MiMo V2 6 Pro　特徴と評判の調査レポート

公開情報による性能評価と導入上の論点

調査基準日 2026 年 9 月 22 日

GPT-6 Astra

MiMo-V2.6-Pro は、低い API 単価と高いエージェント作業能力が特徴の Xiaomi 製オープンウェイトモデルである。公開直後の独立評価で高く評価される一方、上位の非公開モデルとの差や、日本語の専門実務・長時間運用の未検証部分が残る。[1]

本書は、公式発表、44 ページの技術報告書、公開モデル、第三者ベンチマーク、報道および初期利用者の反応を整理したものである。メーカーの公表値、独立評価、利用者の個別体験、本書の考察を区別する。API を用いた独自の実測試験は行っていない。[2, 3]

本文の角括弧内の番号は文末の参考文献に対応する。価格、ランキング、提供条件は調査時点の情報であり、更新され得る。

1 基本仕様とモデルの位置づけ

Pro は、テキストだけでなく画像・動画・音声を入力できる大規模な推論モデルである。多数の専門部分を必要に応じて使う MoE 構造を採用し、総パラメータは約 1.02 兆、トークンごとに計算に使うアクティブパラメータは 420 億とされる。[4]

表 1 MiMo V2 6 Pro の基本仕様

項目	内容
開発元	Xiaomi MiMo
モデル構造	MoE 総パラメータ約 1.02 兆 アクティブ 420 億
入力と出力	入力はテキスト・画像・動画・音声 出力はテキスト
コンテキスト長	100 万トークン
最大出力	128K トークン
主な機能	推論 ツール呼び出し 構造化出力 ストリーミング コンテキストキャッシュ
公開ウェイト	XiaomiMiMo/MiMo-V2.6-Pro-RL
ライセンス表示	公式モデルカードでは MIT
主な利用経路	Xiaomi の API OpenRouter 公開ウェイトによる自前運用

出典：公式モデルカード、公式モデルページ、OpenRouter。[4, 5, 6]

「マルチモーダル」は、ここでは画像・動画・音声を理解してテキストを出力できることを指す。後述する映像や音楽の制作例は、コードや外部ツールを組み合わせた作業として理解する必要がある。[5]

公開日には表記差がある。Xiaomi の公式発表は 9 月 22 日付だが、Artificial Analysis と OpenRouter はモデル公開日を 9 月 21 日と表示している。本書は 9 月 22 日付の公式発表を基準とし、表示差の原因までは断定しない。[1, 2, 6]

表 2 同時公開モデルの違い

モデル	規模と位置づけ
MiMo-V2.6-Pro	最大性能を狙う旗艦モデル 総 1.02 兆 アクティブ 420 億
MiMo-V2.6-Flash	より低い単価を重視するモデル 総 3090 億 アクティブ 150 億
MiMo-V2.6-Distill-Qwen-9B	Qwen3.5-9B を基に MiMo 生成データで追加学習した 90 億パラメータの小型モデル

出典：各公式モデルカード。9B 版の公開チェックポイントは教師あり追加学習（SFT）によるもので、Pro と同等の能力を保証するものではない。[4, 7, 8]

2 第三者評価による性能の位置づけ

Artificial Analysis（以下 AA）の Intelligence Index v4.3.2 は、知識、推論、科学、コーディング、業務作業などの複数評価を統合した指数である。正答率や IQ ではなく、異なる版の指数をそのまま比較することも適切ではない。以下は調査時点の同一版による比較である。[9]

表 3 Artificial Analysis の総合指数

モデルと評価設定	指数	ウェイト
GPT-6 Astra max [10]	53	非公開
Grok 4.7 xhigh [11]	46	非公開
MiMo-V2.6-Pro [1]	46	公開
GLM-5.3 max [12]	45	公開
Kimi K3 max [13]	44	公開
MiMo-V2.5-Pro [14]	26	公開

注：指数はページの整数表示に合わせた。Grok 4.7 と MiMo-V2.6-Pro は整数では同じ 46 だが、小数値では Grok がわずかに高い。設定が異なれば評価も変わり得る。

