Executive Summary

GLM-4.5: A Breakthrough Open-Source AI Model Excelling at Complex Tasks

The Problem

Large language models have struggled to achieve excellence across all three critical domains: agentic abilities (interacting with real-world tools), complex reasoning (solving multi-step problems), and advanced coding (tackling real software engineering challenges). While proprietary models like OpenAI's o1/o3 and Anthropic's Claude have shown promise in specific areas, a single open-source model that excels across all these capabilities has been missing.

The Breakthrough

GLM-4.5 introduces a revolutionary **hybrid reasoning approach** that combines both thinking and direct response modes, enabling the model to adapt its processing style based on task complexity. Built as a Mixture-of-Experts (MoE) model with 355B total parameters but only 32B activated parameters, GLM-4.5 achieves remarkable efficiency while delivering top-tier performance across agentic, reasoning, and coding tasks.

How It Works

The model employs a sophisticated multi-stage training process on 23T tokens, followed by expert model iteration and reinforcement learning. GLM-4.5 automatically determines when to use extended reasoning versus direct responses, achieving **91.0% on AIME 24 mathematics problems** and **70.1% on TAU-Bench agentic tasks**. The architecture uses deeper networks rather than wider ones, with 96 attention heads that specifically enhance reasoning capabilities.

Why This Matters

This breakthrough democratizes access to state-of-the-art AI capabilities, allowing researchers and developers to build sophisticated applications

without relying on expensive proprietary APIs. The model's strong performance across diverse tasks—from scoring **64.2% on SWE-bench Verified** coding challenges to achieving **77.8% on BFCL v3 function calling**—makes it suitable for real-world deployment in production environments.

The Business Opportunity

Organizations can now deploy a single, efficient model that handles customer service agents, code development, mathematical reasoning, and complex problem-solving, reducing infrastructure costs while maintaining high performance across all domains.