Executive Summary

GLM-4.5: A Breakthrough in Agentic, Reasoning, and AI Foundation Models

The Problem

Current large language models face a critical challenge: while some excel at specific tasks like mathematical reasoning or coding, no single open-source model demonstrates mastery across all three essential domains for real-world problem-solving: Agentic abilities (interacting with external tools and environments), complex Reasoning (multi-step problem-solving), and advanced Coding (real-world software engineering). This fragmentation limits their effectiveness as comprehensive problem-solvers.

The Breakthrough

GLM-4.5 introduces a revolutionary **hybrid reasoning approach** that combines both thinking and direct response modes in a single Mixture-of-Experts (MoE) architecture. With 355B total parameters but only 32B activated parameters, the model achieves remarkable efficiency while delivering state-of-the-art performance across agentic, reasoning, and coding tasks, ranking **3rd globally** among all evaluated models including proprietary ones.

How It Works

The model leverages a multi-stage training process on 23T tokens, followed by expert model iteration and reinforcement learning. Its hybrid reasoning allows it to engage in deliberative thinking for complex problems while providing instant responses for simpler queries. Key innovations include loss-free balance routing, 2.5x more attention heads for improved reasoning, and the ability to process up to 128K tokens of context, achieving **91.0% on AIME 24** and **64.2% on SWE-bench Verified**.

Why This Matters

This breakthrough unifies the three critical capabilities needed for advanced AI systems in a single open-source model. Developers can now rely on one model that excels at web browsing agents, mathematical problem-solving, and software engineering tasks, dramatically simplifying deployment and reducing costs. The availability of both GLM-4.5 (355B) and GLM-4.5-Air (106B) versions makes this technology accessible to researchers and companies with different resource constraints.

The Business Opportunity

The release of GLM-4.5 creates new possibilities for autonomous AI systems that can handle complex real-world tasks—from customer service agents that can browse the web for answers to development tools that can write and debug production code. With performance competitive with closed-source models but available as open-source, it enables companies to build sophisticated AI applications without vendor lock-in while maintaining cutting-edge capabilities.