この比較は、MiMo-V2.6-Pro が公開ウェイトの先頭集団に入り、前世代から大きく改善したことを支持する。一方、指数 46 と 26 の差から「約 2 倍賢い」とは言えず、全用途で上位の非公開モデルを超えたとも言えない。業務ごとの品質は、個別評価と実案件に近い試験で確認すべきである。

3 メーカー公表値に見る改善と限界

Xiaomi の比較表では、ソフトウェア開発、業務自動化、複数ツールを使う作業に大きな改善が見られる。以下はメーカー公表値であり、AA の独立評価とは区別して読む必要がある。[4]

表 4 Xiaomi 公表のベンチマーク比較

評価項目	V2.5-Pro	V2.6-Pro	Opus 5
DeepSWE 1.1	19.0	71.9	74.0
ProgramBench	12.5	26.5	37.0
AutomationBench 1.0.6	16.0	53.1	50.3
Toolathlon Verified	49.1	76.9	80.6
Terminal-Bench 2.1	65.2	89.9	89.1
Terminal-Bench 4.0	1.5	34.9	49.0

出典：公式モデルカードの比較表。高いほどよい。Opus 5 は Claude Opus 5。評価環境、実行用プログラム、時間やトークンの予算などの条件に留意する。[4]

業務自動化や一部のターミナル作業では Opus 5 を上回るが、ProgramBench や、より難しい Terminal-Bench 4.0 では差が残る。本書では、今回の進歩を、作業の分解、ツール実行、途中結果の確認、修正を経て成果物を完成させる能力の改善と評価する。これらの成績には、モデル自体に加えて実行環境や再試行の設計も影響する。

4 強化学習と公開姿勢の特徴

技術報告書の中心は、強化学習（RL）の規模と報酬の精度を高めることである。第 1 に、一度の更新に使う課題と試行を増やす。第 2 に、コーディング、一般業務、視覚、サイバー分野などを混ぜ、複数のエージェント実行環境で訓練する。第 3 に、成功・失敗だけでなく成功した解法同士も比較し、品質や効率を評価する。[3]

報告書は、MoE のルーターを固定して学習を安定させる工夫や、採点の抜け道を利用する reward hacking への対処も説明する。ここでの「自己改善」は、人が設計した課題・環境・採点系を利用する訓練ループの高度化として理解できる。実運用中の無制限な自動再学習や、汎用人工知能の実現を立証した資料ではない。[3]

Xiaomi は、今回の RL 工程について、Pro が約 262 万ドル、Flash が約 85 万ドル、いずれも 6 日未満で 30 ステップを完了したと公表している。この金額と期間は追加の強化学習工程に関するもので、事前学習を含むモデル全体の開発費・開発期間ではない。[2]

学習の進捗を示すダッシュボード、技術報告書、関連コード、学習環境を公開している点も特徴である。公式 GitHub では関連リポジトリを確認できる。ただし、公開ウェイトやコードがあることと、第三者が全学習過程を同じ条件で再現できることは区別する必要がある。[2, 15]

5 制作と科学研究のデモをどう評価するか

制作分野では 3D ゲーム、映像、音楽などの事例を示している。ロボットの例はシミュレーション内の作業である。数学では、研究者が示した探索方針とその後の修正・統合を交え、既知の Li–Yorke の定理を Lean 4 で形式化した。既存の定理の厳密な形式化支援として興味深いが、未知の難問を完全自律で解決した例ではない。[2]

材料研究では、PFAS を捕捉する MOF（金属有機構造体）の候補探索が紹介されている。公式事例は、文献・特許調査から仮説形成、候補設計、計算環境の準備、結合に関するシミュレーション、候補の絞り込みまでを接続したと説明する。[16]

同事例は公開時点で実験室での検証待ちであり、候補材料の性能が実物で確認された段階ではない。「従来約 1 か月の工程を 2〜3 日に短縮」という説明も研究チームの見積もりである。本書では、計算に基づく研究候補の探索を加速した事例と評価する。[16]

