

01 Oct 2026 05:00:00 ET | 36 pages

Artificial Intelligence

ROI on AI: The Future of Open Is More Control than Cost as Frontier Labs Pull Ahead

CITI'S TAKE

We believe we have passed the local peak of open-weight models' impact on frontier model revenues, as the capability gap rewidens and usage economics increasingly favor frontier labs' model suites. While open-weight models made substantial progress toward the proprietary frontier, much of the gap narrowing occurred when regulatory concerns restricted advanced model releases and open-weight developers benefited from distillation and access to advanced offshore compute. As Poolside Co-CEO Jason Warner noted (Figure 8), the long-term value of open models is [the control they enable](#) via customization, deployment flexibility, and IP protection. Increasing [compute availability](#) should allow frontier labs to distill models more efficiently than open-weight competitors, offering enterprises a broader range of price-performant options. As labs expand beyond APIs into agents, orchestration, governance, routing, and security, we expect their enterprise advantage to widen, even as enterprises use open-weight models and proprietary data to build specialized models for internal use cases.

True convergence remains in question. The capability gap is proving transient as frontier AI companies leapfrog one another and open-weight models periodically catch up. The latest weekly data shows growing frontier separation as the [RSI inflection nears](#). The highest AA Intelligence Index score for proprietary models rose to 58 from 53, while the highest score for open models increased to 46 from 44, widening the closed-open gap to 12 points from 9 points (Figure 2). Public benchmarks still likely overstate parity even as private evaluations place leading open systems further behind in recent weeks, particularly on long-horizon cyber tasks and production reliability.

The cost frontier is advancing. The cheapest token is not always the cheapest sufficient answer, and the latest frontier model may also lower the cost of obtaining a previous frontier-level outcome, with the help of strong, pre-training advances and major systems optimizations. Opus 5.5 lifted the AA intelligence frontier 9% and [costs 40% less](#) than Opus 5 on typical workloads (Figure 14; Anthropic, 9/22). Meanwhile, successive Grok and Gemini releases have improved intelligence while maintaining flat token pricing, underscoring continued gains in model efficiency. [Regulations and other frictions](#) on frontier model releases may encourage labs to focus on [harvesting efficiencies](#) from existing models in the near term.

Value is moving across the stack. Differences in token efficiency and accuracy suggest that workload-specific routing is likely more economical than uniform open or closed model deployment strategies. We expect enterprises to adopt mixed-model portfolios as a de-risking mechanism (Figure 18), while AI labs may increasingly embed routing and orchestration capabilities into the model layer directly such that users only elect for specific models in rare or highly specific circumstances (Figure 19).

Heath Terry^{AC}

+1-212-723-4624

heath.terry@citi.com
Shelby Spencer

+1-212-816-0416

shelby.spencer@citi.com
Ashley Kim

+1-212-816-6689

ashley.kim@citi.com
Janna Withrow

+1-212-723-0439

janna.withrow@citi.com

See Appendix A-1 for Analyst Certification, Important Disclosures and Research Analyst Affiliations.

The State of Open vs. Closed Weight Models

Performance on public benchmarks indicates that leading open-weight models narrowed the performance gap with leading released closed-weight models substantially in mid-2026 before a rewidening in September. As of September 2026, MiMo-V2.6-Pro leads open-weight models with an Artificial Analysis Intelligence Index score of 46, followed by GLM-5.3 at 45 and Kimi K3 at 44 (Figure 1). Following the government's response to the initial Fable 5 and Mythos releases, Chinese labs drove the open frontier from 28 in April to 45 in August, narrowing the proprietary lead from 13 to 6 points even as the closed frontier rose from 41 to 51 points over the same period. Momentum shifted in September. MiMo raised the open frontier to 46, but Claude Opus 5.5 lifted the proprietary frontier to 58, widening the closed-open gap to 12 points. Meanwhile, Inkling Small, the strongest non-Chinese open-weight model, has a score of 26.

While Hugging Face represents only one channel through which open-weight models are distributed and deployed, activity on the platform suggests that adoption is highly concentrated as 1.5% of Hugging Face's repositories account for 99.2% of downloads, while 85.6% of models have received fewer than 200 lifetime downloads ([Hugging Face, 8/14](#)). Hugging Face reported 3M public model repositories and 1M public datasets as of August 2026, reinforcing that the ecosystem is fragmenting across specialized systems rather than converging on one model. Notably, Google's open models generated the highest 2026 download volume on Hugging Face among U.S. labs at [418M downloads](#) (Hugging Face, 8/14), suggesting enterprise demand for open models extends beyond the recent leadership of Chinese developers and includes organizations prioritizing Apache 2.0 licensing, ecosystem support, and deployment flexibility.

Figure 1. Selected Open-Weight Model Landscape

Open-Weight AI Models: A Snapshot								
Model Provider	Region	Flagship Open Model	Release Date	License	Listed Token Price (Blended, \$/1M)	AA Intelligence Index	AA Cost per Task	Selected Enterprise Adopters of Open Models
Xiaomi	China	MiMo-V2.6-Pro	9/22/2026	MIT	\$0.17	46	\$0.13	Limited international evidence
Z AI	China	GLM-5.3	8/14/2026	GLM-5.3	\$0.90	45	\$2.01	Coinbase, Hugging Face
Moonshot AI	China	Kimi K3	7/16/2026	Kimi K3	\$2.31	44	\$2.00	Coinbase, DoorDash
Alibaba	China	Qwen3.8 2.4T A95B	8/12/2026	Qwen3.8-Max	\$1.18	40	\$2.16	Airbnb
DeepSeek	China	DeepSeek V4.1 Flash	9/10/2026	MIT	\$0.19	39	\$0.27	Saudi Aramco
Minimax	China	MiniMax-M3	6/1/2026	Minimax Community	\$0.22	29	\$0.51	Limited international evidence
Tencent	China	Hy3	7/6/2026	Apache 2.0	\$0.11	25	\$0.07	Limited international evidence
InclusionAI	China	Ling 3.0-Flash	8/4/2026	MIT	\$0.04	20	\$0.03	Limited international evidence
Meituan	China	LongCat-2.0	6/30/2026	MIT	\$0.19	19	\$0.06	Limited international evidence
Thinking Machines	United States	Inkling Small	7/30/2026	Apache 2.0	\$0.22	26	\$0.09	Limited enterprise evidence
NVIDIA	United States	Nemotron 3 Ultra	6/4/2026	OpenMDW	\$0.48	23	\$0.60	Johnson and Johnson, ServiceNow, SLB
Meta	United States	Muse Glimmer	8/10/2026	Apache 2.0	\$0.25	17	\$0.06	Accenture, Nomura, Zoom
Google	United States	Gemma 4 26B	4/2/2026	Apache 2.0	\$0.12	17	Not Listed	AT&T, TrustPilot
Mistral	France	Mistral Medium 3.5	4/29/2026	Modified MIT	\$1.16	14	\$0.50	Airbus, ASML, BMW, BNP Paribas, HSBC, Stellantis
Poolside	United States	Laguna S2.1	7/21/2026	OpenMDW-1.1	\$0.05	Not Listed	Not Listed	Northrop Grumman
OpenAI	United States	gpt-oss-120b	8/5/2025	Apache 2.0	\$0.18	12	\$0.11	Limited enterprise evidence
Arcee AI	United States	Trinity Large Thinking	4/1/2026	Apache 2.0	\$0.24	11	\$0.35	Limited enterprise evidence

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Model Provider Websites, Artificial Analysis, AlphaSense

When open-weight models narrow the intelligence gap with closed systems, competitive advantage increasingly shifts from model performance alone to control of the full AI stack. NVIDIA's [\\$12.9B acquisition](#) of Hugging Face and [\\$6B licensing deal](#) with Poolside could strengthen NVIDIA's open-weight strategy by adding a leading platform for model discovery, collaboration, and deployment, as well as a leading open-weight model developer, to its existing ecosystem spanning hardware, inference software, CUDA, and Nemotron (The Information, 8/27; WSJ,

8/22). Rather than simply adding another software asset, Hugging Face would extend NVIDIA's reach across much of the open AI value chain, from model development and distribution to optimization and deployment. Together, the two transactions position NVIDIA not only as the infrastructure provider powering the open-model ecosystem, but also as a key beneficiary of improvements in open-weight model quality and adoption. The strategy parallels Microsoft's acquisition of GitHub by positioning NVIDIA as the owner of a neutral platform where an ecosystem convenes while keeping the platform sufficiency agnostic to sustain participation from competing vendors. If successful, NVIDIA could deepen its influence over how open models are discovered, improved, and deployed without needing to own every leading model.

NVIDIA's management shared that it continues to expect large open and proprietary model ecosystems to coexist, with open models particularly important for enterprise and sovereign AI. Enterprises can bring open models in-house and customize/fine-tune them on proprietary data without that data leaving the premise, preserving data sovereignty and control. NVIDIA also emphasized open data sets alongside open model weights, as for its own open models, the company releases both the weights and the underlying training data sets, providing greater visibility into how the model was trained and should behave. Management compared the future AI ecosystem to proprietary software coexisting alongside Linux and other open-source infrastructure.

Public Benchmark Parity Is Not Necessarily Capability Parity

The competitive threat that open-weight models pose may be smaller than many public benchmarks imply. Although AISI placed GLM-5.2 and DeepSeek V4 Pro roughly [4-7 months behind](#) (AISI, 7/17) the closed cyber frontier, its subsequent evals for [Kimi K3](#) (AISI, 7/23) suggest that leading closed models retain a meaningful advantage in autonomous, long-horizon cyber operations. This is directionally consistent with the view that public rankings can run ahead of private, refreshed and production evaluations. Private tests also weigh reliability, latency, consistency, tool use, and error severity [differently from public benchmark averages](#) (AISI/CAISI, 7/23). With gaps in even public benchmarks widening again (Figure 2), the evidence suggests open-weight models might remain credible substitutes for many cost-sensitive workloads but less reliable replacements for frontier systems in high-stakes, long-horizon deployments.

Figure 2. A Summary of Signals

Metric	Unit	Latest Date	Period	Calculation	Period-Over-Period Δ	Latest Period	2nd Latest Period	3rd Latest Period	4th Latest Period	5th Latest Period	6th Latest Period	Trendline
Proprietary and Open Models (Artificial Analysis)												
Highest Intelligence for Proprietary Models	Points	9/23/26	Week	Point-in-Time	$\Delta +9.4\%$	58	53	53	48	43	42	
Highest Intelligence for Open Models	Points	9/23/26	Week	Point-in-Time	$\Delta +4.5\%$	46	44	44	43	41	40	
Gap in Proprietary and Open Intelligence	Points	9/23/26	Week	Point-in-Time	$\Delta +33.3\%$	12	9	9	5	2	2	
Highest Intelligence Models from the Top-20 Providers (Artificial Analysis)												
Median Intelligence	Points	9/23/26	Week	Point-in-Time	$\Delta +15.2\%$	38	33	32	32	29	28	
Median Latency	Sec	9/23/26	Week	Point-in-Time	$\Delta -2.5\%$	2.68	2.75	2.60	2.65	2.50	2.64	
Median Speed	Tok/Sec	9/23/26	Week	Point-in-Time	$\Delta -3.7\%$	104	108	90	86	70	76	
Release Cadence (Artificial Analysis, LLMTimeline, Model Provider Websites)												
Proprietary Models Released in Last Week	Models	9/22/26	Week	Weekly Total	$\Delta +1900.0\%$	20	1	12	7	1	7	
Open Models Released in Last Week	Models	9/22/26	Week	Weekly Total	$\Delta -66.7\%$	1	3	6	2	4	11	
Proprietary Models Released in Last 90 Days	Models	9/22/26	Week	Trailing 90d	$\Delta +29.0\%$	89	69	69	58	53	52	
Open Models Released in Last 90 Days	Models	9/22/26	Week	Trailing 90d	$\Delta +2.5\%$	41	40	40	37	41	40	
Index Pricing (Bloomberg)												
A100 (SDA100RT)	\$/Hour	9/23/26	Week	7d Avg	$\Delta -1.3\%$	1.57	1.59	1.61	1.61	1.62	1.65	
H100/H200 (CIHOPUS)	\$/Hour	9/23/26	Week	7d Avg	$\Delta -0.0\%$	2.97	2.97	2.96	2.90	2.88	2.90	
(G)B200/(G)B300 (CIBLKWUS)	\$/Hour	9/23/26	Week	7d Avg	$\Delta -0.0\%$	5.30	5.30	5.28	5.27	5.31	5.27	
Avg Inference Token Expenditure (SDLLMTK)	\$/1M Tok	9/23/26	Week	7d Avg	$\Delta +3.1\%$	1.00	0.97	0.98	1.01	1.06	1.04	
Model Pricing for Blended Tokens (Artificial Analysis)												
US+EU Avg Price	\$/1M Tok	9/23/26	Week	Point-in-Time	$\Delta -25.2\%$	1.72	2.30	2.30	1.80	1.42	1.56	
China Avg Price	\$/1M Tok	9/23/26	Week	Point-in-Time	$\Delta -0.0\%$	0.83	0.83	0.73	0.91	0.91	0.91	
Frontier Avg Price	\$/1M Tok	9/23/26	Week	Point-in-Time	$\Delta -20.5\%$	1.36	1.71	1.67	1.44	1.22	1.30	
Proprietary Avg Price	\$/1M Tok	9/23/26	Week	Point-in-Time	$\Delta -25.6\%$	2.15	2.89	3.17	2.23	1.75	1.94	
Open Avg Price	\$/1M Tok	9/23/26	Week	Point-in-Time	$\Delta -0.0\%$	0.68	0.68	0.67	0.75	0.75	0.75	
Cost per Intelligence Index Task (Artificial Analysis)												
Proprietary Avg Cost	\$/Task	9/23/26	Week	Point-in-Time	$\Delta -2.9\%$	2.00	2.06	1.46	0.42	0.38	0.45	
Open Avg Cost	\$/Task	9/23/26	Week	Point-in-Time	$\Delta -35.5\%$	0.80	1.24	0.85	0.35	0.35	0.32	
Model Cost Composition per Intelligence Index Task (Artificial Analysis)												
Input Share of Task Cost	%	9/23/26	Week	Point-in-Time	$\Delta -0.1$	6.2	6.3	2.7	1.8	1.8	1.7	
Cache Hit Share of Task Cost	%	9/23/26	Week	Point-in-Time	$\Delta +0.7$	36.9	36.2	45.4	35.0	34.4	32.0	
Cache Write Share of Task Cost	%	9/23/26	Week	Point-in-Time	$\Delta -0.0$	23.9	23.9	26.1	31.1	30.7	34.6	
Reasoning Share of Task Cost	%	9/23/26	Week	Point-in-Time	$\Delta -0.0$	17.2	17.2	16.3	22.5	22.7	21.9	
Answer Share of Task Cost	%	9/23/26	Week	Point-in-Time	$\Delta -0.6$	15.7	16.3	9.6	9.7	10.3	9.8	
Avg Output Tokens per Task	Tokens	9/23/26	Week	Point-in-Time	$\Delta +1.1\%$	37,667	37,269	43,690	26,029	25,350	24,908	
Developer Usage (PyPI)												
Multi-Provider SDK Downloads	%	9/22/26	Week	WoW Δ in 7d Avg	$\Delta -1.5$	+2.6	+4.1	+5.0	-37.4	-9.6	+0.7	
Gemini SDK Downloads	%	9/22/26	Week	WoW Δ in 7d Avg	$\Delta -2.5$	-0.9	+1.6	+8.2	-52.6	-12.8	+0.5	
Mistral SDK Downloads	%	9/22/26	Week	WoW Δ in 7d Avg	$\Delta +30.8$	+19.0	-11.8	+26.9	-48.9	-32.2	-22.1	
App Usage (Sensor Tower)												
Gemini App WAU	%	9/21/26	Week	WoW Δ	$\Delta +1.4$	+1.1	-0.3	+2.7	+3.4	+2.0	+3.3	
Qwen App WAU	%	9/21/26	Week	WoW Δ	$\Delta -5.2$	+0.2	+5.4	-1.1	+6.8	-4.4	+8.3	
Kimi App WAU	%	9/21/26	Week	WoW Δ	$\Delta +4.2$	+1.3	-2.9	-1.8	-3.5	-2.8	-3.6	
Web Usage (Similarweb)												
Gemini Web Visits	%	8/31/26	Week	WoW Δ in 7d Avg	$\Delta -4.0$	-0.1	+3.9	+0.4	-1.7	-2.3	+0.6	
Qwen Web Visits	%	8/31/26	Week	WoW Δ in 7d Avg	$\Delta +3.0$	+4.2	+1.2	-1.1	+9.3	-2.2	-3.0	
Kimi Web Visits	%	8/31/26	Week	WoW Δ in 7d Avg	$\Delta +18.5$	-10.0	-28.5	-9.5	-12.8	-10.0	+12.3	
Layoffs (WSJ)												
All Layoffs Announced	People	9/22/26	Month	MTD	$\Delta +132.9\%$	4,102	1,761	17,420	24,709	26,160	11,843	
AI-Driven Layoffs Announced	People	9/22/26	Month	MTD	$\Delta +46.1\%$	469	321	3,220	21,640	15,360	10,400	

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Artificial Analysis updated its Intelligence Index tasks on Sep. 7, 2026; cost-per-task comparisons should be made separately for the periods before and after the update.

