ChatGPT

匿名AIモデル「Ox Alpha」と智譜の未発表GLMフラッグシップ説に関する徹底調査報ート

エグゼクティブサマリ

調査基準日：2026年8月22日（JST）

本調査の結論を先に述べる。

指定記事が扱う匿名モデルは、OpenRouter上で2026年8月20日に公開された `stealth/ox-alpha`（**Ox Alpha**）である。Ox Alphaの存在、匿名の第三者プロバイダーが運用していること、**1,048,576トークンのコンテキスト、最大131,072トークン出力、テキスト・画像・動画入力、ツール呼び出し、現在無料**という基本仕様はOpenRouter自身が確認しており、ここは極めて信頼度が高い。OpenRouterは同時に「自社は開発者・所有者・プロバイダーではない」と明記している。現在観測されるOpenRouter上のP50値はおおむね初期レイテンシ5.3秒、出力24 token/sだが、これは配信条件で変動し得る。[^OR1] [1]

一方、記事の最も刺激的な二つの主張――**「Ox Alphaは智譜（Zhipu/Z.ai）の未発表GLM-5.x系フラッグシップである」**、および**「プログラミングテストでGPT-5.6とClaudeを上回った」**――は、現時点では同じ強さでは支持できない。

まずモデル帰属については、相応に説得力のある状況証拠がある。調査の発端となった開発者Ben Davisは、

> “99% sure it's GLM-5.x, all the evidence points to it (same video encoder, same tokenizer, style matches, same audio rejection, etc.)”

と投稿している。和訳すると、**「99% GLM-5.xだと確信している。証拠はすべてそちらを指している。同じ動画エンコーダー、同じトークナイザー、文体の一致、同じ音声拒否などだ」**という意味になる。原文リンクは `https://x.com/davis7/status/2090669483740279155`。ただし、これは独立した一次技術検証ではなく投稿者による帰属推定である。[^BD1] [2]

それでもGLM説を単なる憶測として片付けにくい最大の理由は、**智譜自身にOpenRouterで未発表モデルを匿名評価した前例がある**からだ。GLM-5 Technical Reportは、OpenRouter上の匿名モデル「Pony Alpha」が実際にGLM-5だったことを公式に明らかにしている。したがって「智譜が次期モデルをOpenRouterでステルス試験する」という行為パターン自体には一次資料上の前例がある。[^G5P] [3]

さらに、Ox Alphaで報告された特徴は、公開済みの智譜技術を非常に自然に組み合わせたように見える。GLM-5は**744B総パラメータ／40B active、256 experts、80 layers、DeepSeek Sparse Attention（DSA）、28.5Tトークン規模のbase-model training、Reasoning RL→Agentic RL→General RL、非同期RL基盤**を採用する。GLM-5.2は公式GitHubで**1Mコンテキスト**と長文推論最適化を掲げる。一方GLM-5V-Turboは**CogViT、MTP、画像・動画・テキスト入力、30種類超のタスクを用いたjoint RL、multimodal coding/GUI agent**に特化している。Ox Alphaはちょうど、**GLM-5.2/5.3側の1M・coding/agent能力とGLM-5V-Turbo側のvideo/multimodal能力を一つにしたような表面特性**を示す。[^G5][^G52][^G5V] [4]

しかも智譜の2025年12月30日付香港上場目論見書には、2026〜2028年の研究計画として、**「新しいモデルアーキテクチャ」「新しいattention/memory mechanisms」「infinity context」「test-time and**

online learning」「deep reasoning」「self-refinement/self-evolution」「AI agent」への投資が明記されている。この資料はOx Alpha出現の約8か月前、GLM-5公開より前に提出された企業一次資料であり、Oxの1M長文・agentic coding・新しいマルチモーダル構成との技術的整合性はかなり高い。もっとも、それは「Ox＝GLM」を直接証明するものではない。[^IPO] 5

しかし、**「GPT-5.6とClaudeを上回った」というヘッドラインは、そのまま一般化してはならない**。指定記事の80%はOx Alphaを**10問**で試して8問成功した小規模試験であり、比較対象に使われたGPT-5.6 Sol 52%、Claude Fable 5 65%、GLM-5.3 62%は同条件・同母数の統制されたhead-to-headとは限らないと記事自体が認めている。8/10という観測値の単純なWilson 95%信頼区間は約**49.0〜94.3%**で、標本誤差が非常に大きい。 6

さらに決定的なのは、8月20日更新の**DeepSWE公式113タスク・v1.1 leaderboard**では、Ox Alphaはまだ掲載されておらず、同じmini-swe-agent条件下で**Claude Opus 5 = 74±4%、GPT-5.6 Sol = 73±3%、Claude Fable 5 = 70±4%、GLM-5.3 = 69±3%**となっていることである。したがって、現時点の公開証拠は「Oxが10問の特定サブセットで8/10だった」ことを示唆するにとどまり、**「GPT-5.6やClaudeを一般的に上回るcoding model」であることを立証していない**。[^DS1] 7

第三者のArtificial Analysisも、GPT-5.6 Sol（max）をCoding Agent Indexで**80点・首位**と評価し、DeepSWE、Terminal-Bench v2、SWE-Atlas-QnAの三系統を統合している。したがって、少なくとも現在公開されている広範なcoding evaluationでは「Ox Alphaの優位」が確立したとはいえない。[^AA1] 8

本調査の総合信頼度評価は以下の通りである。

判断	本調査の信頼度	理由
Ox Alphaという匿名モデルが実在し、1M・画像/動画・coding/agentic対応	**極めて高い 98%**	OpenRouter一次資料
記事の「10問中8問成功」という観測が実際に報告された	**高い 90%**	記事と投稿ベース。ただし独立再現なし
「OxがGPT-5.6/ClaudeよりcodingでOverall優秀」	**低い 20〜30%**	n=10、統制不足、公式DeepSWEと整合しない
OxがGLM技術系統に属する	**中〜やや高い 65〜75%**	tokenizer/video fingerprint、技術整合性、Pony Alpha前例
Oxが「智譜の次期正式フラッグシップ」そのもの	**中程度 45〜60%**	GLM派生実験版・VLM派生・評価checkpointの可能性を排除できない
Oxが744B/40B active級MoE	**低〜中 30〜40%**	GLM-5系との整合性はあるが直接証拠なし
100兆tokens/dayの能力	**低い 15〜25%**	一次定義・実測根拠なし。「capacity」の意味も未指定

重要なのは、**「GLM説」と「ベンチマーク勝利説」を分けて評価すること**である。前者にはかなり面白い技術的状況証拠が存在する。後者は、現段階では見出しが証拠を大きく先取りしている。