6 API 料金と速度

公式の従量課金は表 5 の通りである。単位はいずれも 100 万トークン当たりの米ドルであり、キャッシュが利用できる繰り返し処理では入力料金が大きく下がる。[5, 17, 18]

表 5 公式 API 料金 米ドル毎 100 万トークン

モデル	入力 キャッシュなし	入力 キャッシュあり	出力
MiMo-V2.6-Flash	0.14	0.0028	0.28
MiMo-V2.6-Pro	0.435	0.0036	0.87
MiMo-V2.6-Pro-UltraSpeed	4.35	0.036	8.70

出典：各公式モデルページ。料金は 2026 年 9 月 22 日の確認値。無料枠、定額プラン、外部ツール料金は含まない。[5, 17, 18]

通常版 Pro で、キャッシュなし入力 10 万トークンと課金対象の出力 1 万トークンを使う場合、指定したトークン数だけの費用は 0.0522 ドルとなる。これは本書の単純計算であり、長い推論や再試行による追加消費、外部ツール費用を含む実作業の見積もりではない。

AA の指数評価におけるタスク当たりコストは、前版が約 0.05 ドル、今回が約 0.13 ドルである。同じ単価でも、処理に使うトークン数などにより総費用は変わる。取得時点の通常版 Pro の出力速度は約 125 トークン／秒だった。これは AA の測定条件下の値で、個々の業務の完了時間を意味しない。[1, 14]

UltraSpeed は通常版の 10 倍の単価で、Xiaomi は最大 20 倍の出力速度をうたう。「最大」はメーカー公称の上限であり、あらゆる処理の完了時間が常に 20 分の 1 になることを保証するものではない。[18]

7 公開直後の評判と情報の確かさ

初期の反応は好意的だが、独立評価、報道、短時間の利用談では証拠としての重みが異なる。特に、Pro、Flash、アプリケーション側の挙動を混同しないことが重要である。

報道では、PC Watch が公開ウェイトでの高評価と無償公開を取り上げている。これは肯定的なメディア紹介の例だが、記事数の多さを独立した実地検証の多さと同一視することはできない。[19]

Hacker News では、学習の経過や途中の問題を公開する姿勢を評価するコメントが見られた。一方、学習データの識別子だけでは内容がわかりにくいという指摘もある。研究の透明性への評価と、実際の業務性能への評価は分けて読む必要がある。[20]

Reddit の Pro に関するスレッドには、作業の取りまとめを評価する感想や、DeepSeek より出力が多いとの感想がある。ただし少数利用者の観察であり、管理された比較実験ではない。[21]

別の Reddit 投稿では、Flash を OpenCode で使用した際のループ、ツールの重複実行、停止などが報告されている。コメントではアプリ側の処理を疑う意見もあり、原因は確定していない。これらを Pro 全体やモデル単体の欠陥として一般化することはできない。[22]

日本語の個人試用記事では、Flash で会計用 HTML を作成できた一方、日本語として不自然な漢字が混じったと報告されている。これも Flash についての一例であり、Pro の日本語性能を直接示すものではない。[23]

8 知識の信頼性と未検証の領域

AA-Omniscience では知識の正答率が改善する一方、誤答を控える能力を測る指標は悪化している。単一の総合指数だけでは、この違いを捉えられない。[1, 14, 24]

表 6 AA Omniscience における前版との比較

指標	V2.5-Pro	V2.6-Pro
正答率	約 22.4%	約 34.9%
Hallucination Rate 低い方がよい	約 24.7%	約 40.6%

出典：AA のモデル別ページと指標の定義。[1, 14, 24]

ここでの 40.6%は、全回答の 40.6%が虚偽という意味ではない。AA は Hallucination Rate を、誤答数を「誤答数・部分正答数・回答を試みなかった数の合計」で割った値と定義する。正答以外の応答が分母となり、未知の問題で不適切に回答してしまう傾向を見る指標である。[24]

この結果から、答えられる問題が増えたことと、わからないときに適切に保留できることを別々に評価すべきだと考えられる。専門調査では、回答の流暢さだけでなく、参照先の実在、引用箇所との一致、重要事項の見落としを確認する必要がある。