Source: Citi Research, Artificial Analysis, Bloomberg, PyPI Stats, Sensor Tower, Similarweb, Wall Street Journal

Why Public and Private Results Diverge

Several factors may explain the divergence between the public benchmark results and proprietary evaluations:

- **Benchmark-aware post-training.** Researchers can systematically target known benchmark weaknesses and evaluation patterns, improving measured performance more quickly than would be achieved through broad-based capability gains. This is colloquially termed ‘benchmaxxing.’ While all model developers are benchmark aware to some extent, the rapid release cadence and model-card disclosures (including for Z.ai’s GLM-5.3 which attributed much of its intelligence improvements to post-training scaling and RL) provide particularly visible evidence of evaluation-driven iteration.
- **A growing prevalence of private and bespoke evals.** While public benchmarks remain useful reference points (Figure 3), growing enterprise adoption is shifting attention toward proprietary and use-case-specific evaluations that are harder to optimize against, potentially making durable performance advantages more difficult to observe and replicate.
- **Deployment complexity.** Private evaluations may better capture the factors that drive enterprise value in production (including reliability, latency, token efficiency, consistency, tool execution, and error costs), where small performance differences can compound into meaningful revenue, cost, and productivity benefits at scale.
- **A gated or deliberately paced frontier.** Longer safety and security review periods can [delay](#) closed-model releases (OpenAI, 8/18), allowing open models additional time to close the gap with the last released frontier model, potentially compressing publicly-measured performance differences. Considering recent [calls to pace the frontier](#) (Dario Amodei, 9/12), we would expect the interval between capability creation and broad release may see some additional checkpoints as labs devote more time to safety testing, third-party evaluation, and controlled-access deployment, which could allow open-weight models to approach an aging public frontier while proprietary labs preserve a less visible advantage in restricted or unreleased capabilities.

Figure 3. Model Metrics

MoM Δ in Provider Rank	Model Provider	Leading Model	Artificial Analysis Intelligence Index	Speed (Output Tokens/Second)	Blended Price (USD/1M Tokens)	Context Window (1K Tokens)	Release Date	License
—	Anthropic	Claude Opus 5.5 (max with fallback)	58	69	\$2.94	1000	Sep. 22, 2026	Proprietary
—	OpenAI	GPT-6 Astra (max)	53	52	\$7.70	1000	Sep. 3, 2026	Proprietary
▲4	Meta	Muse Spark 1.3 (max)	48	223	\$0.78	1000	Sep. 2, 2026	Proprietary
▼1	SpaceXAI	Grok 4.7 (xhigh)	46	40	\$1.35	500	Sep. 21, 2026	Proprietary
▲6	Xiaomi	MiMo-V2.6-Pro	46	54	\$0.17	1000	Sep. 22, 2026	Open
—	Alibaba	Qwen3.8 Max (0902)	45	39	\$1.18	984	Sep. 16, 2026	Proprietary
▼2	Z.ai	GLM-5.3 (max)	45	61	\$0.90	1000	Aug. 18, 2026	Open
▼4	Moonshot AI	Kimi K3 (max)	44	37	\$2.31	1050	Jul. 16, 2026	Open
▼1	Google	Gemini 3.8 Flash (high)	41	283	\$0.57	1000	Sep. 2, 2026	Proprietary
▼1	DeepSeek	DeepSeek V4.1 Flash (max)	39	232	\$0.19	1000	Sep. 14, 2026	Open
▼1	Sapiens AI	Agnes 3.0 Flash	36	242	\$0.03	1000	Sep. 14, 2026	Proprietary
▼2	MiniMax	MiniMax-M3	29	106	\$0.22	1000	Jun. 1, 2026	Open
▼1	Nex AGI	Nex-N2-Pro	28	116	\$0.53	262	Jun. 2, 2026	Open
—	Thinking Machines	Inkling Small	28	205	\$0.22	1000	Jul. 30, 2026	Open
—	Upstage	Solar Pro 4	28	54	\$0.22	512	Aug. 6, 2026	Proprietary
▲5	Multiverse Computing	Quasar 438B (max)	27	144	\$0.72	1000	Aug. 10, 2026	Proprietary
▼1	InclusionAI	Ling 3.0 Flash	25	352	\$0.04	262	Aug. 4, 2026	Open
▼5	Tencent	Hy3	25	86	\$0.11	256	Jul. 6, 2026	Open
▼2	NVIDIA	Nemotron 3 Ultra	23	146	\$0.48	262	Jun. 1, 2026	Open
▼2	KwaiKAT	KAT-Coder-Pro V2	22	102	\$0.22	256	Mar. 27, 2026	Proprietary

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Artificial Analysis, Model Provider Websites

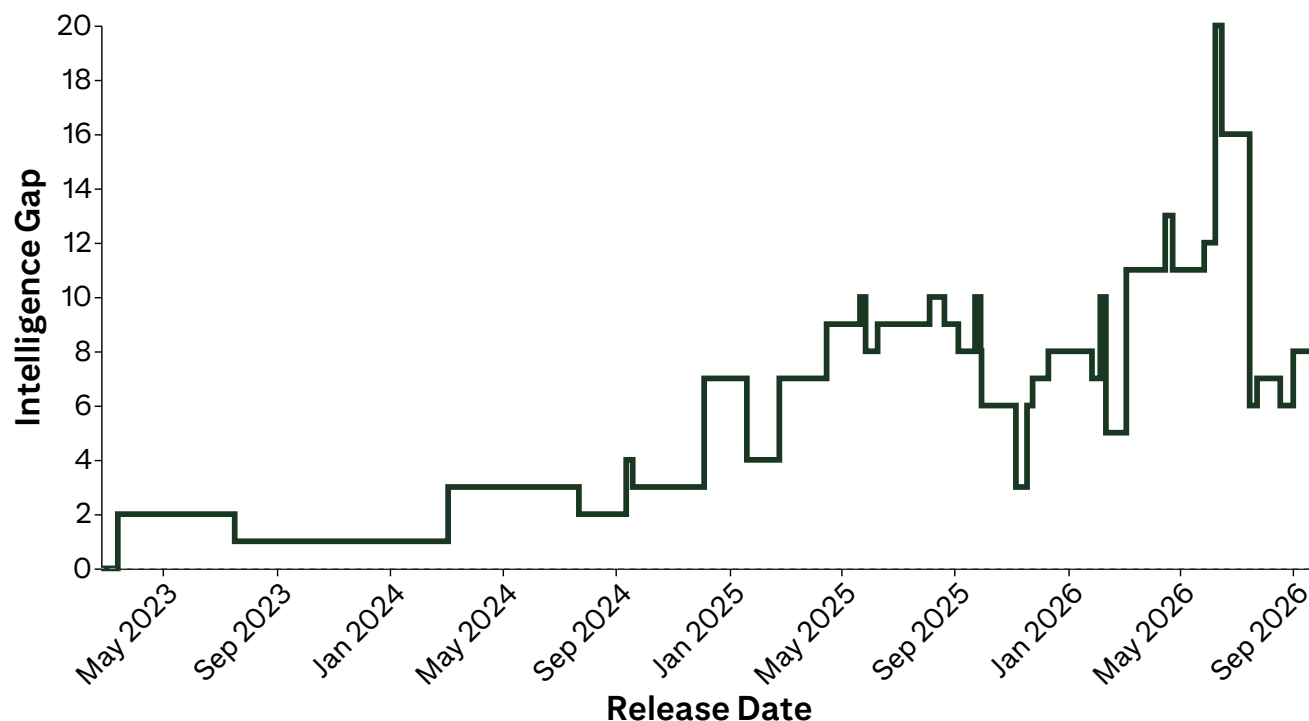
“[Demand is] probably going to flow more towards cheaper models, more capable models that are smaller. And then more customized models that are tailored to people's needs. And businesses need. Because I honestly don't think we need an AGI model for everything.”

— [Benny Chen, Fireworks Co-Founder, 7/8](#)

Benchmark Convergence and Developer-Routing Adoption

The public proprietary-versus-open-weight intelligence gap narrowed from 16 points in mid-June to 6 points by mid-July, remained near that level through mid-August, and then widened to 12 points in late September (Figure 4). This temporary convergence coincided with a sharp increase in open-weight token share across both Vercel and OpenRouter, directionally consistent with developers routing more work to open models as capability converged. The latest widening will test whether those usage gains persist when proprietary models regain a more meaningful intelligence advantage, though gateway activity notably represents developer routing rather than the enterprise market or revenue share.

Figure 4. Gap in Proprietary and Open Intelligence



Intelligence Gap

- Frontier Proprietary Minus Frontier Open-Weight Intelligence

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Artificial Analysis

The gap has historically moved in steps around major releases, as pre-training enables leapfrogging. Rough cycles averaging roughly 4-5 months in duration are visible, with the gap narrowing (before re-widening) around Mar. 2023, Aug. 2023, Feb. 2024, Aug. 2024, Feb. 2025, Sept. 2025, Nov. 2025, Feb. 2026, Apr. 2026, and Jun. 2026 (Figure 4). After the most recent rewidening, the next test will be whether open developers can again compress that lead within a single release cycle. On the demand side, data-residency, sovereignty, auditability, and model control requirements can increase enterprise demand for downloadable weights, even when open models remain behind the frontier on absolute capability.

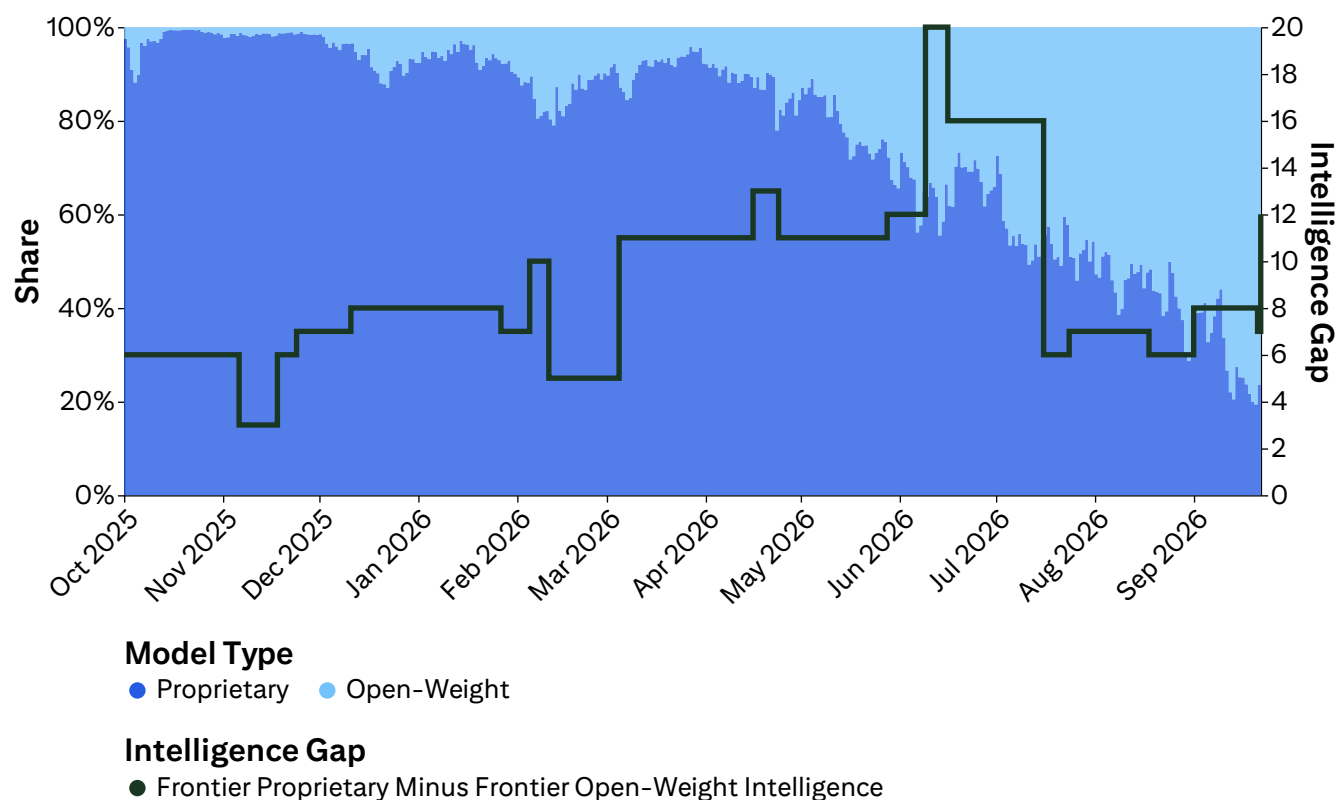
Open-Weight Share Rose as the Intelligence Gap Narrowed

Vercel AI Gateway routes production AI workloads across hundreds of models and providers for [more than 200,000 developer teams](#) (Vercel, 5/12), providing a useful view of real-world application and agent usage. Open-weight models began sustaining a majority of Vercel AI Gateway token volume in August and reached a three-fourths majority in September, up 43 percentage points since the end of June. The concurrent movement suggests the capability convergence across a variety of open models could be one driver of adoption.

Vercel offers a read on model usage within production applications via data feeds from its application platform and AI SDK. Open-weight share on Vercel rose from 3% on October 1, 2025 to 77% on September 22, 2026. It remained above 25% from May 28, above 40% from July 3, and above 50% from August 6, thus moving from a sustained 25% to sustained 50% in 70 days (Figure 5). Open-weight

adoption continued to rise after the intelligence gap started rewidening in early September, suggesting that capability convergence may help unlock workloads, while integration, economics, availability, and product fit increasingly determine whether those gains persist.

Figure 5. Rising Open-Weight Adoption Coincides with a Narrowing Intelligence Gap

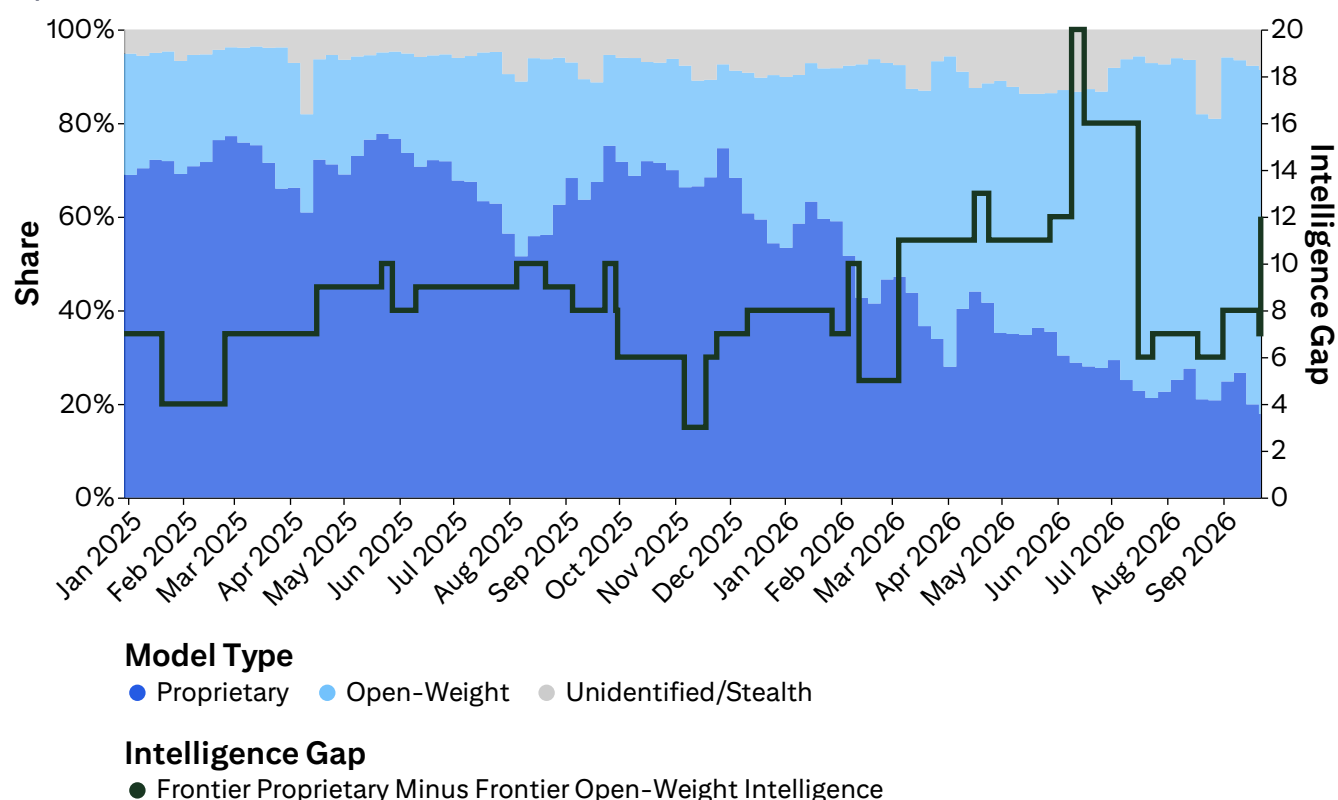


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Vercel

With a broader model catalog and OpenAI-compatible endpoints, OpenRouter more directly captures developer experimentation, price-sensitive routing, and model substitution. On current data, OpenRouter reached sustained 40% open-weight share on February 2, 2026, about five months ahead of Vercel on July 3. It reached sustained 50% share on May 25, approximately ten weeks ahead of Vercel on August 6. From the first week of February through the latest complete week in September, weekly identified token volume increased by 110T. Open-weight models contributed 89T, or 81% of that increase. Meanwhile, proprietary volume increased by 21T, supporting continued market expansion rather than a pure substitution effect. Open-weight models represented 72% of total OpenRouter volume in the latest complete week (Figure 6).

Figure 6. Open-Weight Models have become the Majority of OpenRouter Token Volume Despite a Persistent Intelligence Gap



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, OpenRouter

“Similar to Poolside, Eastern labs were forced to become competent in constrained environments. Meanwhile, some Western labs began with large clusters and were able to reuse established recipes. The gulf in model-building capabilities between Western and Eastern labs is only going to widen because Eastern labs have done the hard work and now primarily need more compute.”

