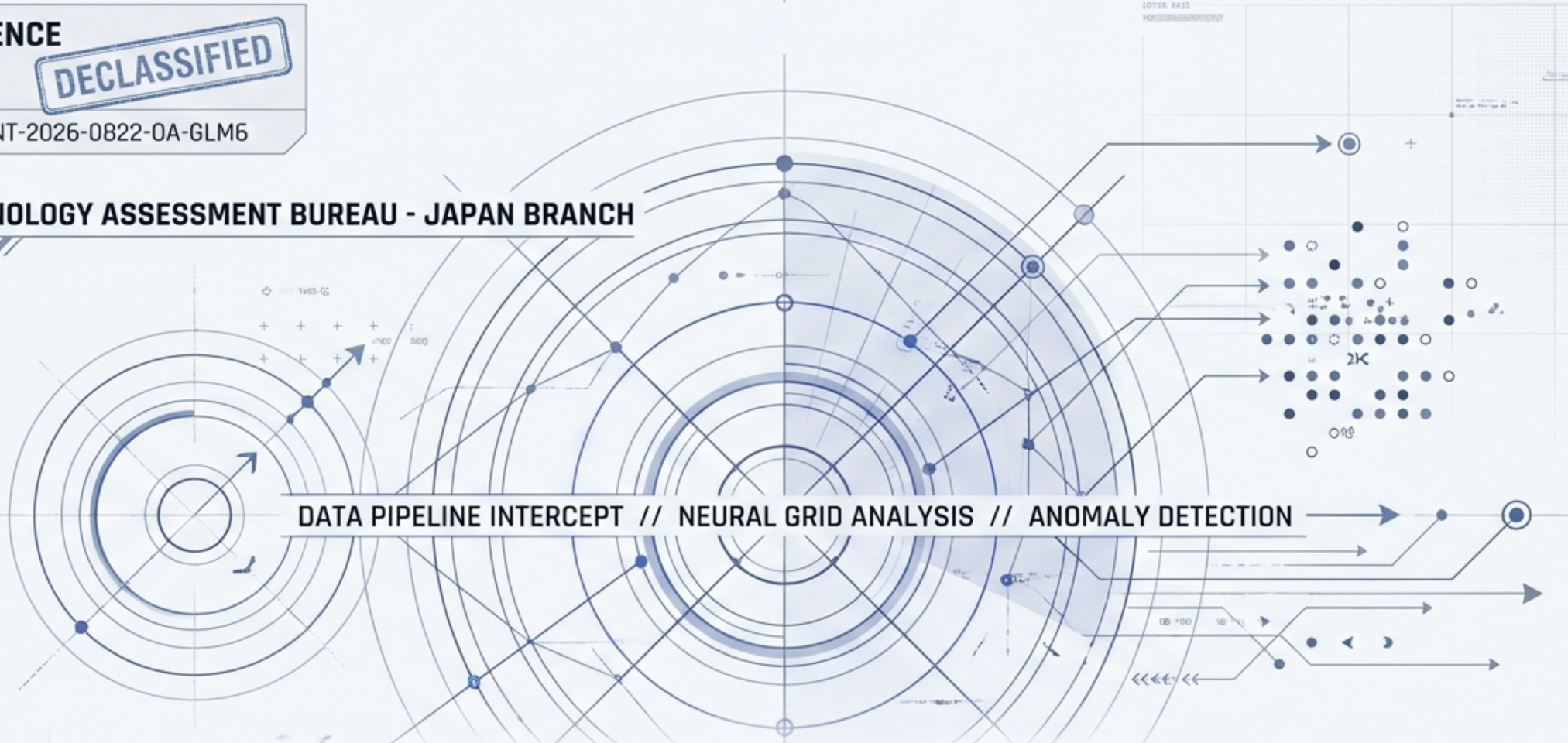
INTELLIGENCE DOSSIER

DECLASSIFIED

DOC NO: AI-INT-2026-0822-OA-GLM6

AI TECHNOLOGY ASSESSMENT BUREAU - JAPAN BRANCH



匿名AI「Ox Alpha」の正体と実力

Zhipu次期旗艦モデル「GLM-6」説とGPT-5.6超えの真相に迫る徹底調査レポート

DATE:	2026-08-22 (JST)
STATUS:	INTELLIGENCE DOSSIER / DECLASSIFIED

The Verdict：2つの仮説を切り離す

98%

Confirmed

「Ox Alpha」は実在し、OpenRouter上で1Mコンテキストのマルチモーダルモデルとして稼働している。

65-75%

High Probability

トークナイザーや動画処理の挙動、そして既存技術との整合性から、智譜の未発表モデル（またはその派生）である可能性は極めて高い。

20-30%

Low / Unverified

10問の極小サンプルに依存した主張であり、統計的優位性は全く立証されていない。

What We Know：確認済みのファクト

Context Window	1,048,576 tokens
Max Output	131,072 tokens
Input Modalities	Text, Image, Video
Core Capabilities	Tool Calling, Reasoning, Structured Output
Current Cost	無料（Free preview, 配信条件で変動）
Observed Performance	初期レイテンシ 5.3s / スループット 24 tok/s (P50値)

