

Zhipu releases GLM-5, a 744-billion-parameter open model for agentic engineering

Source copy kept by wheresthe.ai for the story "Zhipu releases GLM-5, a 744-billion-parameter open model for agentic engineering" (<https://wheresthe.ai/d/cn-zhipu-glm-5-release-2026-02/>).

Source: <https://z.ai/blog/glm-5>

Text saved 21 Oct 2026 (the page text after its scripts ran in a browser).

SHA-256 of the saved text: f4d46d7987f042ba100cb066f7bb609d37a8156c68dbc954937268da163984dc

2026-02-12 · Research

GLM-5: From Vibe Coding to Agentic Engineering

Try it at Z.ai Call it at Z.ai Z.ai Coding Plan GitHub HuggingFace  Tech Report

We are launching GLM-5, targeting complex systems engineering and long-horizon agentic tasks. Scaling is still one of the most important ways to improve the intelligence efficiency of Artificial General Intelligence (AGI). Compared to GLM-4.5, GLM-5 scales from 355B parameters (32B active) to 744B parameters (40B active), and increases pre-training data from 23T to 28.5T tokens. GLM-5 also integrates DeepSeek Sparse Attention (DSA), significantly reducing deployment cost while preserving long-context capacity.

Reinforcement learning aims to bridge the gap between competence and excellence in pre-trained models. However, deploying it at scale for LLMs is a challenge due to RL training inefficiency. To this end, we developed slime, a novel asynchronous RL infrastructure that substantially improves training throughput and efficiency, enabling more fine-grained post-training iterations. With advances in both pre-training and post-training, GLM-5 delivers significant improvement compared to GLM-4.7 across a wide range of academic benchmarks and achieves best-in-class performance among all open-source models in the world on reasoning, coding, and agentic tasks, closing the gap with frontier models.

GLM-5 is designed for complex systems engineering and long-horizon agentic tasks. On our internal evaluation suite CC-Bench-V2, GLM-5 significantly outperforms GLM-4.7 across frontend, backend, and long-horizon tasks, narrowing the gap to Claude Opus 4.5.

On Vending Bench 2, a benchmark that measures long-term operational capability, GLM-5 ranks #1 among open-source models. Vending Bench 2 requires the model to run a simulated vending machine business over a one-year horizon; GLM-5 finishes with a final account balance of \$4,432, approaching Claude Opus 4.5 and demonstrating strong long-term planning and resource management.

GLM-5 is open-sourced on Hugging Face and ModelScope, with model weights released under the MIT License. GLM-5 is also available on developer platform api.z.ai and BigModel.cn, with compatibility with Claude Code and OpenClaw. You can also try it for free on Z.ai.

Benchmark

GLM-5

(Thinking)

GLM-4.7

(Thinking)

DeepSeek-V3.2

(Thinking)

Kimi K2.5

(Thinking)

Claude Opus 4.5

(Extend Thinking)

Gemini 3.0 Pro

(High Thinking Level)

GPT-5.2

(xhigh)

REASONING Humanity's Last Exam 30.5 24.8 25.1 31.5 28.4 37.2 35.4

Humanity's Last Exam

w/ Tools

50.4 42.8 40.8 51.8 43.4* 45.8* 45.5* AIME 2026 I 92.7 92.9 92.7 92.5 93.3 90.6 - HMMT Nov. 2025 96.9 93.5
90.2 91.1 91.7 93.0 97.1 IMOAnswerBench 82.5 82.0 78.3 81.8 78.5 83.3 86.3 GPQA-Diamond 86.0 85.7 82.4
87.6 87.0 91.9 92.4 CODING SWE-bench Verified 77.8 73.8 73.1 76.8 80.9 76.2 80.0 SWE-bench Multilingual
73.3 66.7 70.2 73.0 77.5 65.0 72.0

Terminal-Bench 2.0

Terminus-2

56.2 /

60.7†

41.0 39.3 50.8 59.3 54.2 54.0

Terminal-Bench 2.0

Claude Code

56.2 /

61.1†

32.8 46.4 - 57.9 - - CyberGym 43.2 23.5 17.3 41.3 50.6 39.9 - GENERAL AGENT BrowseComp 62.0 52.0 51.4
60.6 37.0 37.8 -

BrowseComp

w/ Context Manage

75.9 67.5 67.6 74.9 67.8 59.2 65.8 BrowseComp-Zh 72.7 66.6 65.0 62.3 62.4 66.8 76.1 τ²-Bench 89.7 87.4 85.3
80.2 91.6 90.7 85.5

MCP-Atlas

Public Set

67.8 52.0 62.2 63.8 65.2 66.6 68.0 Tool-Decathlon 39.2 23.8 35.2 27.8 43.5 36.4 46.3 Vending Bench 2
\$4,432.12 \$2,376.82 \$1,034.00 \$1,198.46 \$4,967.06 \$5,478.16 \$3,591.33

*: refers to their scores of full set.

