25 June 2026 (No. of pages: 2)

Japanese report:15 June 2026

The Rise of Agentic AI and the Polarization of AI Costs

Moving toward an era where the more advanced the AI is—and the more complex the tasks it handles—the higher the cost

Economic Research Dept. **Takayuki Nitta**
Miho Tanabe

Summary

- The pricing structure for generative AI is currently becoming increasingly polarized. While free and low-cost chat functions remain, there is a growing trend toward premium plans and pay-as-you-go pricing for high-performance models and business automation features. Most recently, US-based Anthropic had planned to transition "Claude Fable 5," which it launched on June 9, to a pay-as-you-go model (however, service has been suspended since June 12 following an order from the US government). The industry is shifting toward a structure where costs are kept low for light-duty applications, while demanding appropriate compensation for high-load, high-value-added uses.

- One factor behind this trend is the rapid expansion in the use of AI agents. Because AI agents internally repeat processes such as task planning, search, code execution, and retries, the volume of information processed by the AI tends to grow significantly, even if the user sees only a single output. Furthermore, AI service providers are under pressure to recoup their massive investments in AI infrastructure, which is likely why they decided to implement this pricing structure change.

- Meanwhile, it appears that companies—as users—are increasingly willing to pay more for AI services. As the range of tasks that AI can handle expands, companies are shifting their comparison framework from viewing AI as a "tool" to viewing it as a "human worker," making it easier for them to justify paying premium prices for high-performance models and agent capabilities.

- Japanese companies need to view generative AI not merely as a chat tool, but as an external computing resource that supports business processes. Building on this understanding, they must establish a cost-effective management framework by visualizing usage, selecting models based on specific use cases, and optimizing the combination of external AI services and their own management infrastructure.

Attention
This report is a summary translation. The official document is only in Japanese.