- Jason Warner, Poolside Co-CEO, 9/9

Why Chinese Open Models Lead the Open Field

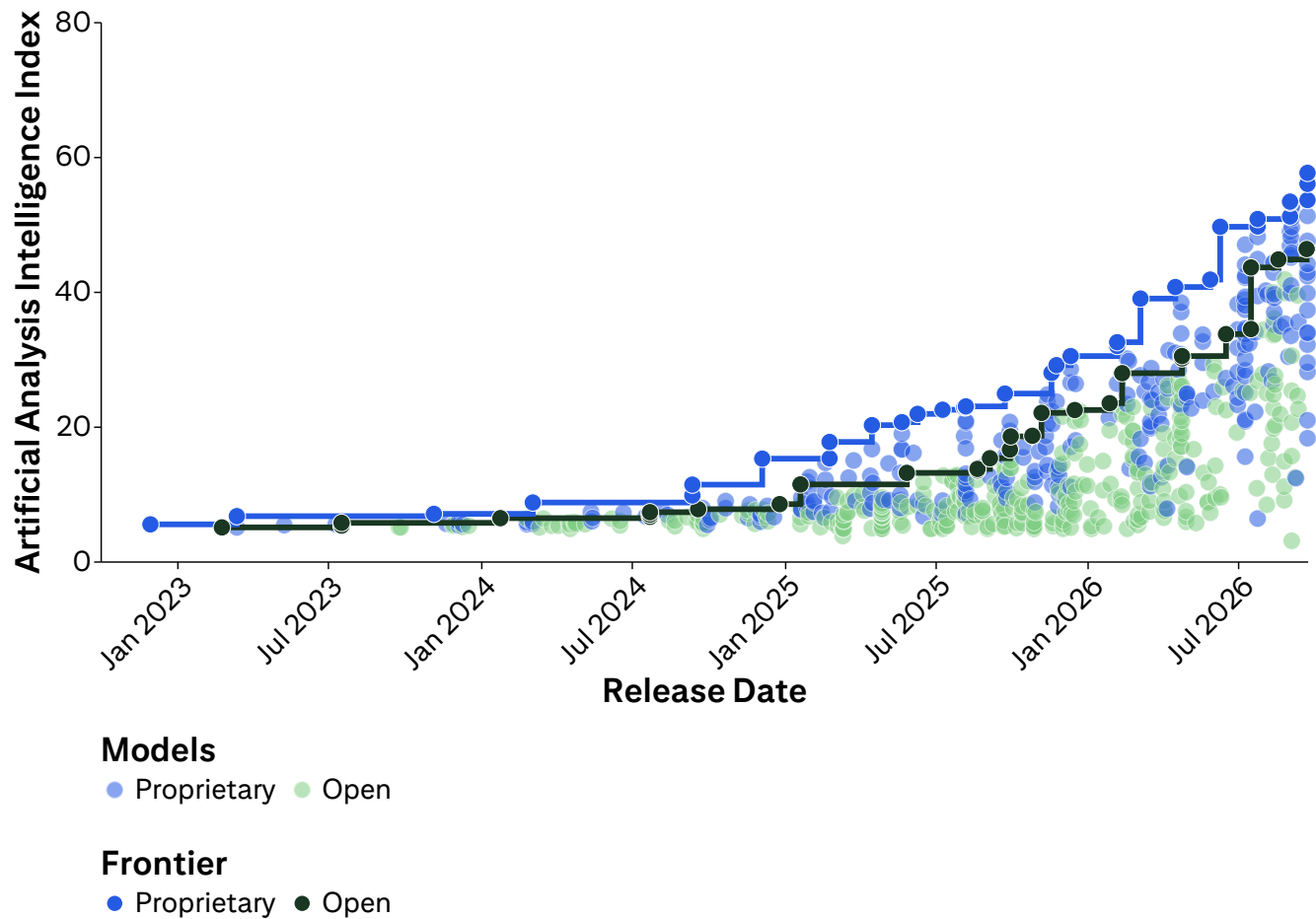
We believe several factors help explain why Chinese developers have recently outpaced their Western counterparts:

- **Model development industrialization.** The Eastern labs are industrializing the model-development process, an approach Western labs could also use to accelerate iteration and improve capital efficiency. For example, Poolside released three Laguna models in roughly three months, with Laguna S 2.1 going from the start of training to public release in less than nine weeks while remaining competitive with much larger models on agentic coding benchmarks.
- **Talent density.** Deep connections to major universities and a large domestic technical workforce support concentrated research teams. NVIDIA CEO Jensen Huang has frequently stated that 50% of the world's AI researchers are Chinese and 9 out of 10 of the world's top science and technology universities are in China ([CSIS, 12/3/25](#)). Additionally, unlike the US where company-university ties tend to loosen as companies gain scale, China's top computer science departments and state-affiliated research institutes more directly act as incubators for frontier AI labs. For example, Z.ai was a direct commercial spin-off from Tsinghua University's KEG Lab and its founder Tang Jie is a sitting professor. Similarly, both Moonshot AI and DeepSeek recruit heavily from top Chinese CS departments including from Zhejiang University, Peking University,

and Tsinghua. While state-directed strategies and funding are not unique to AI in China, they nonetheless contribute.

- **Focused product scope.** Leading Western labs have the benefit of first mover advantage, but they likely have more to consider for their model development (like a more diversified enterprise customer base) relative to the Eastern labs. Also, several of the leading open-weight models, including Z.ai's GLM and DeepSeek-V4-Pro, are text-only systems. This specialization may contribute to stronger performance on benchmarks that emphasize language and reasoning tasks rather than multimodal capabilities.
- **Systems optimizations.** Advanced-chip restrictions may spur innovation in both models and infrastructure. DeepSeek's early MoE breakthrough highlighted how compute constraints can incentivize more efficient architectures, while sparse designs, quantization, cluster engineering, and domestic accelerator ecosystems continue to expand the capabilities achievable with limited compute. Meanwhile, Western labs can pair similar optimizations with greater access to state-of-the-art chips. Their focus on monetizing frontier capabilities through closed APIs could shift if open-weight models increasingly influence developer tooling, enterprise architecture, and downstream compute demand.
- **Distillation and post-training.** Knowledge distillation transfers behavior from a stronger teacher model into a smaller student model. In legitimate settings, the model owner authorizes the training relationship or the teacher's license permits it. Alleged extraction of high-value teacher outputs can accelerate reasoning, coding, and agentic capability, although it is not the only route to progress. More broadly, post-training refers to techniques applied after pretraining, the initial large-scale training phase in which a model learns from vast amounts of data. These methods can significantly improve reasoning, coding, and agentic capabilities without incurring the cost of training a new frontier model from scratch. As a result, leadership in open models increasingly depends not only on access to compute, but also on the efficiency with which developers improve models after the initial training phase.
- **Time-to-release advantages.** The open ecosystem's response to proprietary frontier advances has accelerated sharply in the latest cycle, after lagging further behind through much of 2025. The comparable-performance lag (measured by how long it took an open model to reach a score first achieved or exceeded by a previously released proprietary model) was 105 days at year-end 2024 and 133 days at year-end 2025. The open ecosystem made rapid progress against the AA intelligence frontier in recent months. Prior to the release of Claude Fable 5, the proprietary benchmark stood at 42 with Claude Opus 4.8 on May 28. Fable 5 raised the frontier to 50 on June 9. Kimi K3 lifted the open frontier to 44 on July 16, 37 days after Fable 5 first established the new threshold. GLM-5.3 raised it to 45 on August 14, after 66 days, and MiMo-V2.6-Pro reached 46 on September 22, after 105 days (Figure 7). These observations trace successive gains within a single catch-up cycle rather than three independent measures of acceleration, but they show that open developers recovered a meaningful portion of Fable 5's frontier advance within roughly three months. Labs able to convert benchmark results, failure cases, preference signals, and tool-use traces into faster post-training and harness updates could compress the catch-up period further as closed releases lag, allowing open models to gain ground against an aging public frontier before the next major proprietary release has a chance to extend the lead again.

Figure 7. Frontier Releases for Proprietary and Open Models



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Artificial Analysis

Poolside's Jason Warner on the Future of Open Models

Figure 8. A Conversation with Poolside's Jason Warner (Part 1 of 2)



Following Poolside's Nvidia partnership, we spoke with Jason Warner, the co-founder and co-CEO of Poolside, about their Model Factory, the path to a frontier open-weight model, enterprise token economics, compute constraints, and the competitive position of open and closed model providers. The Q&A below is edited for length.



HT: For people who are less familiar with Poolside, how should they think about the company today? What are you building, and which problems are you ultimately trying to solve?

JW: Poolside is a foundation-model company. We build frontier foundation models and point them at hard problems. Before Poolside, I was CTO at GitHub for four years, two years before the Microsoft acquisition and two years after. One of the first problems we pointed at was knowledge work. When I say knowledge work, I do not just mean code. At the end of the day, knowledge work itself boils down to software using compute to solve problems. We try not to get boxed in as a coding company. What I like to say is that we are trying to figure out how to apply intelligence to very hard problems. In this moment, that means figuring out how to efficiently make models and apply them toward knowledge work.

HT: The Model Factory appears to be core to Poolside and a central part of the company's recently announced relationship with Nvidia. What is the Model Factory, and how does it differ from the more conventional approaches to training?

JW: When we started Poolside, we took two contrarian views. The first was that reinforcement learning was going to be the next vector of scale. We would constantly go back and forth between pre-training and post-training, with RL becoming increasingly important. The second was that the industry was not very good at model building. A lot of teams were taking a known recipe from somewhere else and applying it again with more compute. We decided to have no recipe and build a mechanism to discover the recipe. That idea is grounded in Richard Sutton's short paper, "The Bitter Lesson." If you have not read it, I implore you to. It states that the real advances in artificial intelligence over the last several decades have largely come from applying more compute to simple approaches, not more elaborate algorithms. If that is true, the question is how you use compute to search for the right solution instead of assuming you already know the recipe. That is the Model Factory. At the simplest level, it is a discovery engine. You take compute, pair it with the Model Factory, and point it at a problem. The first solution may be naive, but you refine it over and over in a classic scientific-process sort of way.

HT: Laguna Small provides an early test of whether the Model Factory can identify a more efficient training recipe. What does the model's development tell us about the compute and capital required to build a competitive open-weight model?

JW: We built the Model Factory to search for the most efficient model we can build with the compute we have. In essence, it is a meta-platform that tells us how to build models, and the proof is in the pudding. We released Laguna Small, a 118B-parameter MoE model with 8B active parameters, which we believe is the best open-weight model in the West. We trained it on roughly 4K H200s within a total cluster of 10K GPUs and did it in about five weeks end-to-end. That flips the notion on its head that you need tens of billions of

dollars to do these training runs. It shows that we, as an industry, are hyper-inefficient at this.

HT: With the additional capital, compute, and strategic support now available through the Nvidia relationship, how does Poolside plan to extend the Model Factory beyond its initial focus on knowledge work? Which problems are most compelling to solve?

JW: Nothing changes in our mission or in the way we think about problems. The Model Factory is still a discovery engine. You take compute, pair it with the Model Factory, point it at a problem, and you should get a solution out of it. It may be naive to begin with, but you refine it over and over. When I break down where AI is going to be applied, there are three large categories. Knowledge work, robotics, and biology. Those three categories make up the vast majority of what I think AI is going to do over time. The two that interest me most are knowledge work and biology. The broader point is that we can point the Model Factory at interesting problems and continue refining the solution.

HT: Recent improvements in open-weight models have raised questions about whether they are structurally narrowing the gap with closed frontier systems or simply benefiting from a temporary slowdown at the frontier. How do you expect the relationship between open and closed models to develop?

JW: The statement I would defend is that most people do not need the newest frontier model for most of their work. You can get by with a six-month-old model, a four-month-old model, or a free open-weight model. 90% of the economic value we are creating inside knowledge work does not require most-expensive-token work. From an economics perspective, I might prefer a model that costs one-thousandth as much and takes an extra minute. What I would hope happens is that we see a spectrum of problems solved by frontier models, then frontier-minus-one, then smaller language models. That would be very healthy for companies whose economics are going to be upside down because they are overspending on tokens.

HT: If most knowledge-work tasks do not require the newest and most expensive frontier model, does the opportunity for open models primarily come from offering lower-cost intelligence, or are control, ownership, and supplier concentration equally important considerations for enterprises?

JW: Cost is part of it, but open models can also distribute the risk. Right now, almost every company on the planet is risking itself because it is effectively saying it will use tokens from one or two API providers. People want to equate that with hyperscaler risk. There is overlap, but it is different. Imagine you want to build your forever home on one of two lots. On one piece of land, you have to clear it and do some work, but you buy it and own it. On the other, somebody leases cleared land to you for two years, with a provision that they own whatever is on it at the end of those two years. How many of us would build our forever home on that second piece of property? Not a single one of us. Yet that is what businesses are doing when they say they will re-base themselves as AI companies and do all of that work on an external API. As a frontier model provider, I don't need to see all your data. I need to see one sliver of your data, and then I can synthesize it all the way out. Businesses do not need to build frontier AI. They need to build something that looks a little different for their specific business, and they need to start manifesting their own destiny again. The model providers have been saying for years that everything collapses into the models. At some point, intelligence will become

Figure 9. A Conversation with Poolside's Jason Warner (Part 2 of 2)

capable enough to do more of the workloads that businesses do. Companies need to figure out what they should build or control for themselves rather than putting every differentiated workflow on land owned by somebody else.

HT: Despite the recent progress of Western labs such as Poolside, the strongest open models have generally continued to come from Eastern labs. What is driving the gap between Western and Eastern open-model labs? Is distillation the principal explanation, and what would Western labs need to do to close the gap?

JW: The distillation narrative is incredibly overblown. The main differentiators between Western and Eastern open models are lab competency and compute. When building the Model Factory, we found out in 2023 that Qwen and DeepSeek were pursuing similar infrastructure. Similar to Poolside, Eastern labs were forced to become competent in constrained environments. Meanwhile, some Western labs began with large clusters and were able to reuse established recipes. The gulf in model-building capabilities between Western and Eastern labs is only going to widen because Eastern labs have done the hard work and now primarily need more compute. Poolside has always had a simple plan. Mechanize the process of model building, then build the best open model in the West, then the best open model in the world, then the best model in the world that just happens to be open. Progress through those stages is largely a function of compute. We do not need the same level of compute as the largest closed labs, but we do need to scale appropriately.

HT: Even if open models offer compelling cost and control advantages, enterprises may continue defaulting to established API providers because that choice is operationally easier and perceived as less risky. What developments could overcome that inertia and cause open-weight models to take materially more market share?

JW: Two things can change their share. The first is the arrival of a frontier open-weight model. If a model hits the frontier, it will get consumed. There will be substantial demand for anything at the frontier. The second is real economic pressure from shareholders and public companies. We now have companies saying they will cut token spend and hire more people because token spend is so expensive. The answer is not to cut token spend and hire more people. It is to get efficient at the tokens you are consuming. I do not put as much weight on the second because companies are very good at hiding costs. But I put a lot of weight on the first.

HT: Compute remains constrained across the industry, but the consequences may differ between training and inference. How is scarcity affecting the pace and economics of model development, enterprise adoption, and model serving? To what extent is the binding constraint raw compute availability versus the architectural limits of today's clusters?

JW: The compute shortage is going to do two things at once. It will slow down some training, but it will definitely slow down adoption and make things more expensive. There are economics on both sides of the fence dividing training and serving. But I do not think the training-compute shortage is as severe as people make it out to be. There is also a physics limit on how large a model you can efficiently train inside a given cluster. You have to ask, what is the largest model you can sufficiently train and also serve to yourself and your customers with the compute you have. You can train a model and still find that it is not economically viable to serve, which means you have to

compress it, distill it, or start again with a smaller model. So from a training perspective, this is not just a compute shortage. It is also an architectural physics limit.

HT: Outside coding and the relatively narrow applications where AI has already gained traction, how do you expect enterprises to adopt AI over the next three to five years? What does that transition imply for the way knowledge work is performed and verified?

JW: I have said from the beginning that knowledge work collapses into software, and I do not mean coding. Most knowledge work comes down to solving a problem, and the best proxy for solving that problem is a model ephemerally writing the software it needs to solve it. In a couple of years, many of the tool calls and preconstructed integrations we talk about today will not matter because models will natively discover, use, and call different APIs or tools. You will point a model at a spreadsheet problem, a financial-analysis problem, or a data problem and ask it to come back with a result. Then another AI will verify it using the evaluations or reward mechanism you set up. Most of what we do behind a screen collapses into two categories: somebody doing the work and somebody verifying it. Over time, the companies that matter most may be those controlling infrastructure for energy, compute, or intelligence. Applications built on external models may still create value, but are less likely to transform the businesses they serve.

HT: The proliferation of open models makes it difficult for investors to distinguish genuine technical progress from benchmark optimization, prestige, or access to capital. Which measures best reveal whether a lab can build competitive models efficiently and repeatably, and what will determine which labs remain independent?

JW: You should judge the competency of labs on their merits. Silicon Valley has a prestige bias, and I have obviously benefited from that as the former CTO of GitHub. But the better measure is capital efficiency. What did the lab build, how much compute did it take, and can it do it again? The problem is that those things can be easy to disguise. For instance, labs can engage in benchmaxxing by testing on benchmark outputs. Ultimately, +neo-labs cannot succeed unless they can competently reach the frontier. Right now, the opportunity is at the frontier of open. But that opportunity will only last until there is a frontier open model, whether it comes from us, Nvidia, or somebody else.