†: A verified version of Terminal-Bench 2.0 that fixes some ambiguous instructions.

See footnote for more evaluation details.

Office

Foundation models are moving from “chat” to “work,” much like Office tools for knowledge workers and programming tools for engineers.

GLM-4.5 is our first step for reasoning, coding, and agent, enabling the model to complete complex tasks. With GLM-5, we further enhance complex systems engineering and long-horizon agent capabilities. GLM-5 can turn text or source materials directly into .docx, .pdf, and .xlsx files—PRDs, lesson plans, exams, spreadsheets,

financial reports, run sheets, menus, and more—delivered end-to-end as ready-to-use documents.

Our official application, Z.ai is rolling out an Agent mode with built-in skills for PDF / Word / Excel creation, supporting multi-turn collaboration and turning outputs into real deliverables.

Westbrook High School Football Sponsorship Proposal NVIDIA Equity Research Report Google Earnings Review Prompt + task contextView full trajectory at Z.ai

You are writing a visually engaging and well-structured sponsorship proposal intended to be delivered as a DOC document.

Author background: The proposal is written on behalf of a U.S. high school student council.

Purpose of the document: The goal of this document is to present a clear and compelling proposal to potential sponsors in order to secure financial sponsorship for an upcoming school football game or football season.

The proposal should:

Introduce the football event and its significance within the school and local community

Explain how sponsorship funds will be used

Clearly outline sponsorship opportunities and benefits for sponsors

Demonstrate why sponsoring this event provides meaningful visibility and community engagement

Target audience: Local businesses, community organizations, and potential corporate sponsors interested in youth sports, education, and community involvement.

————— Overall positioning:

This is a formal but youth-led sponsorship proposal. The tone should be:

Positive, energetic, and respectful

Professional but approachable

Community-oriented and sincere

Avoid exaggerated claims or overly commercial language.

————— Required structure and content:

Introduction

Brief introduction of the school, student council, and football program

Purpose of the sponsorship request

About the Football Event

Description of the football game or season

Importance of football to school spirit, teamwork, and student life

Expected attendance (students, families, community members)

Use of Sponsorship Funds

How sponsorship money will support the event (equipment, facilities, uniforms, event operations, etc.)

Emphasis on student benefit and community impact

Sponsorship Opportunities

Different sponsorship levels (e.g., Gold, Silver, Bronze)

What sponsors receive at each level (logo placement, announcements, banners, programs, social media mentions, etc.)

Benefits to Sponsors

Brand visibility within the school and local community

Positive association with youth development and education

Opportunities for long-term partnership

Conclusion and Call to Action

Expression of appreciation

Clear next steps for interested sponsors

————— Visual and design requirements (very important):

The document must be visually rich and engaging. Include and reference visual elements such as:

Photos or image placeholders of football games, players, or school spirit events

Tables comparing sponsorship levels and benefits

Highlight boxes or callouts for key information

Use captions such as: "Image: Our school football team during a home game" "Table: Sponsorship levels and benefits overview"

Visuals should support clarity and excitement, not decoration only.

————— Color and style guidelines:

Use a colorful, energetic, and school-friendly visual style.

Suggested color palette (can be adapted to school colors):

Primary color (section titles): deep school color (e.g., navy blue or maroon)

Secondary color (subsections): lighter complementary color

Accent colors: bright but tasteful tones (e.g., gold, orange, or light blue)

Body text: dark gray or black

Table headers / highlight boxes: light, cheerful background colors

Color usage rules:

Use color to create visual hierarchy and excitement.

Avoid overly dark or dull designs.

Ensure good contrast for readability.

————— Writing and layout constraints:

Use clear, simple, and friendly language.

Paragraphs should be short and easy to read.

Do NOT insert line breaks in the middle of sentences.

Use bullet points and tables where appropriate.

Ensure the document reads well both on screen and when printed.

Quality bar:

The document should look like a well-prepared student council sponsorship proposal.

