

An In-Depth Comparative Analysis of GLM-5, MiniMax M2.5, and Kimi K2.5

TechRounder PDF Edition

Live article:

<https://www.techrounder.com/ai/an-in-depth-comparative-analysis-of-glm-5-minimax-m2-5-and-kimi-k2-5/>

By Vipin PG | Published February 14, 2026 | Updated February 15, 2026 | Format: Analysis | 5 min read

In brief

GLM-5, MiniMax M2.5, and Kimi K2.5 are 2026 open-weight frontier models built for long-horizon, agentic work, not just chat. Choose GLM-5 for heavy systems reasoning with sparse attention, M2.5 for fast, low-cost high-volume execution, and K2.5 for strongest native multimodal (text+vision) performance with a huge 256k context.

Key points

The article argues that the early 2026 release cycle marks a structural shift in AI, with open-weight flagships like Zhipu AI's GLM-5, MiniMax M2.5, and Moonshot AI's Kimi K2.5 pushing the industry beyond simple conversational copilots toward autonomous, long-horizon agentic engineering. It says these aren't incremental upgrades but a coordinated leap in architecture, reinforcement learning, multimodal integration, and cost efficiency, with benchmark performance that can match or even surpass proprietary systems such as Claude Opus 4.6 and GPT-5.2. The piece frames itself as a technical and strategic comparison spanning architecture, post-training reinforcement systems, benchmark results, deployment economics, and enterprise use cases. As a concrete example of the new scale-and-efficiency focus, it highlights GLM-5 as a 744B-parameter Mixture-of-Experts model with 40B active parameters per token, a 200,000-token context window, 28.5T training tokens, and DeepSeek Sparse Attention to make massive contexts practical without extreme latency or memory cost.

The early 2026 AI release cycle has triggered a structural shift across the global artificial intelligence field. Three flagship open-weight models, GLM-5 from Zhipu AI, MiniMax M2.5, and Kimi K2.5 from Moonshot AI, have redefined what frontier intelligence looks like. This is no longer about conversational "copilots." The industry is transitioning toward autonomous, long-horizon agentic engineering.

These models are not incremental upgrades. They represent a coordinated leap in architecture, reinforcement learning, multimodal integration, and economic efficiency. Across several standardized benchmark suites, they match or surpass proprietary systems such as Claude Opus 4.6 and GPT-5.2.

This article presents a complete technical and strategic comparison of GLM-5, MiniMax M2.5, and Kimi K2.5, covering architecture, post-training reinforcement systems, benchmarking results, deployment economics, and enterprise use cases.

Architectural Foundations and Pre-Training Scale

The February 2026 generation is defined not just by larger training corpora, but by deep architectural optimization. Massive context windows must now be served efficiently without extreme latency or memory cost.

GLM-5: Massive-Scale Systems Engineering

- Total Parameters: 744 billion (Mixture-of-Experts)
- Active Parameters per Token: 40 billion
- Context Window: 200,000 tokens
- Training Data: 28.5 trillion tokens
- Attention Mechanism: DeepSeek Sparse Attention (DSA)

GLM-5 more than doubles the parameter count of GLM-4.5. It replaces standard attention with DeepSeek Sparse Attention to dramatically reduce memory overhead when handling large contexts. The result is a model optimized for complex systems engineering and long-horizon reasoning.

MiniMax M2.5: Efficiency-First Intelligence

- Total Parameters: 230 billion
- Active Parameters per Token: 10 billion
- Context Window: 205,000 tokens
- License: MIT

M2.5 takes a different path. Instead of brute scaling, it focuses on throughput, task decomposition efficiency, and cost optimization. It is built for speed and affordability, enabling high-volume autonomous execution at scale.

Kimi K2.5: Native Multimodal Swarm Intelligence

- Total Parameters: 1 trillion (MoE)
- Active Parameters per Token: 32 billion
- Context Window: 256,000 tokens
- Architecture: 61-layer Transformer with Multi-head Latent Attention (MLA)
- Experts: 384 experts, 8 activated per token
- Training Data: 15 trillion multimodal tokens
- Vision Encoder: MoonViT (400M parameters)

Kimi K2.5 is natively multimodal. Vision and text were trained jointly from the beginning. Unlike adapter-based systems, this avoids degradation when reasoning across images, video, and text simultaneously.

Post-Training Reinforcement Learning Frameworks

The 2026 generation solves one of the biggest historical limitations of large models: long-horizon credit assignment.

GLM-5: Slime Asynchronous Reinforcement Learning

The Slime infrastructure separates generation and evaluation into asynchronous pipelines. This significantly increases training throughput and reduces hallucination rates. The model is trained to abstain when uncertain, especially in complex codebases.

MiniMax M2.5: Forge Framework

Forge is an agent-native reinforcement system that decouples training engines from agent logic. It uses tree-structured merging for samples, achieving up to 40x training acceleration. A process reward mechanism ensures optimal task decomposition.

Kimi K2.5: Parallel-Agent Reinforcement Learning (PARL)

PARL trains Kimi K2.5 to function as an orchestrator rather than a single agent. It can autonomously spawn up to 100 sub-agents and execute 1,500 parallel tool calls. Reward shaping prevents sequential fallback and useless parallelism.

Hardware Infrastructure and Geopolitical Resilience

GLM-5 is a landmark shift in hardware independence. It was trained entirely on Huawei Ascend chips using the MindSpore framework, without NVIDIA GPUs.