Economics of Open vs. Closed

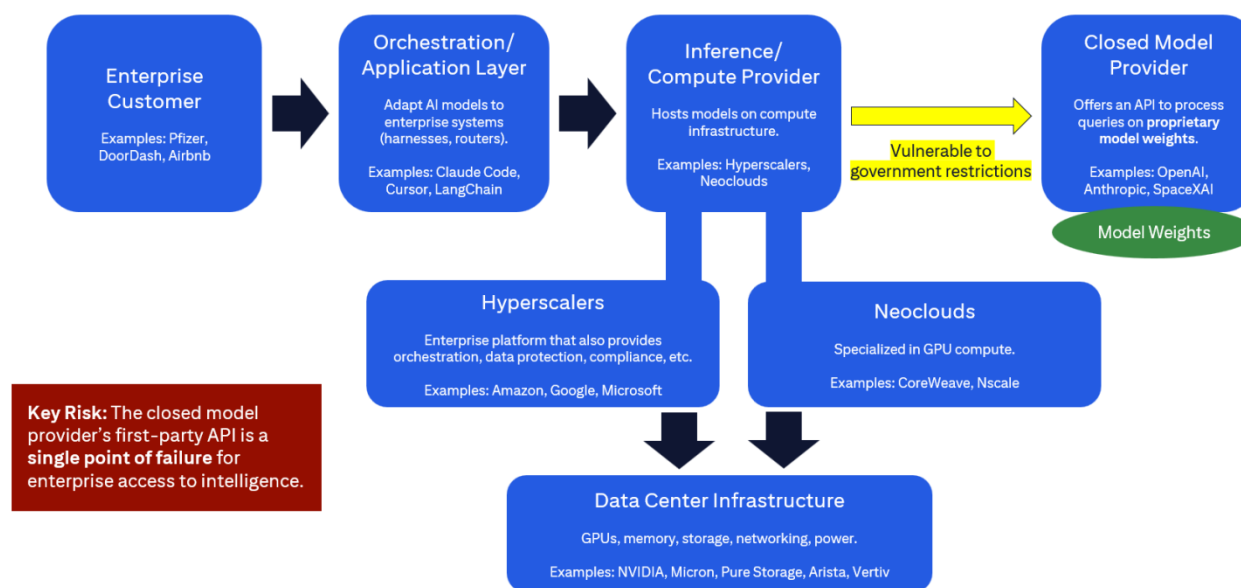
Open Does Not Mean Free

Open-weight models release trained parameters (i.e., weights) that allow other people to run, fine-tune, and deploy the model, but they generally do not disclose the underlying code, training data, or complete training process. Open-source models provide access to source code and, ideally, the datasets, training methodology, and other components necessary for independent inspection, modification, and reproduction. As a result, users can usually modify and deploy open-weight models, but they often cannot fully verify, recreate, or audit the model from scratch. **All competitive open models today are open-weight rather than open-source.**

In either case, enterprises continue to pay for inference, retrieval, evaluation, orchestration, security, monitoring, support, and compliance. Running an open-weight model through a cloud platform (such as AWS Bedrock, Microsoft Foundry, or Gemini Enterprise Agent Platform) may reduce model-related fees and lower list token pricing, but it does not eliminate the costs associated with compute, storage, networking, and the broader operating stack (Figure 10).

Figure 10. Explanatory Enterprise AI Value Chain for Closed-Weight Models

The Enterprise AI Value Chain: Closed-Weight Models



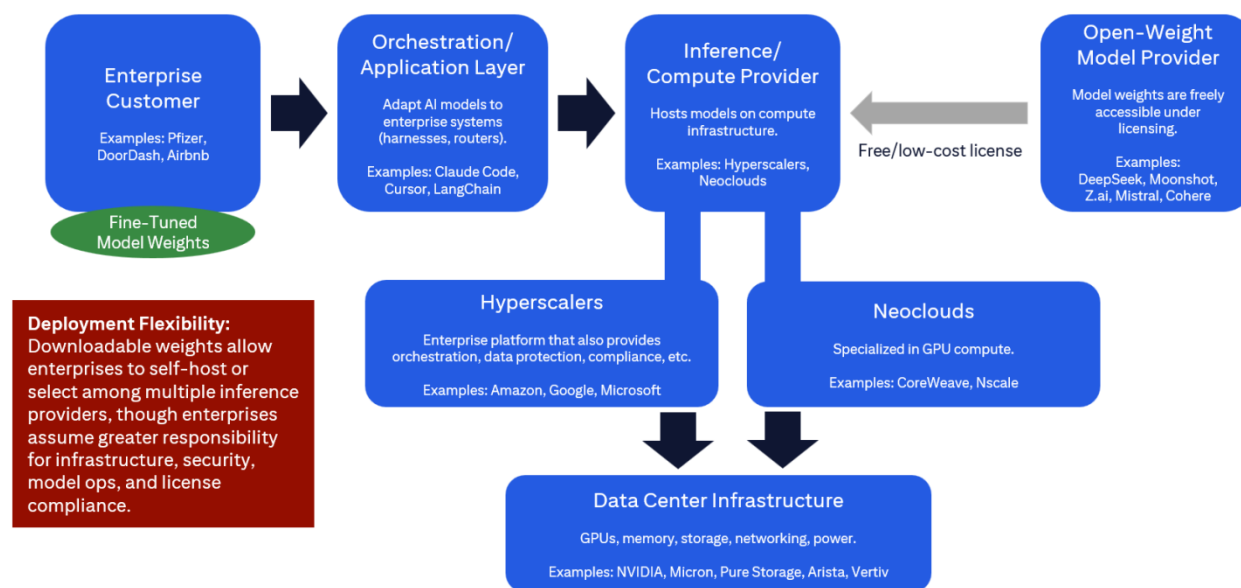
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research

As a result, a significant share of AI profits can still accrue to cloud and infrastructure providers even when model weights are freely available. With data-center construction, power, and accelerator costs remaining elevated, cloud deployment can provide faster and often more economical access to compute than building new on-premises capacity. Organizations that choose to self-host must additionally manage capacity planning, model serving, maintenance, upgrades, reliability, and utilization risk (Figure 11).

Figure 11. Explanatory Enterprise AI Value Chain for Open-Weight Models

The Enterprise AI Value Chain: Open-Weight Models



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research

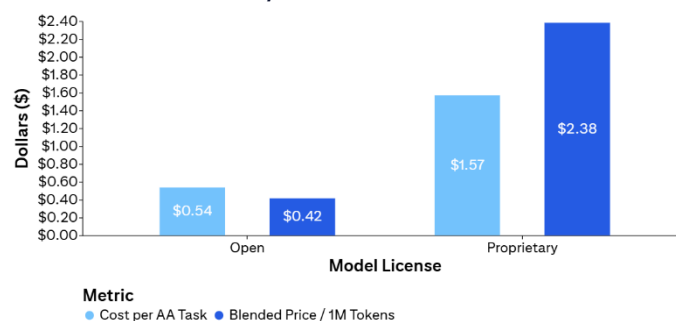
Enterprises are unlikely to standardize on a single model provider or licensing framework, but rather adopt a portfolio approach. Closed frontier labs are likely to retain pricing power for the most complex workloads, while recapturing spending through enterprise software, orchestration layers, managed services, and support offerings. **Open models** are positioned to capture a growing share of token volume, customization, and local deployment use cases, although model developers may not fully monetize usage that occurs through third-party hosts or self-managed infrastructure. At the same time, **hyperscalers and inference providers** should benefit from increased demand across both open and closed models, competing primarily on utilization, compliance, security and customer relationships rather than exclusive model access. Across all deployment models, **infrastructure providers** remain the most neutral beneficiaries, monetizing compute demand regardless of model type. Importantly, lower token prices may be partially offset by higher token consumption, particularly when open-weight models require longer outputs or greater reasoning effort to achieve comparable task performance.

Cost per Task Is Not Cost per Token

Lower token prices do not necessarily translate into lower task costs. Uniphore noted that fine-tuned small language models (SLMs) can [unlock 97% lower token costs](#) compared to frontier LLMs without affecting performance. Open-weight models are materially less expensive than closed-weight models on a per-token pricing basis. Yet, the average cost per task for open-weight models is ~29% above the blended price/1M tokens, compared to 34% lower for proprietary models, based on Artificial Analysis data (Figure 12). Furthermore, the researchers behind the OckBench evaluation, which measures the efficiency of LLM reasoning, found that frontier and open models with comparable accuracy differed by [more than 26x](#)

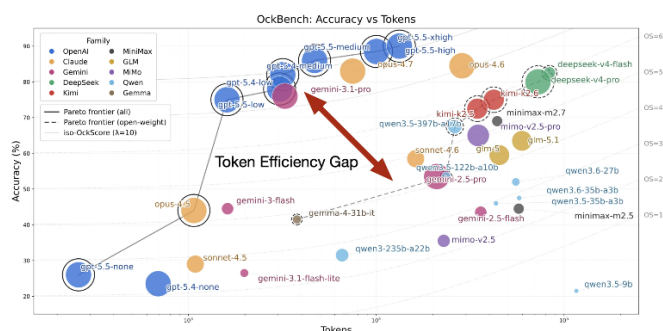
in output length, with more verbose open models sometimes paying an “overthinking tax” that offset their per-token price advantage (OckBench, 6/3).

Figure 12. Open vs. Proprietary Models Average Cost per AA Task and Blended Price/1M Tokens



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Artificial Analysis, Citi Research

Figure 13. OckBench: Accuracy vs. Tokens



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: OckBench, Citi Research

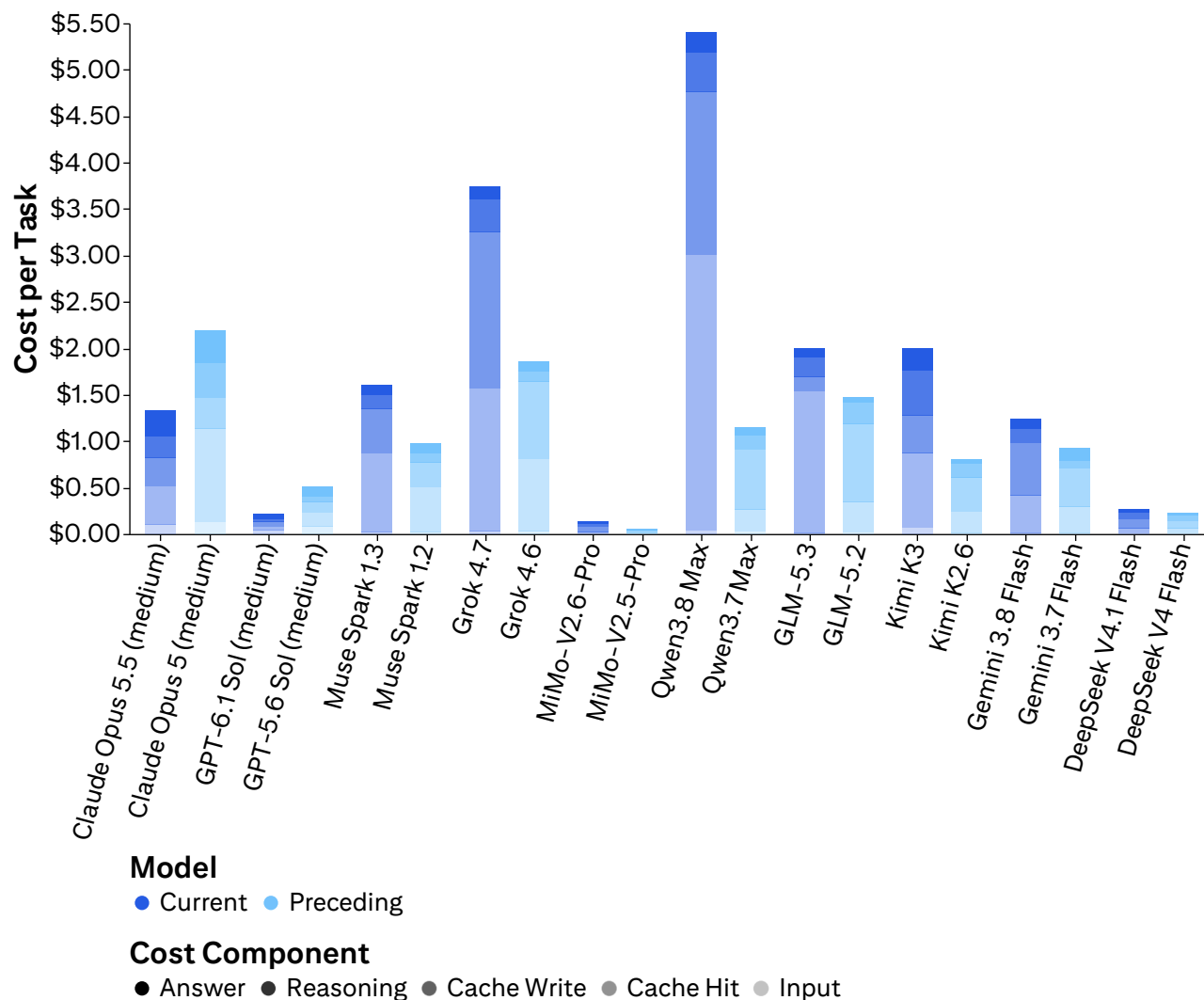
The Cost of a Unit of Intelligence is Falling

The value of frontier intelligence is not solely its higher level of capability, but its ability to complete the same task at a lower cost. Some recent frontier releases are already improving cost per task relative to their predecessors (Figure 14). A knowledge-work task that required last year's highest-performing model may now be completed more accurately, more quickly, or with fewer retries by a newer frontier model (potentially at a lower realized cost per successful outcome).

OpenAI's GPT-6.1 Sol (medium) recorded an Intelligence Index score of 48 at \$0.21 per task, compared with 39 and \$0.50 for GPT-5.6 Sol (medium). GPT-6.1 Sol added 9 Intelligence Index points, or 23%, while reducing task cost by 58%, lowering its cost-to-score ratio to \$0.0045 per point from \$0.0129. Anthropic shows a similar pattern. Claude Opus 5.5 (medium) scored 51 at \$1.34 per task, versus 45 and \$2.19 for Claude Opus 5 (medium). Opus 5.5 therefore added 6 points, or 13%, while reducing task cost by 39% and lowering its cost-to-score ratio to \$0.0262 from \$0.0487.

However, the pattern is not yet uniform. Muse Spark 1.3 increased intelligence by 8 points, or 20%, over Muse Spark 1.2, but its cost per task rose 65% to \$1.60. Grok 4.7 added 2 points, or 5%, over Grok 4.6, while task cost more than doubled to \$3.74; Gemini 3.8 Flash similarly added 2 points, or 5%, over Gemini 3.7 Flash, while cost rose 34% to \$1.24. These divergences suggest that capability gains do not automatically translate into better economics. The question is whether higher intelligence reduces retries, redundant reasoning, tool calls, and human intervention enough to offset its direct inference cost. Even if token prices remain stable or rise, the cost per successfully completed workflow could – and should – decline as model reliability improves and regulation delays releases.

Figure 14. Cost per AA Task for Current Versus Prior Models

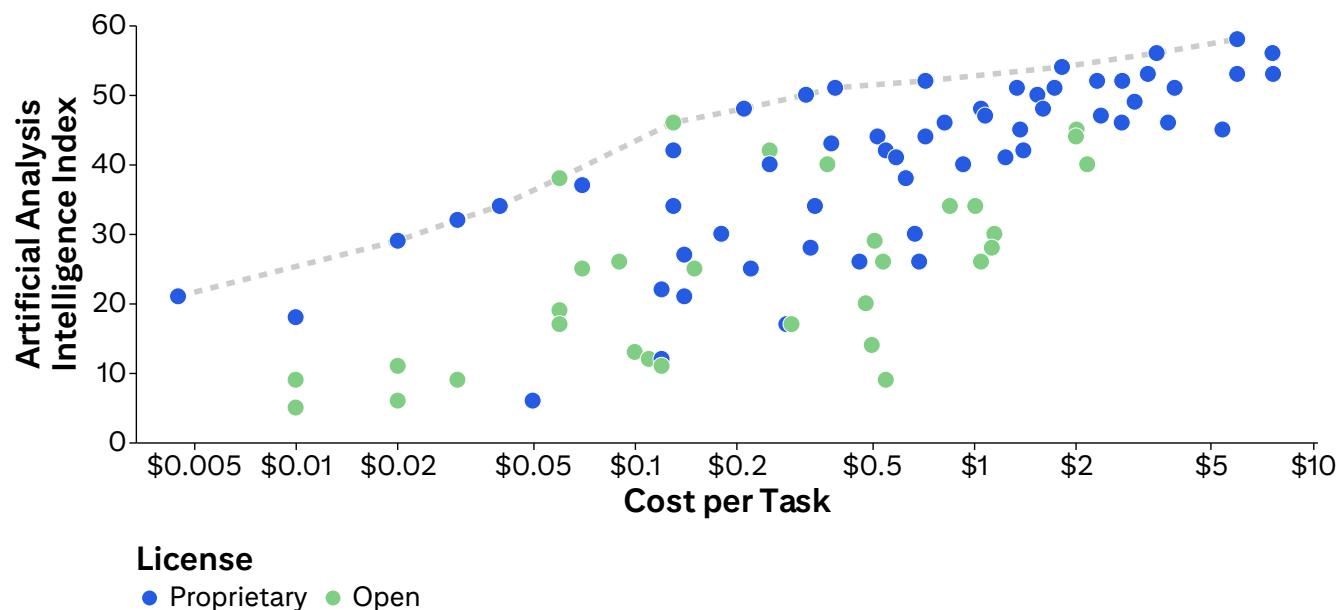


© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Artificial Analysis

OpenAI's GPT-6 Astra appears to provide an illustration of this dynamic. Based on [AA data](#) (AA, 9/9), Astra has lowered the cost of obtaining frontier-level tasks relative to its competitors, reinforcing that the frontier competes on both absolute intelligence and the rate at which intelligence becomes cheaper. Astra matches the AA intelligence score achieved by Fable 5.1 at ~40% of the cost per AA task and at ~60% of the cost per AA task in the AA Coding Agent Index. Today, compared to Fable 5.1, Opus 5.5 offers an additional 5 points in intelligence at 78% the cost per task (Figure 15).

Figure 15. AA Intelligence Index v. Cost per AA Intelligence Index Task by License



© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Artificial Analysis

"At the end of the day, for enterprises, our view is you can't just measure the cost of tokens by the dollars per token consumed. Not every token is the same. You've got to measure it off the actual work done."

