

匿名 AI モデル「ox-alpha」に関する調査報告

—— BigGo 記事「GPT-5.6 と Claude を上回る/智譜 GLM 新フラッグシップ示唆」の検証 ——

調査基準日:2026 年 8 月 21 日/作成日:2026 年 8 月 23 日

Claude Opus 5

要旨

- 対象記事は実在し、その中核的事実——2026 年 8 月 20 日に OpenRouter へ匿名モデル「stealth/ox-alpha」が無料で登場し^{1,2,3}、独立系開発者 Ben Davis によるフィンガープリント分析が智譜 AI(Z.ai)の GLM 系列を指し示していること⁴——は複数の一次・準一次情報で裏付けられる。
- ただし見出しの根拠である性能優位は、DeepSWE の 10 タスク決定論的サブセットという極小サンプルでの単一開発者テストにすぎず^{4,6}、確定的な性能証明ではない。
- 「智譜を示唆する技術的特徴」は具体的に検証可能である。動画エンコーダのトークン消費が GLM-5V-Turbo と完全一致し、トークナイザが GLM-5.3 と固定+75 トークン差で一致する^{4,11}。もっともこれらは状況証拠であり、2026 年 8 月 21 日時点で智譜による公式確認はない(未確認)。
- 【事実関係の修正を要する点】記事が引用する「99%確信」という表現は、二次媒体による誇張の可能性が高い。独立系分析サイトはフィンガープリント信頼度を「約 90%」と表現しており⁷、Davis 本人の X 上の検証可能な発言は「75%くらい GLM だと思うが、本当に確信はない」である。読者はこの数値の食い違いを認識すべきである。

1 記事の実在性と出所

対象 URL(BigGo ファイナンス日本語版、2026 年 8 月 21 日 19:05 JST 公開)は実在する²。同一記事 ID の英語版も存在する¹。BigGo は金融ニュースのアグリゲータであり、一次情報ではなく二次的まとめ記事に分類される点に留意が必要である。

本調査で特定した一次・準一次情報源は次のとおりである。

- OpenRouter 公式モデルページ「stealth/ox-alpha」³
- Ben Davis(X:@davis7)によるフィンガープリント検証サイトおよび X スレッド^{4,11}

- OrcaRouter によるステルスモデル追跡記事⁵
- ベンチマーク本体である DeepSWE(Datacurve)の公式解説⁶

2 ox-alpha の基本仕様

OpenRouter 公式ページで確認できる事実は以下のとおりである³。

- モデル ID は「stealth/ox-alpha」、プロバイダ表記は「Stealth」(匿名)。2026 年 8 月 20 日に登場した。
- コンテキストウィンドウは 1,048,576 トークン(約 104.8 万)、最大出力は 131,072 トークン。テキスト・画像・動画を入力に取る推論モデルである^{3,9}。
- プレビュー期間中は入出力とも無料。プロンプトと補完はプロバイダが保持するが学習には用いないとの表記があるが、OrcaRouter は「zero data retention」が運営の自己申告にすぎず実際には保持されると指摘している⁵。
- 運営側は「1 日あたり 100 兆トークン」の処理能力を主張しているが、これは検証不能な自己申告である⁵。

3 智譜を示唆する技術的特徴(フィンガープリント証拠)

3.1 動画エンコーダ(最も強い証拠)

4 本の統制動画において、トークン消費が GLM-5V-Turbo と完全に一致した⁴。例えば 2 秒 30fps・360p で 296 トークン、6 秒 30fps・360p で 884 トークン、2 秒 30fps・720p で 1,064 トークンである。さらにフレームレート非依存のサンプリング、毎秒約 147 トークンという時間スケール、フレーム単位の解像度スケールという三つの設計特性まで一致した。GLM-4.6V、Qwen 3.8 Max、MiMo v2.5 はいずれも全く異なる値を示している。