指定記事の内容、原文追跡、出典信頼性

指定されたBigGo Finance記事は2026年8月21日19:05（GMT+9）付で、Ox Alphaが8月20日にOpenRouterへ登場し、約1Mコンテキスト、画像・動画入力、reasoning、coding/agent機能を持つと説明する。そのうえでBen Davisによるtokenizer/video encoder/style/audio rejectionの比較と、小規模DeepSWE試験を組み合わせ、「智譜の未発表GLM新フラッグシップである可能性」を提示している。記事本文自身も、**これは個別試験からの推論で、智譜による公式確認はない**と明記している。[^BG1] [6]

記事の主張を分解すると次のようになる。

記事の主要主張	検証結果	判定
Ox Alphaは8月20日公開のステルスモデル	OpenRouterが8月20日公開と明記	**確認**
1,048,576 context / 131,072 output	OpenRouter一次資料と一致	**確認**
text/image/video入力	OpenRouter一次資料と一致	**確認**
reasoning、tool calling、structured output	OpenRouter一次資料で概ね確認	**確認**
GLM-5.xと同一系統のtokenizer	Davis/記事によるblack-box比較のみ	**未独立検証**
GLM-5V-Turboと動画token behaviorが一致	記事内controlled testのみ	**未独立検証**
audio rejectionがGLMと一致	記事/Davisのみ	**弱い状況証拠**
10 DeepSWE tasksで8/10	記事による報告。生ログ未公開	**報告事実は確認、結果は未再現**
GPT-5.6 Sol 52%、Fable 5 65%、GLM-5.3 62%を上回る	同条件比較でない。公式full DeepSWEは異なる順位	**一般化は不当**
meriyah課題をone-shotで解き51,469 regression tests通過	記事のみ。trajectory/verifier artifact未公開	**未独立検証**
Oxは智譜の次期GLM旗艦	直接確認なし	**有力仮説の一つ**
100T tokens/day serving capacity	OpenRouter現行一次ページでは確認できず	**未確認**

OpenRouterの現行ページはOxについて「third-party model provider」「provider has chosen to remain anonymous」と明記しているため、**OpenRouter自身も2026年8月22日時点で開発元を公開していない**。また、Oxへのprompt/completionはprovider側に保持される一方、「trainingには使わない」と説明されている。この点は機密コードを投入する企業利用では重要である。[^OR1] [9]

プログラミングテストの扱いは特に注意が必要である。記事の80%は8/10にすぎず、95% Wilson intervalは約49〜94%。一方、DeepSWE公式113問ではGPT-5.6 Solが73±3%、Fable 5が70±4%、GLM-5.3が69±3%。この三者はかなり接近しており、隣接モデルの信頼区間も重なる。DeepSWE自身もfrontier modelsが狭い帯域に集中すると説明している。Oxが真に80%を維持するなら大きな成果だが、**113問で80%前後を再現するまでは「勝利」と呼ぶべきではない**。 [7]

原文と情報伝播経路

BigGoの指定ページ： `https://finance.biggo.jp/news/9dc856ba-634d-467a-bea2-6ba70233113c@Deep Research` [^BG1]

Ben Davisの原文： `https://x.com/davis7/status/2090669483740279155` [^BD1]

記事の中心的帰属根拠に最も近い英語一次表現は前述の「99% sure it's GLM-5.x...」であり、**これは「智譜が確認した」という意味ではなく、Davis個人のblack-box fingerprintingに基づく確信度表現**である。 2

BigGo記事と極めて近い内容は同日、Sina/Sina FinanceやSohuにも出ており、それらは**「来源：网易科技（出典：NetEase Tech）」**と表示している。したがって、BigGoの文章は独自取材というより、NetEase Tech系の記事が複数媒体へ配信・再編集されたチェーンとみるのが合理的である。ただし、本調査では検索可能な範囲から**NetEaseの真正なオリジナル記事URLそのものを確定できなかったため、直接原稿のURLは「未確認」**とする。 10

BigGoの信頼性評価

BigGoは本来、台湾発の価格比較サービスとして展開され、2019年の日本展開時の資料では**FunmulaがBigGoを運営**し、日本展開でBlue Gooseと提携したと説明されている。ただし、これは歴史資料であり、**2026年8月時点のBigGo Financeの最終的な法人運営主体を指定記事ページだけから確定することはできなかったため、現行法人主体は未指定**とする。 11

BigGo Financeは株式・市場データ・ニュースを集約するサービスであり、指定記事には強い独自取材を示す記者名・一次データセット・実験ログが付いていない。加えてサイト自身が情報の遅延や中断、正確性・完全性を保証しない趣旨の免責を掲げる。したがって、**匿名モデルの存在を知る二次情報源としては有用だが、モデル帰属やベンチマーク優位を確定する証拠としての重みは低い**。 6

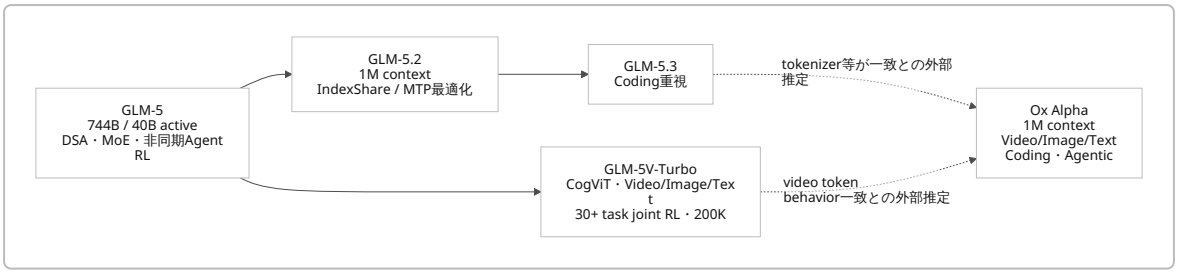
「類似記事が複数媒体にある」ことも、独立した複数社の検証を意味しない。同じNetEase Tech由来の記事をSina/Sohu/BigGoが再配信しているなら、**ソース数は増えても独立証拠数は増えていない**。過去の記事訂正率・誤報率などを定量評価できる公開監査データは本調査では確認できず、BigGo Financeの**長期的なニュース精度は未指定**である。

総合すると、指定記事の情報源評価は **「速報・仮説発見：B-、未発表モデル帰属の確証：C〜D」** とみるのが妥当である。記事が自ら不確実性を明記している点はプラスだが、見出しは本文より強い。