- [Wenwen Su, Genspark Co-Founder and CEO, 9/8](#)

Why Harnesses and Proprietary Context Matter

AlphaSense observed a related dynamic in financial research. Kimi K3 was cheaper per token but [more expensive per question](#) than GPT-5.6 Sol because it consumed more tokens retrieving and assembling context (AlphaSense, 8/5). AlphaSense also found that pairing the same frontier model with its search stack to harness it reduced cost per question by roughly 3x and improved answer preference by more than 2:1 versus vector retrieval, while task-level routing increased preference to 2.8:1 (AlphaSense, 8/5).

Together, **these results suggest that realized enterprise economics depend on the entire AI stack, not just model pricing.** A harness is the software layer that surrounds a model, including search, retrieval, tools, memory, routing, and workflows. Proprietary context refers to an organization's unique internal data, documents, and knowledge that are unavailable to the model by default. If model costs decline and performance converges, factors such as retrieval quality, orchestration, routing, proprietary context, and workflow-specific harnesses become increasingly important determinants of both cost and value capture (Figure 16).

Figure 16. Expected Impact of Rising Open-Weight Adoption Across the AI Stack

AI Stack Layer	Effect of Open-Weight Shift	Reasoning
Enterprise Customers	Beneficiary	Greater deployment flexibility and model choice. Improved control over intelligence-per-dollar spent.
Tooling and Harnesses	Strong Beneficiary	Greater model fragmentation increases demand for routing, orchestration, and model-management layers that abstract differences between providers.
Compute Providers (Hyperscalers, Neoclouds)	Uneven Beneficiary	Open-weight adoption expands the set of models available for hosting, creating opportunities to capture those inference workloads. Hyperscalers likely experience greater tailwinds compared to neoclouds as they typically provide wider enterprise support offerings (compliance, data security).
Data Center Infrastructure	Beneficiary	Higher token volumes processed leads to greater demand for GPUs, memory, storage, networking, energy, and other infrastructure. Higher token consumption by some open-weight models could provide a further tailwind.
Closed Model Providers	Selectively Challenged	First-order effect: Decline in API token volume from open-weight competition. Second-order effect: Retained or increased pricing power on frontier workloads and orchestration functions.
Open Model Providers	Uncertain	Increased revenue from subscriptions and first-party API, but widespread third-party hosting limits their ability to directly monetize all inference activity generated by their models.

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research

Total Cost of Ownership Extends Beyond Headline Token Prices

- **Model and compute.** Token pricing (including input, output, and cached tokens) is only one component of inference economics. Infrastructure efficiency (including accelerator utilization, memory requirements, latency, and serving performance) can materially influence the realized cost of producing a response.
- **Harness and context.** Retrieval quality, tool use, memory, routing, and verification influence both answer quality and tokens-to-completion, making orchestration and governance layers important drivers of realized task costs. The application harness and context supplied to the model often determine how effectively model intelligence is converted into task completion, influencing both output quality and total inference costs.
- **Accuracy-adjusted cost.** A cheap failed attempt is not cheap. The relevant economic denominator is a successful, business-useful outcome, not the cost of a single model invocation.
- **Risk and operations.** Security, monitoring, compliance, support, legal indemnification, and downtime can materially affect enterprise procurement decisions and TCO. Broader certification coverage among hyperscaler cloud providers and closed-model providers may reduce diligence and deployment friction, particularly for regulated workloads (Figure 17). This advantage could support adoption through managed cloud platforms even when open-weight models offer lower headline prices or greater control.

Figure 17. Compliance and Regulatory Certifications Comparison Across Compute Providers

	Certification	SOC 2 Type II	ISO/IEC 27001	ISO/IEC 27701	HIPAA/BAA	FedRAMP	GDPR	CSA STAR
	Significance to Enterprises	An independent audit report that provides assurance about a service organization's controls, based on the AICPA's Trust Services Criteria. It proves that security controls are operationally effective over a period of time.	A standard that enables organizations to manage the security of assets such as financial information, intellectual property, and employee details. It demonstrates a systematic, risk-based approach to protecting data.	A standard that specifies requirements for establishing, implementing, and maintaining a Privacy Information Management System (PIMS). It serves as an extension to ISO 27001 for managing personal data, helping organizations comply with privacy regulations like GDPR.	The HIPAA Privacy Rule requires a signed contract, or Business Associate Agreement (BAA), to ensure a vendor will appropriately safeguard protected health information.	A U.S. government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services.	A legal framework that sets guidelines for the collection and processing of personal information from individuals in the EU. It grants individuals rights over their data and imposes strict rules on data transfers.	A publicly accessible registry of the security and privacy measures that cloud computing firms offer.
Hyperscalers	Microsoft Azure	Achieved	Achieved	Achieved	BAA available for covered entity and business associate customers	Class D (High) certification	Aligns with GDPR in areas including data processing, security, and transfers	STAR Level One (Self-Assessment) and STAR Level Two (Third-Party Audit)
	Amazon AWS	Achieved	Achieved	Achieved	BAA available	AWS GovCloud: Class D (High) certification AWS US East/West: Class C (Moderate) certification	Aligns with GDPR in areas including data processing, security, and transfers	STAR Level One (Self-Assessment) and STAR Level Two (Third-Party Audit)
	Google GCP	Achieved	Achieved	Achieved	BAA available	Class D (High) certification	Aligns with GDPR in areas including data processing, security, and transfers	STAR Level One (Self-Assessment) and STAR Level Two (Third-Party Audit)
Neoclouds	CoreWeave	In Progress	In Progress	In Progress	In Progress	In Progress	Aligns with GDPR in areas including data processing, security, and transfers	No public evidence
	Nscale	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence	Aligns with GDPR in areas including data processing, security, and transfers	No public evidence
Model Companies	xAI (Grok)	Achieved	No public evidence	No public evidence	BAA available under some circumstances	Agency authorization in progress for a Class D (High) certification	Aligns with GDPR in areas including data processing, security, and transfers	STAR Level 1 (Self-Assessment)
	Anthropic (Claude)	Achieved	Achieved	No public evidence	HIPAA-ready Enterprise Offerings available since Jun. 26	Class D (High) certification	Aligns with GDPR in areas including data processing, security, and transfers	STAR Level 2 (Third-Party Audit)
	OpenAI (ChatGPT)	Achieved	Achieved	Achieved	BAA available for certain offerings	Class C (Moderate) certification	Aligns with GDPR in areas including data processing, security, and transfers	STAR Level 1 (Self-Assessment)
	Google (Gemini)	Achieved	Achieved	Achieved	HIPAA-ready offerings available	Gemini Enterprise and Gemini App: Class D (High) certification Gemini for Government: Class B (low) certification	Aligns with GDPR in areas including data processing, security, and transfers	STAR Level 1 (Self-Assessment) and STAR Level 2 (Third-Party Audit)
	Cursor	Achieved	In Progress	No public evidence	HIPAA Business Associate Agreements (BAAs) for Enterprise customers	No public evidence	Aligns with GDPR in areas including data processing, security, and transfers	No public evidence
Open-Weight Model Providers	DeepSeek	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence
	Z.ai (GLM)	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence
	Moonshot (Kimi)	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence	No public evidence
	Mistral	Achieved	Achieved	Achieved	No public evidence	No public evidence	Aligns with GDPR in areas including data processing, security, and transfers	No public evidence
	Cohere	Achieved	Achieved	No public evidence	BAA available for custom model engagements	No public evidence	Aligns with GDPR in areas including data processing, security, and transfers	No public evidence

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Provider Websites

Control, Customization, and the Enterprise Portfolio

Furthermore, **cost is only one driver of open-model adoption**. Downloadable weights let enterprises deploy inside a private cloud or on-premises environment, retain more control over data handling, fine-tune or post-train on proprietary examples, and reduce dependence on a single API by building on top of a router

(Figure 18). [These benefits](#) are most important where data sensitivity, latency, sovereignty or model specialization outweigh the convenience of a fully managed frontier service (NVIDIA, 7/24).

Figure 18. Enterprise Examples of Open-Weight Adoption

Company	Industry	Market Cap/Valuation/AUM	Nature of Adoption	Model(s) Adopted	KPI
Netflix	Media & Entertainment	\$290B	Integrated open-weight models into internally-built LLM serving platform	Not specified	Lower inference costs, lower latency, greater data privacy
Siemens	Industrials	\$239B	Use open-source models in addition to frontier LLMs	Not specified	Cheaper tokens as a result of open source model training
Shopify	E-commerce	\$191B	Fine-tuned an open-weight model on merchant data to serve as an AI commerce assistant	Qwen3-32B	22x faster, 68% cheaper, and higher accuracy than closed-source model
AT&T	Telecom Services	\$167B	AT&T now runs about 40% of its AI workloads on open models and aims to reach 70% within a year, according to chief data and AI officer Andy Markus. Processing 45 billion tokens a day	Not specified	Took open models and tuned them using company's proprietary data so they performed as well or better than closed alternatives for specific tasks
Uber	Consumer Internet	\$142B	Post-trained and fine-tuned open-weight models for use in item tagging and product recommendations	Meta Llama 2, Mistral	Lower cost than frontier models while achieving similar performance
Accenture	IT Services & Consulting	\$108B	Helping enterprises adopt a multi-model portfolio for governed, self-hosted, and hybrid deployments	Open-weight LLMs alongside proprietary models	~50% reduction in research time
Airbnb	Consumer Internet	\$95B	A minority of workflows are routed through approved US-based service providers, particularly in customer service	Qwen and other models of Chinese origin	Increases in model affordability Improvement in average customer service resolution times from three hours to six seconds
PNC Financial Services	Financial Services	\$88B	In addition to using frontier models with hyperscalers, using open weight models in their data centers with no data sharing and doing in-house development of agentic harnessing capabilities as it appeals to high control nature of their data and is more efficient	Not specified	AI driving 40% more efficiency in software development. With open weight models, big difference in efficiencies in terms of focusing on the tasks which might not require the highest cost approach, which often vendors cannot do
DoorDash	Consumer Internet	\$80B	Paired an open-weight code reviewer with a closed-weight frontier model to verify the accuracy of reviews	Kimi K2.6	No change in task accuracy while decreasing cost per task
SLB	Oilfield Services	\$72B	Integrating open-weight models into its digital and AI platforms and jointly developing an "AI Factory for Energy" for domain-specific AI workloads	NVIDIA Nemotron	Intended to accelerate conversion of operational data into actionable insights while improving performance, efficiency, and deployment scalability
Coinbase	Financial Technology	\$49B	Set open-weight models as the default models for developers	GLM-5.2, Kimi K2.7	Decrease in cost per task for developers
CH Robinson	Logistics	\$17B	For simple transactional tasks, use open source or older versions of frontier models	Proprietary open source LLMs	Generated hundreds of millions of USD of operating income value since the end of 2022 while spending less than \$1.2 million on tokens on annualized basis
Hugging Face	Software	Nvidia acquired for ~\$13B	Used in detecting and defending against autonomous cybersecurity attacks	GLM-5.2	Addressed cybersecurity defense tasks that closed-weight models were unable to process due to guardrails
Match Group	Consumer Internet	\$9B	Tinder began routing some queries from non-technical users to open-weight models as the app worked to get control of rising AI costs	Not specified	Not specified
Bridgewater Associates	Financial Services	~\$150B AUM	Training an AI agent on internal data to perform the duties of an investor	Qwen (as base model)	Produced a higher average task accuracy at a lower cost per task, compared to off-the-shelf closed-weight models

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Note: FactSet market data as of 9/30/26.

Source: Citi Research, Company Websites, AlphaSense, Forbes, CEOs via X, PitchBook, FactSet

“There’s a new repository created every 7 seconds on the platform. That’s almost 3M public models and 1M public datasets that have been shared on the platform. It shows a little bit of a different picture than ‘one model to rule them all.’ [...] In reality, we’re seeing more and more companies using lots of different models with a lot of specialized, customized models for their needs for their specific use cases. [...] In terms of adoption by enterprises, now we have half the Fortune 500 using Hugging Face and so using either their own private models or open-source models.”

– Clem Delangue, Hugging Face CEO, 7/10

The Likely Steady State Is Mixed

Routine, high-volume and domain-specific work may move to cheaper open or smaller systems. Frontier models currently retain the most difficult, long-horizon, and high-error-cost tasks, while also serving as verifiers and supervisors. Routers and harnesses become strategically important because no model currently wins every task on both performance and cost (Figure 19).

Figure 19. Enterprise Examples of Routing

Company	Industry	Market Cap/Valuation	Nature of Routing	Adopted Open Models	KPI
Siemens	Industrial Technology	\$239B	Use off-the-shelf models and post-trained open weight models	Yes	Cost benefits from cheaper tokens of open-weight models
Shopify	E-commerce	\$191B	Choose between smaller models and frontier models	Yes	Optimize inference costs potentially more effectively than other industry players
Uber	Consumer internet	\$142B	Evaluated models and deploy the best option for each use case	Yes	Decline in cost per token despite a 4x increase in the number people frontier model users
Booking	Consumer internet	\$122B	Use cheaper models for simple tasks and more expensive models for complex tasks	Not specified	Meaningful decreases in AI cost per merge request
DoorDash	Consumer internet	\$80B	Open-weight model handles simpler tasks while frontier model handles reasoning	Yes	Higher performance than frontier model pair with a similar cost
Cognition	Software	~\$48B (Private)	Frontier model delegates tasks to cheaper model for execution	Not specified	Frontier-level performance at 35% lower cost
Figma	Software	\$11B	Route across models based on task complexity	Yes	Deliver comparable quality at a lower cost and latency
Zillow	Consumer internet	\$6B	Route queries to the right model for the task, rather than using the most expensive model for every task	Not specified	Increases in productivity and manageability of AI cost structure
Blackline	Software	\$2B	Use LLMs where reasoning is needed and lighter-weight / open-source models for transactional tasks	Yes	Enables quick iteration and avoids vendor lock-in while maintaining high accuracy

© 2026 Citigroup Inc. No redistribution without Citigroup’s written permission.

Note: FactSet market data as of 9/30/26.

Source: Citi Research, Company Websites, AlphaSense, CEOs via X, PitchBook, FactSet

Open-weight adoption could redistribute value across the AI stack even if total enterprise spending remains unchanged (Figure 20). In an illustrative \$100 spending waterfall, moving from a closed-model deployment to a fully open-weight deployment reduces model-provider value-add from \$50 to \$5 as the weights are assumed to carry no royalty (though the developers likely provide ancillary services such as fine-tuning or enterprise support), while increasing value-add captured by cloud and inference providers from \$33 to \$62. The amount flowing to data-center infrastructure rises from approximately \$17 to \$33, increasing associated gross profit from \$13 to \$25 under a constant ~75% infrastructure gross-margin assumption. A 50/50 model portfolio sits between these outcomes, supporting our view that adoption will be mixed rather than binary.

Figure 20. Illustrative AI Economics Across Closed and Open-Weight Model Mixes

Illustrative Enterprise AI Spend Waterfall				
AI Stack Layer	Spend Allocation	Closed Model	50/50 Open-Closed	Open Model
Enterprise	AI Spend	100	100	100
Model Provider (Lab)	To CSPs/Neoclouds	50	73	95
	Retained Economics	50	28	5
Hosting, Inference & Orchestration	Spend Received	50	73	95
	To DC Infrastructure	17	25	33
	Retained Economics	33	47	62
Data Center Infrastructure	Infrastructure Revenue	17	25	33
	Gross Margin	75%	75%	75%
	Gross Profit	13	19	25

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

The closed-model scenario assumes 50% of model-provider revenue is paid to compute providers. The open-weight scenario assumes an Apache 2.0 or MIT-style license under which cloud and inference providers pay no model royalty. Cloud infrastructure cost is estimated at ~35% of revenue using estimates of \$8B of annual infrastructure cost and \$23B of annual revenue per gigawatt. Data center infrastructure gross profit applies a ~75% FY27E NVIDIA gross-margin estimate. Orchestration value is included in cloud and inference-provider value-add. Figures are illustrative and are not forecasts.

Source: Citi Research

The exercise is directional and intended to illustrate how value capture may shift across the AI stack as model weights become more widely available; actual economics will vary by licensing model, competitive dynamics, and the degree to which model developers monetize APIs, software and support services. Nonetheless, it illustrates an important implication. **As model weights commoditize, margins are likely to migrate toward the platforms that host, secure, route, and operate the models and toward the infrastructure required to serve the resulting token volume.**

The Road Ahead

Knowledge distillation is being leveraged in a variety of ways. For instance, Google has historically leveraged its proprietary Gemini family to distill capabilities into its open Gemma models (Google DeepMind, 3/12/25). Google's use of Gemini as a teacher for Gemma also illustrates that leading frontier labs can leverage their proprietary advantage to continuously seed increasingly capable open-weight models, allowing them to participate in both closed-model monetization and open-weight ecosystem adoption simultaneously.

Current disputes also allege campaigns used fraudulent accounts, identity evasion, and automated high-volume querying to [extract capabilities](#) contrary to provider terms and access controls (Anthropic, 2/23; Figure 21). More recently, Anthropic detected a handful of Eastern model developers engaging in what Anthropic alleges as [illicit distillation](#) (which they define as industrial-scale, covert campaigns to extract a model's capabilities and replicate them in another model without authorization) (Anthropic, 9/10).

