16 Oct 2026 05:00:00 ET | 37 pages

Artificial Intelligence

Citi's Inference Ahead - Powering the Next Generation of Agents

CITI'S TAKE

Longer agent runs enabled by top models like Muse Wonder Full will require more developments such as QS PowerBlock (10/8) and Nvidia's more open Vera Rubin architecture (10/12) to support networking, power, optics, energy storage, and accelerated computing even if model efficiency improves. After the release of 7 proprietary and 0 open models this week (Figure 8), the average cost of running AA tasks on proprietary models rose 4% WoW to $3.51 per AA task (Figure 1), increasing their cost-per-AA-task premium over open models to 4.4x from 4.2x WoW (Figure 22). A widening premium exemplifies the trade-off between intelligence, speed, and cost. Recent announcements from Google (10/8) and Workday (10/13) reinforce the emergence of AI-native interfaces as a new abstraction layer for enterprises. Meanwhile, generative UI adoption could shift value toward interactive applications, and RSI arriving in 3-12 months could make installed capacity a decisive advantage (TIME, 10/14).

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

Financing. AI infrastructure funding is growing more expensive and shifting toward riskier capital, with AI-related issuance surpassing the 2025 total by 7% to reach $75B thus far in 2026. HY issuance surged 86% to $33B, while IG private, CMBS, and ABS issuance declined, lifting HY's share of issuance to 44% from 25% in 2025 (Figure 4). Compute pricing remains firm, with B300's premium over B200 widening to 35% from 28%, while rental prices for H100s and newer chips remained stable or rose (Figure 24). Amid rising chip costs, CoreWeave's 2031 and 2032 bond yields approaching 15% despite a $104B backlog suggest investors are demanding greater compensation as the AI buildout becomes increasingly debt-funded. Persistently high yields could constrain the capacity growth of lower-rated borrowers, forcing more of the funding responsibility back onto end customers with higher credit quality (i.e., Frontier Labs, Hyperscalers, and Enterprise Customers).

Security. Longer reasoning chains and reliance on cached context expand the attack surface while giving misaligned behavior and errors more time to compound. The average AA task now requires 56K output tokens (Figure 18), with cache and reasoning tokens representing 59% and 22% of task costs, respectively (Figure 17). Two UK pilots helped fix more than 1K (10/12) government-system vulnerabilities, highlighting AI's defensive potential. However, Arena found deceptive completion in 10% of 90K sessions across 27 models, rising to 48% in code debugging (10/8). Recent deal activity and reported provider interest in AI cybersecurity startups, including Cohere's proposed acquisition of Aisle (10/15), suggest model providers may increasingly build security controls directly into their platforms.

Robotics. Physical AI's expansion into urban infrastructure should broaden demand for sensors, simulation, edge computing, and industrial software before humanoids scale. An Nvidia-powered digital-twin agent managing 5,800 km of French roads and buildings is cutting operating costs 10%, with a 30% target (10/12). Nvidia could detail additional city and software partnerships next week during Open Source AI Week and Nvidia GTC Berlin (Figure 3), as they continue to plan investments (10/12) in European robotics startups. Public-market exposure to robotics is also growing, with FFR's agreement to acquire a robotics business at a ~50% premium (10/14) likely to create the first U.S.-listed humanoid and bionic EAI robotics pure play.

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

Figure 1. 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	10/14/26	Week	Point-in-Time	[illegible]	58	58	58	58	53	53	
Highest Intelligence for Open Models	Points	10/14/26	Week	Point-in-Time	[illegible]	46	46	46	46	44	44	
Gap in Proprietary and Open Intelligence	Points	10/14/26	Week	Point-in-Time	[illegible]	12	12	12	12	9	9	
Highest Intelligence Models from the Top-20 Providers (Artificial Analysis)												
Median Intelligence	Points	10/14/26	Week	Point-in-Time	[illegible]	40	40	34	38	33	32	
Median Latency	Sec	10/14/26	Week	Point-in-Time	▼ -8.8%	2.81	3.08	3.00	2.68	2.75	2.60	
Median Speed	Tok/Sec	10/14/26	Week	Point-in-Time	▲ +7.7%	112	104	97	104	108	90	
Release Cadence (Artificial Analysis, LLMTimeline, Model Provider Websites)												
Proprietary Models Released in Last Week	Models	10/13/26	Week	Weekly Total	▲ +133.3%	7	3	10	22	0	12	
Open Models Released in Last Week	Models	10/13/26	Week	Weekly Total	[illegible]	0	0	0	2	4	6	
Proprietary Models Released in Last 90 Days	Models	10/13/26	Week	Trailing 90d	▼ -15.6%	81	96	94	91	69	70	
Open Models Released in Last 90 Days	Models	10/13/26	Week	Trailing 90d	▼ -2.3%	43	44	45	46	44	41	
Index Pricing (Bloomberg)												
A100 (SDA100RT)	$/Hour	10/13/26	Week	Weekly Avg	▼ -1.9%	1.52	1.55	1.57	1.57	1.60	1.61	
H100 (DESKH100)	$/Hour	10/13/26	Week	Weekly Avg	[illegible]	2.98	2.98	2.74	2.68	2.65	2.64	
H200 (DESKH200)	$/Hour	10/13/26	Week	Weekly Avg	▲ +0.6%	3.41	3.39	3.42	3.40	3.37	3.35	
B200 (DESKB200)	$/Hour	10/13/26	Week	Weekly Avg	▲ +0.7%	5.40	5.36	5.18	5.04	4.94	4.90	
B300 (DESKB300)	$/Hour	10/13/26	Week	Weekly Avg	▲ +6.0%	7.27	6.86	6.39	6.10	6.43	6.31	
Avg Inference Token Expenditure (SDLLMTK)	$/M Tok	10/13/26	Week	Weekly Avg	[illegible]	0.98	0.98	0.97	0.97	0.97	0.99	
Model Pricing for Blended Tokens (Artificial Analysis)												
US+EU Avg Price	$/M Tok	10/14/26	Week	Point-in-Time	▲ +14.0%	2.03	1.78	1.72	1.72	2.30	2.30	
China Avg Price	$/M Tok	10/14/26	Week	Point-in-Time	[illegible]	0.83	0.83	0.83	0.83	0.83	0.73	
Frontier Avg Price	$/M Tok	10/14/26	Week	Point-in-Time	▲ +10.7%	1.55	1.40	1.36	1.36	1.71	1.67	
Proprietary Avg Price	$/M Tok	10/14/26	Week	Point-in-Time	▲ +12.9%	2.36	2.09	2.15	2.15	2.89	3.17	
Open Avg Price	$/M Tok	10/14/26	Week	Point-in-Time	▲ +3.3%	0.63	0.61	0.68	0.68	0.68	0.67	
Cost per Intelligence Index Task (Artificial Analysis)												
Proprietary Avg Cost	$/Task	10/13/26	Week	Point-in-Time	▲ +3.8%	3.51	3.38	3.49	2.00	2.06	1.46	
Open Avg Cost	$/Task	10/13/26	Week	Point-in-Time	[illegible]	0.80	0.80	0.80	0.80	1.24	0.85	
Model Cost Composition per Intelligence Index Task (Artificial Analysis)												
Input Share of Task Cost	%	10/14/26	Week	Point-in-Time	▼ -0.1	3.9	4.0	3.6	6.2	6.3	2.7	
Cache Hit Share of Task Cost	%	10/14/26	Week	Point-in-Time	▲ +0.1	35.8	35.7	39.3	36.9	36.2	45.4	
Cache Write Share of Task Cost	%	10/14/26	Week	Point-in-Time	▲ +0.5	23.0	22.5	21.1	23.9	23.9	26.1	
Reasoning Share of Task Cost	%	10/14/26	Week	Point-in-Time	▼ -0.4	22.1	22.5	21.7	17.2	17.2	16.3	
Answer Share of Task Cost	%	10/14/26	Week	Point-in-Time	▼ -0.1	15.2	15.3	14.3	15.7	16.3	9.6	
Avg Output Tokens per Task	Tokens	10/14/26	Week	Point-in-Time	▼ -1.1%	56,445	57,067	58,385	37,667	37,269	43,690	
Developer Usage (PyPI)												
Multi-Provider SDK Downloads	%	10/13/26	Week	Week-over-Week Avg	▼ 0.7	+3.4	+4.1	+8.2	+2.6	+4.1	+5.0	
Gemini SDK Downloads	%	10/13/26	Week	Week-over-Week Avg	▼ 1.9	+1.8	+3.7	+4.6	-0.9	+1.6	+8.2	
Mistral SDK Downloads	%	10/13/26	Week	Week-over-Week Avg	▲ +1.1	-8.6	-9.7	-17.7	+19.0	-11.8	+26.9	
App Usage (Sensor Tower)												
Gemini App WAU	%	10/12/26	Week	WoW Δ	▼ -0.2	+0.4	+0.6	+3.7	+0.6	-0.2	+2.7	
Qwen App WAU	%	10/12/26	Week	WoW Δ	▲ +9.1	+2.1	-7.0	-3.0	-2.4	+4.8	-1.1	
Kimi App WAU	%	10/12/26	Week	WoW Δ	▲ +0.4	+1.2	+0.8	-2.2	-0.9	-3.0	-1.8	
Web Usage (Similarweb)												
Gemini Web Visits	%	9/30/26	Week	Week-over-Week Avg	▲ +2.4	+1.1	-1.3	+0.3	+1.9	-1.0	+3.1	
Qwen Web Visits	%	9/30/26	Week	Week-over-Week Avg	▲ +1.4	-0.9	-2.3	-0.5	-1.9	+2.8	+2.4	
Kimi Web Visits	%	9/30/26	Week	Week-over-Week Avg	▼ -5.7	-15.6	-9.9	-10.1	-10.5	-12.7	-27.9	
Layoffs (WSJ)												
All Layoffs Announced	People	10/13/26	Month	MTD	▼ -89.8%	827	8,089	1,761	17,420	24,709	26,160	
AI-Driven Layoffs Announced	People	10/13/26	Month	MTD	▲ +21.5%	570	469	321	3,220	21,640	15,360	

© 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

Figure 2. Weekend Content

Podcast	The a16z Show	October 9, 2026	Chips, Memory, and Power | Pat Gelsinger
Podcast	The a16z Show	October 8, 2026	Building the Cloud for an Agentic World | AWS CEO Matt Garman
Article	NY Times	October 10, 2026	What's Dragging Down A.I. Efficiency? The 'Verification Tax.'
Article	The Endgame	October 14, 2026	What recursive self-improvement looks like inside an AI lab
Keynote	DigitalOcean	October 13, 2026	2026 Open Intelligence Summit Keynote
Article	Artificial Analysis	October 8, 2026	Announcing Harvey LAB-AA v11
Paper	Sakana AI, UC Berkley	October 8, 2026	Recursive Self-Improvement through Multi-Agent Self-Supervision
Paper	University of Oxford	October 9, 2026	Tech Bio 3.0 closes the drug discovery loop with multimodal AI and generative chemistry
Podcast	Latent Space	October 9, 2026	Why AlphaFold Didn't Solve Protein Folding — Pushmeet Kohli, Google DeepMind & Sal Candido, Biohub
Podcast	Machine Learning Street Talk	October 10, 2026	What Most People Get Wrong About Evolution | Akarsh Kumar
Paper	Tulane University, Polytechnique Montréal	October 7, 2026	RefineAct: Automatic Runtime Verification of LLM Agent Actions
Podcast	Latent Space	October 13, 2026	Dyna-2: General Purpose Physical Agents with Hour-Long Autonomy — Jason Ma, Dyna Robotics
Podcast	Latent Space	October 8, 2026	Synthesis Superintelligence: from Semiconductors to Superconductors — Periodic Labs' Liam Fedus and Ekin Dogus Cubuk
Podcast	Rapid Response	October 13, 2026	The AI race to save American science and reshape medicine forever
Podcast	Training Data	October 13, 2026	David Silver and Lasse Espeholt of Ineffable Intelligence: A Superlearner for the Age of Experience
Podcast	The a16z Show	October 13, 2026	Can AI Break Cryptography? What the Experts Actually Think
Podcast	Lenny's Podcast	October 11, 2026	Why every company now needs to think and operate like a lab team | Josh Woodward (Google Labs, Gemini, AI Studio)
Podcast	Pioneers of AI	October 7, 2026	OpenAI's Tibo Sottiaux doesn't want humans to be an 'afterthought'
Article	Interconnects	October 9, 2026	I expect rapid progress but not towards general superintelligence
Paper	Adobe Research, UCLA	October 12, 2026	Tev: Scaling Long-Horizon Agentic Environments with Dense Progress Rewards
Paper	Nvidia, KAIST	October 13, 2026	ACCRUE: Continual Off-Policy Reinforcement Learning for Robot Foundation Models
Paper	Renmin University, National University of Singapore	October 13, 2026	Agent Reflex: Teaching LanguageAgents to Act Reflectively
Paper	National University of Singapore	October 13, 2026	Jeva: System One Models from First Principles