Sponsors should clearly understand the event, the value of sponsorship, and how to get involved.

The final output should be ready to be shared as a DOC file without further editing.

Image should be in the center.

Document (.docx) generated by GLM-5 Getting started with GLM-5 Use GLM-5 with GLM Coding Plan

Try GLM-5 in your favorite coding agents—Claude Code, OpenCode, Kilo Code, Roo Code, Cline, Droid, and more. <https://docs.z.ai/devpack/overview>

For GLM Coding Plan subscribers: Due to limited compute capacity, we're rolling out GLM-5 to Coding Plan users gradually.

Max plan users: You can enable GLM-5 now by updating the model name to "GLM-5" (e.g. in ~/.claude/settings.json for Claude Code). Other plan tiers: Support will be added progressively as the rollout expands. Quota note: Requests to GLM-5 consume more plan quota than GLM-4.7.

Prefer a GUI? We offer Z Code —an agentic development environment that lets you control (even remotely) multiple agents and have them collaborate on complex tasks.

Start building now: <https://z.ai/subscribe>

Use GLM-5 with OpenClaw

Beyond coding agents, GLM-5 also supports OpenClaw—a framework that turns GLM-5 into a personal assistant that can operate across apps and devices, not just chat.

OpenClaw is included in GLM Coding Plan. See the guidance.

Chat with GLM-5 on Z.ai

GLM-5 is accessible through Z.ai. Manually change the model option to GLM-5, if the system does not automatically do so. We offer both Chat and Agent mode for GLM-5:

Chat Mode: Instant response, interactive chat, lightweight delivery Agent Mode: Multiple tools, diverse skills, delivering results directly Serve GLM-5 Locally

The model weights of GLM-5 are publicly available on HuggingFace and ModelScope. For local deployment, GLM-5 supports inference frameworks including vLLM and SGLang. Comprehensive deployment instructions are available at the official GitHub repository.

We also support deploying GLM-5 on non-NVIDIA chips, including Huawei Ascend, Moore Threads, Cambricon, Kunlun Chip, MetaX, Enflame, and Hygon. Through kernel optimization and model quantization, GLM-5 can achieve a reasonable throughput on those chips.

Footnote Humanity's Last Exam (HLE) & other reasoning tasks: We evaluate with a maximum generation length of 131,072 tokens (temperature=1.0, top_p=0.95, max_new_tokens=131072). By default, we report the text-only subset; results marked with * are from the full set. We use GPT-5.2 (medium) as the judge model. For HLE-with-tools, we use a maximum context length of 202,752 tokens. SWE-bench & SWE-bench Multilingual: We run the SWE-bench suite with OpenHands using a tailored instruction prompt. Settings: temperature=0.7, top_p=0.95, max_new_tokens=16384, with a 200K context window. BrowseComp: Without context management, we retain details from the most recent 5 turns. With context management, we use the same discard-all strategy as DeepSeek-V3.2 and Kimi K2.5. Terminal-Bench 2.0 (Terminus 2): We evaluate with the Terminus framework using timeout=2h, temperature=0.7, top_p=1.0, max_new_tokens=8192, with a 128K context window. Resource limits are capped at 16 CPUs and 32 GB RAM. Terminal-Bench 2.0 (Claude Code): We evaluate in Claude Code 2.1.14 (think mode) with temperature=1.0, top_p=0.95, max_new_tokens=65536. We remove wall-clock time limits, while preserving per-task CPU and memory constraints. We fix environment issues introduced by Claude Code and also report results on a verified Terminal-Bench 2.0 dataset that resolves ambiguous instructions (see: <https://huggingface.co/datasets/zai-org/terminal-bench-2-verified>). Scores are averaged over 5 runs. CyberGym: We evaluate in Claude Code 2.1.18 (think mode, no web tools) with (temperature=1.0, top_p=1.0, max_new_tokens=32000) and a 250-minute timeout per task. Results are single-run Pass@1 over 1,507 tasks. MCP-Atlas: All models are evaluated in think mode on the 500-task public subset with a 10-minute timeout per task. We use Gemini 3 Pro as the judge model. r²-bench: We add a small prompt adjustment in Retail and Telecom to avoid failures caused by premature user termination. For Airline, we apply the domain fixes proposed in the Claude Opus 4.5 system card. Vending Bench 2: Runs are conducted independently by Andon Labs. Legal Privacy Policy Terms of Service © 2026 Z.ai Inc.