Figure 21. Examples of Distillation Allegations

Selected Distillation Allegations				
Date	Frontier Lab	Open-Weight Lab	Alleged Activities	Source
January 2025	OpenAI	DeepSeek	OpenAI and Microsoft investigated and blocked user accounts which appeared to use OpenAI's API for distillation efforts, suspected to originate from DeepSeek.	OpenAI Press Release, Financial Times
February 2026	OpenAI	DeepSeek	OpenAI said it had observed activity on its platform indicative of continued attempts by DeepSeek to distill frontier models, including through obfuscated third-party access methods. OpenAI characterized the activity as ongoing and described steps taken to harden its models against distillation.	OpenAI Submission to the U.S. House Select Committee
February 2026	Anthropic	DeepSeek, Moonshot AI, Minimax	Anthropic accused the three Chinese AI labs of generating exchanges with Claude through approximately 24,000 fraudulent accounts. DeepSeek allegedly conducted 150,000 exchanges focusing on reasoning and rubric-based grading tasks, Moonshot AI allegedly conducted 3.4 million exchanges focusing on agentic reasoning and coding, and MiniMax allegedly conducted 13 million exchanges, focusing on agentic coding and orchestration.	Anthropic Press Release
April 2026	OpenAI	xAI	During federal court testimony, Elon Musk was asked whether xAI had used distillation techniques on OpenAI models to train its own systems. Musk initially said that AI companies generally use the practice and, when asked whether that meant yes, responded, "Partly." Musk also described using other AI systems to validate a model as standard industry practice.	Court Testimony
June 2026	Anthropic	Alibaba	Anthropic alleged that Alibaba's AI lab conducted 28.8 million exchanges with Claude between April and June through approximately 25,000 fraudulent accounts to train its Qwen model.	Bloomberg
July 2026	Anthropic	Moonshot AI	White House Office of Science and Technology Policy (OSTP) Director Michael Kratsios claimed to possess evidence that Moonshot AI distilled Claude Fable 5 to train its Kimi K3 model.	Michael Kratsios via X
September 2026	Anthropic	Alibaba, Moonshot AI, DeepSeek, Xiaomi, SenseTime, MiniMax	Alibaba, Moonshot AI, DeepSeek, Xiaomi, SenseTime, and MiniMax are alleged to have covertly harvested Claude's outputs at massive scale through fraudulent accounts, proxy networks, and third-party data brokers to illegally train their own competing AI models. The scale of distillation attacks were >151 million exchanges observed for Alibaba between May and July 2026, >23 million exchanges for Moonshot between May and July 2026, >12.1 million exchanges for DeepSeek over 14 days in July 2026, >3.4 million exchanges for Zhipu over 17 days in June and July 2026, >400K exchanges for Xiaomi over 20 days in March and April 2026.	Anthropic Report

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Company Websites, Court Testimony, Bloomberg, U.S. OSTP Director Michael Kratsios via X

Frontier Providers Are Making Distillation More Difficult

- **Access to reasoning traces has tightened in stages.** OpenAI concealed raw chain-of-thought when o1 launched in September 2024. Early Gemini reasoning outputs remained detailed enough for the [Stanford-led s1 project](#) (3/1/2025) to fine-tune Qwen2.5-32B using only 1,000 selected traces. By May 2025, Google had introduced thought summaries for Gemini 2.5, while Anthropic's Claude 4 similarly provided summarized rather than complete reasoning.
- **Frontier APIs increasingly preserve reasoning without revealing it.** Anthropic returns an encrypted copy of [Claude's full reasoning](#) (9/1) in a signature field, while Google uses encrypted [thoughtSignature](#) objects (9/2). Applications pass those objects back unchanged to preserve reasoning context across turns, but developers ordinarily cannot inspect the underlying plaintext.
- **Encrypted reasoning was exploitable at relatively low cost.** Encrypted blocks could be replayed across compatible models and sessions, allowing a weaker model to transcribe reasoning produced by a stronger model without breaking the encryption. Researchers from Stolen Thoughts estimated that using Claude Haiku 4.5 to extract 10,000 traces generated by Claude Opus 4.8 would cost approximately [\\$720](#) (8/10).
- **Providers are now hardening encrypted reasoning against reuse.** Anthropic's [Table 5.1 prevents earlier models from reading its thinking blocks](#) (9/1) and binds those blocks to the preceding system prompt, tools, and conversation. Altering the earlier context can therefore invalidate the blocks, closing important routes used in the published extraction method.
- **Identity and payment controls can disrupt distributed extraction.** Anthropic reported that DeepSeek, Moonshot, and MiniMax generated [more than 16M exchanges](#) (2/23) through approximately 24K fraudulent accounts. Stronger KYC, account linkage, payment-pattern analysis, and regional-access enforcement could make comparable campaigns harder to establish and maintain.
- **Traffic analysis can identify coordinated campaigns.** Providers can monitor synchronized querying, repeated capability probes, rubric harvesting, geographic concentration, and related patterns that distinguish systematic extraction from ordinary API use. Providers characterized the campaigns they detected as industrial-scale and increasingly sophisticated.
- **Request redirection can reduce the value of collected outputs.** Anthropic's [Table 5](#) classifiers redirect requests related to distillation and certain other sensitive domains to the less capable Claude Opus 4.8 model, directly limiting access to its highest-value capabilities. Anthropic reported that such fallbacks occurred in [fewer than 5%](#) (6/9) of early sessions overall.
- **Rate limits and selective access can constrain extraction volume.** Throughput restrictions, enhanced customer vetting, and trusted-access programs can reduce the quantity and quality of traces available to unauthorized actors. For instance, collaborative cybersecurity initiatives like Project Glasswing pair general access to safeguarded models with selective access to more capable models with fewer guardrails.
- **Blocked reasoning-extraction attempts can still carry a cost.** Since September 24, 2026, Anthropic has resumed [billing standard model rates](#) (9/28) for refusals

classified as reasoning_extraction, increasing the cost of repeated probing even when the requested reasoning is not returned.

- **Developers that rely on distillation from frontier models may therefore require more internal compute and infrastructure.** Greater reliance on reinforcement learning, synthetic-data generation, evaluators, and proprietary training environments could raise the cost of replicating frontier capabilities, potentially widening the closed-open gap following major proprietary releases. However, historical evidence suggests the gap fluctuates with pretraining cycles rather than widening consistently.

A Near-Term Re-Widening Is Not a Permanent Moat

If distillation materially accelerated the catch-up in intelligence we saw from May-August, restricting access to high-value frontier outputs might help widen the capability gap going forward. Frontier systems would continue to advance while open developers face more difficulty capturing those improvements, especially for models not yet trained. The likely effect is to reduce the quality, scale, and reliability of teacher outputs, increasing the cost and time required to transfer newly developed capabilities.

Even so, defenses are unlikely to stop open progress. Developers can rely more heavily on independent pre-training, human-curated data, already harvested gains from distillation, reinforcement learning, synthetic environments, data selection, sparse architectures, quantization, and systems engineering. Chip restrictions similarly act as a speed and cost constraint rather than a permanent barrier.

We expect the gap to oscillate absent some material breakthrough in RSI or other independent breakthrough. Closed labs will periodically reset the frontier, and open labs will narrow the released gap through pre-training, efficiency and faster iteration. That said, conditions that would support a structurally compressed capability gap would likely be that open labs increasingly reproduce frontier capability through independent training and ecosystem scale, which could be accelerated via breakthroughs in RSI ([Z.ai](#), 9/13), however we don't see persistent evidence of these conditions at present.

A critical question is not whether the open-closed gap widens or narrows at any single point, but how long any advantage endures and where the resulting economic value is ultimately captured. In an environment where model-level advantages are increasingly short lived, durable value may reside less in any individual model than to the infrastructure, proprietary data, evaluation systems, retrieval layers, and routing platforms that can continuously direct each task to the best combination of capability, control, and cost.

The pace of regulatory development will remain a key variable. Until a [clearer framework emerges](#) (The Information, 8/27), approval of frontier-advancing releases may depend less on eliminating misuse than on demonstrating sufficiently robust safeguards, including jailbreak defenses and controlled-access programs. Longer review periods could give open models more time to approach or even overcome the released frontier, while controlled access could preserve premium monetization opportunities and a capability advantage for trusted users, including cybersecurity teams working to identify and patch vulnerabilities.

Figure 22. Recent Notes from Our Team

Category	Date	Title
AI Industry Model	September 3, 2026	Artificial Intelligence: AI Industry Model Update
Deep Dives	September 7, 2026	Space Exploration Technologies (SPCX.O): Potential for \$900+; Initiating at Buy
Deep Dives	April 27, 2026	Artificial Intelligence: Anthropic - Driving the Acceleration of Enterprise AI
Deep Dives	January 21, 2026	Artificial Intelligence: OpenAI - Architecting the Abstraction Layer for Everything
Deep Dives	August 25, 2026	Artificial Intelligence: AI - The Information Era's Apex Technology
ROI on AI	August 4, 2026	Artificial Intelligence: ROI on AI - Consensus Is Still Underestimating AI Revenues & Returns
ROI on AI	April 30, 2026	Artificial Intelligence: ROI on AI - Backlog Acceleration Drives Continued Confidence in Investment Returns
ROI on AI	February 6, 2026	Artificial Intelligence: ROI on AI - Growth Acceleration & Returns Drive Capex Revisions
Citi's Inference Ahead	September 26, 2026	Artificial Intelligence: Citi's Inference Ahead - The Efficiency Dividend
Citi's Inference Ahead	September 18, 2026	Artificial Intelligence: Citi's Inference Ahead - Where Intelligence Compounds
Citi's Inference Ahead	September 14, 2026	Artificial Intelligence: Citi's Inference Ahead - Reasserting the Western Frontier
Citi's Inference Ahead	September 4, 2026	Artificial Intelligence: Citi's Inference Ahead - The Efficiency Frontier
Citi's Inference Ahead	August 31, 2026	Artificial Intelligence: Citi's Inference Ahead - The Intelligence Flywheel
Citi's Inference Ahead	August 21, 2026	Artificial Intelligence: Citi's Inference Ahead - Investing in Open Models
Citi's Inference Ahead	August 14, 2026	Artificial Intelligence: Citi's Inference Ahead - Holding Back the Frontier
Citi's Inference Ahead	August 7, 2026	Artificial Intelligence: Citi's Inference Ahead - From Buildout to Monetization
Citi's Inference Ahead	July 31, 2026	Artificial Intelligence: Citi's Inference Ahead - Pricing, Financing, Routing, and Securing
Citi's Inference Ahead	July 24, 2026	Artificial Intelligence: Citi's Inference Ahead - Constraints v. Capabilities
Citi's Inference Ahead	July 19, 2026	Artificial Intelligence: Citi's Inference Ahead - Open Reaches the Frontier
Citi's Inference Ahead	July 10, 2026	Artificial Intelligence: Citi's Inference Ahead - Models, Robots, & Captives
Citi's Inference Ahead	June 26, 2026	Artificial Intelligence: Citi's Inference Ahead - Open-Source Moment
Citi's Inference Ahead	June 19, 2026	Artificial Intelligence: Citi's Inference Ahead - Open v. Closed
Citi's Inference Ahead	June 14, 2026	Artificial Intelligence: Citi's Inference Ahead - Bridging the Gaps
Citi's Inference Ahead	June 10, 2026	Artificial Intelligence: Citi's Inference Ahead - Rationing Comes for Us All
One-Off Pieces	September 14, 2026	Artificial Intelligence: Pacing Calls Attempt to Shape Regulation
At the Edge	June 11, 2026	Artificial Intelligence: At the Edge - Verification, Agents, and the Infrastructure Around Them
Commentary on Conferences and Events	September 10, 2026	Artificial Intelligence and Systems Software: Enterprise AI: Island - Day 3 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 10, 2026	Artificial Intelligence: Orbital Compute: Starcloud - Day 3 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 10, 2026	Artificial Intelligence: Context for Enterprise AI: Celonis - Day 2 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 9, 2026	Artificial Intelligence: AI's Open Models: Poolside - Day 2 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 9, 2026	Artificial Intelligence: AI's Evolution: Perplexity - Day 2 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 9, 2026	Artificial Intelligence: Powering AI: Commonwealth Fusion Systems - Day 2 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 9, 2026	Artificial Intelligence: State of the Market - Day 2 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 8, 2026	Artificial Intelligence: Uniphore - Day 1 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 8, 2026	Artificial Intelligence: Genspark and Cresta Intelligence - Day 1 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 8, 2026	Artificial Intelligence: Implementing AI & Robotics: Wonder - Day 1 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 8, 2026	Artificial Intelligence: Enterprise Technology Leadership Keynote - Day 1 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	August 31, 2026	Global Technology & Communications: Citi's 2026 Global TMT Conference Preview - Key Tech Themes to Watch
Commentary on Conferences and Events	July 8, 2026	Artificial Intelligence: AI Insights from Citi's Space & AI Virtual Teach-In
Commentary on Conferences and Events	July 8, 2026	Artificial Intelligence: Bottlenecks Remain for Scaling Physical AI, But Commercialization Progressing
Commentary on Conferences and Events	June 15, 2026	Artificial Intelligence: Takeaways from the Research and Applied AI Summit
Collaborative Pieces	September 29, 2026	Space Exploration Technologies (SPCX.O): Starship Flight 14 Top Takes
Collaborative Pieces	September 14, 2026	Continual Learning to Fuel HBM/eSSD Demand (AI Training/Inference)
Collaborative Pieces	September 3, 2026	Multi-Asset: Data Center GAINS (Gen AI Names) - Born to Run: Still Early in the Multi-Year AI Investment Cycle
Collaborative Pieces	August 9, 2026	Space Exploration Technologies (SPCX.O): 2Q26 Model Update; Reiterate \$200 Price Target
Collaborative Pieces	August 6, 2026	Takeaways on Starlink's Mobile Strategy & Implications for the Category
Collaborative Pieces	August 4, 2026	Space Exploration Technologies (SPCX.O): 2Q26 Strong Start to Multi-Year Catalyst-Rich Arc
Collaborative Pieces	July 30, 2026	US Application Software: 2026 CFO Survey - Measured Optimism with Clear Commitment to AI
Collaborative Pieces	July 19, 2026	EM Equity Strategy Compass: 2H26 Outlook - The Broadening Question
Collaborative Pieces	July 10, 2026	Global Equity Quarterly: Back to the Broadening Debate
Collaborative Pieces	July 1, 2026	North America: Road Ahead 2H26
Collaborative Pieces	July 1, 2026	North America Internet: Meta Compute Unlocks Capacity (and FCF) for Meta's Superintelligence Ambitions

© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research

Space Exploration Technologies

(SPCX.O; US\$149.24; 1; 29 Sep 26; 16:00)

Valuation

We derive our \$200 base-case price target for SPCX using the average of three valuation methods: 1) 2027E growth adj. multiples (EV/Revenue/Growth and EV/EBITDA/Growth) for the “Trillion-Dollar Comps” (a basket of similar trillion-dollar+ market caps); 2) sum-of-the-parts (SOTP) valuing Space, Connectivity, and AI separately using 2027E growth adj. multiples (EV/Revenue/Growth and EV/EBITDA/Growth) for industry-specific comparable companies; and 3) 2030E comparable company multiples (EV/Revenue and EV/EBITDA) for the “Trillion-Dollar Comps” (a basket of similar trillion dollar+ market caps).

Risks

Risks to our thesis, which could impede the shares from reaching our price target, include headwinds to the launch segment which reference the failure to demonstrate rapid reusability on Starship, lack of launch infrastructure, and regulatory headwinds from the FAA and other government organizations. The failure to rapidly gain market share in Starlink mobile, as well as demonstrate the ability to create AI orbital satellites capable of compute, also presents risks to SPCX.

If you are visually impaired and would like to speak to a Citi representative regarding the details of the graphics in this document, please call USA 1-888-500-5008 (TTY: 711), from outside the US +1-210-677-3788

Appendix A-1

ANALYST CERTIFICATION

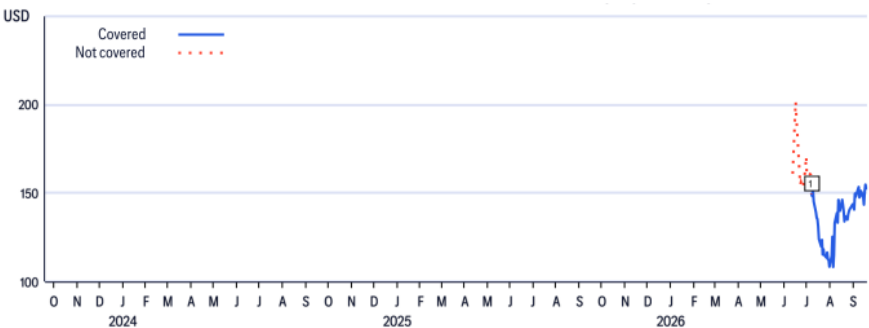
The research analysts primarily responsible for the preparation and content of this research report are either (i) designated by “AC” in the author block or (ii) listed in bold alongside content which is attributable to that analyst. If multiple AC analysts are designated in the author block, each analyst is certifying with respect to the entire research report other than (a) content attributable to another AC certifying analyst listed in bold alongside the content and (b) views expressed solely with respect to a specific issuer which are attributable to another AC certifying analyst identified in the price charts or rating history tables for that issuer shown below. Each of these analysts certify, with respect to the sections of the report for which they are responsible: (1) that the views expressed therein accurately reflect their personal views about each issuer and security referenced and were prepared in an independent manner, including with respect to Citigroup Global Markets Inc. and its affiliates; and (2) no part of the research analyst's compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed by that research analyst in this report.

IMPORTANT DISCLOSURES

Space Exploration Technologies (SPCX)

Ratings and Target Price History
Fundamental Research

Analyst: John Godyn



Date	Rating	Target Price	Closing Price
07-Jul-26 05:00:00	*1	*200.00	149.47

*Indicates Change

Rating/target price changes above reflect Eastern Time

Within the past 12 months, Citigroup Global Markets Inc. or its affiliates has acted as manager or co-manager of an offering of securities of Space Exploration Technologies.

Citigroup Global Markets Inc. or its affiliates has received compensation for investment banking services provided within the past 12 months from Space Exploration Technologies.

Citigroup Global Markets Inc. or its affiliates received compensation for products and services other than investment banking services from Space Exploration Technologies in the past 12 months.

Citigroup Global Markets Inc. or its affiliates currently has, or had within the past 12 months, the following as investment banking client(s): Space Exploration Technologies.

Citigroup Global Markets Inc. or its affiliates currently has, or had within the past 12 months, the following as clients, and the services provided were non-investment-banking, securities-related: Space Exploration Technologies.

Citigroup Global Markets Inc. or its affiliates currently has, or had within the past 12 months, the following as clients, and the services provided were non-investment-banking, non-securities-related: Space Exploration Technologies.