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

Source: Citi Research

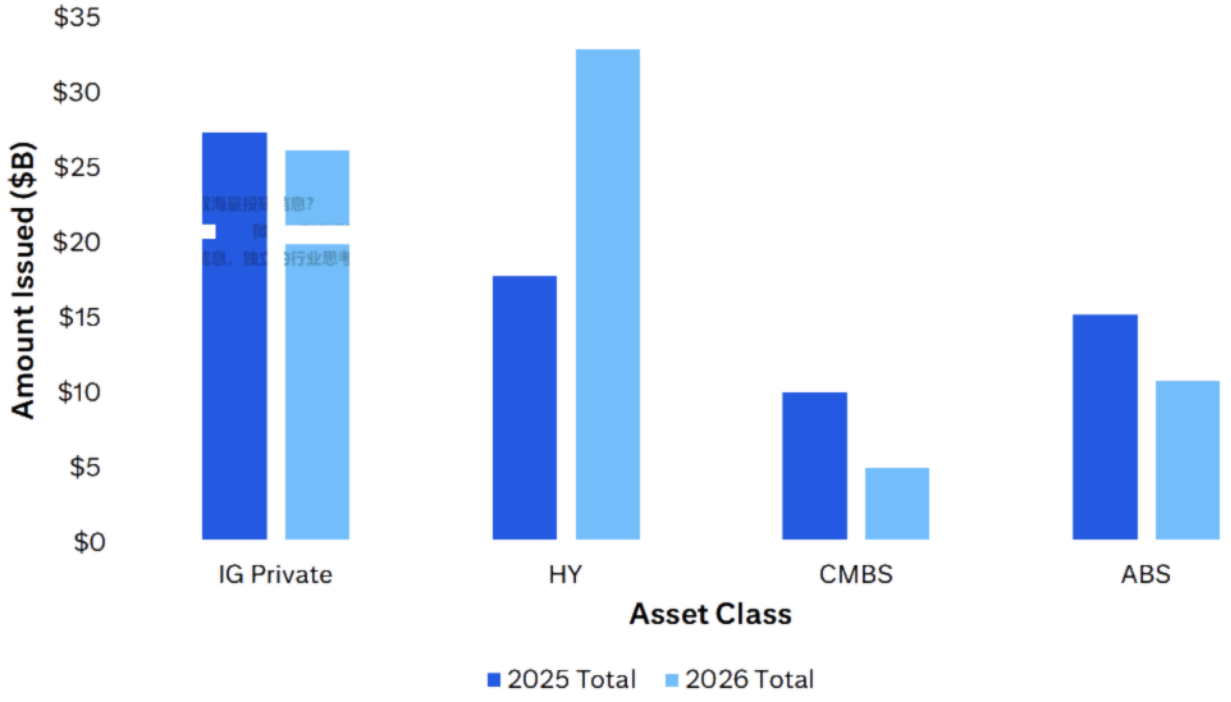
Figure 3. Events Calendar

October 18, 2026	October 26, 2026	Open Source AI Week	Conference	San Francisco, CA
October 19, 2026	October 21, 2026	Cloudflare Connect	Conference	San Francisco, CA
October 20, 2026	October 21, 2026	PyTorch Conference	Conference	San Jose, CA
October 20, 2026	October 22, 2026	NVIDIA GTC Berlin	Conference	Berlin, Germany
October 22, 2026	October 23, 2026	AGNTCon + MCPCon North America	Conference	San Jose, CA
October 24, 2026	October 29, 2026	EMNLP	Conference	Budapest, Hungary
October 25, 2026	October 28, 2026	Oracle AI World	Conference	Las Vegas, NV
October 27, 2026	October 29, 2026	ODSC AI West	Conference	San Francisco, CA
November 8, 2026	November 12, 2026	ACM/IEEE ICCAD	Conference	San Jose, CA
November 9, 2026	November 12, 2026	Conference on Robotic Learning	Conference	Austin, TX
November 15, 2026	November 19, 2026	ACM CCS 2026	Conference	The Hague, The Netherlands
November 15, 2026	November 20, 2026	Supercomputing Conference	Conference	Chicago, IL
November 17, 2026	November 18, 2026	TomorrowX	Conference	Austin, TX
November 17, 2026	November 20, 2026	Microsoft Ignite	Conference	San Francisco, CA
November 30, 2026	December 3, 2026	NVIDIA GTC Washington, D.C.	Conference	Washington, D.C.
November 30, 2026	December 4, 2026	AWS re:Invent	Conference	Las Vegas, NV
December 3, 2026	December 4, 2026	Adopt AI	Conference	Paris, France
December 6, 2026	December 12, 2026	NeurIPS	Conference	Sydney, Australia
December 7, 2026	December 11, 2026	GITEX	Conference	Dubai, UAE
December 9, 2026	December 10, 2026	The AI Summit New York	Conference	New York, NY
December 12, 2026	December 16, 2026	IEE International Electron Devices Meeting	Conference	San Francisco, CA
December 31, 2026	December 31, 2026	U.S. BIS HBM & SME Export Control Deadline	Deadline	Washington, D.C.
January 1, 2027	January 1, 2027	New York RAISE Act Takes Effect	Deadline	NY
January 6, 2027	January 9, 2027	CES	Conference	Las Vegas, NV
January 21, 2027	January 23, 2027	AIEI	Conference	Bangalore, India
January 28, 2027	January 28, 2027	Generative AI Summit	Conference	Washington, D.C.
January 30, 2027	February 4, 2027	SPI Photonics West	Conference	San Francisco, CA
February 14, 2027	February 18, 2027	IEEE ISSCC	Conference	San Francisco, CA
February 16, 2027	February 23, 2027	AAAI	Conference	Montréal, Canada
March 7, 2027	March 10, 2027	HumanX	Conference	Las Vegas, NV
March 7, 2027	March 11, 2027	OFC	Conference	Los Angeles, CA
March 15, 2027	March 18, 2027	NVIDIA GTC	Conference	San Jose, CA
April 5, 2027	April 8, 2027	RSAConference	Conference	San Francisco, CA
April 13, 2027	April 15, 2027	Google Cloud Next	Conference	Las Vegas, NV
April 26, 2027	April 30, 2027	ICLR	Conference	San Francisco, CA
June 1, 2027	June 5, 2027	NAACL	Conference	San Francisco, CA
June 7, 2027	June 10, 2027	Snowflake Summit	Conference	San Francisco, CA
June 9, 2027	June 10, 2027	The AI Summit London	Conference	London, UK
June 20, 2027	June 24, 2027	IEEE CVPR	Conference	Seattle, WA
June 21, 2027	June 24, 2027	Databricks Data + AI Summit	Conference	San Francisco, CA