智譜・GLM技術系譜と未発表モデルを示唆する一次情報

智譜（北京智譜華章科技股份有限公司、Zhipu/Z.ai）は清華大学系の研究を背景にGLMを発展させてきた企業で、初期GLMはautoregressive blank infillingを中心とするGeneral Language Modelとして論文化され、その後GLM-130B、ChatGLM、GLM-4.x、GLM-5系列へ発展した。GLM-5 technical reportでは、単なるdense Transformerのスケールアップではなく、MoE、sparse attention、長文mid-training、大規模agent RLを中核に置く方向へ明確に移行している。[^GLM0][^G5] 12

公開済み技術の関係を簡略化すると、Ox仮説は次のように位置付けられる。



実線部分は公開されたGLM技術の系譜、点線部分は**本件の帰属仮説**であって、智譜公式発表ではない。 13

GLM-5ではbase-model trainingを27T-token corpusから開始し、全base training stageで28.5T tokens、contextを4Kから200Kへmid-trainingで延伸した。post-trainingはReasoning RL、Agentic RL、General RLを順次実施し、On-Policy Cross-Stage Distillationで能力忘却を抑制する。アーキテクチャは256 experts、80 layers、**744B total / 40B active**、DSAを採用する。さらに中国国内7種GPUプラットフォームへのfull-stack optimizationを公式報告している。[^G5] 14

GLM-5.2について公式GitHubは、長時間タスク能力と**solid 1M-token context**を強調している。リポジトリではIndexShareによる1M長文時のper-token FLOPs削減やMTP speculative decoding改善も説明されている。したがって、Oxの1,048,576 contextは智譜にとって技術的な飛躍ではなく、**既にtext-only系で持つ1Mコンテキストをmultimodalモデルへ統合したもの**と考えれば自然である。[^G52] 15

GLM-5V-Turboはさらに重要である。公式ドキュメント上、**video / image / text / file入力、text output、200K context、128K max output**を持つ「multimodal coding foundation model」で、CogViT vision encoder、推論向けMTP、pretrainingからpost-trainingまでのnative multimodal fusion、STEM・grounding・video・GUI agents・coding agentsを含む**30+ task joint reinforcement learning**を採用している。[^G5V] 16

Oxの公開仕様は1M contextであり、GLM-5V-Turboの200Kを大きく上回る。そのため、仮にOxが智譜製なら、単なる「GLM-5V-Turboの名称隠し」よりも、**5.2/5.3のlong-context stackと5V系multimodal stackを統合した後継checkpoint**という解釈の方が技術的には説明しやすい。ただし、そのcheckpointが正式名称「GLM-5.4」「GLM-5V」「GLM-6」等の何になるかは**未指定**であり、「GLM-5.x」というラベルすらDavisの推測である。 17

企業ロードマップという強い状況証拠

2025年12月30日の香港上場目論見書は、未発表技術を推測する上でSNSよりはるかに重要である。智譜はIPO純調達額の約70%、**HK$2.9214 billion**をgeneral-purpose large AI modelの研究開発へ振り向け、2026〜2028年に段階投入すると記載している。その内部ではpretrained models、deep reasoning、agents、emerging technologies、GLM framework、data processing/corporaへ資金を割り当てている。[^IPO] 5

特にdeep reasoning領域の説明は、本件との整合性が高い。目論見書は今後のR&D対象として、**新しいmodel architecture、attention/memory mechanism、infinity context、test-time/online learning、deep reasoning algorithms、self-refinement/self-evolution**を列挙している。これはOxを名指しする資料ではないが、「1M以上の文脈」「test-time compute」「long-horizon agent」という方向が偶然の記事推測ではなく、智譜が上場時に公表していた正式R&Dロードマップに含まれていることを示す。 5

同目論見書によれば2024年のR&D費は約RMB2.195 billion、2025年上期だけでも約RMB1.595 billionで、外部compute providerへの研究用compute費も大きく増加していた。したがって、GLM-5以降もfrontier modelを短い周期で反復するための資本投入は公開財務情報とも整合する。[^IPORD] 18

特許・ソースコード・研究論文

目論見書は、基準日時点で**中国登録特許86件（うち発明特許84件）、出願232件**を保有するとしている。特に未発表GLMを考えるうえで興味深いのは、「LLM instruction-following最適化」「web navigation AI agent training」「LLMのdepth動的調整」「decision-tree based MoE training」「LLM safety evaluation」などである。[^IPOPT] [19]

「LLMのdepth動的調整」に対応するCN118153551Bは、2024年3月11日出願、2024年10月1日grantで、Google Patents上でも智譜がassigneeとして確認できる。ただし、**特許は実施製品を証明しない**。動的深度がOxに実装されていると判断できるblack-box証拠は現時点でない。[^PAT1] [20]

公開ソースでは、`zai-org/GLM-5` が744B-A40B系列や1M-context関連の実装・設定を公開している一方、**Ox Alphaそのもののweights、model config、tokenizer files、source codeは公開されていない**。Nous ResearchのHermes Agent等にはOx Alphaのmodel catalog/integrationを示すコード更新が見つかるが、これはendpointの存在を裏付けるだけで、モデル内部の由来を示すソースコードではない。 [21]

GLM-5V-Turboに関しては2026年4月の公式リリースと論文があり、native multimodal agent foundationとして設計されたことが論文化されている。したがって、「動画を理解できるGLM系フラッグシップ」という発想自体は未発表技術ではない。Oxで新しいのは、その能力を1M context・最新coding能力と組み合わせているように見える点である。[^G5VP] [22]

採用・実運用

智譜の一次資料ではMaaS/API、cloud、on-premise、device deploymentを広く提供している。上場目論見書には2025年初め、ある「leading lifestyle-focused social media app」が英語ユーザー急増時に智譜モデルを使った自動翻訳機能を導入した例が記載されているが、企業名は開示されていないため**採用企業名は未指定**である。 [23]

GLM-5V-Turboの公式資料はClaude CodeやOpenClawと組み合わせたGUI/coding workflowを明示しており、少なくともagent ecosystemとの統合は実運用を意識した設計である。Ox自身についてもOpenRouterはHermes Agent、Claude Code、DeepSeek Harness、ZCodeなどから大量の利用が生じていることを表示しているが、**これはOxの開発元を示す証拠ではない**。 [24]

採用求人については、今回の検索で**「次世代GLM」「Ox Alpha」を直接示す、日付・職種・要件を検証可能な智譜公式求人票を確認できなかった**。したがって求人情報からの帰属推論は「未指定」とし、SNSの求人転載等を証拠として採用しない。

