

OpenRouter 上の匿名モデル「stealth/ox-alpha」に関する技術分析： GLM 系モデルとの関連性および初期コミュニティ評価の検証

最終更新：2026 年 8 月 23 日
Gemini 3.7 Flash

2026 年 8 月 20 日、AI モデル配信プラットフォーム OpenRouter 上に「stealth/ox-alpha」（以下、ox-alpha）と称される未知の匿名モデルが登場し、期間限定の無料プレビューが提供されました。OpenRouter は、本モデルについて、匿名の第三者が開発・運用し、OpenRouter 自身は開発者、所有者または提供者ではないと説明しています。[1] 本モデルは、プログラミングや長時間のエージェント型作業を想定したモデルとして紹介され、AI 開発者コミュニティを中心に関心を集めています。[1][2]

OpenRouter 上で確認できる公開仕様によれば、ox-alpha は 1,048,576 トークン（約 100 万トークン）のコンテキストウィンドウと、最大 131,072 トークンの出力能力を備えています。テキスト、画像および動画を入力し、テキストを出力するほか、ツール呼び出し（function calling）と構造化出力に対応します。ただし、JSON Schema の厳格な強制には対応していません。[1]

評価項目	確認された仕様およびコミュニティ観測情報
モデル識別名	stealth/ox-alpha [1]
提供形態／状態	OpenRouter 上の匿名プレビュー。匿名の第三者が開発・運用 [1]
コンテキストウィンドウ	1,048,576 トークン（1M Context） [1]
最大出力トークン数	131,072 トークン [1]
サポートモダリティ	入力：テキスト、画像、動画／出力：テキスト。音声入力は非対応 [1][2]
推論速度／初回応答時間	外部計測時点で約 49～78 tok/s との観測例あり。負荷・計測期間で変動するため固定値ではない [2]

推定アーキテクチャ（未確認）	GLM-5V-Turbo との速度比較等から 744B 総／40B アクティブ級 MoE との推測あり。ただし直接確認された仕様ではない [2]
----------------	--

外部検証では、計測時点のデコード速度が智譜 AI（Zhipu AI、海外ブランド名：Z.ai）の「GLM-5V-Turbo」と約 6%の差にとどまったことなどを根拠に、ox-alpha が総パラメータ数 7,440 億、アクティブパラメータ数約 400 億級の Mixture-of-Experts（MoE）である可能性が推測されています。[2] ただし、これは既知の GLM-5 系モデルの規模と速度比較から導かれた推測であり、速度だけからパラメータ規模を一意に確定することはできません。また、匿名の提供者が 1 日 100 兆トークン規模の供給能力を掲げたと報告されていますが、OpenRouter の標準仕様として確認された数値ではなく、実処理量、予約容量または理論値のいずれを意味するのかも不明です。[2]

技術的指紋による推定根拠と留意点

独立した技術検証者らは、ox-alpha に対するブラックボックス検証を実施し、複数の技術的特徴から、本モデルが智譜 AI の GLM 系モデル、特に未発表のマルチモーダル版である可能性が高いと推定しました。[2][3] 主な根拠は以下のとおりです。なお、これらは公式確認ではなく、外部から観測可能な挙動に基づく推定です。判断主体を明確にするため、本稿では以下を「コミュニティ検証」として扱います。

第一に、トークナイズにおける整合性です。集約された検証結果では 25 種類のプロンプトについて、ox-alpha のトークン数が、智譜 AI が 2026 年 8 月 14 日に発表したテキストモデル「GLM-5.3」と固定的な「+75 トークン」の差を除き一致したと報告されています。[2] 公開された個別検証投稿の一つでは、英語、ドイツ語、中国語、コード、絵文字等を含む 6 テキストで同じ固定差が確認されています。[3] これは同一または極めて近いトークナイザーの使用を強く示唆しますが、+75 トークンは、隠れたシステムプロンプトまたは API ラッパーに由来すると推定されるもので、公式に説明されたものではありません。

第二に、動画入力におけるエンコーダー挙動の類似性です。制御された 4 条件のテスト動画を入力した際のトークン消費量を解析したところ、ox-alpha と GLM-5V-Turbo は各条件で同じトークン数を示したと報告されています。具体的には、フレームレートに依存しないフレームサンプリング、動画長 1 秒当たり約 147 トークンの増加、およびフレーム解像度に応じたトークン変動パターンが一致しました。[2]