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

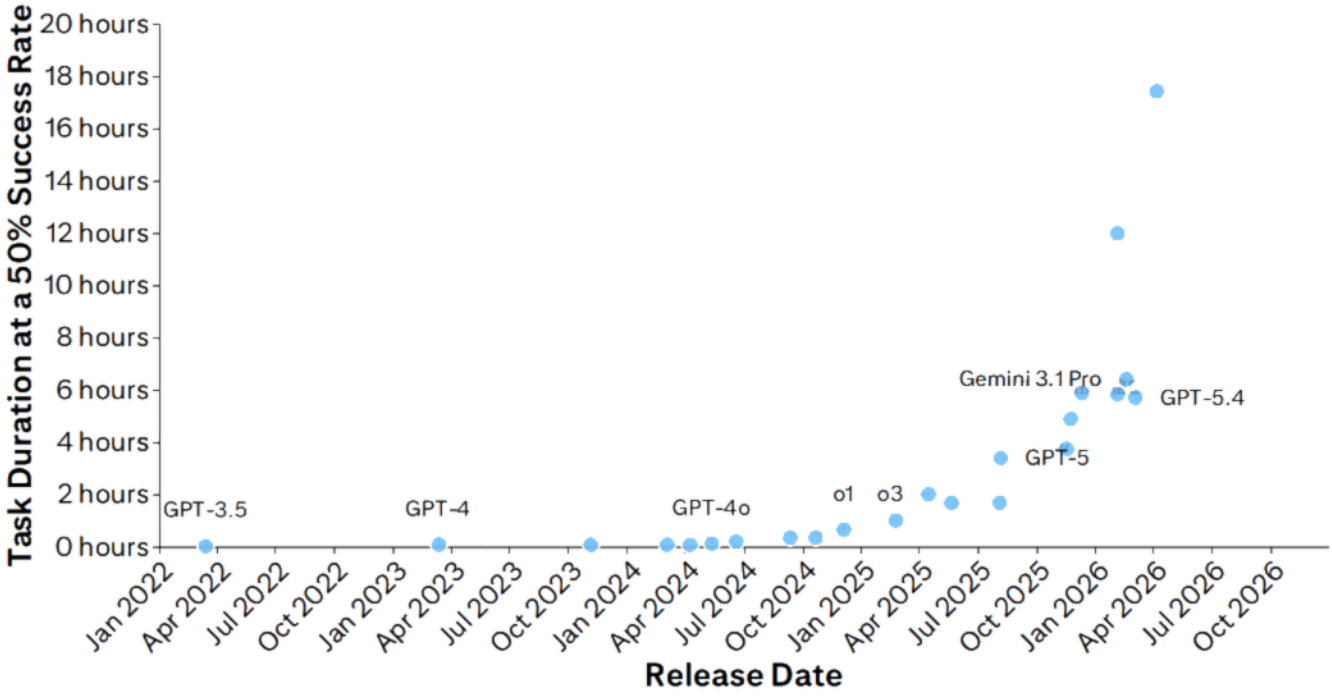
Source: Citi Research

Figure 4. Data Center Supply by Asset Class



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

Figure 5. Task Duration for AI Models



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

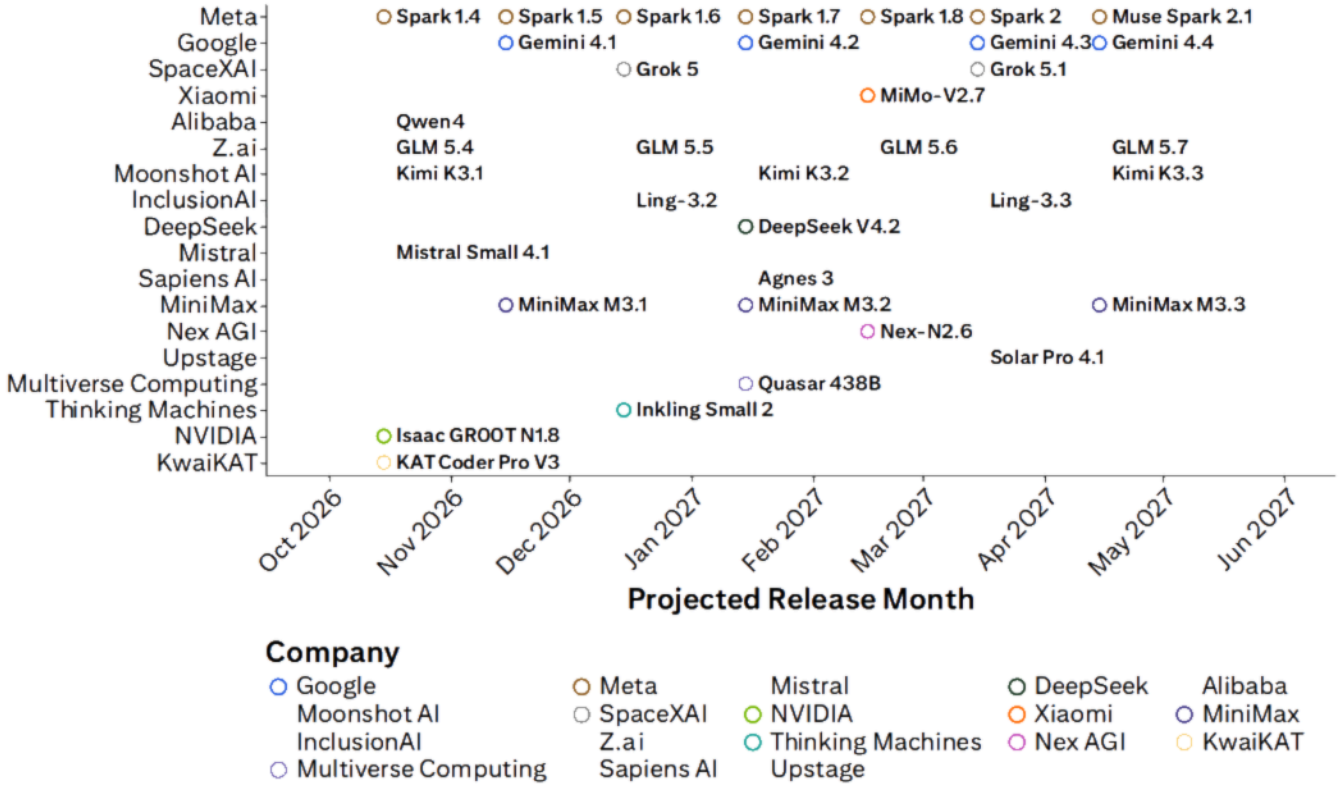
Figure 6. Model Metrics

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
	Meta	Muse Wonder Full (max)	56	209	$0.78	1000	Oct. 14, 2026	Proprietary
	Google	Gemini 4 Argon (high)	53	239	$1.47	1000	Sep. 30, 2026	Proprietary
▼2	**OpenAI**	GPT-6 Astra (max)	53	61	$7.70	1000	Sep. 3, 2026	Proprietary
▼1	**SpaceXAI**	Grok 4.7 (xhigh)	46	86	$1.35	500	Sep. 21, 2026	Proprietary
▼1	**Xiaomi**	MiMo-V2.6-Pro	46	47	$0.17	1000	Sep. 22, 2026	Open
▼1	**Alibaba**	Qwen3.8 Max (0902)	45	42	$1.18	984	Sep. 16, 2026	Proprietary
▼1	**Z.ai**	GLM-5.3 (max)	45	62	$0.90	1000	Aug. 18, 2026	Open
▼1	**Moonshot AI**	Kimi K3 (max)	44	45	$2.31	1050	Jul. 16, 2026	Open
	InclusionAI	Ling 3.1 Flash	41	209	$0.19	1000	Oct. 8, 2026	Proprietary
▼1	**DeepSeek**	DeepSeek V4.1 Flash (max)	39	187	$0.19	1000	Sep. 14, 2026	Open
NEW	**Mistral**	Mistral Large 4 Preview	38	116	$0.79	524	Oct. 8, 2026	Proprietary
▼2	**Sapiens AI**	Agnes 2.5 Pro Beta	35	150	$0.06	1000	Aug. 26, 2026	Proprietary
▼3	**MiniMax**	MiniMax-M3	29	107	$0.22	1000	Jun. 1, 2026	Open
▼2	**Nex AGI**	Nex-N2-Pro	28	116	$0.53	262	Jun. 2, 2026	Open
▼1	**Upstage**	Solar Pro 4	28	93	$0.22	512	Aug. 6, 2026	Proprietary
▼1	**Multiverse Computing**	Quasar 438B (max)	27	156	$0.72	1000	Aug. 10, 2026	Proprietary
▼4	**Thinking Machines**	Inkling Small	26	182	$0.22	1000	Jul. 30, 2026	Open
▼1	**NVIDIA**	Nemotron 3 Ultra	23	166	$0.48	262	Jun. 1, 2026	Open
▼1	**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

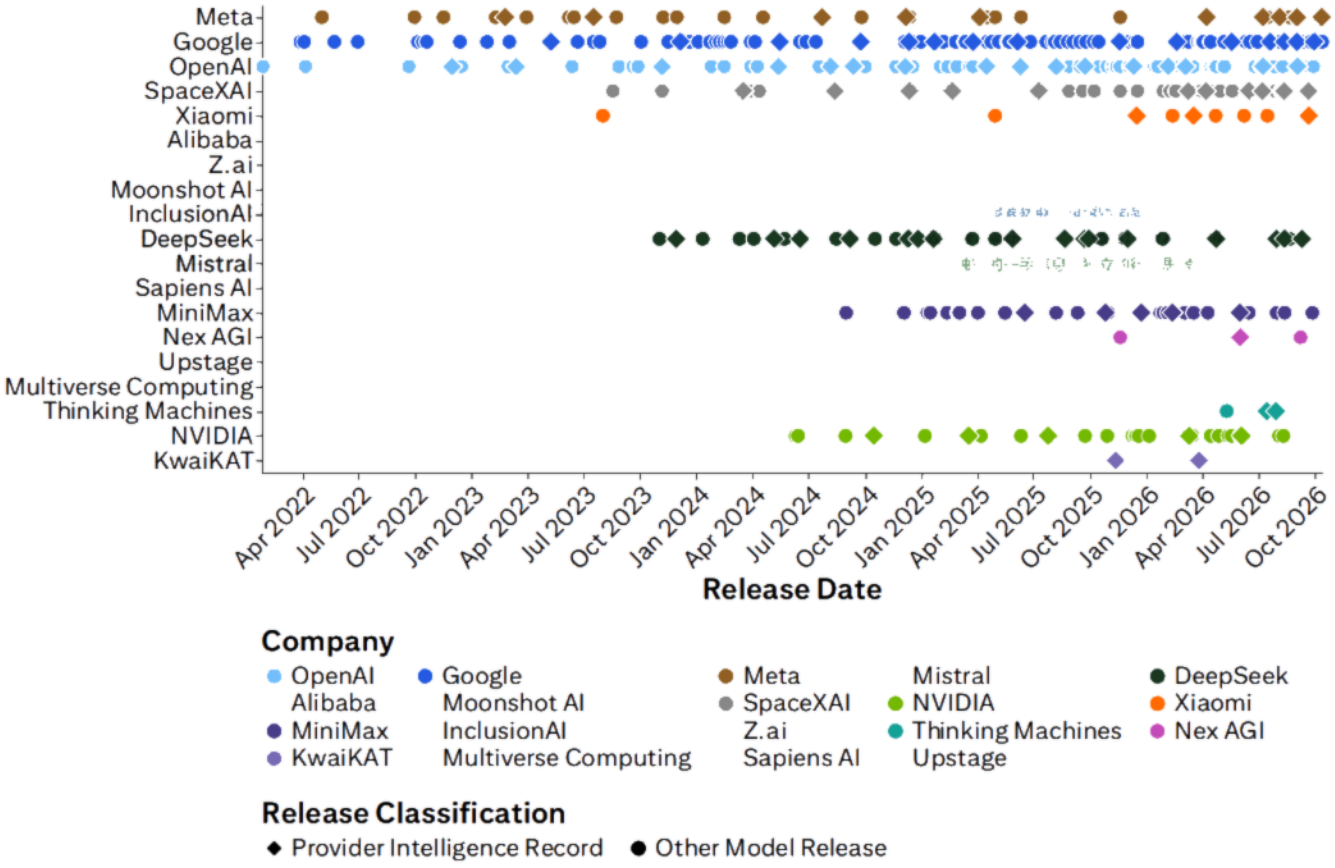
Figure 7. Projected Model Releases



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

Source: Citi Research, OpenRouter, Artificial Analysis, LLMTimeline, Model Provider Websites

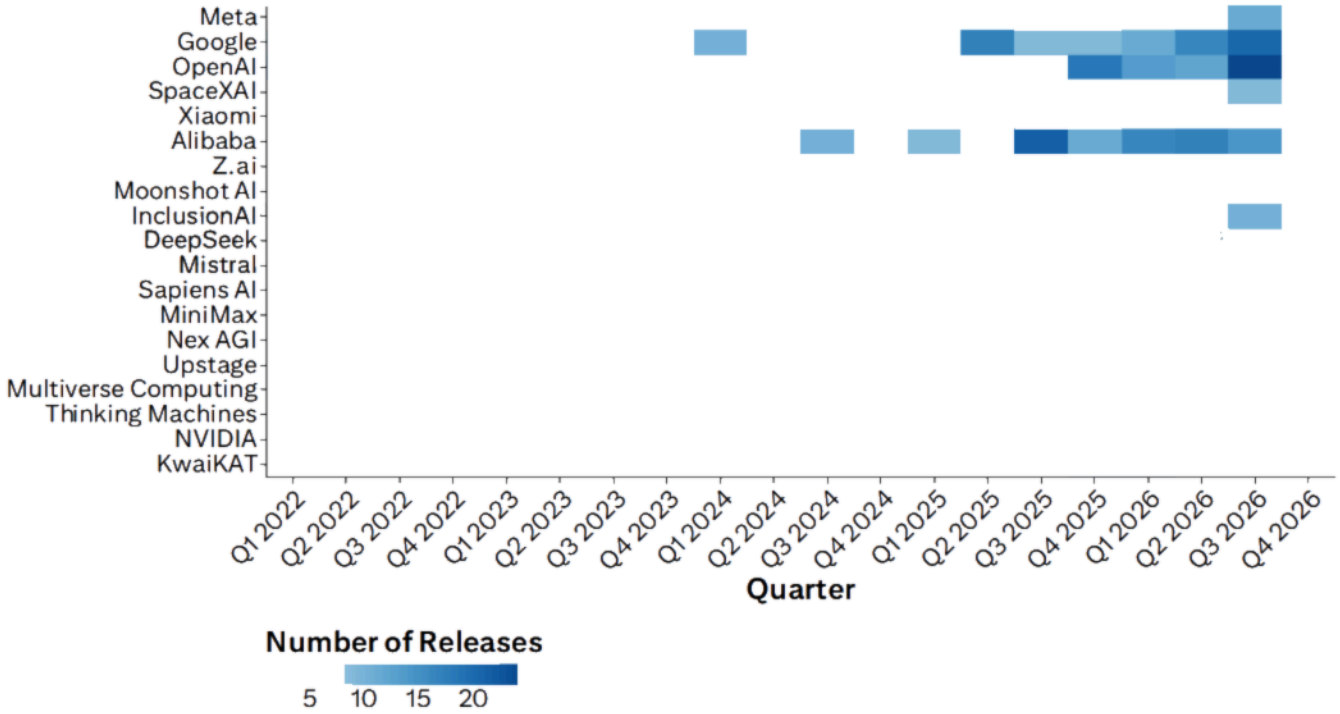
Figure 8. Historical Model Releases



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

Source: Citi Research, OpenRouter, Artificial Analysis, LLM Timeline, Model Provider Websites