GPT-5.6・Claudeとの仕様とcoding性能比較

比較対象としては、記事が直接挙げる**GPT-5.6 Sol**と**Claude Fable 5**を中心にする。Claudeには2026年7月公開のOpus 5もあるため、現在のClaude最高性能層を論じる際はOpus 5も補助的に参照すべきである。DeepSWEでもOpus 5が74%、Fable 5が70%と別々に掲載されている。 [25]

比較軸	Ox Alpha	GLM公開系から推定できるもの	GPT-5.6 Sol	Claude Fable 5
開発者	**匿名・未指定**	Zhipu/Z.ai	OpenAI	Anthropic
総パラメータ	**未指定**	GLM-5: 744B	**未指定**	**未指定**

比較軸	Ox Alpha	GLM公開系から推定できるもの	GPT-5.6 Sol	Claude Fable 5
Active params	**未指定**	GLM-5: 40B	**未指定**	**未指定**
architecture	**未指定**	GLM-5: MoE+DSA/MLA系。5V: CogViT+MTP	**未指定**	**未指定**
training tokens	**未指定**	GLM-5 base: 28.5T	**未指定**	**未指定**
training data内容	**未指定**	code/reasoning優先、long-context agentic data等を公開	詳細**未指定**	詳細**未指定**
post-training	**未指定**	Reasoning RL→Agent RL→General RL、cross-stage distillation等	詳細recipe**未指定**	詳細recipe**未指定**
context	**1,048,576**	GLM-5.2: 1M、5V-Turbo: 200K	**1,050,000**	**1M**
max output	**131,072**	5V-Turbo: 128K	**128K**	**128K**
text	✓	✓	✓	✓
image	✓	5V系✓	✓ input	✓
video	✓	5V系✓	**非対応**	native videoは**未指定**
audio	記事では拒否	別ASRモデルあり	**非対応**	本比較では**未指定**
tool/function calling	✓	✓	✓	✓
reasoning control	reasoning model	GLM-5.3等effort制御	none〜max	adaptive reasoning
公開価格	**現在無料**	GLM-5.3 APIは低価格帯	$4 input / $20 output per MTok	$10 / $50 per MTok
モデル内部公開度	完全非公開	GLM-5系open weights多数	closed	closed
DeepSWE full	**未掲載**	GLM-5.3 69±3%	**73±3%**	**70±4%**
記事の10問試験	**80%**	62%と記事報告	52%と記事報告	65%と記事報告

Ox仕様はOpenRouter、GLMはZ.ai/technical report、GPT-5.6はOpenAI公式、Fable 5はAnthropic公式に基づく。[^OR1][^G5][^G5V][^OAI][^ANT] 26

GPT-5.6

OpenAI公式モデルページによればGPT-5.6 Solは1,050,000 context、128,000 max output、knowledge cutoffは2026年2月16日。reasoning effortはnone / low / medium / high / xhigh / maxで、textは入出力、

imageは入力対応、audio/videoは非対応。API価格は現在1M tokensあたり**$4 input、$0.40 cached input、$20 output**で、272Kを超えるinputではrequest全体にlong-context surchargeが適用される。[^OAI] [27]

OpenAIはモデルのparameter count、expert count、training token count、具体的training dataset compositionを公開していないため、それらは**未指定**である。「GPT-5.6は何Bか」といった数字を第三者推測でOxとの比較表に入れるのは不適切である。

独立評価ではArtificial AnalysisがGPT-5.6 Sol（max）をCoding Agent Index **80点**としており、そのindexはDeepSWE、Terminal-Bench v2、SWE-Atlas-QnAを組み合わせる。AA Intelligence IndexではFable 5に1点差の59と評価され、cost/taskはFableより低いと報告されている。つまり**記事の「GPT-5.6 = 52%」だけを見ると、現在公開されているGPT-5.6 coding性能をかなり過小評価する可能性がある**。[^AA1] [8]

Claude Fable 5

Anthropic公式ではFable 5は2026年6月9日公開、1M context、128K max output、adaptive thinkingを持ち、価格は**$10/MTok input、$50/MTok output**。Fable 5とMythos 5は基礎能力を共有しつつ、Fable側ではcyber/biology領域のより強い安全制御を適用するという特殊な位置付けである。[^ANT] [28]

Anthropicもparameter count、MoE expert構成、training tokens、raw training corpusを公開していないため、これらは**未指定**である。したがって「Ox/GLMは40B activeだからClaudeより効率的」といった比較も、Claude側の内部規模が分からない以上成立しない。

現在のDeepSWEではFable 5 maxが70±4%、平均task cost $21.63、119K output tokens、88 agent steps。GPT-5.6 Solは73±3%、$8.39、60K、61 steps、GLM-5.3は69±3%、$3.99、80K、124 stepsである。**この公開値を見る限り、GLM-5.3は絶対性能が3〜4ポイント低い代わりに大幅に安く、GPT-5.6はscore/costのバランスが良い**という姿が見える。Oxはfull leaderboardにないため比較不能である。 [7]

安全性・制御

GPT-5.6はOpenAIがsystem card/Preparedness評価を公開し、cyber・biology等を含むlayered safeguard評価を行っている。Fable 5はそもそもMythos 5に強いcyber/biology safeguardを適用した公開版という設計思想である。対してOx AlphaについてOpenRouterが公開しているのはStealth Model Termsやdata-retention条件程度で、**model-level safety training、red-team、CBRN/cyber評価、拒否policy、system-card相当資料はすべて未指定**である。 [29]

智譜については、GLM-5.3のweights公開を安全評価のため後ろ倒しする方針が報道されており、またIPO特許一覧にはLLM safety evaluation特許も含まれる。しかし、これもOxの具体的安全機構を意味しない。 [30]

コストとレイテンシの公平な読み方

Oxは現在OpenRouter上で無料で、P50 throughput約24 tok/s、latency約5.3秒と表示されている。ただし無料previewは補助金付き・rate-limitedな実験配信である可能性があり、**商用価格や専有GPU時の実力を推定する根拠にはならない**。[^OR1] [9]

GPT-5.6の公式API単価とFableの公式単価は上表の通りだが、実タスクコストはreasoning effort、cache hit、tool calls、output lengthで大きく変わる。したがって、単純な$/MTokより、DeepSWEのような「一つのissueを解決するまでのcost/task、wall-clock time、steps、success probability」を同時に測るべきである。 [31]

「未発表GLM新フラッグシップ」仮説の技術評価と一次情報タイムライン

Oxが智譜製だと仮定した場合、最も合理的な技術像は「完全に未知の新architecture」ではなく、**GLM-5/5.2/5.3のlanguage-agent stackとGLM-5V-Turboのmultimodal stackを統合し、long-context servingを強化したcheckpoint**である。