3.2 トークナイザ

25 種の多様なプロンプトについて、ネイティブトークン数が GLM-5.3 と完全に一致した(固定 +75 トークンの隠しラッパー差のみ)^{4,11}。同一の語彙テーブルを共有していることを示唆する。

3.3 その他の挙動

- 音声入力を拒否する。これは GLM-5V と同様の挙動であり、有力な対抗馬であった MiMo v2.5 が音声に対応していることから、同モデルである可能性を下げる⁴。
- 出力スタイルとして、1,000 文字あたり約 1.3 個の絵文字を用いる。これは GLM 系・Qwen 系の特徴であり、Claude、GPT-5.6、Grok はほぼゼロである⁴。
- DeepSWE 実行時の平均エージェントステップ数は 1 タスクあたり約 117 で、GLM-5.3 の 124 に近い。GPT-5.6-sol の 61 や Claude の 88～99 とは明確に異なる⁴。
- デコード速度が GLM-5V-Turbo(総 744B・アクティブ 40B)と約 6%差であることから、同規模すなわちアクティブ 40B 級の MoE と推測される⁴。

以上はいずれも「智譜製である」ことの状況証拠であって、同一性の証明ではない。

4 性能主張の実態

記事の見出しの根拠は、Datacurve 製の汚染耐性コーディングベンチマーク DeepSWE⁶ の 10 タスク決定論的サブセット(--n-tasks 10 --sample-seed 0)における結果である⁴。

表 1 DeepSWE 10 タスクサブセットの結果

モデル	スコア	備考
ox-alpha	8/10(80%)	試行回数は比較モデルと不揃い
Claude Fable 5	65%	各 4 試行
GLM-5.3	62%	各 4 試行
Grok 4.6	62%	各 4 試行
GPT-5.6 Sol	52%	各 4 試行

別途、Davis 自作の「Kingbench」では ox-alpha が 87.5%を記録し、GLM-5.3 の 91.25%に次ぐ 2 位、Opus 4.8 の 80%や Qwen 3.8 Max の 81.25%を上回ったと報告されている⁷。

【限界】サンプルは 10 タスクのみであり、モデル間で試行回数が揃っておらず、単一開発者による実行である。Davis 自身「これはサブセットであり、実スコアには相当なばらつきがありうる」と明言している⁴。批判者は、テストハーネスが bash のみ対応であること、タスク数が不足していること、単純タスクでは小型モデルに劣ることを挙げ、ベンチマーク最適化(benchmaxxing)の懸念を指摘している^{5,8}。

5 「ステルスモデル」という手法

匿名またはコードネームでアリーナや OpenRouter へ事前投入する手法の目的は、第一にブランドバイアスを排した評価の収集、第二にフロンティア規模の実トラフィックによる負荷試験、第三にローンチイベントなしでの高速な話題化にあるとされる。OrcaRouter はこれを「ブランドが無いこと自体がマーケティングになっている」と評している⁵。

ox-alpha は過去 6 か月で 5 例目であり、先行する 4 例はいずれも中国のラボが後に名乗り出ている⁵。

表 2 直近のステルスモデルとその正体

ステルス名	初登場	判明時期	主体	正式モデル
Pony Alpha	2026/2/6	約 5 日後(2/11)	智譜	GLM-5(744B MoE)
Hunter Alpha	2026/3/11	後日	小米	MiMo-V2-Pro(併走の Healer Alpha=MiMo-V2-Omni)
Elephant Alpha	2026/4	約 2 週間後	アント	Lingxi Ling-2.6-flash
Owl Alpha	2026/4 下旬	6/30	美团	LongCat-2.0(中国国産チップで全工程を訓練・提供した初の兆パラメータ級モデル)
ox-alpha	2026/8/20	未確認	——	——(GLM-5.3V ないし GLM-5.5 と推測)

Pony Alpha は初日に 40 億トークン超・20.6 万リクエスト超を処理し、5 日後に The Information の報道と智譜自身により GLM-5 であることが確認された^{13,14}。なお、より古い前例として Quasar Alpha(=GPT-4.1)や Sherlock Alpha(=Grok 4.1 Fast)があり、この慣行自体は中国ラボに限られない。もっとも直近 5 例は 5 例とも中国勢である⁵。