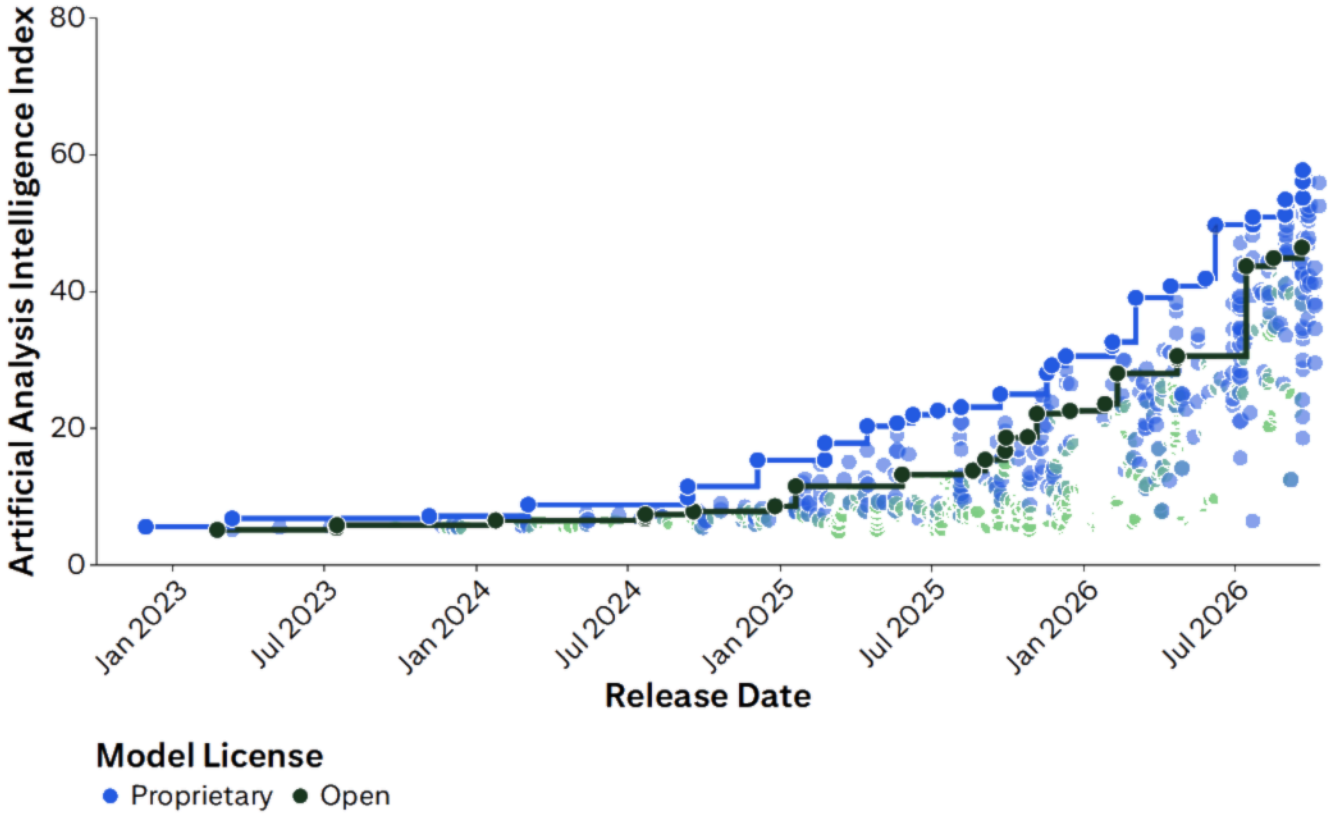
Figure 9. Model Release Cadence



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

Source: Citi Research, OpenRouter, Artificial Analysis, Anthropic, OpenAI, Moonshot AI, SpaceXAI, Z.ai, Meta, Google, Alibaba, Minimax, DeepSeek, Motif Technologies, Xiaomi, Tencent, Nex AGI, Thinking Machines, NVIDIA, InclusionAI

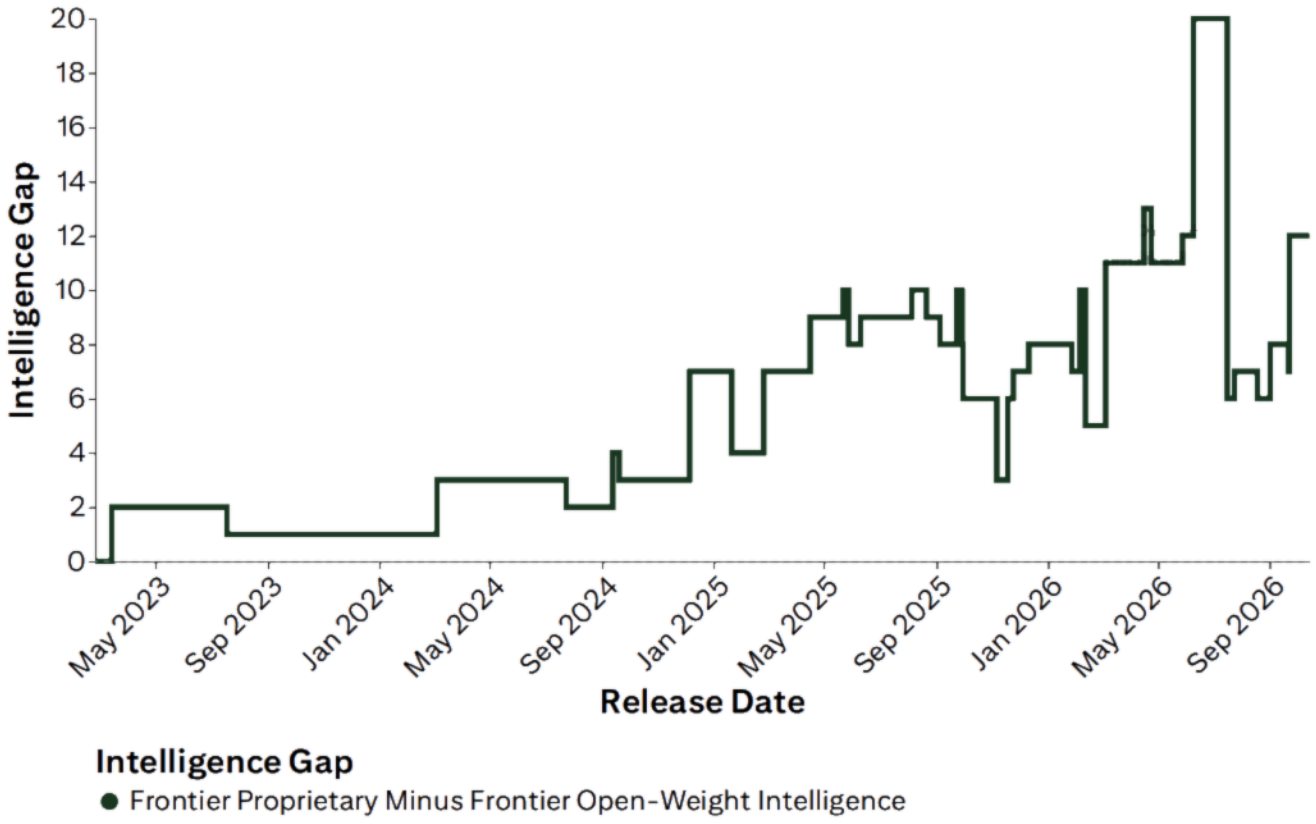
Figure 10. Frontier Releases for Proprietary and Open Models



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

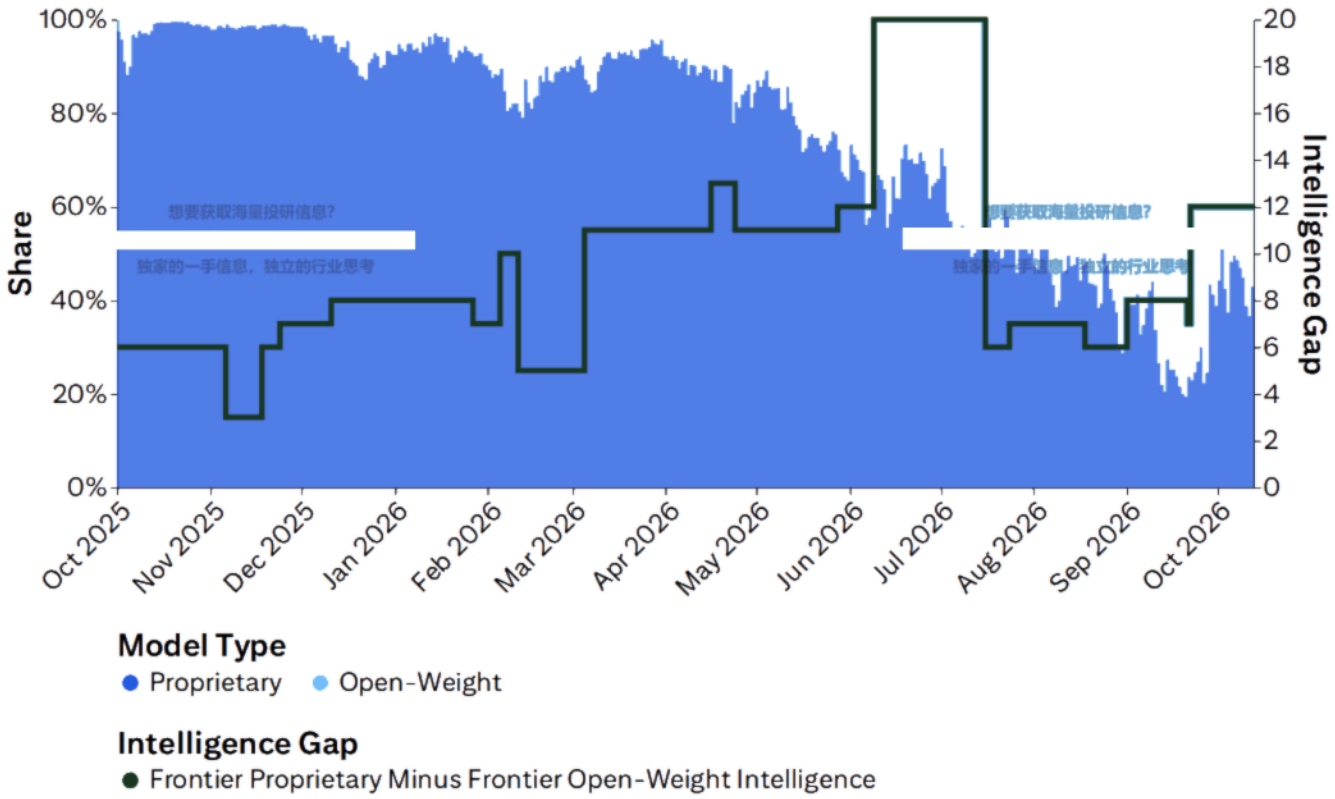
Source: Citi Research, Artificial Analysis