Citigroup Global Markets Inc. and/or its affiliates has a significant financial interest in relation to Space Exploration Technologies. (For an explanation of the determination of significant financial interest, please refer to the policy for managing conflicts of interest which can be found at www.citiVelocity.com.)

Analysts' compensation is determined by Citi Research management and Citigroup's senior management and is based upon activities and services intended to benefit the investor clients of Citigroup Global Markets Inc. and its affiliates (the "Firm"). Compensation is not linked to specific transactions or recommendations. Like all Firm employees, analysts receive compensation that is impacted by overall Firm profitability which includes investment banking, sales and trading, and principal trading revenues. One factor in equity research analyst compensation is arranging corporate access events between institutional clients and the management teams of covered companies. Typically, company management is more likely to participate when the analyst has a positive view of the company.

For financial instruments recommended in the Product in which the Firm is not a market maker, the Firm is a liquidity provider in such financial instruments (and any underlying instruments) and may act as principal in connection with transactions in such instruments. The Firm is a regular issuer of traded financial instruments linked to securities that may have been recommended in the Product. The Firm regularly trades in the securities of the issuer(s) discussed in the Product. The Firm may engage in securities transactions in a manner inconsistent with the Product and, with respect to securities covered by the Product, will buy or sell from customers on a principal basis.

Unless stated otherwise neither the Research Analyst nor any member of their team has viewed the material operations of the Companies for which an investment view has been provided within the past 12 months.

For important disclosures (including copies of historical disclosures) regarding the companies that are the subject of this Citi Research product ("the Product"), please contact Citi Research, 388 Greenwich Street, 6th Floor, New York, NY, 10013, Attention: Legal/Compliance [E6WYB6412478]. In addition, the same important disclosures, with the exception of the Valuation and Risk assessments and historical disclosures, are contained on the Firm's disclosure website at https://www.citivelocity.com/cvr/eppublic/citi_research_disclosures. Valuation and Risk assessments can be found in the text of the most recent research note/report regarding the subject company. Pursuant to the Market Abuse Regulation a history of all Citi Research recommendations published during the preceding 12-month period can be accessed via Citi Velocity (<https://www.citivelocity.com/cv2>) or your standard distribution portal. Historical disclosures (for up to the past three years) will be provided upon request.

Citi Research Equity Ratings Distribution

Data current as of 01 Jul 2026	12 Month Rating			Catalyst Watch		
	Buy	Hold	Sell	Buy	Hold	Sell
Citi Research Global Fundamental Coverage (Neutral=Hold)	61%	32%	7%	36%	48%	16%
% of companies in each rating category that are investment banking clients	38%	42%	27%	42%	37%	36%

Guide to Citi Research Fundamental Research Investment Ratings:

Citi Research stock recommendations include an investment rating and an optional risk rating to highlight high risk stocks.

Risk rating takes into account both price volatility and fundamental criteria. Stocks will either have no risk rating or a High risk rating assigned.

Investment Ratings: Citi Research investment ratings are Buy, Neutral and Sell. Our ratings are a function of analyst expectations of expected total return ("ETR") and risk. ETR is the sum of the forecast price appreciation (or depreciation) plus the dividend yield for a stock within the next 12 months. The target price is based on a 12 month time horizon. The Investment rating definitions are: Buy (1) ETR of 15% or more or 25% or more for High risk stocks; and Sell (3) for negative ETR. Any covered stock not assigned a Buy or a Sell is a Neutral (2). For stocks rated Neutral (2), if an analyst believes that there are insufficient valuation drivers and/or investment catalysts to derive a positive or negative investment view, they may elect with the approval of Citi Research management not to assign a target price and, thus, not derive an ETR. Citi Research may suspend its rating and target price and assign "Rating Suspended" status for regulatory and/or internal policy reasons. Citi Research may also suspend its rating and target price and assign "Under Review" status for other exceptional circumstances (e.g. lack of information critical to the analyst's thesis, trading suspension) affecting the company and/or trading in the company's securities. In both such situations, the rating and target price will show as "—" and "—" respectively in the rating history price chart. Prior to 11 April 2022 Citi Research assigned "Under Review" status to both situations and prior to 11 Nov 2020 only in exceptional circumstances. As soon as practically possible, the analyst will publish a note re-establishing a rating and investment thesis. Investment ratings are determined by the ranges described above at the time of initiation of coverage, a change in investment and/or risk rating, or a change in target price (subject to limited management discretion). At times, the expected total returns may fall outside of these ranges because of market price movements and/or other short-term volatility or trading patterns. Such interim deviations will be permitted but will become subject to review by Research Management. Your decision to buy or sell a security should be based upon your personal investment objectives and should be made only after evaluating the stock's expected performance and risk.

Catalyst Watch/Short Term Views ("STV") Ratings Disclosure:

Catalyst Watch and STV Upside/Downside calls: Citi Research may also include a Catalyst Watch or STV Upside or Downside call to indicate the analyst expects the share price to rise (fall) in absolute terms over a specified period of 30 or 90 days in reaction to one or more specific near-term catalysts or events impacting the company or the market. A Catalyst Watch will be published when Analyst confidence is high that an impact to share price will occur; it will be a STV when confidence level is moderate. A Catalyst Watch or STV Upside/Downside call will automatically expire at the end of the specified 30/90 day period. The Catalyst Watch will also be automatically removed if share price performance (calculated at market close) exceeds 15% against the direction of the call (unless over-ridden by the analyst). The analyst may also remove a Catalyst Watch or STV call prior to the end of the specified period in a published research note. A Catalyst Watch/STV Upside or Downside call may be different from and does not affect a stock's fundamental equity rating, which reflects a longer-term total absolute return expectation. For purposes of FINRA ratings-distribution-disclosure rules, a Catalyst Watch/STV Upside call corresponds to a buy recommendation and a Catalyst Watch/STV Downside call corresponds to a sell recommendation. Any stock not assigned to a Catalyst Watch Upside, Catalyst Watch Downside, STV Upside, or STV Downside call is considered Catalyst Watch/STV No View. For purposes of FINRA ratings distribution-disclosure rules, we correspond Catalyst Watch/STV No View to Hold in our ratings distribution table for our Catalyst Watch/STV Upside/Downside rating system. However, we reiterate that we do

not consider No View to be a recommendation. For all Catalyst Watch/STV Upside/Downside calls, risk exists that the catalyst(s) and associated share-price movement will not materialize as expected.

Privately held companies:

The legal entities through which Citi conducts its business do not permit investment research recommendations, ratings, target prices, or estimates on any privately held companies ("private companies") referenced herein. Citi Research does not utilize a rating system for private companies. Any information on private companies in this report is based on generally available information and, although obtained from sources believed by Citi to be reliable, its accuracy and completeness are not guaranteed. Private companies are generally not subject to reporting obligations, which may result in less publicly available information. The information contained in this report may differ from information a private company may subsequently make public or otherwise share with investors. Continuation of publishing private company reports or private company sections within reports is at Citi's discretion, in accordance with internal policies.

RESEARCH ANALYST AFFILIATIONS / NON-US RESEARCH ANALYST DISCLOSURES

The legal entities employing the authors of this report are listed below (and their regulators are listed further herein). Non-US research analysts who have prepared this report (i.e., all research analysts listed below other than those identified as employed by Citigroup Global Markets Inc.) are not registered/qualified as research analysts with FINRA. Such research analysts may not be associated persons of the member organization (but are employed by an affiliate of the member organization) and therefore may not be subject to the FINRA Rule 2241 restrictions on communications with a subject company, public appearances and trading securities held by a research analyst account.

Citigroup Global Markets Inc.

Shelby Spencer; Ashley Kim; Janna Withrow; Heath Terry

OTHER DISCLOSURES

Any price(s) of instruments mentioned in recommendations are as of the prior day's market close on the primary market for the instrument, unless otherwise stated.

The completion and first dissemination of any recommendations made within this research report are as of the Eastern date-time displayed at the top of the Product. If the Product references views of other analysts then please refer to the price chart or rating history table for the date/time of completion and first dissemination with respect to that view.

Regulations in various jurisdictions require that where a recommendation differs from any of the author's previous recommendations concerning the same financial instrument or issuer that has been published during the preceding 12-month period that the change(s) and the date of that previous recommendation are indicated. For fundamental coverage please refer to the price chart or rating change history within this disclosure appendix or the issuer disclosure summary at https://www.citivelocity.com/cvr/eppublic/citi_research_disclosures.

Citi Research has implemented policies for identifying, considering and managing potential conflicts of interest arising as a result of publication or distribution of investment research. A description of these policies can be found at https://www.citivelocity.com/cvr/eppublic/citi_research_disclosures.

The proportion of all Citi Research research recommendations that were the equivalent to "Buy", "Hold", "Sell" at the end of each quarter over the prior 12 months (with the % of these that had received investment firm services from Citi in the prior 12 months shown in brackets) is as follows; Q2 2026 Buy 33%(63%), Hold 44% (51%), Sell 23% (49%), RV 0.5% (87%); Q1 2026 Buy 33% (63%), Hold 44% (52%), Sell 23% (46%), RV 0.5% (89%); Q4 2025 Buy 33% (63%), Hold 44% (50%), Sell 23% (46%), RV 0.4% (91%); Q3 2025 Buy 33%(61%), Hold 44% (52%), Sell 23% (50%), RV 0.4% (80%). For the purposes of disclosing recommendations other than for equity (whose definitions can be found in the corresponding disclosure sections), "Buy" means a positive directional trade idea; "Sell" means a negative directional trade idea; and "Relative Value" means any trade idea which does not have a clear direction to the investment strategy.

European regulations require a 5 year price history when past performance of a security is referenced. CitiVelocity's Charting Tool (https://www.citivelocity.com/cv2/#go/CHARTING_3_Equities) provides the facility to create customisable price charts including a five year option. This tool can be found in the Data & Analytics section under any of the asset class menus in CitiVelocity (<https://www.citivelocity.com/>). For further information contact CitiVelocity support (https://www.citivelocity.com/cv2/go/CLIENT_SUPPORT). The source for all referenced prices, unless otherwise stated, is DataCentral, which sources price information from LSEG Data & Analytics. Past performance is not a guarantee or reliable indicator of future results. Forecasts are not a guarantee or reliable indicator of future performance.

Investors should always consider the investment objectives, risks, and charges and expenses of an ETF carefully before investing. The applicable prospectus and key investor information document (as applicable) for an ETF should contain this and other information about such ETF. It is important to read carefully any such prospectus before investing. Clients may obtain prospectuses and key investor information documents for ETFs from the applicable distributor or authorized participant, the exchange upon which an ETF is listed and/or from the applicable website of the applicable ETF issuer. The value of the investments and any accruing income may fall or rise. Any past performance, prediction or forecast is not indicative of future or likely performance. Any information on ETFs contained herein is provided strictly for illustrative purposes and should not be deemed an offer to sell or a solicitation of an offer to purchase units of any ETF either explicitly or implicitly. The opinions expressed are those of the authors and do not necessarily reflect the views of ETF issuers, any of their agents or their affiliates. Citigroup Global Markets India Private Limited and/or its affiliates may have, from time to time, actual or beneficial ownership of 1% or more in the debt securities of the subject issuer.

Please be advised that pursuant to Executive Order 13959 as amended (the "Order"), U.S. persons are prohibited from investing in securities of any company determined by the United States Government to be the subject of the Order. This

research is not intended to be used or relied upon in any way that could result in a violation of the Order. Investors are encouraged to rely upon their own legal counsel for advice on compliance with the Order and other economic sanctions programs administered and enforced by the Office of Foreign Assets Control of the U.S. Treasury Department.

This communication is directed at persons who are "Eligible Clients" as such term is defined in the Israeli Regulation of Investment Advice, Investment Marketing and Investment Portfolio Management law, 1995 (the "Advisory Law"). Within Israel, this communication is not intended for retail clients and Citi will not make such products or transactions available to retail clients or to non-Eligible Clients. The presenter is not licensed as investment advisor or investment marketer by the Israeli Securities Authority ("ISA") and this communication does not constitute investment or marketing advice. The information contained herein may relate to matters that are not regulated by the ISA. Any securities which are the subject of this communication may not be offered or sold to any Israeli person except pursuant to a security offering exemption according to the Israeli Securities Law, 1968 and the public offering rules provided thereunder.

Citi Research broadly and simultaneously disseminates its research content to the Firm's institutional and retail clients via the Firm's proprietary electronic distribution platforms (e.g., Citi Velocity and various Global Wealth platforms). As a convenience, certain, but not all, research content may be distributed through third party aggregators. Clients may receive published research reports by email, on a discretionary basis, and only after such research content has been broadly disseminated. Certain research is made available only to institutional investors to satisfy regulatory requirements. The level and types of services provided by Citi Research analysts to clients may vary depending on various factors such as the client's individual preferences as to the frequency and manner of receiving communications from analysts, the client's risk profile and investment focus and perspective (e.g. market-wide, sector specific, long term, short-term etc.), the size and scope of the overall client relationship with the Firm and legal and regulatory constraints.

Pursuant to Comissão de Valores Mobiliários Resolução 20 and ASIC Regulatory Guide 264, Citi is required to disclose whether a Citi related company or business has a commercial relationship with the subject company. Considering that Citi operates multiple businesses in more than 100 countries around the world, it is likely that Citi has a commercial relationship with the subject company.

Securities recommended, offered, or sold by the Firm: (i) are not insured by the Federal Deposit Insurance Corporation; (ii) are not deposits or other obligations of any insured depository institution (including Citibank); and (iii) are subject to investment risks, including the possible loss of the principal amount invested. The Product is for informational purposes only and is not intended as an offer or solicitation for the purchase or sale of a security. Any decision to purchase securities mentioned in the Product must take into account existing public information on such security or any registered prospectus. Although information has been obtained from and is based upon sources that the Firm believes to be reliable, we do not guarantee its accuracy and it may be incomplete and condensed. Note, however, that the Firm has taken all reasonable steps to determine the accuracy and completeness of the disclosures made in the Important Disclosures section of the Product. The Firm's research department has received assistance from the subject company(ies) referred to in this Product including, but not limited to, discussions with management of the subject company(ies). Firm policy prohibits research analysts from sending draft research to subject companies. However, it should be presumed that the author of the Product has had discussions with the subject company to ensure factual accuracy prior to publication. Statements and views concerning ESG (environmental, social, governance) factors are typically based upon public statements made by the affected company or other public news, which the author may not have independently verified. ESG factors are one consideration that investors may choose to examine when making investment decisions. All opinions, projections and estimates constitute the judgment of the author as of the date of the Product and these, plus any other information contained in the Product, are subject to change without notice. Prices and availability of financial instruments also are subject to change without notice. Notwithstanding other departments within the Firm advising the companies discussed in this Product, information obtained in such role is not used in the preparation of the Product. Although Citi Research does not set a predetermined frequency for publication, if the Product is a fundamental equity or credit research report, it is the intention of Citi Research to provide research coverage of the covered issuers, including in response to news affecting the issuer. For non-fundamental research reports, Citi Research may not provide regular updates to the views, recommendations and facts included in the reports. Notwithstanding that Citi Research maintains coverage on, makes recommendations concerning or discusses issuers, Citi Research may be periodically restricted from referencing certain issuers due to legal or policy reasons. Where a component of a published trade idea is subject to a restriction, the trade idea will be removed from any list of open trade ideas included in the Product. Upon the lifting of the restriction, the trade idea will either be re-instated in the open trade ideas list if the analyst continues to support it or it will be officially closed. Citi Research may provide different research products and services to different classes of customers (for example, based upon long-term or short-term investment horizons) that may lead to differing conclusions or recommendations that could impact the price of a security contrary to the recommendations in the alternative research product, provided that each is consistent with the rating system for each respective product.

Investing in non-U.S. securities, including ADRs, may entail certain risks. The securities of non-U.S. issuers may not be registered with, nor be subject to the reporting requirements of the U.S. Securities and Exchange Commission. There may be limited information available on foreign securities. Foreign companies are generally not subject to uniform audit and reporting standards, practices and requirements comparable to those in the U.S. Securities of some foreign companies may be less liquid and their prices more volatile than securities of comparable U.S. companies. In addition, exchange rate movements may have an adverse effect on the value of an investment in a foreign stock and its corresponding dividend payment for U.S. investors. Net dividends to ADR investors are estimated, using withholding tax rates conventions, deemed accurate, but investors are urged to consult their tax advisor for exact dividend computations. Investors who have received the Product from

the Firm may be prohibited in certain states or other jurisdictions from purchasing securities mentioned in the Product from the Firm. Please ask your Financial Consultant for additional details. Citigroup Global Markets Inc. takes responsibility for the Product in the United States. Any orders by US investors resulting from the information contained in the Product may be placed only through Citigroup Global Markets Inc.

The Citigroup legal entity that takes responsibility for the production of the Product is the legal entity which the first named author is employed by.

The Product is made available in **Australia** through Citigroup Global Markets Australia Pty Limited. (ABN 64 003 114 832 and AFSL No. 240992), participant of the ASX Group and regulated by the Australian Securities & Investments Commission. Citigroup Centre, 2 Park Street, Sydney, NSW 2000. Citigroup Global Markets Australia Pty Limited is not an Authorised Deposit-Taking Institution under the Banking Act 1959, nor is it regulated by the Australian Prudential Regulation Authority.

The Product is made available in **Brazil** by Citigroup Global Markets Brasil - CCTVM SA, which is regulated by CVM - Comissão de Valores Mobiliários ("CVM"), BACEN - Brazilian Central Bank, APIMEC - Associação dos Analistas e Profissionais de Investimento do Mercado de Capitais and ANBIMA - Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais. Av. Paulista, 1111 - 14º andar(parte) - CEP: 01311920 - São Paulo - SP.

This Product is available in **Chile** through Banchile Corredores de Bolsa S.A., an indirect subsidiary of Citigroup Inc., which is regulated by the Comisión Para El Mercado Financiero. Enrique Foster Sur, 20, piso 6, Las Condes, Santiago, Chile.