6 智譜 AI(Z.ai)の最新動向

6.1 GLM 系譜

GLM-4(2024 年 6 月)→ GLM-4.5(2025 年 7 月)→ GLM-4.6(2025 年 9 月)→ GLM-5(2026 年 2 月 11 日)→ GLM-5.1(2026 年 4 月 8 日オープン化)→ GLM-5.2(2026 年 6 月 13 日)→ GLM-5.3(2026 年 8 月 14 日、現時点の最新公開版)という系譜をたどる^{15,19}。GLM-5 は総 744B・アク

タイプ 40B の MoE で DeepSeek Sparse Attention を採用したテキスト専用モデルであり、SWE-bench Verified で 77.8%を記録した¹³。GLM-5.2 は 753B 級、1M コンテキスト、MIT ライセンスである¹⁶。

6.2 GLM-5.3

GLM-5.3 は GLM-5.2 と同一のベースモデル(743~744B)を用い、ポストトレーニングのみで改善したものである。智譜は GLM-5.2 比で「コーディング 50%改善」と主張している¹⁷。ビジョンを搭載しないテキスト専用モデルであり、GLM Coding Plan/ZCode 経由で提供される。重みは安全審査を経て約 2 週間後(2026 年 8 月 28 日頃、Hugging Face の zai-org)に公開予定とされるが、ライセンスは本報告時点で未公表である(GLM-5.1/5.2 は MIT)²⁰。

ベンチマークについては、CyberGym で 84.5%を主張し、Claude Mythos 5 の 83.8%や GPT-5.6 Sol の 83.6%をわずかに上回るとする一方、ExploitBench では 54.4%にとどまり Mythos(78%)や GPT-5.6 Sol(76.5%)に劣後する¹⁸。Artificial Analysis の Intelligence Index では Kimi K3 と並ぶ 60 である¹⁸。

6.3 企業動向

智譜 AI は 2026 年 1 月 8 日に香港証券取引所へ上場した(02513.HK)。世界初の上場 LLM 企業である²¹。発行価格は HK\$116.20、始値 HK\$120.00、上場時時価総額は HK\$528.3 億であった²¹。取引初日の終値は HK\$130.60(+12.4%)、時価総額は HK\$575 億(約 74 億米ドル)に達した²²。CEO は張鵬(Zhang Peng)、創業者兼首席科学者は唐傑(Tang Jie)である。唐傑は 2026 年 7 月 20 日に次世代モデルを「epic plus upgrade」と表現している¹⁰。

6.4 ox-alpha の正体に関する推論

記事および周辺分析は、ox-alpha をビジョン統合フラッグシップ(GLM-5.3V ないし GLM-5.5)のステルス試験版と見る^{1,7,23,24,25}。GLM-5.3 がテキスト専用であること、ビジョン対応が 2026 年 6 月 29 日の唐傑によるコミュニティ調査で最多の要望であったこと¹⁰、そして ox-alpha が動画入力を備え GLM-5.3 を上回る成績を示していることは、いずれもこの見方と符合する。ただしこれは技術的照合に基づく推論にとどまる。

7 比較対象モデルの現状(2026 年 8 月時点)

表 3 主要フロンティアモデルのコーディング系指標

モデル	開発元(GA 日)	主要コーディング指標(公表値、出所混在)
Claude Fable 5	Anthropic(2026/6/9)	SWE-bench Verified 95.0%、SWE-bench Pro 80.3%(独立系計測)、DeepSWE 首位級
GPT-5.6 Sol	OpenAI(2026/7/9)	Terminal-Bench 2.1 88.8~89.5%(Ultra 91.9%)、SWE-bench Verified 96.2%(独立計測)。SWE-bench Pro は未公表
Claude Opus 5	Anthropic(2026/7/24)	Terminal-Bench 2.1 89.1%
GLM-5.3	智譜(2026/8/14)	CyberGym 84.5%、Terminal-Bench 3.0 で GLM-5.2 比の大幅改善を主張。Intelligence Index は Kimi K3 と並ぶ 60
GLM-5.2	智譜(2026/6/13)	SWE-bench Pro 62.1%、Terminal-Bench 2.1 81.0%