示唆される特徴	実現可能性	GLM既存技術との整合性	必要な追加R&D	帰属証拠としての強さ
GLM系tokenizer	**非常に高い**	既存tokenizer共有は自然	tokenizer/version管理	**中〜強**
GLM-5V系video encoder	**高い**	CogViT・video native既存	1M language backboneとのfusion	**強いが未再現**
1M multimodal context	**高い**	5.2=1M、5V=200K	sparse attention、memory、KV/cache最適化	**中**
744B/40B active級MoE	**中**	GLM-5にその構成	expert routing、distributed serving	**弱〜中**
DSA/sparse attention	**高い**	GLM-5公式	multimodal tokenへの拡張	**中**
IndexShare系long-context最適化	**高い**	GLM-5.2公開	kernels、prefill/decode optimization	**中**
MTP/speculative decoding	**非常に高い**	5.2/5Vで既存	multimodal decoding tuning	**中**
強力なagentic coding RL	**非常に高い**	GLM-5 Agent RL、5V 30+ task RL、5.3 coding強化	sandbox、verifiable reward、trajectory data	**強い整合性**
test-time compute拡張	**高い**	IPO roadmapに明記	verifier/reward、inference orchestration	**中**
self-refinement/self-evolution	**中〜高**	IPO roadmapに明記	online safety、reward hacking対策	**弱〜中**
音声非対応	**高い**	GLM側は音声を別モデルで扱う構成と矛盾せず	なし	**弱い**
絵文字/文体fingerprint	**容易**	どのモデルでも模倣可能	なし	**非常に弱い**
75-token hidden wrapper	**容易**	provider gatewayでも付加可能	なし	**弱い**

示唆される特徴	実現可能性	GLM既存技術との整合性	必要な追加R&D	帰属証拠としての強さ
100T tokens/day	技術的には可能だが**要検証**	大規模inference投資とは整合	巨大accelerator fleet、batching、cache、network	**現状ほぼ証拠にならない**

GLM-5/5Vの技術基盤は既に公開され、IPOロードマップにもlong-context、attention/memory、test-time/online learningが明記されるため、上表の多くは「研究上実現可能か」という意味では高い。[32]

ただし**744B/40B active説は慎重に扱うべき**である。GLM-5がそのサイズであることは確実だが、GLM-5V-Turbo公式資料はむしろ「smaller parameter size」で高性能を実現したと説明しており、Oxが744B級とは限らない。5V系列に別サイズのmultimodal backboneを採用し、5.3由来のpost-trainingだけ共有している可能性もある。[33]

仮に744B weightsなら、単純計算で全weight保存だけでも8-bitで約744GB、16-bitなら約1.49TBが必要で、MoEの40B activeは「1 tokenあたり実際に計算するweight量」を減らす一方、全expertをどこかに保持・分散する必要は残る。このため高速配信にはexpert parallelism、high-bandwidth interconnect、cache、quantization、batch schedulingが重要になる。Oxの実パラメータ数は**未指定**であるため、この計算はあくまでGLM-5級だった場合の規模感である。[14]

記事の「100兆tokens/day」が文字通りaggregate processed tokensを意味するなら、平均で約**11.57億tokens/s**に相当する。これは極めて大きな数字であり、input/prefill/cache hit/outputのどれを含むのか、理論capacityなのか実測throughputなのかが不明である。OpenRouterのOxページには現時点でこの数値が正式仕様として記載されていないため、**定義未指定・未検証のmarketing-like claim**として扱うべきである。[34]

未発表モデルを示唆する情報の時系列

日付	種類	一次情報・URL	内容	Oxとの関連度	信頼度
2024-03-11	特許	`https://patents.google.com/patent/CN118153551B/en`	LLMのdepthを動的調整する智譜特許。別途同日MoE training関連出願もIPO資料に掲載	低〜中。adaptive computeの技術蓄積	**高：特許存在**／Ox適用は低
2025-12-30	上場一次資料	`https://www1.hkexnews.hk/listedco/listconews/sehk/2025/1230/2025123000017.pdf`	2026–28に new architecture、attention/memory、infinity context、test-time/online learning、agentsへ投資	**高い方向性一致**	**非常に高**

日付	種類	一次情報・URL	内容	Oxとの関連度	信頼度
2026-02-24頃	論文・公式報告	`https://arxiv.org/html/2602.15763v2`	GLM-5: 744B/A40B、DSA、28.5T、async Agent RL。Pony AlphaがGLM-5だったことも公式開示	**非常に重要な前例**	**非常に高**
2026-04-02	GitHub/公式	`https://github.com/zai-org/GLM-V`	GLM-5V-Turbo公開系統	video encoder説の基盤	**高**
2026-04-29	論文	`https://arxiv.org/abs/2604.26752`	GLM-5V-Turboをnative multimodal agent foundationとして説明	**高**	**高**
2026年前半	GitHub	`https://github.com/zai-org/GLM-5`	GLM-5.2 1M context、long-horizon、IndexShare/MTP等を公開	Oxの1M + fast inferenceと整合	**高**
2026-07-08	研究論文	SAO系RL研究。直接URLは今回の取得結果では**未指定**	single-rollout系RLをGLM-5.2規模で評価	coding/agent post-trainingの進展	**中〜高**
2026-08-14	公式発表	`https://z.ai/blog/glm-5.3`	GLM-5.3、coding能力を大幅強化と発表	Oxの直前checkpoint候補	**高**
2026-08-20	OpenRouter一次資料	`https://openrouter.ai/stealth/ox-alpha`	Ox公開。1.048M、video/image/text、128K級output、coding/agentic	**直接対象**	**非常に高**

日付	種類	一次情報・URL	内容	Oxとの関連度	信頼度
2026-08-20	独立benchmark	`https://deepswe.datacurve.ai/`	113-task v1.1更新。GPT5.6=73、Fable=70、GLM5.3=69。Oxは未掲載	記事の80%主張を検証する基準	**高**
2026-08-20〜21	SNS/black-box test	`https://x.com/davis7/status/2090669483740279155`	same video encoder/tokenizer/style/audio rejectionを理由に「99% GLM-5.x」	**主要リーク的証拠**	**中**
2026-08-21 19:05 JST	二次記事	指定BigGo URL	上記を統合してGLM新flagship説を報道	二次解釈	**中〜低**
2026-08-22	求人情報	**該当する検証可能な公式一次求人は未確認**	次世代GLM/Oxを直接指す求人要件を確認できず	なし	**未指定**
2026-08-22	Ox source code	**未公開**	Hermes Agent等にmodel slug integrationはあるがOx weights/configではない	帰属証拠にならない	**未指定**