This challenges the assumption that export controls can permanently constrain frontier AI development.

However, GLM-5's open weights require approximately 1,490GB of memory for deployment. Running it locally demands datacenter-grade infrastructure, often requiring eight H200-class accelerators for efficient FP8 inference.

MiniMax M2.5, by contrast, operates with far lower hardware overhead and can sustain 100 tokens per second efficiently.

Comprehensive Benchmarking Methodology

Evaluation was conducted using the Artificial Analysis Intelligence Index v4.0. The index includes:

- Agentic Workflows
- Coding and Software Engineering
- General Knowledge and Hallucination
- Scientific and Mathematical Reasoning

All evaluations used zero-shot prompting in Ubuntu 22.04 LTS with Python 3.12. Scores are based on pass@1 metrics.

Software Engineering Performance (SWE-bench Verified)

Model: Claude Opus 4.6 | SWE-bench Verified: 80.9% | SWE-bench Multilingual: 77.5%

Model: MiniMax M2.5 | SWE-bench Verified: 80.2% | SWE-bench Multilingual: -

Model: GLM-5 | SWE-bench Verified: 77.8% | SWE-bench Multilingual: 73.3%

Model: Kimi K2.5 | SWE-bench Verified: 76.8% | SWE-bench Multilingual: 73.0%

M2.5 reaches near parity with Claude Opus 4.6 while consuming fewer tokens and operating at 10% of the cost. GLM-5 demonstrates strong multilingual robustness. Kimi K2.5 continues to improve in structured code generation and debugging.

Scientific and Mathematical Reasoning

In GPQA-Diamond and advanced mathematics benchmarks, Kimi K2.5 and GLM-5 show competitive results against GPT-5.2 and Claude Opus 4.6.

When tool use is enabled in Humanity's Last Exam (HLE), both GLM-5 and Kimi K2.5 show massive score jumps. This confirms the importance of agentic reinforcement frameworks in boosting practical reasoning.

Agentic Workflows and Office Productivity

MiniMax M2.5

- 59% win rate in Cowork Agent evaluation
- 76.3% on BrowseComp
- Strong performance in Excel World Championship simulations

Kimi K2.5

- 59.3% improvement over predecessor in AI Office Benchmark
- Excels in Word annotations, pivot tables, and long document synthesis

GLM-5

- First among open models on Vending Bench 2
- Demonstrates long-horizon planning and macroeconomic simulation strength

Long Context Processing

- GLM-5: 200K tokens
- MiniMax M2.5: 205K tokens
- Kimi K2.5: 256K tokens

Chinese-origin models demonstrate stronger robustness in Chinese long-context scenarios, while GPT and Claude series show stronger English long-context alignment.

Native Multimodality Advantage

Kimi K2.5 leads in multimodal reasoning:

- MMMU Pro: 78.5%
- MathVision: 84.2%
- VideoMMMU: 86.6%

It supports autonomous UI rendering, visual debugging, and vision-to-code pipelines. GLM relies on GLM-4.5V for visual tasks, while MiniMax M2.5 currently lacks native image input support.

Economic Analysis and API Pricing

Model: MiniMax M2.5 | Input (\$/1M): \$0.30 | Output (\$/1M): \$2.40 | Blended Cost: \$0.82

Model: Kimi K2.5 | Input (\$/1M): \$0.60 | Output (\$/1M): \$3.00 | Blended Cost: \$1.20

Model: GLM-5 | Input (\$/1M): \$1.00 | Output (\$/1M): \$3.20 | Blended Cost: \$1.55

Model: Claude Opus 4.6 | Input (\$/1M): \$5.00+ | Output (\$/1M): \$25.00 | Blended Cost: ~\$10.00

M2.5 is the clear economic leader. It can operate continuously at roughly \$1 per hour. Kimi K2.5 introduces aggressive context caching, reducing repeated input costs by 83%.

Latency and Token Efficiency

GLM-5 generates more reasoning tokens but processes them faster. In tests generating 500 tokens:

- GLM-5: 39.7 seconds
- Kimi K2.5: 62.3 seconds

GLM-5's faster internal reasoning throughput offsets its verbosity.

Enterprise Use Case Mapping

Kimi K2.5: Multimodal Research Orchestrator

- Competitive intelligence synthesis
- Vision-to-code pipelines
- Parallel browser automation

MiniMax M2.5: Persistent Codebase Worker

- CI/CD automation
- Repository-wide refactoring
- Autonomous financial modeling

GLM-5: Backend Architect and Macro Simulator

- Microservice architecture design
- Multilingual enterprise stacks
- Long-horizon operational planning

Strategic and Market Implications

The rise of GLM-5, MiniMax M2.5, and Kimi K2.5 marks the erosion of proprietary dominance. Open-weight models now match or exceed closed systems across coding, reasoning, and agentic execution.

The industry is entering the Swarm Era. Human operators are no longer step-by-step copilots. They become strategic directors of autonomous digital labor.

Finally, GLM-5 proves that frontier AI no longer depends exclusively on a single global semiconductor supply chain. Parallel hardware ecosystems are now viable.

For enterprises in 2026, the message is clear. The advantage no longer lies in owning the largest conversational model. It lies in orchestrating autonomous, cost-efficient agentic systems at scale.