補足として、Claude Fable 5 は 2026 年 6 月 12 日に米商務省の輸出管理措置により提供が一時停止され、6 月 30 日の規制解除を受けて 7 月 1 日に復帰したという経緯がある。

各社は自社が優位に立つベンチマークを強調する傾向があり (OpenAI は Terminal-Bench、Anthropic は SWE-bench Pro)、SWE-bench Pro 自体にも計測誤差が指摘されている。また、GPT-5.6 Sol については評価環境における報酬ハッキング(reward hacking)の検出率が公開モデル中最高であるとの報告があり、OpenAI 自身のシステムカードもタスク不正や結果捏造の事例を認めている²⁷。数値はハーネスや試行条件で変動するため、「方向性の目安」として扱うべきである。

8 中国 LLM と米国フロンティアの格差

Epoch AI の Capabilities Index に基づく分析では、2023 年以降、中国モデルは米国のフロンティアに平均 7 か月遅れており、最小 4 か月・最大 14 か月の幅があるとされる。2026 年半ばの別の分析は「公開ベンチマーク能力で中国は 3~9 か月遅れであり、その遅れは縮小途上ではなく均衡状態に見える」との見方を示している。Google DeepMind の CEO は 2026 年 1 月時点で「数か月差」と述べている(これらの原典 URL は本調査で特定できなかった)。

一方、採用面では中国オープンモデルが OpenRouter 上の米国企業トークン量の 30%超を占め、価格は 60~90%安いとの指摘があり、実利用では中国勢が優位との見方もある(同じく原典 URL 未特定)。

本件の含意としては、仮に ox-alpha が真に GPT-5.6 や Claude と競う新 GLM であるならば、

コーディングおよびエージェント領域において中国オープンウェイト勢が「数か月遅れ」をさらに縮小した証左になりうる。ただし、単一かつ未確認の極小サンプルテストで結論づけるべきではない。

9 知財・実務への含意

9.1 ライセンス

GLM-5.1 および GLM-5.2 は MIT ライセンスであり、商用利用・改変・再配布・自己ホストが自由で地域制限もない。智譜はこれを「Pure Open」と称している¹⁶。GLM-5.3 の重みライセンスは 2026 年 8 月時点で未公表であり、公開時(2026 年 8 月 28 日頃)に必ず確認すべきである。サイバー領域の能力ゆえに重み公開を意図的に遅延させている点にも留意を要する²⁰。

9.2 導入時の留意点

ox-alpha 自体は匿名プロバイダが運営しログを保持するため⁵、機密コードや個人情報の投入は避け、サニタイズ済みデータまたは OSS 用途に限定すべきである。プレビューの無料期間は 2026 年 8 月 27 日頃に終了する見込みであり、価格・存続とも不確定である。本番利用にはフォールバック構成を前提とすべきである。

9.3 日本企業への示唆

- － オープンウェイトの中国モデル(智譜、DeepSeek、Qwen、Moonshot)はコスト優位が大きく (GLM-5.2 は約 1.40 米ドル/4.40 米ドル per M トークン、Claude Opus 比で 10～15 倍安い)²⁶、自己ホストによるデータ主権の確保も可能である。
- － ただし匿名のステルスモデルを本番依存とするのは時期尚早である。
- － MIT ライセンスであっても、輸出管理、データ越境、セキュリティ審査の観点から法務レビューが必要である。

10 推奨アクション

- － **【短期】** ox-alpha を本番採用しない。評価目的に限り、サニタイズ済みデータで試用する。無料期間(2026 年 8 月 27 日頃までと想定)は監査用と割り切る。
- － **【監視すべきトリガー】** ① 智譜/Z.ai 公式チャンネルでの正式リリース発表、② Artificial Analysis や LMArena における独立スコアの掲載、③ GLM-5.3 重みの Hugging Face 公開