重要なのは、**「モデル公開前にしか存在し得ない情報」かどうか**で証拠価値を重み付けすることである。IPO目論見書やGLM-5/5V論文はOx以前に公開されており、智譜の技術方向を示す強い資料である。しかし、それらは他社でも実装可能な技術を記述するため、単独では帰属証明にならない。反対にtokenizerやvideo tokenizationのblack-box fingerprintはOx固有性が高い可能性があるが、実験ログが公開されていない。両者を組み合わせて初めて「GLM説が相当 plausible」という評価になる。[35]

またSNSには「OxはGLMではない」とする反対意見も存在するが、これらも開発元を示す一次証拠を伴っていないため、本調査では低ウェイトとした。つまり現時点のSNS論争は、**「99%説 対 否定説」ではなく、どちらも公式確認前のblack-box speculation**である。[36]

プログラミング優位性を再現検証する実験設計

現状で最も大きな問題は、記事が提示する「8/10」が、モデル性能差なのか、task sampling、agent harness、reasoning budget、provider routing、偶然の揺らぎ、benchmark contaminationのどれによるものか分からないことである。

DeepSWEは現在113 tasks、91 repositories、5 languagesからなり、すべてのleaderboard modelをmini-swe-agent上で比較している。タスクは既存PRを流用せずゼロから作成し、behavioral verifierで判定する。公式の再現手順には10問を固定seedで選ぶモードも用意されているため、記事の10問試験に近い条件をまず再現できる。[^DS1][^DS2] [37]

ただしDeepSWEの「contamination free」という設計思想には時間的な注意点がある。タスクは作成時点ではpretraining solutionの流入を避けているが、DeepSWEは2026年5月以降公開されている。**8月20日に初めて観測されたOxが、公開後のDeepSWE taskをpost-training/data collectionで見ていないことまでは保証できない。Oxのtraining cutoffは未指定だからである。** このため、真のfrontier比較には公開DeepSWEと、新規private holdoutの両方が必要である。 [38]

推奨する再現プロトコルは以下である。

公開benchmark層ではDeepSWE v1.1全113問を使用する。Ox Alpha、GPT-5.6 Sol、Claude Fable 5、Claude Opus 5、GLM-5.3を同一日・同一regionから呼び出し、同じmini-swe-agent commit、同じcontainer image、同じrepository commit、同じtool schema、同一wall-clock upper boundで実行する。provider request ID、model slug、reasoning effort、usage、tool trajectory、git diff、test logsを全保存する。Oxのようなstealth modelはbackendが無告知で変わる可能性があるため、**日時単位の再現性記録が不可欠**である。 [39]

reasoning budgetについては二つの実験を分けるべきである。一つは各社の推奨「max effort」での**best-capability test**、もう一つは例えば最大output tokens、wall-clock、あるいはドルコストを揃える**equal-resource test**である。GPT-5.6 Solには複数reasoning levels、Fableにはadaptive reasoning、GLMにもreasoning controlがあり、「max」というラベルだけでは実計算量は同一でない。 [40]

各taskは最低5回、理想10回実施する。113 tasks × 5 repeatsなら1 modelあたり565 task-runsとなり、単発の偶然成功を大幅に抑えられる。temperature/seedがAPIから固定できないモデルでも反復によってrun-to-run varianceを推定できる。

主要評価指標は**task-level Pass@1**とし、副次指標としてpass probability、pass@k、total output/input tokens、wall-clock、time-to-first-tool、cost/task、agent steps、tool-call error rate、retry count、regression-test failures、patch sizeを記録する。単なる「テストを通したか」だけでなく、**同じ成功率を何トークン・何ドル・何分で得たか**まで比較する。

統計評価は、単回paired outcomeならMcNemar testを用い、model間の**absolute risk differenceと95% CI**を主効果量とする。反復runではtaskをrandom effect、modelをfixed effectとしたmixed-effects logistic regressionが適する。さらにrepository/language/task complexityをstratumとしてbootstrapを10,000回程度行い、task-level 95% CIを求める。複数モデルを同時比較する場合はHolm correctionなどでfamily-wise errorを制御する。

「8/10」のような小標本を今後発表する場合でも、点推定80%だけではなく**Wilson CI ≈49.0〜94.3%**を併記すべきである。これだけ幅がある以上、52%のGPT-5.6との差を「勝った」と断定する統計的根拠にはならない。

非公開holdout層としては、Ox公開後に新規作成した50〜100のsoftware-engineering taskを30以上のreposから用意する。既存公開issue/PRをそのまま使わず、maintainerまたは評価者が新しいbug/featureを人工的に構築し、hidden behavioral testsとimmutable commit hashを用意する。これによりDeepSWE公開後のbenchmark contaminationを回避できる。

さらに、memorization検査として同じ仕様を異なる自然言語表現へparaphraseしたvariant、変数名/ファイル構成だけを変えたsemantically equivalent variant、新規commitに対するvariantを作る。あるモデルだけpublic taskで突出しprivate/paraphrased taskで崩れるなら、training contaminationまたはbenchmark-specific optimizationを疑うべきである。

モデル帰属用fingerprintingはcoding benchmarkから分離する。少なくとも次を事前登録して比較する。

fingerprint試験	方法	帰属価値
tokenizer count	100〜500種類の多言語・code・emoji promptについてusage token count比較	高
hidden wrapper	empty/short promptから固定overheadを推定	中。gateway由来を除外必要
video token scaling	fps、duration、resolution、scene contentを独立に変えtoken countを比較	**高**
audio behavior	silent audio、audio-only、audio+videoを比較	中
tool serialization	schema順序、invalid arguments、parallel calls時のerror behavior	中
refusal/error wording	structured invalid requestsを大量比較	低〜中
style	emoji、句読点、reasoning formatting	低
latency scaling	context length/video durationに対するprefill/decode slope	低〜中

Davisの4動画・25prompt級の試験が正しいとしても、**同じtokenizerをコピーするだけなら他社でも再現できる**。一方、動画のduration、resolution、frame samplingを独立操作した際のtoken scaling curveまでGLM-5V-Turboと一致し、さらにtool serializationやspecial-token overheadまで一致するなら、偶然一致の可能性は大幅に下がる。このデータセットとraw API logsの公開が、現状もっとも価値の高い追加証拠である。 41