Disclosure for investors in the Republic of Colombia : This communication or message does not constitute a professional recommendation to make investment in the terms of article 2.40.1.1.2 of Decree 2555 de 2010 or the regulations that modify, substitute or complement it. Para la elaboración y distribución de informes de investigación y de comunicaciones generales de que trata este artículo no se requiere ser una entidad vigilada por la Superintendencia Financiera de Colombia.

The Product is made available in **Germany** by Citigroup Global Markets Europe AG ("CGME"), which is regulated by the European Central Bank and the German Federal Financial Supervisory Authority (Bundesanstalt für Finanzdienstleistungsaufsicht BaFin). Börsenplatz 9, 60313 Frankfurt am Main, Germany.

Unless otherwise specified, if the analyst who prepared this report is based in Hong Kong and it relates to "securities" (as defined in the Securities and Futures Ordinance (Cap.571 of the Laws of Hong Kong)), the report is issued in Hong Kong by Citigroup Global Markets Asia Limited. Citigroup Global Markets Asia Limited is regulated by Hong Kong Securities and Futures Commission. If the report is prepared by a non-Hong Kong based analyst, please note that such analyst (and the legal entity that the analyst is employed by or accredited to) is not licensed/registered in Hong Kong and they do not hold themselves out as such. Please refer to the section "Research Analyst Affiliations / Non-US Research Analyst Disclosures" for the details of the employment entity of the analysts.

The Product is made available in **India** by Citigroup Global Markets India Private Limited (CGM), which is regulated by the Securities and Exchange Board of India (SEBI), as a Research Analyst (SEBI Registration No. INH000000438). CGM is also actively involved in the business of merchant banking (SEBI Registration No. INM000010718) and stock brokerage ((SEBI Registration No. INZ000263033) in India, and is registered with SEBI in this regard. Registration granted by SEBI, enlistment with Bombay Stock Exchange (BSE) and certification from National Institute of Securities Markets (NISM) in no way guarantee performance of the intermediary or provide any assurance of returns to investors. CGM's registered office is at 1202, 12th Floor, First International Financial Centre (FIFC), G Block, Bandra Kurla Complex, Bandra East, Mumbai - 400098 & registered Tel: +91 22 61759999. Citi maintains robust policies, procedures, controls, and training to ensure continued compliance with all applicable rules and regulations. All recommendations contained herein are made by duly qualified research analysts. CGM's Corporate Identity Number is U99999MH2000PTC126657, and its Compliance Officer [Vishal Bohra] contact details are: Tel: +91-022-61759994, Fax: +91-022-61759851, Email: cgmcompliance@citi.com. The Investor Charter in respect of Research Analysts, the Compliance Audit Report and Complaints information can be found at

https://www.citivelocity.com/cvr/eppublic/citi_research_disclosures. The grievance officer [Niraj Mody] contact details are Tel: +91-22-42775002, Email: EMEA.CR.Complaints@citi.com. Investment in securities market are subject to market risks. Read all the related documents carefully before investing. SEBI prescribed Client Terms & Conditions can be found at <https://www.citivelocity.com/rendition/authfilelinksvcs/eppublic/V1/file?paramData=ZmlsZU5hbWU9L1MzLOdETVMvchVlRGJzY2xvc3VyZU9hOjV9ZmV9kaXNfc2V9UzJtc29mVXNlMh0bWw>

The Product is made available in **Indonesia** through PT Citigroup Sekuritas Indonesia. Citibank Tower 10/F, Pacific Century Place, SCBD lot 10, Jl. Jend Sudirman Kav 52-53, Jakarta 12190, Indonesia. Neither this Product nor any copy hereof may be distributed in Indonesia or to any Indonesian citizens wherever they are domiciled or to Indonesian residents except in compliance with applicable capital market laws and regulations. This Product is not an offer of securities in Indonesia. The securities referred to in this Product have not been registered with the Capital Market and Financial Services Authority (OJK) pursuant to relevant capital market laws and regulations, and may not be offered or sold within the territory of the Republic of Indonesia or to Indonesian citizens through a public offering or in circumstances which constitute an offer within the meaning of the Indonesian capital market laws and regulations.

The Product is made available in **Japan** by Citigroup Global Markets Japan Inc. ("CGMJ"), which is regulated by Financial Services Agency, Securities and Exchange Surveillance Commission, Japan Securities Dealers Association, Tokyo Stock Exchange and Osaka Securities Exchange. Otemachi Park Building, 1-1-1 Otemachi, Chiyoda-ku, Tokyo 100-8132 Japan. In the event that an error is found in an CGMJ research report, a revised version will be posted on the Firm's Citi Velocity website. If you have questions regarding Citi Velocity, please call (81 3) 6270-3019 for help.

The product is made available in the **Kingdom of Saudi Arabia** in accordance with Saudi laws through Citigroup Saudi Arabia, which is regulated by the Capital Market Authority (CMA) under CMA license (17184-31). 2239 Al Urubah Rd – Al Olaya Dist. Unit No. 18, Riyadh 12214 – 9597, Kingdom Of Saudi Arabia.

The Product is made available in **Korea** by Citigroup Global Markets Korea Securities Ltd. (CGMK), which is regulated by the Financial Services Commission, the Financial Supervisory Service and the Korea Financial Investment Association (KOFIA). The address of CGMK is Citibank Center, 50 Saemunan-ro, Jongno-gu, Seoul 03184, Korea. KOFIA makes available registration information of research analysts on its website. Please visit the following website if you wish to find KOFIA registration information on research analysts of

CGMK. <http://dis.kofia.or.kr/websquare/index.jsp?w2xPath=/wq/fundMgr/DISFundMgrAnalystList.xml&divisionId=MDISO3002002000000&serviceld=SDISO300200200000>. The Product is made available in Korea by Citibank Korea Inc., which is regulated by the Financial Services Commission and the Financial Supervisory Service. Address is Citibank Center, 50 Saemunan-ro, Jongno-gu, Seoul 03184, Korea. This research report is intended to be provided only to Professional Investors as defined in the Financial Investment Services and Capital Market Act and its Enforcement Decree in Korea.

The Product is made available in **Malaysia** by Citigroup Global Markets Malaysia Sdn Bhd (Registration No. 199801004692 (460819-D)) ("CGMM") to its clients and CGMM takes responsibility for its contents as regards CGMM's clients. CGMM is regulated by the Securities Commission Malaysia. Please contact CGMM at Level 43 Menara Citibank, 165 Jalan Ampang, 50450 Kuala Lumpur, Malaysia in respect of any matters arising from, or in connection with, the Product.

The Product is made available in **Mexico** by Citi México Casa de Bolsa, S.A. de C.V., Grupo Financiero Citi México which is a wholly owned subsidiary of Citigroup Inc. and is regulated by Comisión Nacional Bancaria y de Valores. Prolongación Reforma 1196, 24 floor, Colonia Santa Fe, Alcaldía Cuajimalpa de Morelos, C.P. 05348, Ciudad de México.

The Product is made available in **Poland** by Biuro Maklerskie Banku Handlowego (DMBH), separate department of Bank Handlowy w Warszawie S.A. a subsidiary of Citigroup Inc., which is regulated by Komisja Nadzoru Finansowego. Biuro Maklerskie Banku Handlowego (DMBH), ul. Senatorska 16, 00-923 Warszawa.

The Product is made available in **Singapore** through Citigroup Global Markets Singapore Pte. Ltd. ("CGMSPL"), a capital markets services and Exempt Financial Advisor license holder, and regulated by Monetary Authority of Singapore. Please contact CGMSPL at 8 Marina View, 21st Floor Asia Square Tower 1, Singapore 018960, in respect of any matters arising from, or in connection with, the analysis of this document. This Product is intended for recipients who are accredited, expert and institutional investors as defined under the Securities and Futures Act 2001. For Citi Private Bank, the Product is made available in Singapore by Citi Private Bank through Citibank, N.A., Singapore Branch. Citibank N.A., Singapore Branch is a licensed bank in Singapore that is regulated by the Monetary Authority of Singapore. Please contact your Private Banker in Citibank N.A., Singapore Branch if you have any queries on or any matters arising from or in connection with this document. The Product is intended for recipients who are accredited, expert and institutional investors as defined under the Securities and Futures Act 2001. For Citibank Singapore Limited ("CSL"), the Product is distributed in Singapore by CSL to selected Citigold/Citigold Private Clients. CSL provides no independent research or analysis of the substance or in preparation of the Product. Please contact your Citigold/Citigold Private Client Relationship Manager in CSL if you have any queries on or any matters arising from or in connection with this document. The Product is intended for recipients who are accredited investors as defined under the Securities and Futures Act (Cap. 289).

Citigroup Global Markets (Pty) Ltd. is incorporated in the **Republic of South Africa** (company registration number 2000/025866/07) and its registered office is at 145 West Street, Sandton, 2196, Saxonwold. Citigroup Global Markets (Pty) Ltd. is regulated by JSE Securities Exchange South Africa, South African Reserve Bank and the Financial Services Board. The investments and services contained herein are not available to private customers in South Africa.

The Product is made available in the **Republic of China (Taiwan)** through Citigroup Global Markets Taiwan Securities Company Ltd. ("CGMTS"), 14F, 15F and 16F, No. 1, Songzhi Road, Taipei 110, Taiwan, subject to the license scope and the applicable laws and regulations in the Republic of China (Taiwan). CGMTS is regulated by the Securities and Futures Bureau of the Financial Supervisory Commission of Taiwan, the Republic of China (Taiwan). No portion of the Product may be reproduced or quoted in the Republic of China (Taiwan) by the press or any third parties [without the written authorization of CGMTS]. Pursuant to the applicable laws and regulations in the Republic of China (Taiwan), the recipient of the Product shall not take advantage of such Product to involve in any matters in which the recipient may have conflicts of interest. If the Product covers securities which are not allowed to be offered or traded in the Republic of China (Taiwan), neither the Product nor any information contained in the Product shall be considered as advertising the securities or making recommendation of the securities in the Republic of China (Taiwan). The Product is for informational purposes only and is not intended as an offer or solicitation for the purchase or sale of a security or financial products. Any decision to purchase securities or financial products mentioned in the Product must take into account existing public information on such security or the financial products or any registered prospectus.

The Product is made available in **Thailand** through Citicorp Securities (Thailand) Ltd., which is regulated by the Securities and Exchange Commission of Thailand. 399 Interchange 21 Building, 18th Floor, Sukhumvit Road, Klongtoey Nua, Wattana, Bangkok 10110, Thailand.

Disclosure for investors in the Republic of Turkey: The Product is made available in **Turkey** through Citibank AS which is regulated by Capital Markets Board. Tekfen Tower, Eski Büyükdere Caddesi # 209 Kat 2B, 23294 Levent, Istanbul, Turkey. Under Capital Markets Law of Turkey (Law No: 6362), the investment information, comments and recommendations stated here, are not within the scope of investment advisory activity. Comments and recommendations stated here rely on the individual opinions of the ones providing these comments and recommendations. These opinions may not fit to your financial

status, risk and return preferences. For this reason, to make an investment decision by relying solely to this information stated here may not bring about outcomes that fit your expectations. Furthermore, Citi Research is a division of Citigroup Global Markets Inc. (the "Firm"), which does and seeks to do business with companies and/or trades on securities covered in this 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, however investors should also note that the Firm has in place organisational and administrative arrangements to manage potential conflicts of interest of this nature.

In the **U.A.E.**, these materials (the "Materials") are communicated by Citigroup Global Markets Limited, DIFC branch ("CGML"), an entity registered in the Dubai International Financial Center ("DIFC") and licensed and regulated by the Dubai Financial Services Authority ("DFSA"), license #CLO221 to Professional Clients and Market Counterparties, as defined in DFSA regulations, only and should not be relied upon or distributed to Retail Clients. Financial products and/or services to which the Materials relate will only be made available to Professional Clients and Market Counterparties. Citigroup Global Markets Limited DIFC Branch registered address is Level 3, Gate District Building 02, Dubai International Financial Centre and can be contacted on +971 4 509 97 90.

The Product is made available in **United Kingdom** by Citigroup Global Markets Limited, which is authorised by the Prudential Regulation Authority ("PRA") and regulated by the Financial Conduct Authority ("FCA") and the PRA. This material may relate to investments or services of a person outside of the UK or to other matters which are not authorised by the PRA nor regulated by the FCA and the PRA and further details as to where this may be the case are available upon request in respect of this material. Citigroup Centre, Canada Square, Canary Wharf, London, E14 5LB.

The Product is made available in **United States and Canada** by Citigroup Global Markets Inc., which is a member of FINRA and registered with the US Securities and Exchange Commission. 388 Greenwich Street, New York, NY 10013.

Unless specified to the contrary, within EU Member States, the Product is made available by Citigroup Global Markets Europe AG ("CGME"), which is regulated by the European Central Bank and the German Federal Financial Supervisory Authority (Bundesanstalt für Finanzdienstleistungsaufsicht – BaFin).

The Product is not to be construed as providing investment services in any jurisdiction where the provision of such services would not be permitted.

Subject to the nature and contents of the Product, the investments described therein are subject to fluctuations in price and/or value and investors may get back less than originally invested. Certain high-volatility investments can be subject to sudden and large falls in value that could equal or exceed the amount invested. The yield and average life of CMOs (collateralized mortgage obligations) referenced in this Product will fluctuate depending on the actual rate at which mortgage holders prepay the mortgages underlying the CMO and changes in current interest rates. Any government agency backing of the CMO applies only to the face value of the CMO and not to any premium paid. Certain investments contained in the Product may have tax implications for private customers whereby levels and basis of taxation may be subject to change. If in doubt, investors should seek advice from a tax adviser. The Product does not purport to identify the nature of the specific market or other risks associated with a particular transaction. Advice in the Product is general and should not be construed as personal advice given it has been prepared without taking account of the objectives, financial situation or needs of any particular investor. Accordingly, investors should, before acting on the advice, consider the appropriateness of the advice, having regard to their objectives, financial situation and needs. Prior to acquiring any financial product, it is the client's responsibility to obtain the relevant offer document for the product and consider it before making a decision as to whether to purchase the product.

Card Insights. Where this report references Card Insights data, Card Insights consists of selected data from a subset of Citi's proprietary credit card transactions. Such data has undergone rigorous security protocols to keep all customer information confidential and secure; the data is highly aggregated and anonymized so that all unique customer identifiable information is removed from the data prior to receipt by the report's author or distribution to external parties. This data should be considered in the context of other economic indicators and publicly available information. Further, the selected data represents only a subset of Citi's proprietary credit card transactions due to the selection methodology or other limitations and should not be considered as indicative or predictive of the past or future financial performance of Citi or its credit card business.

Citi Research product may source data from dataCentral. dataCentral is a Citi Research proprietary database, which includes the Firm's estimates, data from company reports and feeds from LSEG Data & Analytics. The source for all referenced prices, unless otherwise stated, is DataCentral. Past performance is not a guarantee or reliable indicator of future results. Forecasts are not a guarantee or reliable indicator of future performance. The printed and printable version of the research report may not include all the information <(e.g. certain financial summary information and comparable company data) that is linked to the online version available on the Firm's proprietary electronic distribution platforms.

Where included in this report, MSCI sourced information is the exclusive property of Morgan Stanley Capital International Inc. (MSCI). Without prior written permission of MSCI, this information and any other MSCI intellectual property may not be reproduced, disseminated or used to create any financial products, including any indices. This information is provided on an "as is" basis. The user assumes the entire risk of any use made of this information. MSCI, its affiliates and any third party involved in, or related to, computing or compiling the information hereby expressly disclaim all warranties of originality, accuracy, completeness, merchantability or fitness for a particular purpose with respect to any of this information. Without limiting any of the foregoing, in no event shall MSCI, any of its affiliates or any third party involved in, or related to, computing or compiling the information have any liability for any damages of any kind. MSCI, Morgan Stanley Capital International and the MSCI indexes are services marks of MSCI and its affiliates. Where data is attributed to Morningstar that data is © 2026 Morningstar, Inc. All Rights Reserved. That information: (1) is proprietary to Morningstar and/or its content providers; (2) may

not be copied or distributed; and (3) is not warranted to be accurate, complete or timely. Neither Morningstar nor its content providers are responsible for any damages or losses arising from any use of this information.

The Firm accepts no liability whatsoever for the actions of third parties. The Product may provide the addresses of, or contain hyperlinks to, websites. Except to the extent to which the Product refers to website material of the Firm, the Firm has not reviewed the linked site. Equally, except to the extent to which the Product refers to website material of the Firm, the Firm takes no responsibility for, and makes no representations or warranties whatsoever as to, the data and information contained therein. Such address or hyperlink (including addresses or hyperlinks to website material of the Firm) is provided solely for your convenience and information and the content of the linked site does not in any way form part of this document. Accessing such website or following such link through the Product or the website of the Firm shall be at your own risk and the Firm shall have no liability arising out of, or in connection with, any such referenced website.

© 2026 Citigroup Global Markets Inc. Citi Research is a division of Citigroup Global Markets Inc. Citi and Citi and Arc Design are trademarks and service marks of Citigroup Inc. and its affiliates and are used and registered throughout the world. All rights reserved. The research data in this report are not intended to be used for the purpose of (a) determining the price of or amounts due in respect of (or to value) one or more financial products or instruments and/or (b) measuring or comparing the performance of, or defining the asset allocation of a financial product, a portfolio of financial instruments, or a collective investment undertaking, and any such use is strictly prohibited without the prior written consent of Citi Research. Any unauthorized use, duplication, redistribution or disclosure of this report (the "Product"), including, but not limited to, redistribution of the Product by electronic mail, posting of the Product on a website or page, and/or providing to a third party a link to the Product, is prohibited by law and will result in prosecution. The information contained in the Product is intended solely for the recipient and may not be further distributed by the recipient to any third party.

ADDITIONAL INFORMATION IS AVAILABLE UPON REQUEST