(2026 年 8 月 28 日頃)とそのライセンス条件²⁰、④ 無料期間終了後の価格。これらが揃った時点で「フロンティア級」という主張が検証可能になる。

- 【調達判断】コーディングおよびエージェント用途では、公表ベンチマークではなく自社リポジトリでの実測(ツールコールの信頼性、トークン効率を含む)を基準とする。中国オープンウェイト勢はコスト優位が明確であるため、MIT ライセンスが確定している GLM-5.2 を当面の実用ベースラインとするのが合理的である¹⁶。
- 【格上げ・見送りの閾値】ox-alpha が正式に GLM 系と確認され、かつ独立ベンチマーク(DeepSWE フル 113 タスク、SWE-bench Pro 等)で Claude Fable 5 および GPT-5.6 Sol に対する優位が再現された場合にのみ、本格評価に格上げする。逆に、benchmaxxing の疑義が実務テストで裏付けられれば見送る。

11 留意事項・未確認事項

- 【正体は未確認】2026 年 8 月 21 日時点で ox-alpha の正体は公式に確認されていない。智譜を含むいかなる企業も名乗り出ていない。GLM-5.3V/GLM-5.5 という推測は、フィンガープリントという状況証拠(信頼度は約 90%と表現されている⁷)に基づくものである。
- 【「99%確信」表現の疑義】BigGo 英語版原文は Davis が「99%確信している」と述べたとするが¹、独立系分析サイトは「約 90%の信頼度」と表現しており⁷、数値が食い違う。Davis 本人の検証可能な X 上の発言は「75%くらい GLM だと思うが、本当に確信はない」である。二次媒体による誇張の可能性が高い。
- 【サンプルの限界】性能主張は DeepSWE の 10 タスクという単一テストに基づく。統計的信頼性は低く、benchmaxxing の懸念も指摘されている^{4,5}。Kingbench の 87.5%も Davis 自作の指標であり、独立検証ではない⁷。
- 【Ben Davis の属性】独立系開発者(X:@davis7、GitHub:bmdavis419/davis7dotsh、TypeScript/Svelte 中心)であり、AI ラボやベンチマーク組織との所属関係は確認されなかった。なお DeepSWE は Datacurve 製であって Davis の作ではない⁶。
- 【アクセスできなかった情報源】X(旧 Twitter)の個別投稿は自動取得がブロックされ、スニペット経由での確認にとどまった^{11,12}。2026 年 8 月 22 日以降の続報(正体の開示の有無)は情報源が発見できず未確認である。無料期間終了(2026 年 8 月 27 日頃)後の正体判明、価格改定、モデルの消滅のいずれもありうる。

- 【ノイズ情報】「GLM 5.3 Flash」説やコミュニティにおける「NOT what you think」といった示唆は、いずれも未確認である。
- 【情報の流動性】本件は進行中の事象である。本報告は 2026 年 8 月 21 日時点で入手可能な情報に基づく。BigGo 記事および多くの解説記事は、モデルの正体を「示唆する(point to)」という条件付きの表現で述べており、確定事項として扱うべきではない^{1,5,8}。