今回の調査範囲では、日本語の技術調査、契約文書、特許、長文編集について継続的な比較検証を十分に確認できなかった。また、100 万トークンを入力できる仕様だけでは、その全域から漏れなく正

しく推論できるとは判断できない。数時間に及ぶ作業、混雑時、障害時、再試行時の安定性についても、公開直後の感想だけでは結論を出せない。

9 自前運用の現実性

Pro のアクティブ 420 億パラメータは、トークンごとの計算で使う規模を示す。保持する重みの総量は約 1.02 兆パラメータである。[4]

全パラメータを 4bit で保持すると仮定した単純計算でも、重みだけで約 510GB となる。これは本書の理論的概算で、実測メモリ要件ではない。実際にはキャッシュや実行時の管理領域も加わるため、通常のノート PC や単体 GPU で扱う規模ではない。

手元での実験には 9B 蒸留版がより現実的な候補となる。ただし、その能力は別途評価が必要であり、Pro のベンチマークをそのまま当てはめることはできない。ウェイトの公開と、外部 API でのデータの取扱条件も別の論点である。[8]

10 導入判断への示唆

以下は公開評価を踏まえた本書の考察であり、MiMo による各用途の性能を実測した結果ではない。まず検証しやすいのは、成果をテストや照合で確認できるエージェント作業である。

コード生成・修正・テストの反復は、今回の改善幅と成果の検証しやすさから試用候補となる。資料・表・Web ページの作成では、初回の出力だけでなく、完成までの修正回数、仕上がり、人の修正時間を比較したい。文献調査から仮説形成や計算実験へ進む用途は、材料研究の事例が参考になる。

日本語の専門調査では、実案件に近い課題を用い、引用の正確さ、見落とし、専門用語、指示の遵守を評価する必要がある。知財分野への応用を検討する場合も、技術文献の整理、発明候補の比較、明細書草案の作成補助などについて用途ごとに評価し、モデルの総合順位を直接的な実務保証とみなさないことが適切である。

比較試験では、同じ課題を 10〜20 件程度用意し、正しさ、人の修正時間、ツール実行の失敗、総トークン費用、完了までの時間を記録する方法が考えられる。低いトークン単価の利点が、必要な修正や再試行を含めても残るかを確認できる。

参考文献

全資料の閲覧日は 2026 年 9 月 22 日。更新型ページの数値・価格・順位は同日の取得値を使用した。コミュニティ投稿と個人記事は初期反応の資料として扱い、技術仕様や性能の確定的根拠とは区別した。本文の参照番号は各文献の番号に対応する。

[1] Artificial Analysis. MiMo-V2.6-Pro. 更新型ページ. 独立評価のモデル別ページ.
https://artificialanalysis.ai/models/mimo-v2-6-pro

[2] Xiaomi. Introducing MiMo-V2.6 series. 2026 年 9 月 22 日. 公式発表.
https://mimo.xiaomi.com/mimo-v2-6/article

[3] LLM-Core Xiaomi. MiMo-V2.6: Scaling Reinforcement Learning Towards Self-Improvement. 2026 年 9 月公開. 技術報告書 全 44 ページ.
https://huggingface.co/XiaomiMiMo/MiMo-V2.6-Pro-RL/blob/main/MiMo_V2_6_technical_report.pdf

[4] XiaomiMiMo. MiMo-V2.6-Pro-RL. 更新型ページ. Hugging Face 公式モデルカード.
https://huggingface.co/XiaomiMiMo/MiMo-V2.6-Pro-RL

[5] Xiaomi MiMo. MiMo-V2.6-Pro. 更新型ページ. 公式仕様と価格.
https://mimo.mi.com/models/en-US/mimo-v2.6-pro

[6] OpenRouter. Xiaomi: MiMo-V2.6-Pro. 更新型ページ. 提供モデルの情報と公開日表示.
https://openrouter.ai/xiaomi/mimo-v2.6-pro

[7] XiaomiMiMo. MiMo-V2.6-Flash-RL. 更新型ページ. Hugging Face 公式モデルカード.
https://huggingface.co/XiaomiMiMo/MiMo-V2.6-Flash-RL

