

Zhipu releases GLM-5.3, claiming a 50% coding gain and parity with Mythos 5 on code review

Source copy kept by wheresthe.ai for the story "Zhipu releases GLM-5.3, claiming a 50% coding gain and parity with Mythos 5 on code review" (<https://wheresthe.ai/d/cn-zhipu-glm-5-3-release-2026-08/>).

Source: <https://docs.z.ai/release-notes/new-released>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 89481e98c56ea0559e450d1d61888a58b2bba872989ddfccef25f60ca93bcc99

Models

2026-08-26 GLM-5.3-Flash

Native visual capabilities enable the model to observe interfaces, rendering results, and interaction feedback—creating a closed loop across code, browsers, and GUIs.

Efficient hybrid architecture: Combines linear and sparse attention with 320B total parameters and 18B activated, significantly reducing compute and KV-cache requirements.

Beyond coding: Supports office document and financial research workflows, autonomously breaking down goals, using tools, and refining outputs. Learn more in our documentation .*

2026-08-18 GLM-5.3

Stronger Coding Capabilities: GLM-5.3 delivers a significant improvement in coding capabilities, achieving a 50% gain over GLM-5.2 on Z.ai Code Bench and reaching state-of-the-art (SOTA) performance among open-source models on public benchmarks, including Terminal Bench 3.0.

Emergent Cybersecurity Capabilities: GLM-5.3 matches Mythos 5 in white-box code review and vulnerability discovery. In collaboration with multiple cybersecurity teams, it has been tested on real-world targets and has identified a total of 2,436 vulnerabilities, including 1,097 medium- and high-severity vulnerabilities. Learn more in our documentation .*

2026-06-16 GLM-5.2

Supports 1M lossless context, significantly improving long-horizon task capabilities and reducing context drift and goal forgetting in complex tasks

Achieves open-source SOTA performance on coding and long-horizon task benchmarks, delivering more stable results in complex system engineering and deep debugging

Significantly improves the real-world developer experience, with more reliable project-level context handling, adherence to engineering standards, and multi-platform development. Learn more in our documentation .*

2026-04-07 GLM-5.1

Designed for long-horizon tasks, GLM-5.1 can work independently for up to 8 hours in a single run, enabling a full loop from planning and execution to iterative refinement and final delivery.

It demonstrates stronger engineering intelligence across autonomous planning, sustained execution, bug fixing, and strategy iteration, while achieving comprehensive capability alignment with Claude Opus 4.6. Built with multi-turn SFT, RL, and a process-quality evaluation framework, GLM-5.1 further improves stability, consistency, and tool use over extended tasks. Learn more in our documentation .*

2026-02-12 GLM-5

Designed for complex system engineering and long-range Agent tasks, GLM-5 shifts the paradigm from coding to engineering, demonstrating strong deep-reasoning performance in backend architecture, complex algorithms, and stubborn bug fixing.

It directly benchmarks against Claude Opus 4.5 in code-logic density and systems-engineering capability, and integrates DeepSeek Sparse Attention for higher token efficiency while preserving long-context quality. [Learn more in our documentation](#) .*

2026-01-14 GLM-Image

We've launched GLM-Image, a state-of-the-art image generation model built on a multimodal architecture and fully trained on domestic chips, combining autoregressive semantic understanding with diffusion-based decoding to deliver high-quality, controllable visual generation.

The update significantly enhances performance in knowledge-intensive scenarios, with more stable and accurate text rendering inside images, making GLM-Image especially well suited for commercial design, educational illustrations, and content-rich visual applications. [Learn more in our documentation](#) .*

2025-12-22 GLM-4.7

We've released GLM-4.7, our foundation model with significant improvements in coding, reasoning, and agentic capabilities. It delivers more reliable code generation, stronger long-context understanding, and improved end-to-end task execution across real-world development workflows.

The update brings open-source SOTA performance on major coding and reasoning benchmarks, enhanced agentic coding for goal-driven, multi-step tasks, and improved front-end and document generation quality. [Learn more in our documentation](#) .*

2025-09-30 GLM-4.6

We've launched GLM-4.6, the flagship coding model, showcasing enhanced performance in both public benchmarks and real-world programming tasks, making it the leading coding model in China.

The update also expands the context window to 200K, improving its ability to handle longer code and complex agent tasks. [Learn more in our documentation](#) .*

Was this page helpful? Yes No