of Customer Spend	*20%*	--	*20%*
AI Lab Economics:			
AI Lab Revenue	100.0	0.0	80.0
% of Customer Spend	*100%*	*0%*	*80%*
Strategic Partner Revenue Share	0.0	--	16.0
% of Revenue	*0%*	--	*20%*
Inference Compute Cost	30.0	--	30.0
% of Customer Spend	*30%*	--	*30%*
AI Lab Paid Inference Profit	50.0	--	34.0
% Margin	*50%*	--	*43%*
Hyperscaler Economics:			
Hyperscaler Revenue	50.0	100.0	50.0
% of Customer Spend	*50%*	*100%*	*50%*
Indirect Hyperscaler Fees	*20.0*	--	*20.0*
Inference Compute Revenue	*30.0*	--	*30.0*
Hyperscaler Infrastructure Costs	30.0	30.0	30.0
% of Inference Compute Revenue	*100%*	--	*100%*
Hyperscaler Profit	20.0	70.0	20.0
% Margin	*40%*	*70%*	*40%*



Source: Barclays Research Estimates

Putting It All Together

Right now, just about every dollar of AI lab ARR is finding its way to the hyperscaler revenue line, because of the high mix of training relative to inference in '26. Over time, inference profits should eventually surpass training costs, and drive higher profit margins for the AI labs. Hyperscalers likely start to lose market share of AI lab training and inference starting in '28 when backstopped AI infrastructure projects come online and become the first option for AI labs. We generally see Amazon Web Services (AWS), Azure (MSFT, covered by Raimo Lenschow), and Google Cloud Platform (GCP) controlling a similar mix of AI lab compute spend in the next two years, before this shift to backstopped infrastructure kicks in.

FIGURE 7. **Right Now Every $1 of AI Lab Revenue Is Nearly $1 of Hyperscaler Revenue, But the Latter Should Decline Going Forward**

AI Industry Summary ($B)	2024	2025	2026E	2027E	2028E
AI Lab Revenue	7	26	137	376	690
% Y/Y Change		*276%*	*418%*	*175%*	*84%*
Est. Year End AI ARR		*44*	*200*	*453*	*782*
Inference Cost	4	17	58	157	292
% of Revenue	*57%*	*66%*	*42%*	*42%*	*42%*
% Y/Y Change		*334%*	*233%*	*173%*	*85%*
Training Cost	7	19	66	132	210
% of Revenue	*96%*	*70%*	*48%*	*35%*	*30%*
% Y/Y Change		*176%*	*255%*	*100%*	*59%*
Hyperscaler AI Revenue	**11**	**36**	**124**	**289**	**502**
% of AI Lab Revenue	*153%*	*136%*	*90%*	*77%*	*73%*
% Y/Y Change		*235%*	*244%*	*134%*	*73%*
AWS AI Revenue - Est.	3	10	37	87	151
% of Hyperscaler AI Revenue	*31%*	*27%*	*30%*	*30%*	*30%*
GCP AI Revenue - Est.	3	10	37	87	151
% of Hyperscaler AI Revenue	*24%*	*28%*	*30%*	*30%*	*30%*
Other Cloud AI Revenue	5	16	49	116	201
% of Hyperscaler AI Revenue	*45%*	*45%*	*40%*	*40%*	*40%*

Source: Barclays Research Estimates