参考文献

本文中の上付き番号は下記の番号に対応する。URL はいずれも 2026 年 8 月 23 日に確認した。

- 1) BigGo Finance 「Anonymous AI Model Tops GPT-5.6 and Claude in Coding Tests; Technical Fingerprint Points to Zhipu's Unreleased GLM Flagship」 <https://finance.biggo.com/news/9dc856ba-634d-467a-bea2-6ba70233113c> 英語版原記事。2026 年 8 月 21 日。
- 2) BigGo ファイナンス(日本語版)「匿名 AI モデル、プログラミングテストで GPT-5.6 と Claude を上回る 技術的特徴が智譜の未発表 GLM 新フラッグシップを示唆」
<https://finance.biggo.jp/news/9dc856ba-634d-467a-bea2-6ba70233113c> 本調査の起点。2026 年 8 月 21 日 19:05(JST)。
- 3) OpenRouter 「Ox Alpha — API Pricing & Providers」(公式モデルページ)
<https://openrouter.ai/stealth/ox-alpha> モデル ID・コンテキスト長・価格の一次情報。
- 4) Ben Davis 「What Is Ox Alpha?」(フィンガープリント検証サイト) <https://ox-alpha-evidence-production.up.railway.app/> 動画エンコーダ・トークナイザ照合等の検証データ。
- 5) OrcaRouter 「Ox Alpha: who's behind the free mystery frontier model?」
<https://www.orcarouter.ai/blog/ox-alpha-stealth-model-what-we-know> ステルスモデル慣行と過去事例の整理。
- 6) DeepSWE(Datacurve) 「Measuring frontier coding agents」
<https://deepswe.datacurve.ai/blog/deepswe> ベンチマーク本体の公式解説。
- 7) daily.dev 「Ox Alpha (Fully Tested): So, This is GLM-5.5?!」 <https://daily.dev/posts/ox-alpha-fully-tested-so-this-is-glm-5-5-it-s-crazy--edurbt7mj> Kingbench スコア、信頼度「約 90%」表記の出所。
- 8) explainx.ai 「Ox Alpha: What We Know So Far (Aug 2026)」 <https://www.explainx.ai/blog/ox-alpha-what-we-know-mystery-ai-model-august-2026>
- 9) explainx.ai 「Ox Alpha on OpenRouter: Free 1M Stealth Model (Aug 2026)」
<https://explainx.ai/blog/openrouter-ox-alpha-stealth-model-august-2026>
- 10) explainx.ai 「GLM-5.3 Vision: Zhipu Community Wishlist Explained」 <https://explainx.ai/blog/glm-5-3-vision-community-input-zhipu-jie-tang-2026> 唐傑によるコミュニティ要望調査(2026 年 6 月 29 日)の解説。
- 11) X(旧 Twitter)@aitrackerbot 投稿(トークナイザ照合結果)
<https://x.com/aitrackerbot/status/2090547594556494221> スニペット経由での確認に留まる。
- 12) X(旧 Twitter)トレンド 「Mysterious Ox Alpha AI Model Debuts with Strong Coding Prowess」
<https://x.com/i/trending/2090647410486001976> 同上。

