Xiaomi surprises with MiMo-V2-Pro: 1 trillion parameters, performance close to GPT-5.2 but much cheaper.

Xiaomi's MiMo-V2-Pro surprises with performance close to GPT-5.2 but at a much lower cost, ushering in an era of actionable AI.

Chinese tech giant Xiaomi has surprised the global AI community by announcing MiMo-V2-Pro — a platform model with up to 1 trillion parameters.

According to the announcement, the performance of this model is approaching that of leading systems from OpenAI and Anthropic, but the cost of using the API is only about 1/6 to 1/7. In addition, the model is optimized to handle fewer than 256,000 tokens per interaction, significantly reducing operating costs.

The project is led by Fuli Luo, who previously worked on the DeepSeek R1. He calls it a "silent ambush" on the global AI market. Xiaomi also revealed plans to open-source a version of the model once the system is stable enough.

From chatbots to 'action-oriented AI'

Unlike many conversation-focused models, MiMo-V2-Pro aims for an 'action space' — where AI not only creates content but also directly performs complex tasks.

This represents a shift from conversational AI to agent-based AI, capable of acting as a 'brain' for large systems, from supply chains to automated programming agents.

This platform also reflects Xiaomi's long-standing strengths in hardware and the IoT ecosystem. After expanding into electric vehicles with models like the Xiaomi SU7, the company is gradually building an integrated ecosystem encompassing devices, software, and AI.

One of the major challenges of AI today is balancing inference capability with processing cost. MiMo-V2-Pro addresses this problem with its 'sparse' architecture — despite having 1 trillion parameters, only about 42 billion parameters are activated in each processing iteration.

As a result, the model is both significantly more powerful than its predecessor (MiMo-V2-Flash) and more resource-efficient.

In addition, Xiaomi uses a Hybrid Attention mechanism with a 7:1 ratio to handle contexts with up to 1 million tokens. This approach allows the model to 'quickly scan' most of the data while focusing on the most important parts – much like a researcher sifting through information in a massive library.

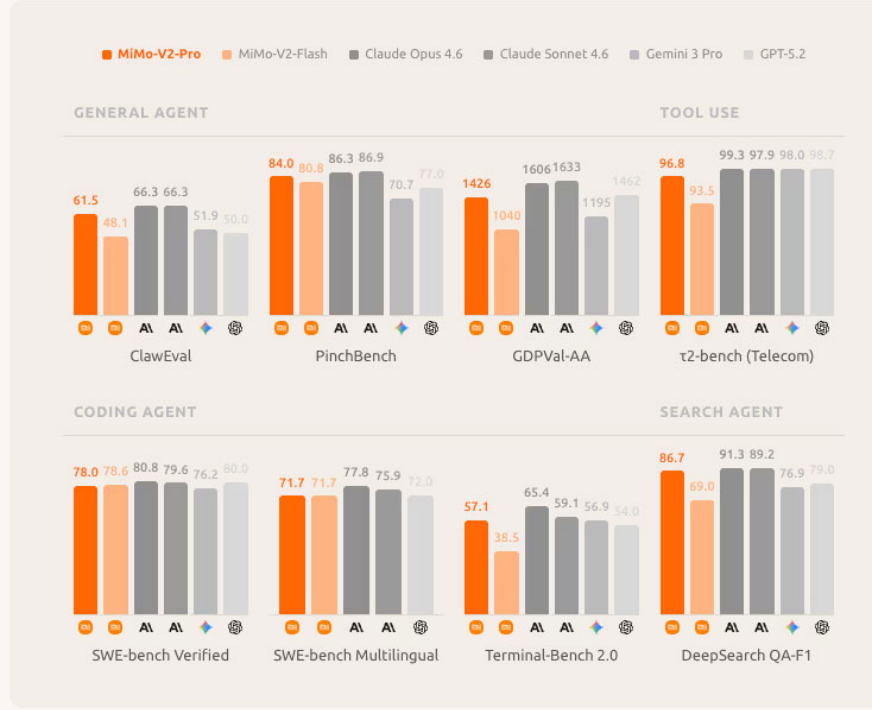
A Multi-Token Prediction layer has also been added, allowing the AI ??to predict multiple tokens simultaneously, significantly reducing processing latency.