Figure 11. Gap in Proprietary and Open Intelligence



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

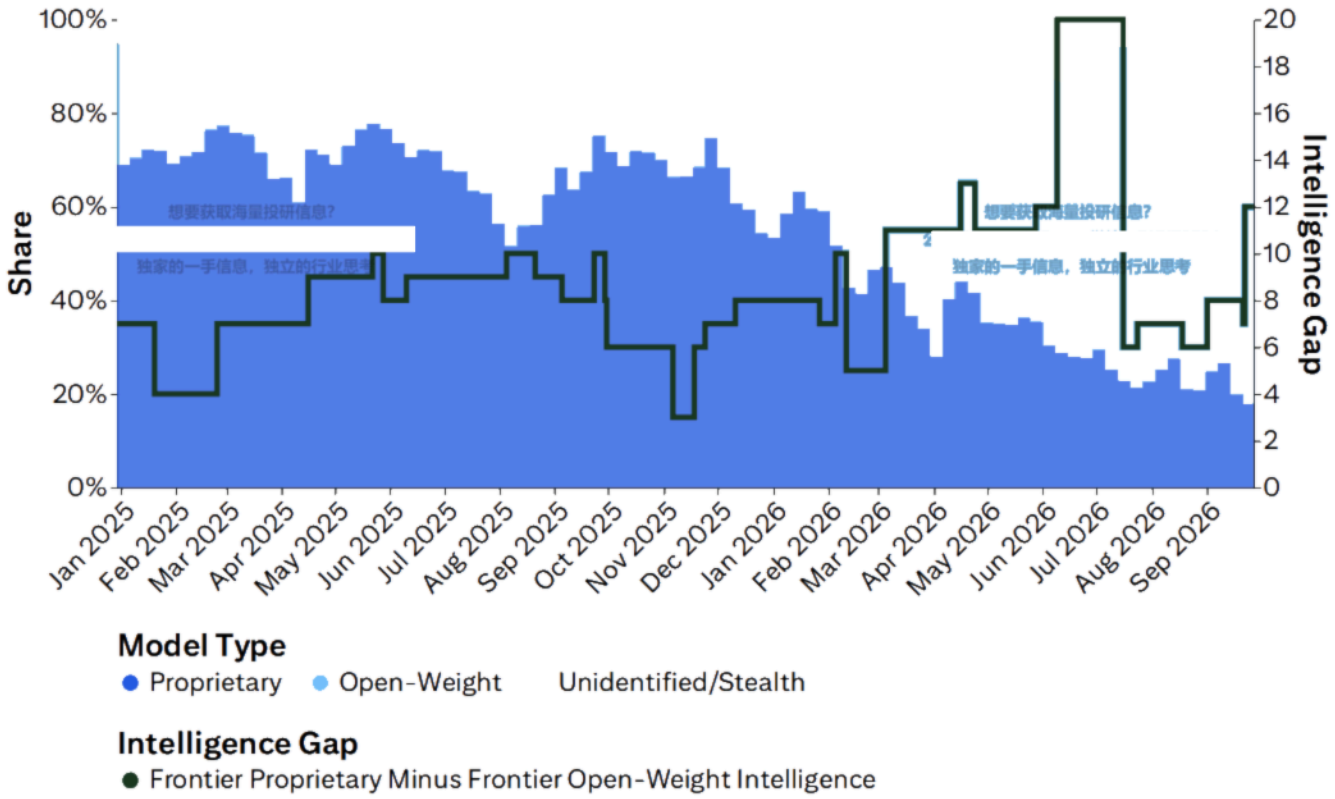
Figure 12. Vercel Token Volume Shares



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

Source: Citi Research, Vercel

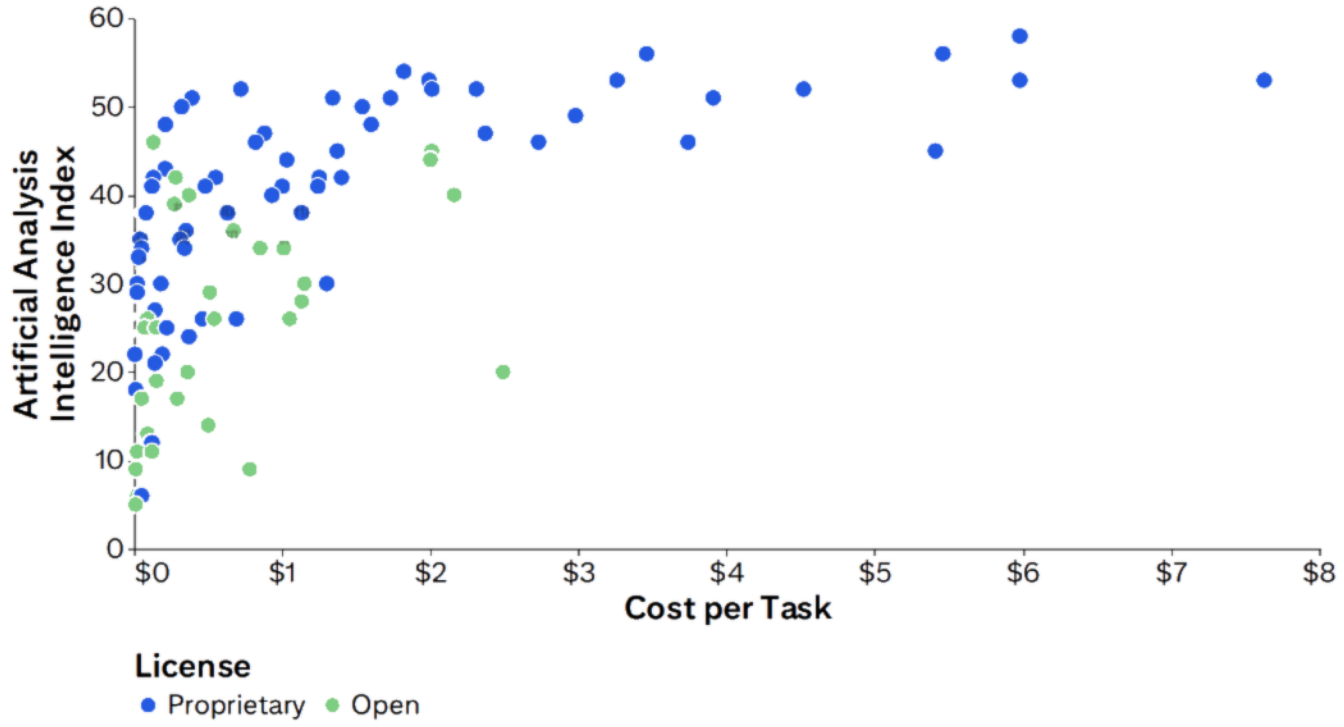
Figure 13. OpenRouter Token Volume Shares



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

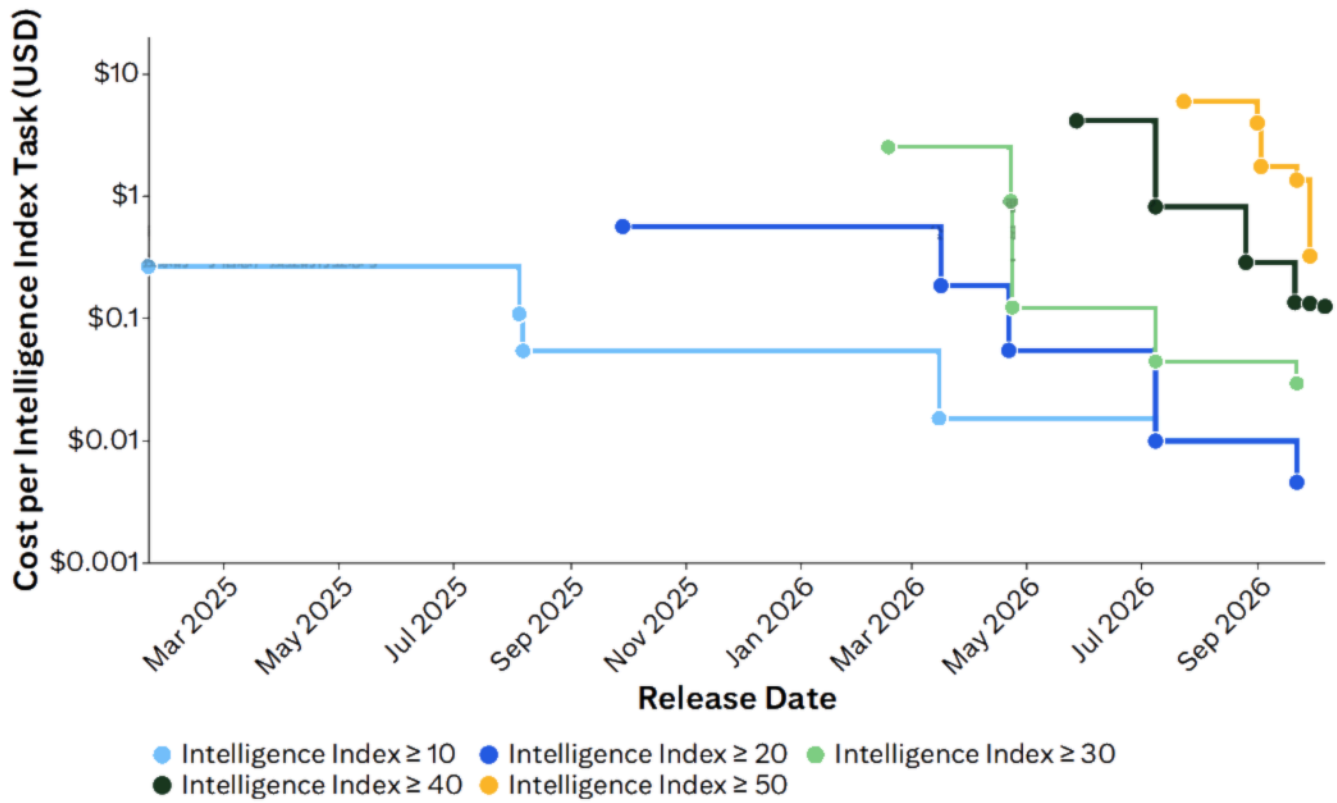
Source: Citi Research, OpenRouter

Figure 14. Intelligence v. Cost per AA Task



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

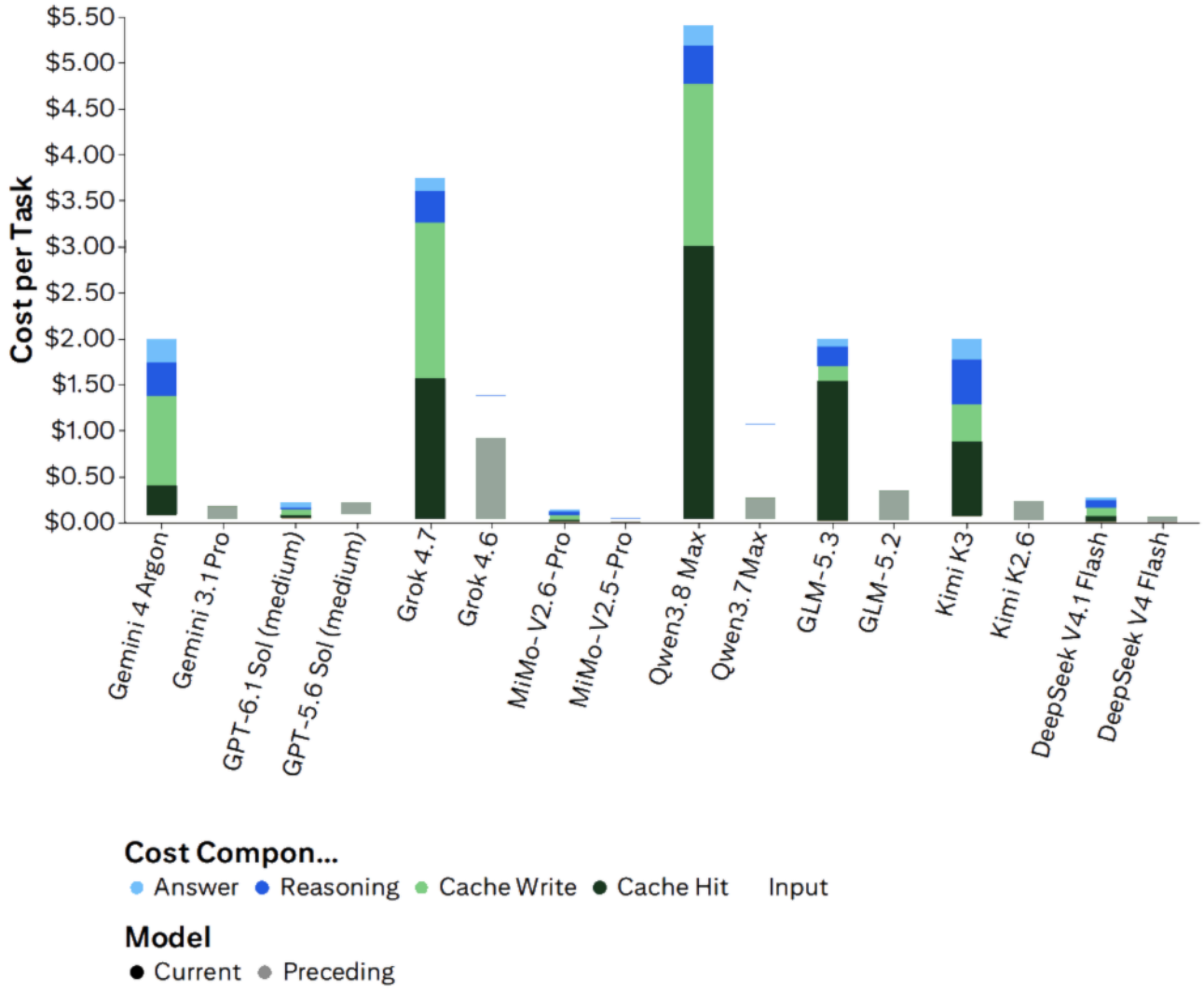
Figure 15. Intelligence Over Time v. Blended Price