リスク評価、シナリオ、最終結論と推奨

誤報リスクは無視できない。 指定記事はBigGo→NetEase系配信という二次・三次情報チェーンにあり、モデル帰属の核心は一人または少人数のblack-box testingに依存する。OpenRouterも智譜も現時点でOxのdeveloper identityを確認していない。記事自身がこの点を留保しているにもかかわらず、見出しだけ読むと「智譜製であることがほぼ確認された」ように見える。この情報構造は、技術ニュースで典型的な**「かなり面白い仮説が、再配信の過程で確定的な見出しになる」リスク**を持つ。 42

リーク偽造・意図的fingerprint mimicryも可能である。匿名providerが既知GLM tokenizerを使い、類似system promptやerror behaviorを実装することは技術的に難しくない。特にstyle、emoji、拒否文は容易に模倣可能である。一方、video encoderのtoken scalingとtokenizer special-token behaviorを多数のcontrolled inputsで同時に再現するのはより難しいため、Davis側がraw resultsを公開すれば証拠価値は大きく上がる。 43

benchmark leakageが最大のナレッジギャップの一つである。Oxのtraining cutoff、post-training dataset、DeepSWE taskへのアクセス歴はすべて未指定。8月公開モデルを5月公開benchmarkだけで評価する以上、公開taskへの過適合を完全には排除できない。 [7]

モデルroutingの問題もある。`stealth/ox-alpha`という一つのendpointの裏でcheckpoint更新、quantization変更、A/B test、provider-side tool logic変更が起きても、ユーザーからは同じslugに見える可能性がある。したがって8月20日の8/10を8月25日に再現できない場合、それが統計揺らぎかmodel updateかを区別できない。OpenRouter上のanonymous previewを学術benchmarkに使う際の根本的制約である。 [9]

法的・倫理的側面では、公開APIのblack-box benchmarking自体と、不正に内部weight/config/trade secretを取得する行為を明確に区別すべきである。特許、公開GitHub、論文、API behaviorからの合法的比較は有用だが、非公開system promptの不正抽出、従業員からのtrade secret取得、漏洩weightの再配布は別問題となる。またOxのproviderはprompt/completionを保持するとOpenRouterが明示しているため、未公開source code、個人データ、企業秘密をfingerprintingのために投入すべきではない。 [44]

商業面では三つのシナリオが考えられる。

シナリオ	内容	商業的含意	現時点評価
A：Ox＝智譜次期正式旗艦	1M native multimodal + frontier codingモデルをステルス評価	中国系open/open-ish frontierがGPT/Claudeへさらに接近。API価格競争とcoding-agent市場への圧力	**十分あり得る**
B：Ox＝智譜の実験checkpoint/5V後継だが正式旗艦ではない	技術fingerprintはGLMだが、product nameやweights構成は評価専用	headlineは半分正しい。技術方向は当たるが製品予測は過剰	**最も保守的なGLM説**
C：Ox＝他社モデル	他社が類似tokenizer/encoder stackまたは共有技術を使用	BigGo/NetEase系の帰属推測は誤報。coding性能そのものは別途価値あり	**無視できない**
D：benchmark-optimized/contaminated model	DeepSWE等に強く最適化	80%が一般coding能力を示さず、商用価値を過大評価	**重要な検証対象**
E：provider-side ensemble/routing	一つの単体modelではなく複数backend/agent logic	parameter/architecture fingerprint推定が根本的に不安定	**証拠不足だが排除不能**

仮にシナリオAが正しく、full DeepSWEでもOxが80%近辺を維持した場合はインパクトが大きい。現在の公式DeepSWE最高はOpus 5 74±4%、GPT-5.6 Sol 73±3%であるため、同一113 tasks・同一harnessで80%を統計的に再現できれば、単なる「安価な中国モデル」ではなくcoding-agent frontierそのものの順位変動になる。 [7]

しかし現時点ではそこまでの証拠はない。むしろArtificial AnalysisではGPT-5.6 Solが複数coding benchmarksを統合したCoding Agent Indexで首位にあり、**公開された広い証拠集合は「GPT-5.6が依然frontier、Fable/Opus/GLMが接近、Oxはまだ評価待ち」という状況を示す。** [45]

最終判定

指定記事の事実部分：概ね妥当。
Ox Alphaの存在と主要公開仕様は一次資料で裏付けられる。**信頼度：95〜98%。**

「GLM系ではないか」という技術仮説：かなり興味深く、合理性がある。
同tokenizer/video behaviorという報告、GLM-5Vの公開architecture、GLM-5.2の1M context、GLM-5.3のcoding focus、Pony Alphaという智譜自身のstealth-model前例、IPO roadmapが同じ方向を指している。ただしfingerprint生データが未公開で、公式確認もない。**GLM技術系統説：65〜75%。**

「智譜の未発表"正式フラッグシップ"」との断定：まだ早い。
智譜製だったとしても、A/B checkpoint、GLM-5V後継、内部evaluation model、正式版候補などがあり、「次のflagship product」という部分は追加推論である。**信頼度：45〜60%。**

「GPT-5.6とClaudeよりプログラミングが強い」という記事見出し：現状では立証されていない。
8/10は小さすぎ、比較条件も揃っていない。公式DeepSWEの113問ではGPT-5.6 Sol 73%、Fable 70%、GLM-5.3 69%で、Oxは未掲載。独立Coding Agent IndexではGPT-5.6が首位。**一般的優位性への信頼度：20〜30%。** 46

追跡調査の優先順位

最優先は、**OpenRouterのprovider identity変更、Ox Alphaページの名称変更・正式モデルへのredirect、model card更新**である。過去のPony Alphaが最終的にGLM-5と公式確認された前例があるため、これは最も情報価値が高い。[^G5P] 47

次に重要なのは、**智譜公式ブログ、`zai-org` GitHub、Hugging Face/OpenRouter catalogでの新model config/tokenizer/weights公開**である。特に1M contextとvideo inputを同時に持つ新GLMが出れば、Oxとの対応関係を直接比較できる。[^G52][^G5V] 48

三番目は、**Davisがtokenizer/video fingerprintのraw prompts、usage counts、API request logsを公開するか**である。現在の「99%」を客観的証拠へ昇格させる最短経路はこれである。 2

四番目は、**DeepSWE全113問でOxを同一mini-swe-agent条件に載せること**。これが行われれば「GPT-5.6/Claude超え」という記事の主要performance claimを直接検証できる。 7

五番目は、**新規private coding benchmarkによるcontamination-resistant test**。Oxが公開DeepSWEで強くても、8月以降に作った完全held-out taskで同じ順位を維持するかが、商業的に最も重要である。