Real-world performance and independent verification.

According to data from Xiaomi, the MiMo-V2-Pro performs better in real-world tasks than in theoretical benchmarks.

In the GDPval-AA test—which assesses the AI's ability to perform in a real-world environment—the model achieved an Elo score of 1426, surpassing many Chinese competitors such as GLM-5 and Kimi K2.5.

The independent review organization Artificial Analysis also confirmed these results, ranking MiMo-V2-Pro in the top 10 globally for AI capabilities, on par with the GPT-5.2 Codex and surpassing Grok 4.20 Beta.



Several key indicators show significant progress:

1. The rate of 'hallucination' decreased to 30% (compared to 48% previously).
2. The ability to reason more concisely and efficiently with significantly fewer tokens.
3. High programming performance, achieving 86.7 points on Terminal-Bench 2.0.

Significant cost advantage

The most noteworthy aspect of the MiMo-V2-Pro lies in its 'price/performance' ratio.

According to Artificial Analysis, the cost of running the entire test is only about $348, while GPT-5.2 costs over $2,300 and Claude Opus 4.6 costs nearly $2,500.

This makes the model an attractive option for businesses looking to deploy AI on a large scale while keeping costs under control.

Additionally, the 1 million token context window allows for processing the entire codebase or enterprise documentation at once, making it ideal for RAG or multi-agent systems.

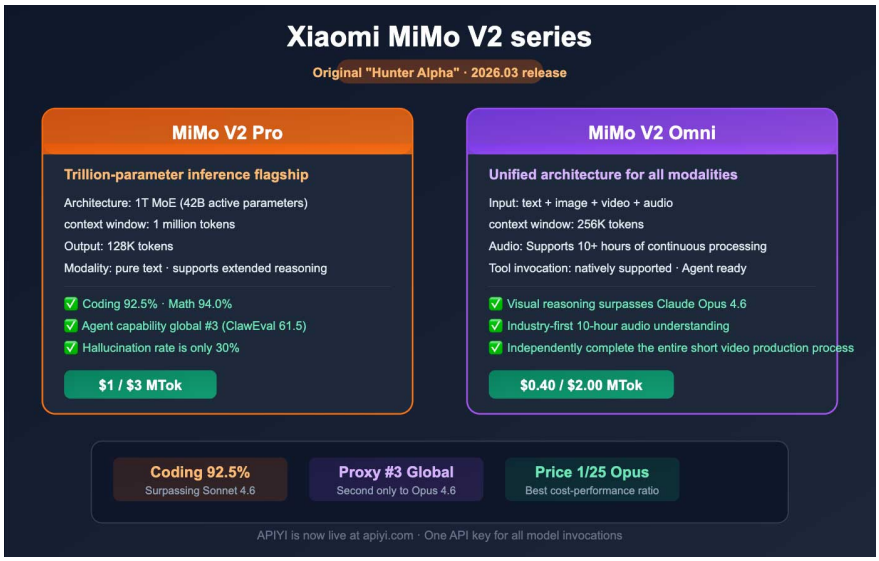
MiMo-V2-Pro opens up the possibility of building more complex AI systems, not just for automation, but also capable of solving multi-step problems.

However, this very ability to 'act' powerfully also increases security risks. The fact that AI can manipulate files, run commands, or access systems makes it more vulnerable to prompt injection attacks or unauthorized access.

Furthermore, because the model's weighting has not been fully unlocked, businesses also find it difficult to conduct in-depth testing at the system level—a crucial factor in sensitive environments.

Xiaomi is offering very competitive pricing to attract developers:

1. Below 256K tokens: $1 USD per million tokens input, $3 USD per million tokens output.
2. From 256K to 1M tokens: $2 USD per million input, $6 USD per million output.
3. Read cache is very cheap, while write cache is free.



Currently, the model only operates through Xiaomi's proprietary API and does not yet support multimodal functionality. However, the company has revealed a future version, MiMo-V2-Omni.

The emergence of MiMo-V2-Pro marks a significant shift in the AI ??industry. The race is no longer about how well a model speaks, but about its ability to execute actions.

If this trend continues, AI will not only be a supporting tool, but will become a direct 'agent' involved in the operational processes of businesses.