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

Source: Citi Research, Artificial Analysis

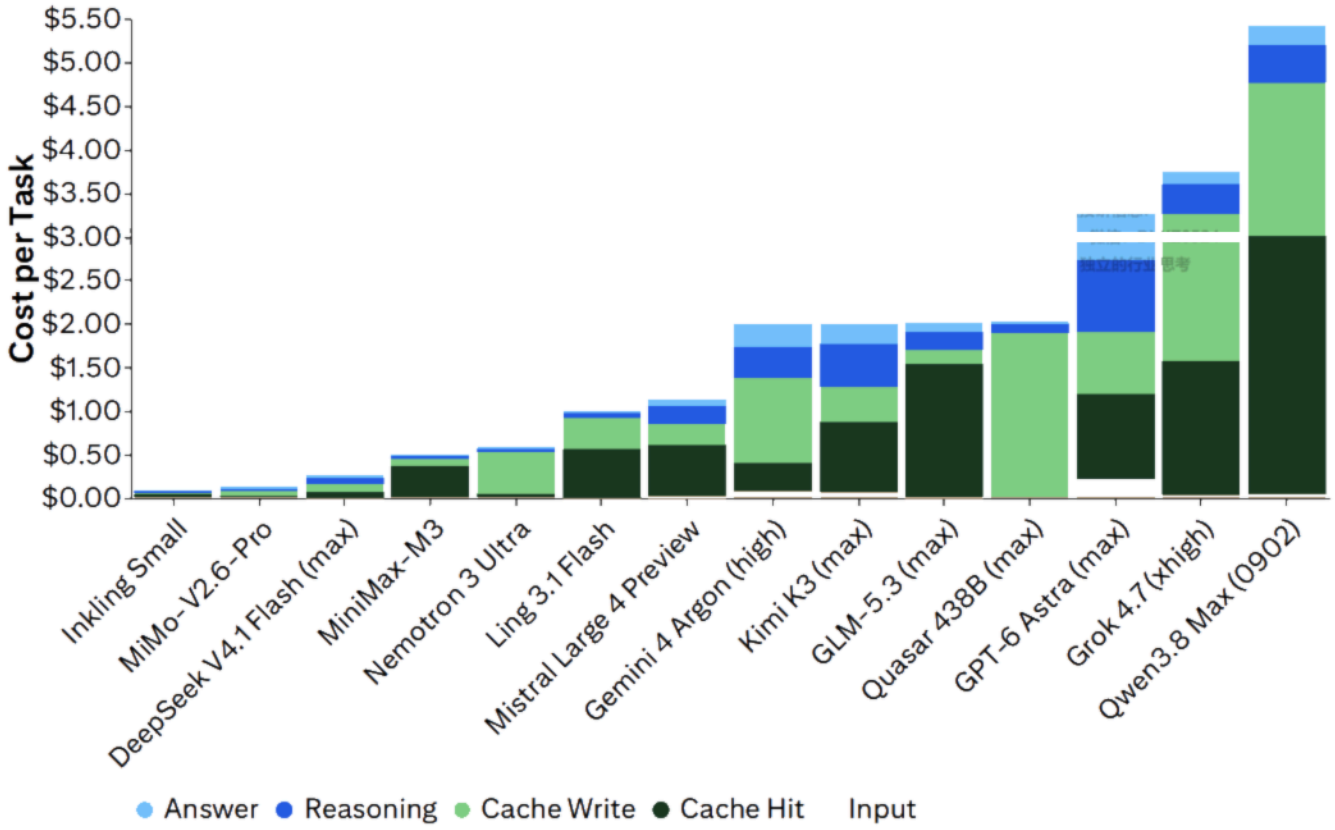
Figure 16. Cost per AA Task for Current Versus Prior Models



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

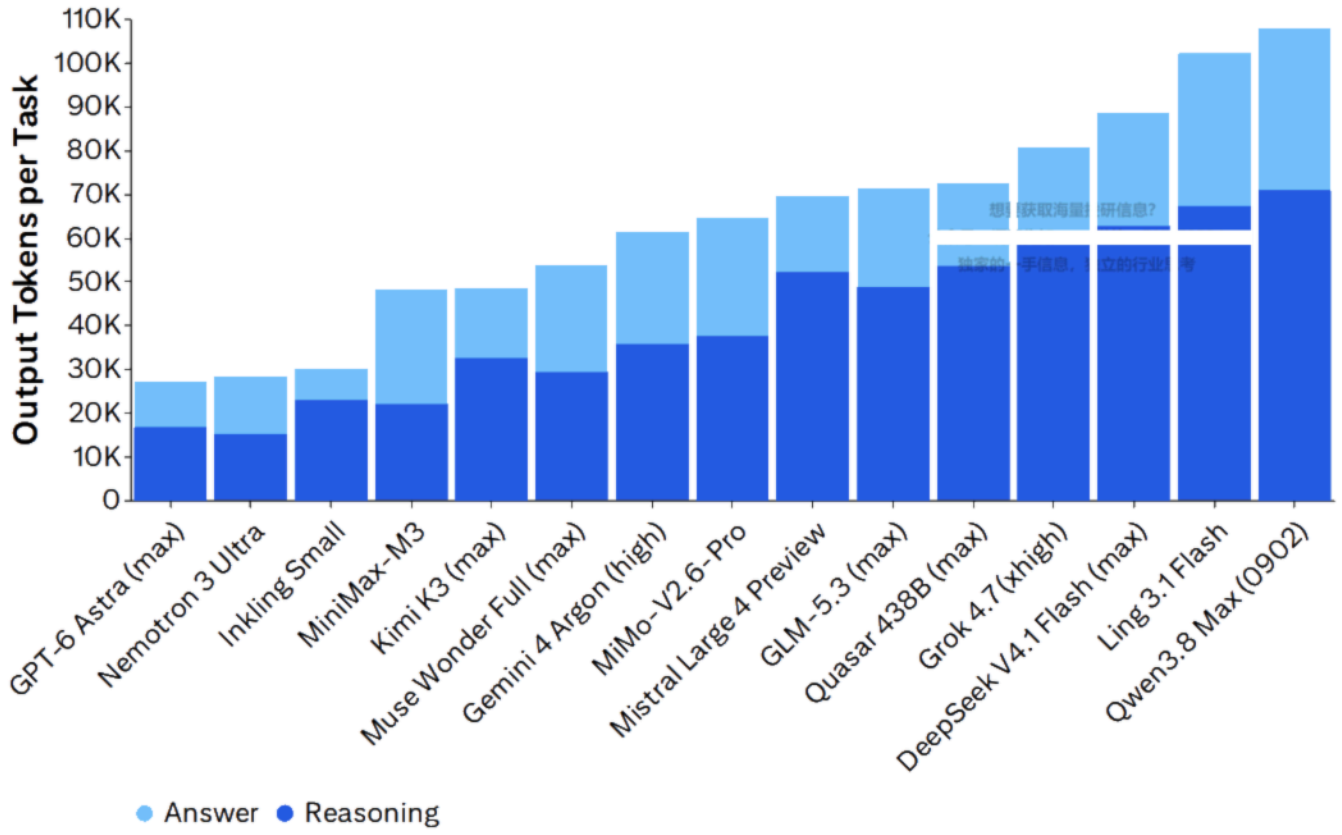
Source: Citi Research, Artificial Analysis

Figure 17. Cost Breakdown per AA Task



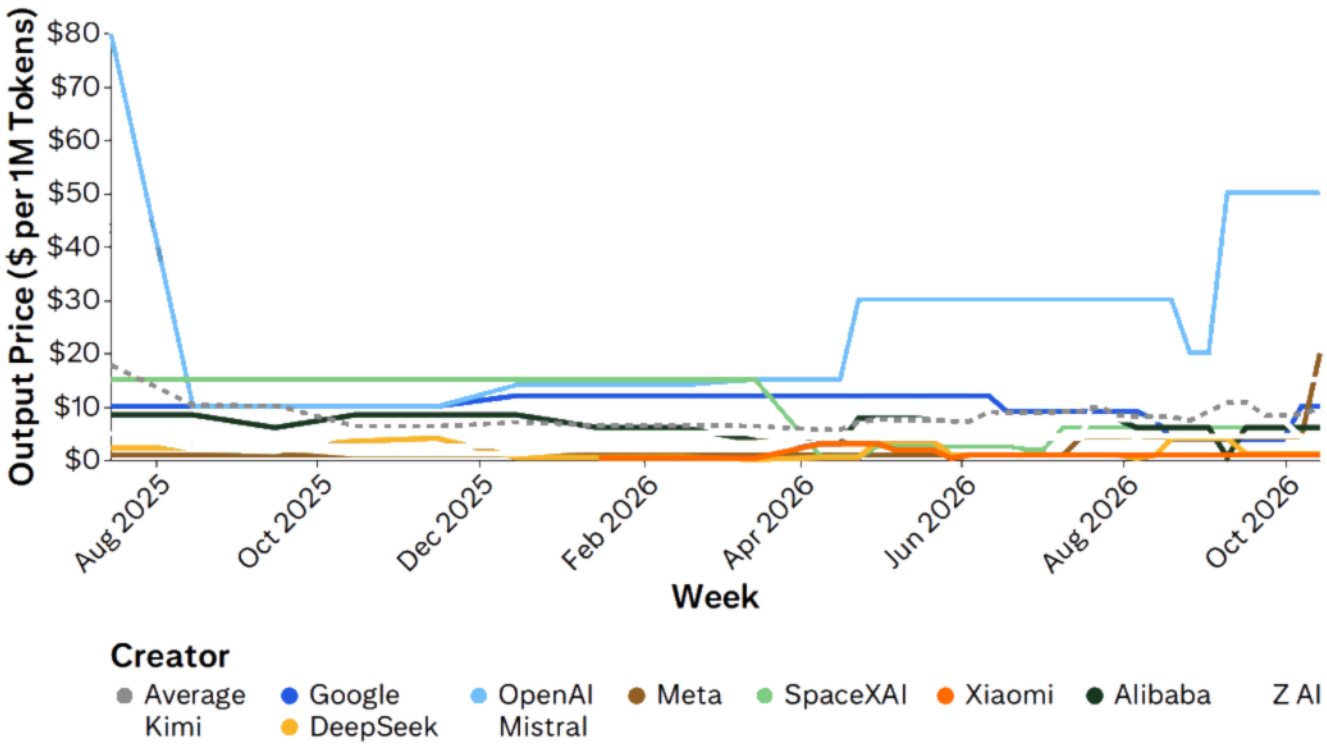
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Artificial Analysis

Figure 18. Output Tokens per AA Task



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

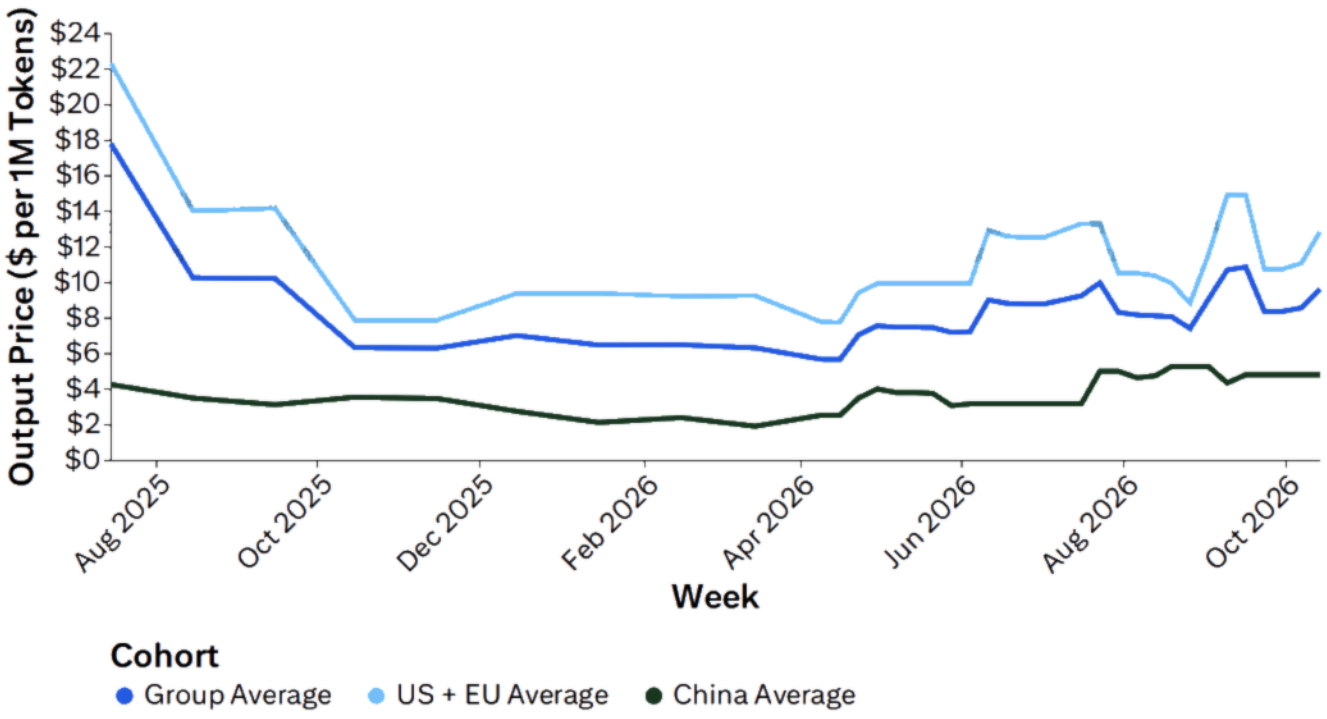
Figure 19. Output Token Pricing by Creator