[8] XiaomiMiMo. MiMo-V2.6-Distill-Qwen-9B. 更新型ページ. Hugging Face 公式モデルカード.
https://huggingface.co/XiaomiMiMo/MiMo-V2.6-Distill-Qwen-9B

[9] Artificial Analysis. Intelligence Benchmarking Methodology. 更新型ページ. Intelligence Index v4.3.2 の評価方法.
https://artificialanalysis.ai/methodology/intelligence-benchmarking

[10] Artificial Analysis. GPT-6 Astra. 更新型ページ. 独立評価のモデル別ページ.
https://artificialanalysis.ai/models/gpt-6-astra

[11] Artificial Analysis. Grok 4.7. 更新型ページ. 独立評価のモデル別ページ.
https://artificialanalysis.ai/models/grok-4-7

[12] Artificial Analysis. GLM-5.3. 更新型ページ. 独立評価のモデル別ページ.
https://artificialanalysis.ai/models/glm-5-3

[13] Artificial Analysis. Kimi K3. 更新型ページ. 独立評価のモデル別ページ.
https://artificialanalysis.ai/models/kimi-k3

[14] Artificial Analysis. MiMo-V2.5-Pro. 更新型ページ. 独立評価のモデル別ページ.
https://artificialanalysis.ai/models/mimo-v2-5-pro

[15] XiaomiMiMo. XiaomiMiMo repositories. 更新型ページ. GitHub 公式組織と公開コード.
https://github.com/XiaomiMiMo

[16] Xiaomi MiMo. MiMo-V2.6 Materials Research. 2026 年 9 月公開. 材料研究の公式ケーススタディ.
https://mimo.xiaomi.com/mimo-v2-6-material-research

[17] Xiaomi MiMo. MiMo-V2.6-Flash. 更新型ページ. 公式仕様と価格.

https://mimo.mi.com/models/en-US/mimo-v2.6-flash

[18] Xiaomi MiMo. MiMo-V2.6-Pro-UltraSpeed. 更新型ページ. 公式仕様と価格.

https://mimo.mi.com/models/en-US/mimo-v2.6-pro-ultraspeed

[19] PC Watch. オープンウェイトの最高評価 AI モデル「MiMo-V2.6」、Xiaomi が無償公開 〜1.02T の Pro は Opus 5 に一部で並び、310B の Flash も好成績. 2026 年 9 月 22 日. 報道記事.

https://pc.watch.impress.co.jp/docs/news/2142608.html

[20] Hacker News 投稿者・コメント投稿者. MiMo v2.6. 2026 年 9 月 21 日投稿 UTC. コミュニティの反応 原スレッド ID 49792730.

https://news.ycombinator.com/item?id=49792730

[21] Reddit r/DeepSeek 投稿者・コメント投稿者. DS 4.1 Flash VS MiMo 2.6 Pro. 2026 年 9 月の投稿. 初期利用談を含む討論.

https://www.reddit.com/r/DeepSeek/comments/1wmr6cp/ds_41_flash_vs_mimo_26_pro/

[22] Reddit r/opencode 投稿者・コメント投稿者. Having a bad experience with mimo 2.6 flash. 2026 年 9 月の投稿. Flash と OpenCode に関する初期利用談.

https://www.reddit.com/r/opencode/comments/1wmsl2p/having_a_bad_experience_with_mimo_26_flash/

[23] ゆいまる. 【期間限定 Opencode 無料枠】Xiaomi が 1T 級 OSS モデル「MiMo-V2.6」を公開！Claude Opus 5 肉薄の Agent 性能とオープン化戦略の全貌を調べながら Opencode でお試しましたー。. 2026 年 9 月 22 日. note 個人による Flash の試用記事.

https://note.com/humble_bobcat51/n/n94012cad5242

[24] Artificial Analysis. AA-Omniscience: Knowledge and Hallucination Benchmark. 更新型ページ. 知識正答率とハルシネーション指標の定義.

https://artificialanalysis.ai/evaluations/omniscience