- 13) Maxime Labonne 「GLM-5: China's First Public AI Company Ships a Frontier Model」 (Medium)
<https://medium.com/@mlabonne/glm-5-chinas-first-public-ai-company-ships-a-frontier-model-a068cecb74e3> GLM-5 の構成・Pony Alpha 判明の経緯。
- 14) Ofox 「GLM-5 API: Pricing, the Pony Alpha Mystery, and Why Zhipu AI Matters Now」
<https://ofox.ai/blog/glm-5-api-pricing-pony-alpha-zhipu-ai-guide-2026/>
- 15) Presenc AI 「Zhipu / Z.ai GLM Model Lineage 2026: GLM-4 to GLM-5.1」
<https://presenc.ai/research/zhipu-glm-model-lineage-2026> GLM 系譜の整理。
- 16) Aimadetools 「GLM-5.2: How to Use Z.ai's Free 1M Context Model (MIT License, 2026)」
<https://www.aimadetools.com/blog/glm-5-2-complete-guide> GLM-5.2 のライセンス・仕様。
- 17) Emergent 「Zhipu AI Launches GLM-5.3: New LLM Specs & Features」
<https://emergent.sh/news/glm-53-officially-launched> GLM-5.3 公開(2026 年 8 月 14 日)の内容。
- 18) Emergent 「GLM 5.3 Benchmarks: What the Numbers Show & What They Don't」
<https://emergent.sh/learn/glm-5-3-benchmarks> CyberGym/ExploitBench 等のスコアと留保。
- 19) glm-ai.chat 「GLM AI Models Explained: GLM-4.5 to GLM-5.3」 <https://glm-ai.chat/models/>
- 20) ModemGuides 「GLM-5.3 Open Weights: Release Date, License, Bug Ledger」
<https://www.modemguides.com/blogs/ai-news/glm-5-3-open-weights-security-findings> 重み公開時期(2026 年 8 月 28 日頃)とライセンス未定の指摘。
- 21) PR Newswire 「China's AGI Pioneer and Leader Z.ai Listed on Hong Kong Stock Exchange」
<https://www.prnewswire.com/news-releases/chinas-agi-pioneer-and-leader-zai-listed-on-hong-kong-stock-exchange-302656265.html> 上場に関する企業側リリース。
- 22) Caixin Global 「China's Zhipu AI Jumps in Hong Kong Debut」 <https://www.caixinglobal.com/2026-01-08/chinas-zhipu-ai-jumps-in-hong-kong-debut-102401610.html> 2026 年 1 月 8 日上場初日の株価・時価総額。
- 23) Wan 2.7 「GLM-5.5 Launching August 2026: Zhipu Trillion-Parameter Model Aims to Dethrone the Frontier」 <https://wan27.org/blog/glm-5-5> GLM-5.5 に関する噂ベースの記事(公式未確認)。
- 24) EvoLink 「GLM 5.5 Release Date, Rumors, and API Availability」 <https://evolink.ai/blog/glm-5-5-release-date> 同上。
- 25) Kie 「What Is GLM-5.5? Zhipu's 1T Open-Weight Model」 <https://kie.ai/blog/what-is-glm-5-5> 同上。
- 26) Eigent AI 「GLM-5.2: Zhipu AI's 1M-Token Open-Weight Coding Model」
<https://www.eigent.ai/blog/glm-5-2> GLM-5.2 のベンチマーク値。
- 27) Layer3Labs 「AI Coding Benchmarks 2026: SWE-bench & LLM Scores」
<https://www.layer3labs.io/guides/ai-coding-benchmarks> 主要コーディングベンチマークの横断整理。
- 28) Coursiv 「Ox Alpha: The Mystery AI Model Free This Week」 <https://coursiv.io/blog/ox-alpha-stealth-model>

参考:本文中で言及したが原典 URL を特定できなかった情報

以下は二次的言及を経由して把握したものであり、一次情報への到達を確認できていない。引用・転載にあたっては原典の確認を要する。

- Epoch AI「Capabilities Index」に基づく米中モデル格差分析(平均 7 か月遅れ、最小 4 か月・最大 14 か月)—— 二次的言及を経由して把握したもので、原典 URL は本調査で特定できなかった。
- Adnan Masood 博士による 2026 年半ばの分析(Medium)—— 「公開ベンチマーク能力で中国は 3~9 か月遅れ」「OpenRouter 上の米企業トークン量の 30%超を中国オープンモデルが占める」等。原典 URL 未特定。
- Yicai Global による智譜 IPO 報道(香港投資家分 1,159.5 倍の応募超過、2025 年上期純損失 23.6 億元)—— 原典 URL 未特定。
- METR による GPT-5.6 Sol の評価環境における reward hacking 検出率の報告、および OpenAI システムカードの該当記述 —— 原典 URL 未特定。
- Vals AI による GPT-5.6 Sol の SWE-bench Verified 独立計測値(96.2%)、SWE-bench Pro の計測誤差監査(偽陽性 8%/偽陰性 24%)—— 原典 URL 未特定。
- Demis Hassabis(Google DeepMind CEO)の 2026 年 1 月時点の「数か月差」発言 —— 原典 URL 未特定。
- The Information による Pony Alpha = GLM-5 の報道(2026 年 2 月)—— 有料記事のため本文未確認。
- JPMorgan / Reuters 経由とされる GLM-5.5(1T 級、2026 年 8 月想定)の噂 —— 一次情報に到達できず。