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

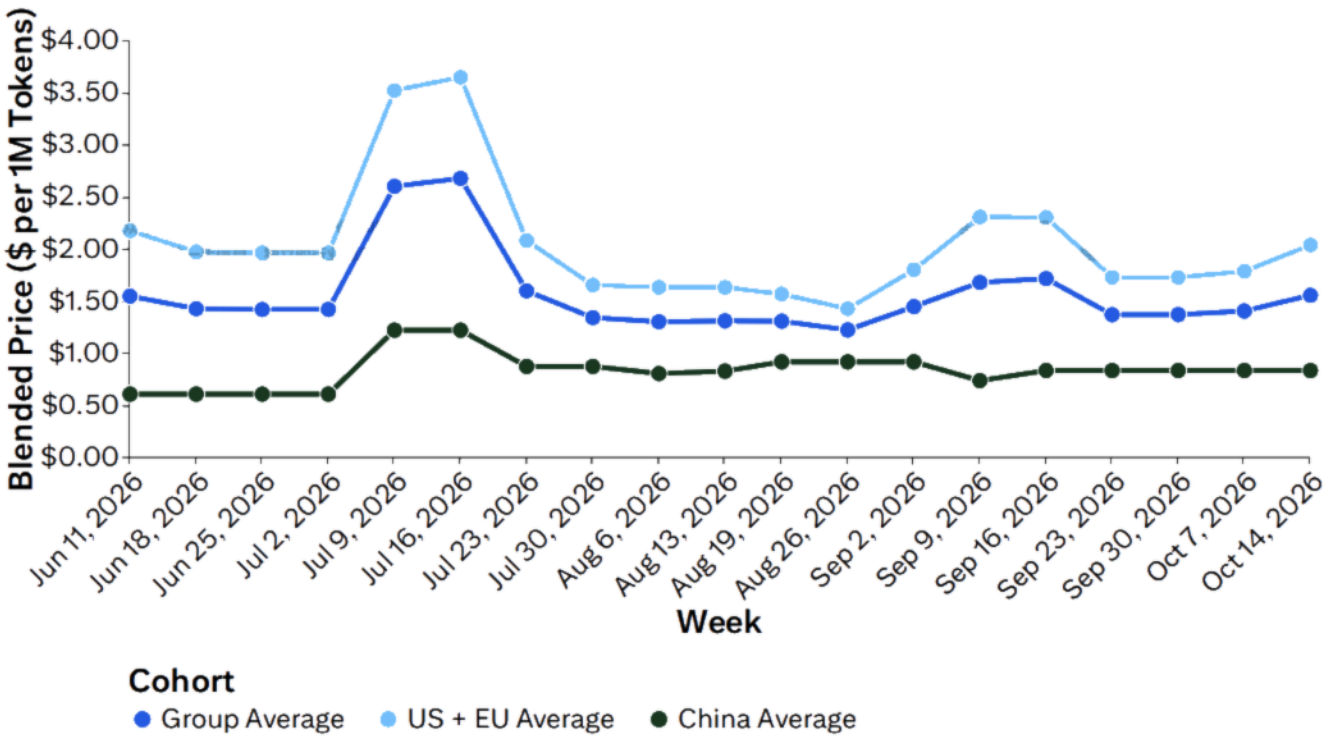
Source: Citi Research, Artificial Analysis

Figure 20. Output Token Pricing



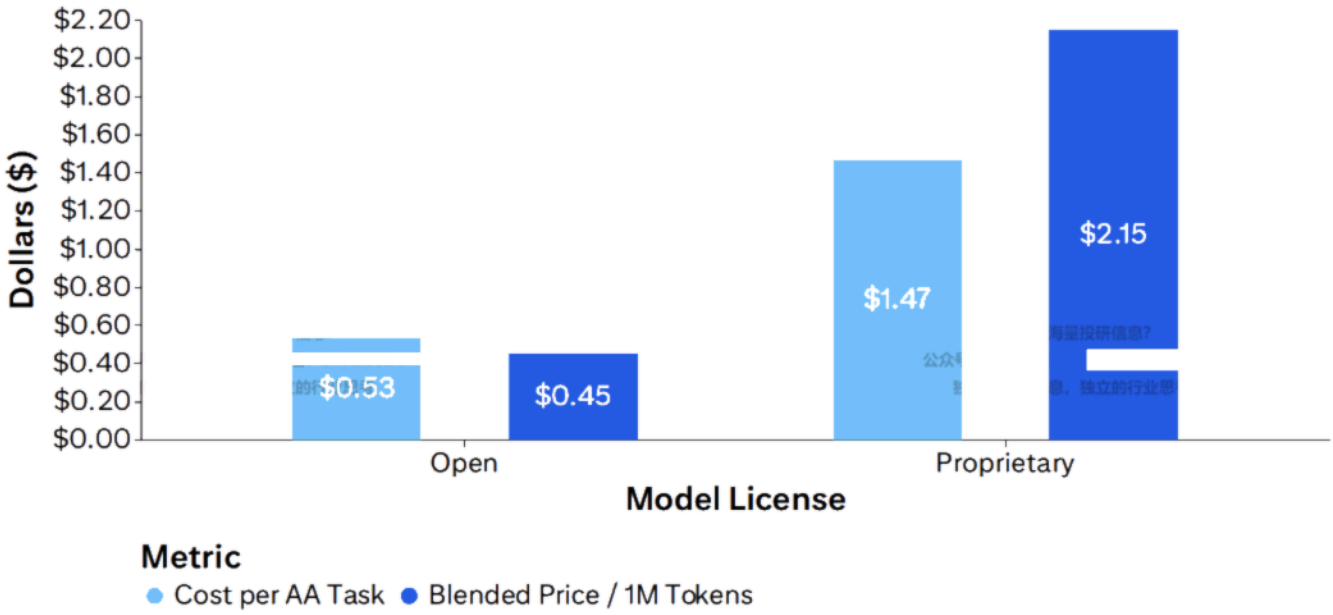
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Artificial Analysis

Figure 21. Blended Token Pricing



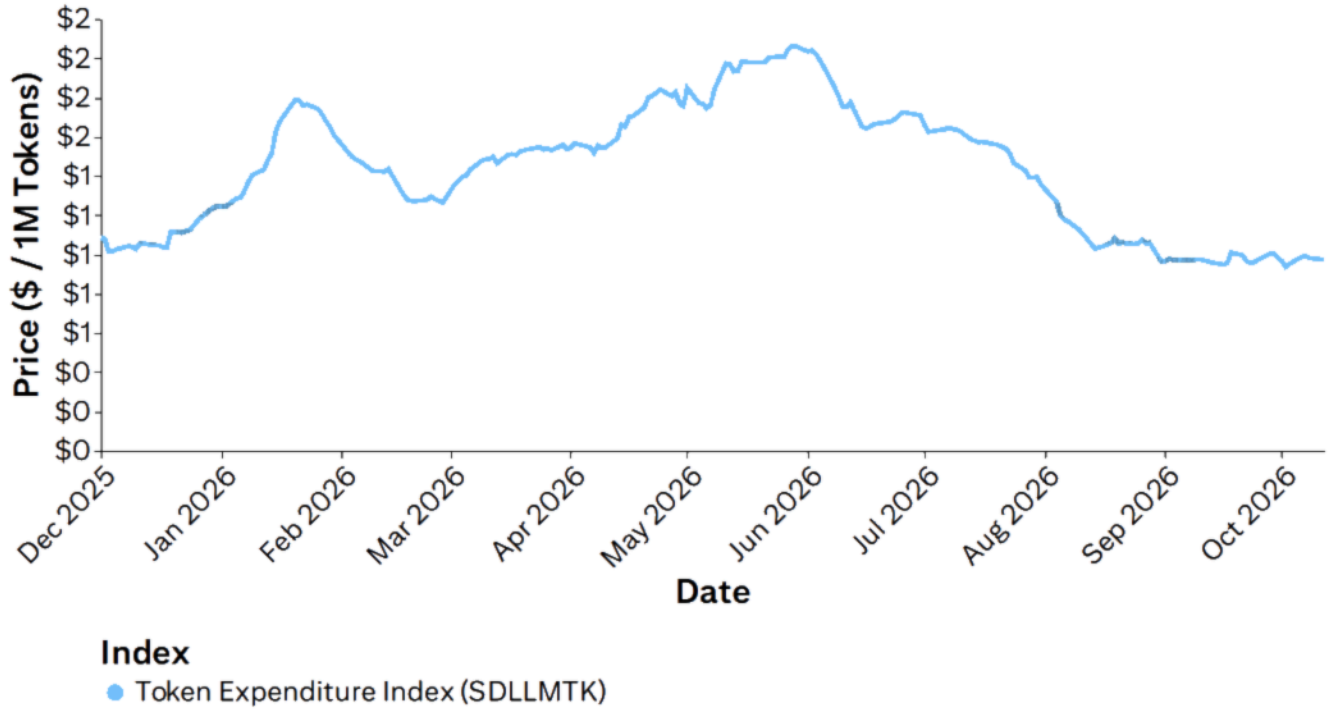
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research

Figure 22. Cost v. Blended Price



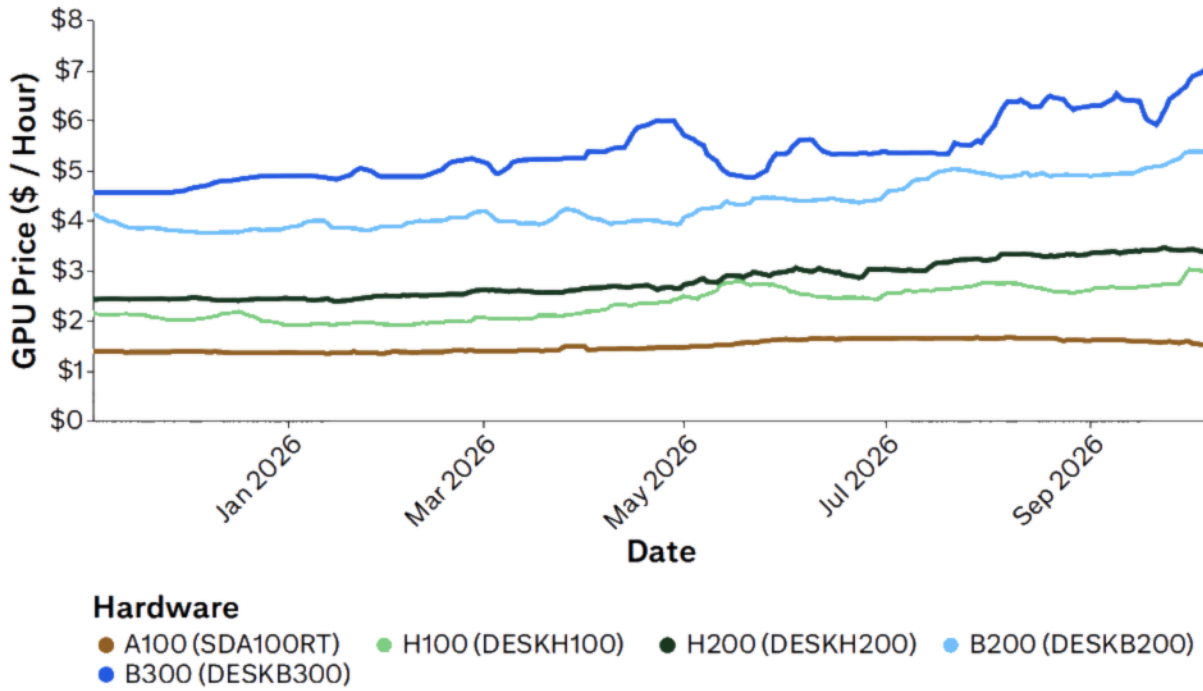
© 2026 Citigroup Inc. No redistribution without Citigroup's written permission.
Source: Citi Research, Artificial Analysis

