FOR IMMEDIATE RELEASE

Fixstars Launches GLM-5.3 on Fixstars Inference Cloud, a High-Speed AI Inference Service Hosted Entirely in Japan

Limited-time campaign: JPY 100,000 in token credits for corporate customers contracting directly with Fixstars

Nov 20, 2026 – Fixstars Corporation (TSE Prime: 3687, US Headquarters: Irvine, CA), a leading company in performance engineering technology, today announced that GLM-5.3 is now available on Fixstars Inference Cloud (FIC), a service that provides API access to high-speed AI models running on GPU servers in Japan.

The GLM-5.3 model offered on FIC has been further optimized by Fixstars and is priced at JPY 10.49 per million input tokens and JPY 210 per million output tokens, among the lowest rates offered by major providers.*

To mark the launch, Fixstars is also running a limited-time campaign that grants JPY 100,000 worth of tokens to corporate accounts that contract directly with Fixstars.

About GLM-5.3

GLM-5.3 is an open-weight large language model (LLM) released in August 2026 by Z.ai, a Chinese AI developer. It excels at complex software development and long-running agentic tasks, and delivers substantial performance gains over its predecessor, GLM-5.2.

Key Features of GLM-5.3 on Fixstars Inference Cloud

The GLM-5.3 model on FIC is quantized to NVIDIA's 4-bit floating-point format (NVFP4) to reduce computational load and memory usage. Fixstars, a company specializing in performance engineering, has further optimized the model, achieving both lower token prices and shorter response times.

Token pricing is among the lowest offered by major providers, and response times are among the fastest.

As of November 20, 2026, GLM-5.3 on FIC is available only to customers with a direct contract with Fixstars. Fixstars plans to make it available through FastMetal at a later date.

Item	Details
Token pricing (per million tokens)	Input: JPY 10.49 / Output: JPY 210 / Cached input: JPY 10.34
Time to first token (TTFT)	Median: 1.127 seconds. Measurement conditions: measured from Tokyo; 64 concurrent requests; agentic workload (average input of approx. 139,000 tokens, cache hit rate of approx. 96%); measured in November 2026

* Z.ai, the developer of GLM-5.3, offers the model through its official API at USD 1.40 per million input tokens and USD 4.40 per million output tokens. FIC's pricing is approximately one-twentieth of that for input and one-third for output. FIC's pricing is also among the lowest of the providers listed on OpenRouter. (As of November 20, 2026)

About Fixstars Inference Cloud

FIC runs open-weight AI models on GPU servers managed by Fixstars in Japan and makes them available through an OpenAI-compatible API. It enables companies restricted from transferring source code or internal documents overseas services to use the latest open-weight models in their business operations and development.

In addition to a direct contract with Fixstars, FIC is also available on a pay-per-token basis through the AI gateway FastMetal.

- **Data stays in Japan**: All inference servers and storage are located in Japan.
- **Operated by Fixstars**: Fixstars, a company listed on the TSE Prime Market, handles everything from optimizing the inference infrastructure to operations and maintenance.
- **Zero Data Retention**: Prompts and outputs are not stored after inference, are not recorded in logs, and are not used to train AI models.
- IDEs, CLIs, and coding agents that support the OpenAI-compatible API can connect to FIC simply by changing the endpoint and API key.

Benefits for corporate customers with direct contracts

- **Technical support in Japanese**: Fixstars engineers provide support in Japanese on topics such as API integration, rate-limit design, and keeping up with model updates.
- **Complimentary seminar invitation**: Customers are invited to attend one of Fixstars' regularly held paid seminars on AI adoption free of charge.

FIC Launch Campaign

Item	Details
Offer	JPY 100,000 worth of tokens for use on FIC, credited to the registered account
Eligibility	Corporations that register for FIC under a direct contract with Fixstars during the campaign period. * Registration is subject to review. * Registration and use via FastMetal are not eligible. * For full terms and conditions regarding token usage, please contact us.
Eligible models	GLM-5.3 and DeepSeek-V4.1-Flash-1017 (as of November 20, 2026). Models added in the future will also be eligible.
Registration period	November 20 – December 15, 2026
How to apply	https://www.fixstars.com/en/products-services/inference-cloud

###

About Fixstars Corporation

Fixstars is a technology company dedicated to accelerating AI inference and training through advanced software optimization solutions. It supports innovation in healthcare, manufacturing, finance, mobility, and other industries. For more information, visit: https://www.fixstars.com/

Media Contact

Public Relations, Fixstars Corporation

Email: press@fixstars.com

Tel: +81-3-6420-0751