GLM-5: Open-Source Mixture-of-Experts Closes the Gap with Frontier Models

Kabui, Charles

2026-02-16

Figure 1: GLM-5

Z.ai released GLM-5, a 744 billion parameter model that uses a mixture-of-experts architecture, meaning it has many specialized sub-networks but only activates a small fraction (40B parameters) for any given task. This keeps it fast despite its size. It was trained on 28.5 trillion tokens of text. GLM-5 uses a memory-efficient attention method from DeepSeek to handle long documents cheaply, and was fine-tuned with "slime," a new reinforcement learning system that lets the model learn from feedback at scale. On benchmarks: 92.7% on AIME 2026 (competition math), 77.8% on SWE-bench (real-world software bug fixing), and 30.5 on Humanity's Last Exam (expert-level questions). On Vending Bench 2, which simulates running a business over a full year, GLM-5 finished with $4,432, ranking first among open-source models. The full weights are MIT-licensed.

The practical significance is straightforward. GLM-5 is the first open-source model to consistently match or approach proprietary models like Claude Opus 4.5 and Gemini 3 Pro across reasoning, coding, and long-running autonomous tasks simultaneously. Because only a fraction of the model activates at once, actual compute costs stay manageable despite the huge total size. It can be deployed locally using popular serving tools, including on non-NVIDIA hardware like Huawei Ascend chips.

This release narrows the open-source vs. proprietary gap to where the distinction is increasingly about ecosystem and API convenience, not raw capability. For teams building AI agents that act autonomously, GLM-5 is now a viable self-hosted alternative to the most capable closed models.

Sources:

- GLM-5 on Hugging Face
- GLM-5 Technical Blog
- GLM-5 GitHub Repository
- slime: Asynchronous RL Infrastructure