第三に、モダリティ制約と出力傾向です。ox-alpha は音声入力を受け付けず、この挙動は比較対象となった GLM-5V 系と整合すると報告されています。[2] また、検証者のサンプルでは、生成テキスト 1,000 文字当たり約 1.3 個の絵文字が現れ、同じ環境で測定した Claude、GPT-5.6、Grok より、GLM や Qwen 等の中国系モデルで観測される出力傾向に近かったとされています。[2] もっとも、文体上の特徴はプロンプトやシステム設定にも左右されるため、単独では弱い補助証拠にとどまります。

以上は、ox-alpha が GLM 系モデルであるとの推定を相当程度補強します。ただし、共通トークナイザーや動画前処理基盤の採用、モデル蒸留、API ラッパーまたは推論基盤の共通性によっても類似は生じ得ます。智譜 AI および匿名提供者から出自を確定する公式発表はなく、モデル

の真の開発元と正式名称は未確定です。[1][2][3]

プログラミングおよびエージェント機能の初期評価

ox-alpha は、ソフトウェア開発分野の限定的なコミュニティ評価で良好な結果を記録したと報告されています。DeepSWE の公式ハーネスを用い、シード 0 で決定論的に選定された 10 タスクを各 1 回実行した小規模試行では、8 タスクを解決しました。[2][7] これは選定された 10 タスクにおける 8/10 という結果であり、DeepSWE 全体に対する 80%達成や、一般的なコーディング性能の優位を意味するものではありません。

モデル名称	ステータス	同一 10 タスクに関する参考値（条件差あり）	対応モダリティ
stealth/ox-alpha	匿名プレビュー	80.0%（8/10、各 1 試行）[2]	テキスト、画像、動画入力
Claude Fable 5	商用 API	65.0%（各 4 試行の平均）[2]	テキスト、画像
GLM-5.3（Text）	商用 API	62.0%（各 4 試行の平均）[2]	テキスト
Grok 4.6	商用 API	62.0%（各 4 試行の平均）[2]	テキスト、画像
GPT-5.6-sol	商用 API	52.0%（各 4 試行の平均）[2]	テキスト、画像

※上表は厳密な同条件比較ではありません。ox-alpha は 10 タスクを各 1 回実行した結果であるのに対し、他モデルの数値は、同じ 10 タスクについて DeepSWE の公開試行データから算出された各 4 試行の平均です。[2] タスク数が少なく統計的不確実性も大きいため、「限定された 10 タスクで、比較モデルの公開試行平均を上回った」という参考情報として解釈する必要があります。包括的かつ独立した大規模ベンチマークでの総合的なコーディング能力を示すものではありません。

個別タスクでは、JavaScript パーサー「meriyah」の「meriyah-explicit-resource-declarations」について、ox-alpha が初回試行でリファクタリングを完遂し、51,469 件の回帰テストを通過したと報告されています。同じタスクで GLM-5.3、GPT-5.6-sol および Grok 4.6 は各 4 試行とも未解決だったとされています。[2] もっとも、これは一つのタスクに関する事例であり、モデル全体の序列を確定するものではありません。

また、個別の利用報告では、69 回のツール呼び出しを伴うマルチステップ・エージェントタス

クを、1 回のエラーと再試行ループなしで完了した例や、ブラウザ上での Ubuntu デスクトップ環境の再現、3D ゲーム生成などのデモが共有されています。[2] 一方、話題となった GPU 加速型 WebGL 流体シミュレーションの一発生成については、既存の著名なオープンソース実装（paveldogreat の WebGL Fluid Simulation）との高い類似が指摘されており、未知の問題解決能力ではなく、学習データに含まれた既存コードを再現した可能性を排除できません。[3] OpenRouter のモデルページでは、Claude Code や Hermes Agent などを通じて同モデルへ送られたトークン量が表示されています。[1] これは、これらのエージェントツール経由で ox-alpha が利用されていることを示しますが、各ツールが常に同モデルへ自動ルーティングすることを意味するものではありません。

智譜 AI の動向と匿名テスト戦略の背景

智譜 AI（Zhipu AI）は、清華大学の技術成果を基盤として 2019 年に設立され、2025 年に海外ブランドを「Z.ai」へ改めました。[5] 2026 年 1 月 8 日に香港証券取引所へ上場しています。

[6]

中国の主要 AI ラボでは、新モデルの正式発表前に、OpenRouter 等の第三者プラットフォーム上で「動物名+Alpha」の匿名コードネームを用いて試験提供する例が観測されています。[2] ただし、匿名モデルと後日公表されたモデルとの対応関係には、企業による公式確認、事後的な表示変更、技術的推定など確認水準の異なるものが含まれるため、同列に扱うべきではありません。