⚠

OpenRouterは「自社は開発者ではない」と明記しており、プロンプトデータはプロバイダ側に保持されるが「学習には使用しない」としている。

The Fingerprint：開発元を示唆する「指紋」

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

Tokenizer
GLM-5.x系統と同一の挙動

Video Encoder
動画トークン処理が
GLM-5V-Turboと一致

Audio Rejection
音声入力の拒否パターンが
一致

「Pony Alphaの前例」 GLM-5公開時、智譜はOpenRouter上で匿名モデル「Pony Alpha」としてステルス試験を行っていた公式前例があり、この行動パターンは企業文化と完全に一致する。

Architectural Convergence：既存技術の自然な融合

GLM-5.2 / 5.3
1M Context (IndexShare)
Agentic RL
Coding特化

＋

GLM-5V-Turbo
CogViT Vision Encoder
Video / Image Input
30+ Task Joint RL

→

Ox Alpha
「1M長文・コーディング能力」と「動画・マルチモーダル能力」を兼ね備えた統合モデル。

Ox Alphaは全く未知のアーキテクチャではなく、智譜が既に個別に公開している先端スタックを見事に統合したチェックポイントとして説明がつく。

Strategic Alignment：上場目論見書との符号

【香港IPO目論見書】
2026-2028年のR&D計画として「Infinity context」「Deep reasoning」「AI agent」への巨額投資を明記。

2025年12月30日

【GLM-5 論文】
744B / 40B Active, 28.5T tokensでの学習を公開。

2026年2月

【GLM-5V-Turbo 公開】
マルチモーダル基盤の確立。

2026年4月

【Ox Alpha 出現】
目論見書の「1M+ Context」「Test-time compute」「Agent」をすべて体現するモデルがステルス公開。

2026年8月20日

Oxの登場は突然変異ではない。智譜が8ヶ月前に公言したロードマップの延長線上に完璧に位置している。

Origin Conclusion：3つの商業的シナリオ

Scenario A
次期正式旗艦モデル (GLM-6)

- Impact: 中国系モデルがAPI価格競争とコーディングエージェント市場に破壊的影響をもたらす。
- Probability: 中（45-60%）。「製品版」と断定するには証拠不足。

MOST PROBABLE

Scenario B
実験用チェックポイント (評価専用)

- Impact: 技術指紋はGLMだが、商用名や最終ウェイト構成とは異なるステルステスト版。
- Probability: 高い。技術方向性は一致するが、見出しの「フラッグシップ」は過剰。