Figure 23. Silicon Data Token Expenditure Index



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

Figure 24. Rental Pricing for GPUs



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

Figure 25. Recent Notes from Our Team

AI Industry Model	September 3, 2026	Artificial Intelligence AI Industry Model Update
Deep Dives	September 7, 2026	Space Exploration Technologies (SPCX Q): Potential for $1B0+ Tribution at $1T
Deep Dives	January 21, 2026	Artificial Intelligence OpenAI - Architecting the Abstraction Layer for Everything
Deep Dives	August 25, 2026	Artificial Intelligence AI - The Information Services Apex Technology
ROI on AI	October 1, 2026	Artificial Intelligence ROI on AI - The Future of OpenAI is More Control than Cost as Frontier Labs Pull Ahead
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	October 9, 2026	Artificial Intelligence Citi's Inference Ahead - Models Race Ahead While Infrastructure Catches Up
Citi's Inference Ahead	October 2, 2026	Artificial Intelligence Citi's Inference Ahead - Building the Rails for Scaled Intelligence
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 Build-out 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 & 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, Pickers, & Explorers
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 - Opens 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 - Returning Curves for Us All
One-Off Pieces	October 12, 2026	The State of AI 2026 - In discussion with Nathan Benaich: Video Webcast
One-Off Pieces	September 14, 2026	Artificial Intelligence Facing 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	October 8, 2026	Artificial Intelligence - inHANDS TRUCTURE Summit 2026: Brownfield, Capital Depth, and the Maturing Neocloud Stack
Commentary on Conferences and Events	October 2, 2026	US Application Software - Decagon Dialogues Showcases A More Connected Platform with Expanding Uses
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: Stackbuild - 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: Pochade - Day 2 Takeaways from Citi's Global TMT Conference
Commentary on Conferences and Events	September 9, 2026	Artificial Intelligence AI Search: 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 Stateful: 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 Genspace and Constellation - 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 Leaders: 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 Tea Talk
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	October 9, 2026	Space Exploration Technologies (SPCX Q): 3Q26 Likely Another Strong Q Along Multi-Year Catalyst-Rich Arc
Collaborative Pieces	October 9, 2026	North America Telecom Operators and Infrastructure - SpaceX Begins to Assemble Spectrum Layer Cake with Low-Band Purchase fro
Collaborative Pieces	October 9, 2026	North America Internet: The Muse Moment: Revenue Analysis by Vertical & Framework to Evaluate Muse's Impact Across Multiple Sec
Collaborative Pieces	September 29, 2026	Space Exploration Technologies (SPCX Q): Starship Flight 11 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 (Get AI Names) - Born to Run: Still Early in the Multi-Year AI Investment Cycle
Collaborative Pieces	August 9, 2026	Space Exploration Technologies (SPCX Q): 2Q26 Model Update, Reiterate $200 Price Target
Collaborative Pieces	August 6, 2026	Takeaways on Stellar's Mobile Strategy & Implications for the Category
Collaborative Pieces	August 4, 2026	Space Exploration Technologies (SPCX Q): 2026 Strong Start to Multi-Year Catalyst-Rich Arc
Collaborative Pieces	July 30, 2026	US Application Software: 2Q26 CIO Survey - Measured Optimism with Clear Commitment to AI
Collaborative Pieces	July 19, 2026	EV 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 Structures Caveats, Sept IC01 for Meta's Superintelligence Ambitions

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

Source: Citi Research

Figure 26. Sector-Specific Notes on AI

Hardware and Semis	August 4, 2026	ON Semiconductor (ON.O): "Beat & Raise" as AI Data Center Revenue Accelerates
Hardware and Semis	August 3, 2026	US Semiconductors and Semiconductor Equipment: Earnings so Far...Estimates Up, Stocks Down; Buyers of AMD, TXN, & AMAT
Hardware and Semis	July 16, 2026	TSMC (2330.TW): AI demand keeps accelerating; Upside surprise on further capex increase
Data Centers	July 23, 2026	US Engineering & Construction - 2Q26 E&C Preview – E&C Growth Cycle Still Durable
Data Centers	July 17, 2026	Asia Communication Infrastructure: 800VDC Architecture for Next-Generation AI Data Centers
Data Centers	July 2, 2026	US Industrial Technology & Mobility: Takeaways from Construction Economist Call
Internet	July 24, 2026	North America Internet - 2Q26 Preview: RDDT, PINS, SNAP, CVNA, and DASH
Internet	July 22, 2026	Alphabet Inc (GOOGL.O) - 1st Impression. AI Momentum Drives Cloud Accel & Backlog. Search & YT Momentum Continues
Internet	July 1, 2026	North America Internet: Meta Compute Unlocks Capacity (and FCF) for Meta's Superintelligence Ambitions
Software	July 31, 2026	US Internet and Software - 2Q26 Cloud Compare: Revenues Accelerate Amid Rising Capex Tide
Software	July 23, 2026	US Application Software - Expecting Diverging Trends Across 2Q Front Office Software Earnings
Software	July 9, 2026	US Software: Themes into 2Q26 Software Earnings - AI Drives Improvement in Spending; Remain Selective
Industrials	July 27, 2026	ITG Inc (ITG.O) - FTTH + Data Center Tailwinds Driving Growth: Initiate Buy/High Risk
Industrials	July 2, 2026	US Industrial Technology & Mobility: Takeaways from Construction Economist Call
Industrials	June 23, 2026	US Industrial Technology & Mobility - Automate 2026 Takeaways: Automation Momentum Remains Constructive
Utilities	June 16, 2026	Overcoming Gridlock 2.0: Solving Power Bottlenecks to Drive AI & Energy Security
Utilities	June 8, 2026	US Diversified Utilities: Crowd Metrics Moving Up; Taking Closer Look at Consensus Trades
Utilities	April 22, 2026	US Electric Utilities: Citi's AI Summit Takeaways for US Power/Utilities
Economics	August 3, 2026	US Economics - ISM Manufacturing boosted by AI tail-wind
Economics	July 9, 2026	Research @ Citi Episode 83: AI's Mid-Year Review - The Token Bottleneck
Economics	June 25, 2026	Global Economic Outlook & Strategy: The Resilient Global Economy - Cascading Supply Shocks
Corporate Credit	July 29, 2026	The Speculative Grade Beat - Underweight AI – Uncertainties Rising, Credit Not Adequately Compensating
Corporate Credit	July 24, 2026	The High-Grade Beat - Hyperscaler High Tide
Corporate Credit	July 7, 2026	US High-Grade Credit: AI and Credit Markets - The Buildout, The Risks, and How to Price Them
Miscellaneous	July 24, 2026	Emerging Markets Economic Outlook & Strategy - Artificial Intelligence and the EM Growth Divide
Miscellaneous	July 8, 2026	Global Theme Machine: Theme Leadership Shifts as Technology Improves but Healthcare Themes Lead Returns
Miscellaneous	June 10, 2026	Global Theme Machine: AI and Industrial Themes Extend Their Lead
CFO Survey	July 30, 2026	US Application Software: 2026 CFO Survey: Measured Optimism with Clear Commitment to AI
CIO Survey	July 9, 2026	Global Technology & Communications: Citi CIO Survey: IT Budgets Improve As AI Moves Higher In Priority
CIO Survey	April 9, 2026	Global Technology & Communications: Citi CIO Survey: IT Budgets Moderate Post 4Q Budget Flush
CIO Survey	January 12, 2026	Global Technology & Communications: Citi CIO Survey: Global IT Budgets Accelerate on Potential 4Q Budget Flush
CIO Survey	October 14, 2025	Global Technology & Communications: Citi CIO Survey: IT Budgets Move Below Historical Levels with US Weakening
CIO Survey	July 15, 2025	Global Technology & Communications: Citi CIO Survey: IT Budgets Moderate As Cyber Regains Top Priority
CISO Survey	June 9, 2026	North America Systems Software: On the Ground at Gartner Security Summit - Fresh Diligence on AI Security Trends
CISO Survey	January 12, 2026	North America Systems Software: State of Enterprise Cybersecurity Spending: Vol. V
CHRO Survey	July 9, 2025	US Application Software: 2025 CHRO Survey: Slow but Steady - HR Budgets Resilient in the Face of Macro Noise
Automation Survey	June 18, 2026	Global Industrials: 7th Proprietary Automation Survey Signals Encouraging Demand Outlook

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

Source: Citi Research

Advanced Micro Devices

(AMD.O; US$562.23; 1; 15 Oct 26; 16:00)

Valuation

Our TP on AMD is $800. We value AMD using SOTP analysis with data center GPU/CPU biz at $370/$325 (35/35x 2028 PE), client $32 (15x), gaming $2 (6x), embedded $30 (20x), and net cash per share ~$41. Our target PE multiples are in-line with respective peer group averages.

Risks

Competition: AMD competes directly with Intel in the microprocessor market. Consequently, any fluctuations in market share between AMD and Intel could result in risk to our estimates. AMD competes directly with NVIDIA in the graphics and AI GPU market. Consequently, any fluctuations in market share between AMD and NVIDIA could result in risk to our estimates.

PC End Market: AMD derives roughly 30% of sales from the PC industry, which is highly dependent on IT spending. Therefore, any major uptick/downtick in IT spending could result in upside/downside to our estimates and rating for AMD.

Valuation: While AMD is currently trading slightly above its historical range, revisions to consensus revenue estimates could result in high volatility to AMD stock.

Tender Offer: Any public or private tender offer for all or part of AMD's business could drive the stock higher and cause us to re-assess our rating.

If the impact on the company from any of these factors proves to be greater than we anticipate, the stock will likely have difficulty achieving our target price. However, should they be less than anticipated, the stock could trade above our target price.

Applied Materials Inc

(AMAT.O; US$430.01; 1; 15 Oct 26; 16:00)

Valuation

We value AMAT at $710 using a 28x P/E multiple on 2028E EPS. We view a 28x multiple, above AMAT's 3-year average multiple of 21x, as appropriate given multi-year WFE upcycle, secular tailwinds, and incremental AI data center investment.

Risks

Downside risks to the achievement of our target price include the following: 1) As fab utilization and capital equipment orders are closely linked to the stock price, any material differences to our supply/demand model (e.g. demand drops suddenly, or supply increases more rapidly than we predict) may cause

our valuation methodology to be inaccurate. 2) As the market share leader in multiple sub-segments of semiconductor equipment, AMAT is uniquely positioned to capitalize on inflections as well as face tough competition from large peers in industry cycles.

ITG Inc

(ITG.O; US$5.94; 1H; 15 Oct 26; 16:00)

Valuation

Our $17 target price reflects a ~10x EV/EBITDA multiple on our ~$252mn FY27E adj. EBITDA estimate. We believe a discount to peer average is warranted given ITG's high customer concentration, elevated net leverage, and limited public track record. That said, we see potential for multiple expansion over time as the company executes its growth strategy and diversifies its revenue base.

Risks

We rate ITG High Risk given the following primary factors: ITG has the highest customer concentration across our broader coverage and recent IPO/limited trading history.

Other risks associated with ITG that could impact the stock performance and its ability to reach our target price include: elevated net leverage could limit near-term incremental M&A; execution risk related to integrating multiple recent acquisitions (7 acquisitions in 2025 out of 22 since 2019); and historical telco capex cyclicality.

Space Exploration Technologies

(SPCX.O; US$148.62; 1; 15 Oct 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.

Texas Instruments Inc

(TXN.O; US$279.12; 1; 15 Oct 26; 16:00)

Valuation

Our target price of $345 is based on 26X C28E EPS, roughly in line with its 5-year trading multiple.

Risks

We note the following upside/downside risks to our target price:

End Market: Texas Instruments derives roughly 40% of sales from the industrial end market. Therefore, any major downtick or increase in the industrial end market could result in downside/upside to our estimates and rating for TI.

Competition: Texas Instruments competes with Analog Devices, Microchip, NXP and others in the analog market. Consequently, any loss of market share could result in downside/upside risk to our estimates and rating to TI.

Product Obsolescence: Texas Instruments spends a smaller percentage of its revenue on R&D compared to its competitors. Consequently, if TI fails to develop future products suitable for its customers, it could result in downside risk to our estimates and rating.

Inventory Risk: Texas Instruments sells roughly 35% of its products through distributors. Thus, any major inventory corrections in the distribution channel could result in downside risk to our estimates and rating.

Semiconductor Cycle: TI has broad-based exposure across multiple products and end markets. Therefore, if the overall semiconductor market enters a downturn, it could result in downside risk to our estimates and rating.

Macroeconomic: Texas Instruments' geographic exposure spans multiple geographies including the U.S., Europe, and Asia. As a result, any prolonged macroeconomic downturn could result in downside risk to our estimates and rating.