匿名モデル名	登場時期	後に対応づけられたモデル／企業	確認水準・留意点
Pony Alpha	2026 年 2 月	GLM-5／智譜 AI	後日 GLM-5 と対応づけられた事例 [2][8]
Hunter Alpha	2026 年 3 月	MiMo 系／Xiaomi との見方	コミュニティによる対応づけ。公式確認の有無を要区別 [2]
Elephant Alpha	2026 年 4 月	Ling 系／Ant Group との見方	コミュニティによる対応づけ。公式確認の有無を要区別 [2]
Owl Alpha	2026 年 4 月下旬	LongCat-2.0／Meituan との見方	コミュニティによる対応づけ。公式確認の有無を要区別 [2]

Ox Alpha	2026 年 8 月 20 日	未公表。未発表 GLM 系との推測	技術的状況証拠に基づく推定であり未確認 [1][2][3]
----------	-----------------	----------------------	-------------------------------

2026 年 2 月に公開された「Pony Alpha」は、後に GLM-5 と対応づけられました。[2][8] これが、ox-alpha についても智譜 AI による正式発表前の匿名試験ではないかと推測される背景の一つです。ただし、過去に類似の事例があることだけで、ox-alpha の出自を確定することはできません。

智譜 AI は 2026 年 8 月 14 日にテキストモデル「GLM-5.3」を発表し、GLM-5.2 と同じベースモデルを用いながら、ポストトレーニングの強化により複雑なコーディングやサイバーセキュリティ等の能力を向上させたと説明しています。[4] 同社は、GLM モデル群を用いた脆弱性発見の成果や、GLM-5.3 のモデル重みを安全性評価後に公開する方針も説明しています。[4] これらは GLM-5.3 に関する情報であり、ox-alpha 自体の正式名称、重み公開または提供継続の予定を示すものではありません。

結論

OpenRouter 上に登場した匿名モデル「stealth/ox-alpha」は、トークン数の一致、動画エンコーダーの処理特性、モダリティ制約、エラー応答および出力スタイルなど、智譜 AI (Z.ai) の GLM シリーズと整合する複数の特徴を示しています。[2][3] とりわけ動画 4 条件におけるトークン消費量の一致は有力な状況証拠です。ただし、これらはブラックボックス検証による推定であり、公式な帰属確認ではありません。

限定された 10 タスクの試行では高いコーディング性能が報告され、同じタスクに関する GPT-5.6-sol や Claude Fable 5 等の公開試行平均を上回りました。[2] しかし、ox-alpha と比較モデルでは試行回数が異なり、タスク数も少ないため、これをもってモデル全般で既存の主要モデルを上回ったと結論づけることはできません。開発元および正式なモデル名も公式には確認されていません。

したがって、現時点で最も妥当な評価は、「ox-alpha は未発表のマルチモーダル GLM 系モデルである可能性が高いが、確定には至らない」というものです。真の能力、安全性、データ取扱いおよび実用性については、提供者による公式発表と、同一条件・十分なタスク数による独立検証を待つ必要があります。なお、OpenRouter は、同モデルについてプロンプトと応答が提供者に保持されると説明しているため、機密情報を含む業務利用には慎重な判断が必要です。[1]

引用文献

1. OpenRouter, “Ox Alpha — stealth/ox-alpha,” <https://openrouter.ai/stealth/ox-alpha> (2026 年 8 月 23 日確認)
2. Ben Davis et al., “What Is Ox Alpha? The evidence, in brief,” <https://ox-alpha-evidence-production.up.railway.app/> (2026 年 8 月 21 日付、同月 23 日確認)
3. FlunkyGraphics, “I fingerprinted Ox Alpha: same tokenizer as GLM-5.3 (+75 token offset), z.ai’s exact error strings, near-identical temp-0 outputs,” Reddit, https://www.reddit.com/r/singularity/comments/1vufbx1/i_fingerprinted_ox_alpha_same_

tokenizer_as_glm53/ (2026 年 8 月 23 日確認)

4. Z.ai, "GLM-5.3: Frontier Coding with Emergent Cyber Capabilities," 2026 年 8 月 14 日, <https://z.ai/blog/glm-5.3>
5. Z.ai, "Company," <https://z.ai/company> (2026 年 8 月 23 日確認)
6. Hong Kong Exchanges and Clearing Limited, "AI IPOs Drive a Strong Start to 2026," https://www.hkexgroup.com/Media-Centre/Insight/Insight/2026/Johnson-Chui/AI-IPOs-Drive-a-Strong-Start-to-2026?sc_lang=en
7. DataCurve, "DeepSWE," <https://deepswe.datacurve.ai/> (2026 年 8 月 23 日確認)
8. OpenRouter, "Pony Alpha," <https://openrouter.ai/openrouter/pony-alpha> (2026 年 8 月 23 日確認)