したがって現時点で最も厳密な表現は、次のようになる。

> **Ox Alphaは、公開仕様とblack-box fingerprintが智譜GLM系列、とりわけ「GLM-5.2/5.3の1M・coding stack」と「GLM-5V-Turboのvideo/multimodal stack」を統合した次世代モデルを強く連想させる匿名モデルである。ただし智譜製であるとの公式確認はなく、「GPT-5.6やClaudeを上回った」という評価は10問の小規模試験から一般化されたもので、現在の公開113問DeepSWEおよび第三者coding leaderboardでは未立証である。**

この二つを混同しないことが、指定記事を読む際の最重要ポイントである。

出典注記

[^BG1]: BigGo Finance「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…」
`https://finance.biggo.jp/news/9dc856ba-634d-467a-bea2-6ba70233113c@Deep Research` 6

[^OR1]: OpenRouter, “Ox Alpha”
`https://openrouter.ai/stealth/ox-alpha` 9

[^BD1]: Ben Davis, X post
`https://x.com/davis7/status/2090669483740279155` 2

[^DS1]: DataCurve, DeepSWE leaderboard
`https://deepswe.datacurve.ai/` 7

[^DS2]: DeepSWE methodology / reproduction documentation. DeepSWEサイトから公式run instructionsを参照。 49

[^GLM0]: THUDM, GLM original repository / ACL 2022 research lineage. 50

[^G5]: Zhipu AI & Tsinghua, “GLM-5: from Vibe Coding to Agentic Engineering”
`https://arxiv.org/html/2602.15763v2` 14

[^G5P]: 同GLM-5 Technical Report。Pony AlphaをGLM-5の匿名OpenRouter evaluationだったと公式に開示。 51

[^G52]: Z.ai official GLM-5 GitHub repository
`https://github.com/zai-org/GLM-5` 15

[^G5V]: Z.ai Developer Documentation, GLM-5V-Turbo
`https://docs.z.ai/guides/vlm/glm-5v-turbo` 16

[^G5VP]: GLM-5V-Turbo research paper
`https://arxiv.org/abs/2604.26752` 52

[^IPO]: Knowledge Atlas Technology / 北京智譜華章科技股份有限公司, Hong Kong IPO Prospectus, 2025-12-30
`https://www1.hkexnews.hk/listedco/listconews/sehk/2025/1230/2025123000017.pdf` 5

[^IPORD]: 同目論見書、R&D支出・compute investment。 18

[^IPOPT]: 同目論見書、Intellectual Property section。 19

[^PAT1]: Google Patents, CN118153551B
`https://patents.google.com/patent/CN118153551B/en` 20

[^OAI]: OpenAI Developers, GPT-5.6 Sol model documentation
`https://developers.openai.com/api/docs/models/gpt-5.6-sol` 27

[^ANT]: Anthropic, Claude Fable 5 / Mythos 5 announcement and model documentation
`https://www.anthropic.com/news/claude-fable-5-mythos-5`
`https://docs.anthropic.com/en/docs/about-claude/models/overview` 53

[^AA1]: Artificial Analysis, “GPT-5.6 benchmarks across Intelligence, Speed and Cost”
`https://artificialanalysis.ai/articles/gpt-5-6-has-landed` 8

navlist関連する直近報道turn0news10turn0news12

---

1 9 17 26 34 44 Ox Alpha - API Pricing & Providers
https://openrouter.ai/stealth/ox-alpha?utm_source=chatgpt.com

2 43 99% sure it's GLM-5.x, all the evidence points to it (same video ...
https://x.com/davis7/status/2090669483740279155?utm_source=chatgpt.com

3 4 13 14 32 33 47 51 GLM-5: from Vibe Coding to Agentic Engineering
https://arxiv.org/html/2602.15763v2

5 18 19 23 35 printmgr file
https://www1.hkexnews.hk/listedco/listconews/sehk/2025/1230/2025123000017.pdf?ref=runtimewire

6 41 42 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 ― BigGo ファイナンス
https://finance.biggo.jp/news/9dc856ba-634d-467a-bea2-6ba70233113c

7 25 31 37 39 46 DeepSWE
https://deepswe.datacurve.ai/

8 45 GPT-5.6 benchmarks across Intelligence, Speed and Cost
https://artificialanalysis.ai/articles/gpt-5-6-has-landed?utm_source=chatgpt.com

10 神秘AI模型编程测试超越GPT-5.6和Claude，技术特征指向智谱 ... - 新浪
https://k.sina.com.cn/article_5953189932_162d6782c06704vtv0.html?from=tech&utm_source=chatgpt.com

11 台湾発の価格比較サイト「BigGo（ビッグゴー）」を運営する ...
https://prtimes.jp/main/html/rd/p/000000003.000029376.html?utm_source=chatgpt.com

12 50 THUDM/GLM: GLM (General Language Model) - GitHub
https://github.com/THUDM/GLM?utm_source=chatgpt.com

15 21 48 GitHub - zai-org/GLM-5: GLM-5: From Vibe Coding to Agentic Engineering · GitHub
https://github.com/zai-org/GLM-5

16 24 GLM-5V-Turbo - Overview - Z.AI DEVELOPER DOCUMENT
https://docs.z.ai/guides/vlm/glm-5v-turbo

20 CN118153551B - A method, device, equipment and medium for ...
https://patents.google.com/patent/CN118153551B/en?utm_source=chatgpt.com

22 52 GLM-5V-Turbo: Toward a Native Foundation Model for Multimodal Agents
https://arxiv.org/abs/2604.26752?utm_source=chatgpt.com

27 40 GPT-5.6 Sol Model | OpenAI API
https://developers.openai.com/api/docs/models/gpt-5.6-sol

28 53 Claude Fable 5 and Claude Mythos 5
https://www.anthropic.com/news/claude-fable-5-mythos-5?utm_source=chatgpt.com

29 Previewing GPT-5.6 Sol: a next-generation model | OpenAI
https://openai.com/index/previewing-gpt-5-6-sol/?utm_source=chatgpt.com

30 China's Z.ai says new model nears Anthropic's Mythos 5 in cyber-defence tests
https://www.reuters.com/technology/chinas-zai-says-new-model-nears-anthropics-mythos-5-cyber-defence-tests-2026-08-14/?utm_source=chatgpt.com

36 So apparently Ox Alpha is NOT a GLM model, not MiMo and also not ...
https://x.com/LexnLin/status/2090760150923497486?utm_source=chatgpt.com

38 49 DeepSWE - Measuring frontier coding agents
https://deepswe.datacurve.ai/blog/deepswe