Scenario C
他社による模倣/最適化版

- Impact: 既知のGLMトークナイザ等を流用した別モデル。ベンチマーク過適合の懸念。
- Probability: 低いが排除不可。

The Claim vs. Reality：コーディング性能の真相

The Media Claim

メディアの主張

「Ox Alpha、プログラミングでGPT-5.6とClaudeを凌駕（DeepSWE 8/10成功）」

- **Flaw:** メディア記事による未統制の小規模テスト。

The Reality

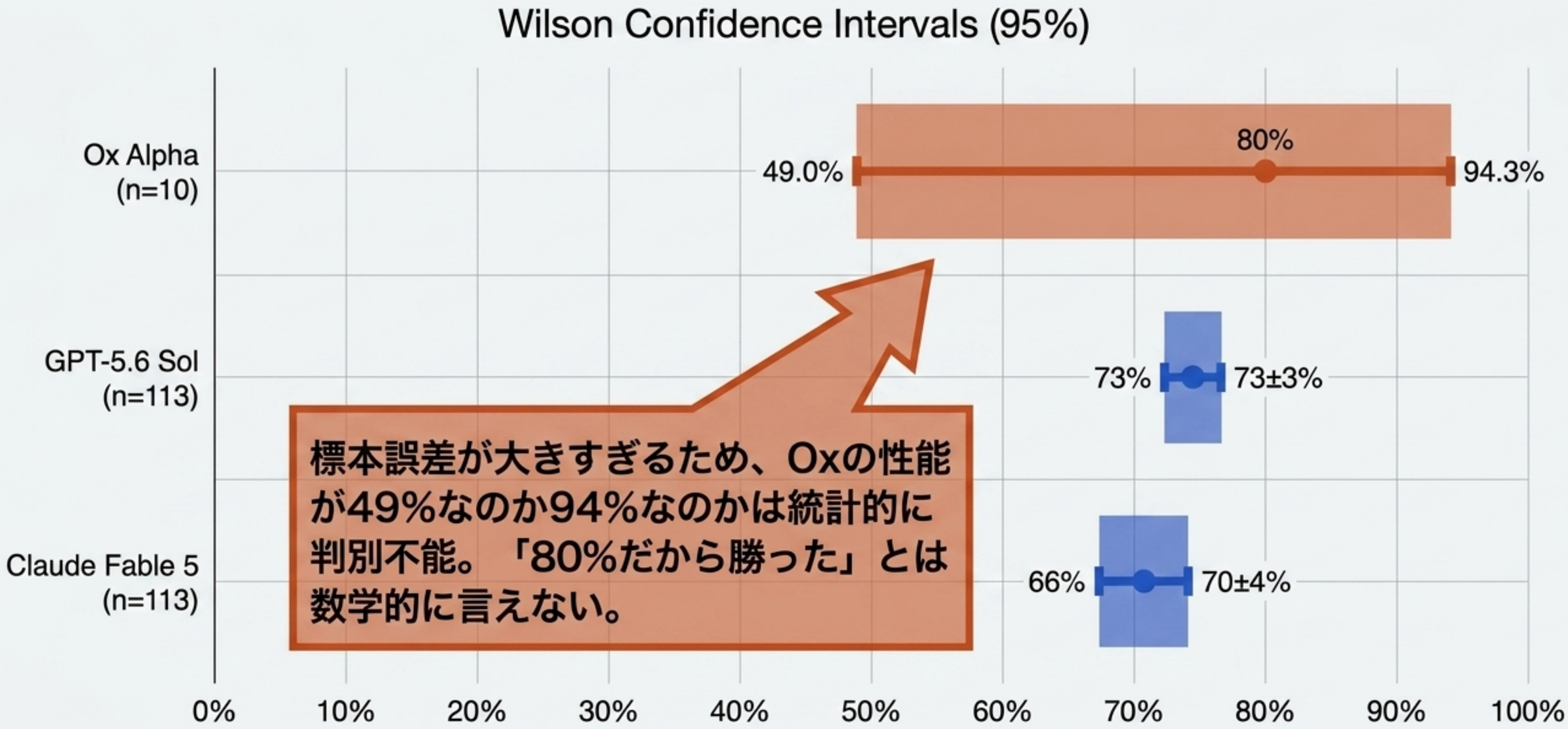
冷徹な事実

公式の「DeepSWE v1.1 113タスク・リーダーボード」にOx Alphaは未掲載。

Current Standings (113 Tasks):	
1.	Claude Opus 5: 74±4%
2.	GPT-5.6 Sol: 73±3%
3.	Claude Fable 5: 70±4%
4.	GLM-5.3: 69±3%

記事の比較対象（GPT-5.6 52%等）は同条件のヘッド・トゥ・ヘッドではなく、一般化は不当である。

The Statistical Trap：小規模サンプルの罠



The Ultimate Comparison：フロンティアモデル比較

Features	0x Alpha	GPT-5.6 Sol	Claude Fable 5
Context	1M	1.05M	1M
DeepSWE Score	80% (n=10)	73±3%	70±4%
API Cost ($/MTok In/Out)	現在無料	$4 / $20	$10 / $50
Parameters	*未指定*	*未指定*	*未指定*
Safety & Control	*未指定*	Layered safeguards	Mythos-level Cyber/Bio eval

第三者評価では依然としてGPT-5.6 SolがCoding Agent Index首位であり、現時点で王座は揺らいでいない。

The Verification Protocol：真実を暴くための推奨テスト

Step 1: 統制下でのフルベンチマーク (Public Layer)

DeepSWE全113タスクを同一日・同一条件（mini-swe-agent）で実行。

単なるPass率だけでなく、Cost/TaskやWall-clock timeも測定。

CRITICAL

Step 2: プライベート・ホールドアウト (Private Layer)

8月以降に新規作成した50〜100の非公開タスクを実行。

Why?: Ox Alphaの学習カットオフが不明なため、公開済みのDeepSWEタスクへの「過適合（Contamination）」を排除するため。

「モデル帰属」の証明には、動画解像度や長さを変化させた際のトークン消費量（Scaling curve）の生データ公開が不可欠。

The Final Takeaway

Ox Alphaは、Zhipuの次世代技術（1M長文＋マルチモーダル）を統合した強力なGLM系モデルのである可能性が極めて高い。

> しかし、『GPT-5.6やClaudeを超えた』と宣言するには、現時点の証拠はあまりにも無防備で統計的に不十分である。

Actionable Advice: 経営層・開発者は、API価格競争のゲームチェンジャーの到来に備えつつも、独立した113タスク評価と新規ホールドアウトテストの結果が出るまでは、既存のフロンティアモデル（GPT-5.6 / Claude）を基準とすべきである。