

17 Jun 2026 01:53:02 ET | 9 pages

Japan IT Services

Microsoft adopts pay-by-use, reportedly mulls DeepSeek V4 for low-cost AI agents – DeepSeek shock 2

CITI'S TAKE

Microsoft announced on the morning of June 17 that it was making M365 Copilot Cowork generally available under a pay-per-use model. In related news, Axios reported that Microsoft is considering adopting DeepSeek as a low-cost model for Cowork ([link](#)). We are not surprised by the adoption of a pay-per-use model as it aligns with our view that rising AI inference costs need to be passed on to customers. We believe the pay-per-use model will increase users' cost consciousness, likely reducing the undisciplined use of frontier AI models (although total usage will likely increase, driven by more disciplined use). Meanwhile, Microsoft's reported adoption of DeepSeek is unexpected. We had thought Chinese models unlikely to be adopted for security reasons, but rising component prices and the consequent increase in frontier AI model fees appear to have led Microsoft to consider adopting DeepSeek.

Moriya Koketsu^{AC}

DeepSeek — We see it as no exaggeration to call Microsoft considering adopting DeepSeek “DeepSeek shock 2.” This further substantiates the commoditization of AI models, and the competitive environment for frontier AI models will likely become even more challenging going forward. We believe the reason for Microsoft's apparent adoption of DeepSeek is rising inference costs due to soaring token fees for frontier AI models. For Microsoft, if customers do not use AI (computing power) due to soaring token fees, revenue will not increase. Meanwhile, if inference costs can be reduced by using DeepSeek, corporate inference demand will rise and Azure's computing power will be utilized. For example, assume frontier model usage fees are \$3 with Azure cloud usage fees for inference processing at \$2, while DeepSeek usage fees are \$1 with cloud usage fees also at \$2. In this case, what matters more to Microsoft than which model customers use is that inference workloads occur on Azure. While revenue sharing with DeepSeek and actual pricing structures are unclear, we believe considering low-cost model adoption aligns with the economic rationale of stimulating inference demand and expanding Azure usage. The important point in this report is that even Microsoft is considering hosting Chinese models. This suggests Chinese models, previously unlikely to be adopted due to security concerns, have room to expand market share in the enterprise AI market if cost competitiveness is sufficiently high.

Implications — As previously noted models are rapidly commoditizing. Chinese-made AI models like Kimi and GLM that approach frontier AI models are on the rise. Microsoft considering adopting DeepSeek suggests companies will implement multi-AI models. In this case, rather than vertically integrating with specific frontier AI models, companies will likely use multiple models according to purpose and cost. Additionally, vendor-free AI implementation will become important. We believe this trend remains positive for NRI, Fujitsu, and NEC, which support AI implementation based on companies' existing systems and business processes. Also, if companies can improve AI RoI through the low-cost DeepSeek model, computing power demand will likely continue.

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



Figure 1. Artificial Analysis ranking

Does not include the recently updated Kimi K2.7 and GLM-5.2. DeepSeek V4 Pro costs around 1/30 as much as GPT-5.5

Artificial Analysis Intelligence Index (5/1)			
Ranking	Company	AI Model	Index
1	Open AI	GPT-5.5 (xhigh)	60
2	Open AI	GPT-5.5 (high)	59
3	Anthropic	Claude-opus-4.7 (max)	57
4	Google	Gemini-3.1-pro-preview	57
5	OpenAI	GPT-5.4 (xhigh)	57
6	OpenAI	GPT-5.5 (medium)	57
7	Moonshot	Kimi-k-2.6	54
8	Xiaomi	Mimo-v2.5-pro	54
9	OpenAI	GPT-5.3-codex (xhigh)	54
10	xAI	Grok-4.3	53
11	Meta	Muse-Spark	52
12	Alibaba	Qwen-3.6-max-preview	52
13	Anthropic	Claude-sonnet-4.6 (max)	52
14	DeepSeek	DeepSeek-v4-pro (max)	52
15	Z.ai	GLM-5.1	51
16	Minimax	MiniMax-M2.7	50
17	OpenAI	GPT-5.4 mini (xhigh)	49
18	DeepSeek	DeepSeek V4 Flash (Max)	47
19	Google	Gemini 3 Flash	46
20	Alibaba	Qwen 3.5 397B	45



Artificial Analysis Intelligence Index (6/17)			
Ranking	Company	AI Model	Index
1	Anthropic	Claude Fable 5 (with fallback)	60
2	Anthropic	Claude Opus 4.8 (max)	56
3	OpenAI	GPT-5.5 (xhigh)	55
4	Anthropic	Claude Opus 4.7(max)	54
5	OpenAI	GPT-5.5 (high)	53
6	Google	Gemini 3.5 Flash	50
7	Anthropic	Claude Sonnet 4.6 (max)	47
8	Google	Gemini 3.1 Pro Preview	46
9	Alibaba	Qwen3.7 Max	46
10	MiniMax	MiniMax-M3	44
11	DeepSeek	DeepSeek V4 Pro (Max)	44
12	Meta	Muse Spark	43
13	Kimi	Kimi K2.6	43
14	Xiaomi	MiMo-V2.5-Pro	42
15	DeepSeek	DeepSeek V4 Flash (Max)	40
16	Z.AI	GLM-5.1	40
17	OpenAI	GPT-5.4 mini (xhigh)	40
18	MiniMax	MiniMax-M2.7	38
19	NVIDIA	Nemotron 3 Ultra	38
20	xAI	Grok 4.3 (high)	38

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

Source: Citi Research, Artificial Analysis