GLM-5.3-Flashの技術構造と市場インパクト：320Bネイティブ・マルチモーダル MoEの総合分析

Gemini 3.7 Flash

2026年8月26日、中国のAI研究機関・企業である智譜AI（Z.ai / Zhipu AI）は、GLM-5シリーズ初となるネイティブ・マルチモーダル基盤モデル「GLM-5.3-Flash」を正式発表し、MITライセンスの下でオープンソース公開した[1]。

本モデルは、正式発表に先駆けて「Ox Alpha（中文コミュニティ呼称：牛来）」というコードネームでOpenRouterやOpenCodeなどの開発者向けプラットフォーム上で匿名プレビュー提供され、驚異的な推論速度と高いコーディング精度によりAIコミュニティで大きな反響を呼んでいた[1]。総パラメータ数320B（3,200億）、1トークンあたりのアクティブパラメータ数18B（180億）という高度な疎性（Sparsity）を備えたMixture-of-Experts（MoE）構造を持ち、アテンション機構の刷新、100万トークン（1M）の長文脈対応、中国国産AIアクセラレータクラスタによる高効率な運用実績を伴って登場した[1]。

本稿では、GLM-5.3-Flashのアーキテクチャ特性、ベンチマーク性能、インフラ最適化技術、経済性、そしてグローバルなAIエコシステムに与えた多面的な反響と戦略的意義について包括的に分析する。

1. 匿名プレビュー「Ox Alpha」の台頭と正式リリースの経緯

1.1 ステルステストにおけるバイラルな反響

GLM-5.3-Flashの公開に先立ち、智譜AIは2026年8月20日頃より、開発者向けアグリゲーション基盤であるOpenRouterおよびOpenCodeにおいて「Ox Alpha（stealth/ox-alpha）」と名付けられた提供元不明のモデルを無料公開した[7]。

このモデルは、事前の予告やモデルカードの公開なしに投入されたにもかかわらず、複雑なプログラミング、推論、エージェントタスクにおいて突出した応答品質とスループットを示した[7]。コミュニティによる小規模なDeepSWE検証では約80%のファーストパス成功率が報告され、Claude Fable 5（65%）やGPT-5.6 Sol（52%）といった主要な商用フロンティアモデルを上回る挙動を見せたことから、Redditのr/LocalLLaMAやX上で瞬く間に議論が沸騰した[7]。OpenRouterを傘下に収めるStripeのCEOであるPatrick Collisonも本モデルの挙動を「非常に印象的（very impressive）」と評し、国際的な注目度は一気に高まった[7]。

プレビュー運用指標	記録・仕様
提供期間（匿名テスト）	2026年8月20日～8月26日（約6日間）[1]
処理トークン総量	23兆トークン以上（1日あたり数兆～100兆トークン規模の要求を処理）[1]

プラットフォーム反響	OpenRouterおよびOpenCodeで週間呼び出し量首位を記録（DeepSeekの2倍以上）[1]
提供形態	1Mコンテキスト、マルチモーダル入力、完全無償アクセス[7]

1.2 フィンガープリント特定と正式公開

コミュニティによるトークナイザー構造の解析や出力トークン数のオフセット挙動の検証により、8月14日にテキスト特化型フラッグシップとして先行発表されていた「GLM-5.3」との一致が確認され、智譜AIの開発モデルであることが確実視されるに至った[7]。そして2026年8月26日、智譜AIは「Ox Alpha」の正体がGLM-5.3-Flashであることを公式に認め、重みデータの一般公開および商用APIの提供を開始した[1]。

2. アーキテクチャの革新と推論効率化のメカニズム

GLM-5.3-Flashは、前世代のGLM-5.2の単純な後継ではなく、モデル知能の向上と推論計算コストの極小化を両立させるためにゼロから再構築された基盤モデルである[17]。入力されたテキスト、画像、動画データは、多様体制約付きハイパーコネクション（mHC）によって安定化された多層構造へと導かれ、局所的な状態を追跡する線形アテンションと大域的な検索を担うスパースアテンションが協調動作するハイブリッド機構を経て、320B-A18Bの疎なMoEフィードフォワードネットワークによって高速に処理される[8]。

2.1 320B-A18B MoE設計とレイヤー削減

GLM-5.3-Flashは、総パラメータ数3,200億（320B）に対して、各トークン処理時に稼働するアクティブパラメータ数を180億（18B）に抑制している[1]。先行するGLM-4.5系列と比較した場合、総パラメータ数は同等水準（355Bから320B）を維持しながら、アクティブパラメータ数は約半分（32Bから18B）に削減され、トランスフォーマーレイヤー数も92層から45層へと半減された[9]。この大幅な薄型化・疎構造化が、モデル全体の容量を保ちつつ1トークンあたりの計算パスを細く絞り込み、超高速な推論実行を可能にする基盤となっている[20]。

2.2 ハイブリッドアテンションとIndexPool機構

長文脈（100万トークン）の処理において、従来のアテンション機構が抱えていた計算量およびKVキャッシュメモリの爆発的増加を抑制するため、本モデルはフロンティアオープンモデルとして初めて線形アテンション（Linear Attention / KDA）とスパースアテンション（Sparse Attention / MLA）のハイブリッド構成を採用した[1]。

線形アテンションは状態空間モデリングを通じて近傍の局所的依存性を極めて低い計算量で追跡し、スパースアテンションは軽量なインデクサーを用いて広範な文脈から必要なトークンを的確に抽出する[18]。さらに、スパースアテンションのインデクサーにおける4つのキャッシュキーベクトルを単一のベクトルに圧縮集約する「IndexPool」機構を導入した[17]。これにより、1Mトークンの超長文脈稼働時におけるインデクサーのメモリオーバーヘッドと検索レイテンシが大幅に圧縮された[18]。

この統合アーキテクチャにより、フルサイズのGLM-5.3と比較してアテンション計算量を3.01倍削減し、KVキャッシュサイズを4.44倍削減することに成功している[1]。

2.3 多様体制約付きハイパーコネクション（mHC）と学習安定化

超深層・大規模ネットワークの学習安定性とスケーリング効率を高めるため、残差接続の進化形で

ある「Manifold-Constrained Hyper-Connections（mHC）」が導入された[17]。従来のハイパーコネクションで問題視されていた勾配の爆発や学習の不安定性を、残差接続の混合を特定の多様体に射影して恒等写像の性質を保持することで克服し、モデルの大規模化と推論時の計算量削減を両立させている[20]。

2.4 ネイティブ・マルチモーダル統合と常時思考モード

GLM-5.3-Flashは、30兆トークンに及ぶマルチモーダル事前学習コーパスを用いて訓練された、GLM-5系列初のネイティブ・マルチモーダルモデルである[1]。テキストに加えて静止画像および長尺動画をネイティブに受け入れ、128Kトークンまでの長文出力をサポートする[8]。また、推論の質を動的に高める思考モード（Thinking Mode）が常時有効化（Always-On）されており、タスクの複雑性に応じて思考深度（low / high / max）を切り替えて高度な論理展開を実行する[17]。

3. インフラストラクチャと国産アクセラレータクラスタの最適化

GLM-5.3-Flashのリリースにおいて技術的・産業的に最も注目された要素の一つが、プレビューおよび本番提供におけるすべての推論トラフィックが中国国産AIチップクラスタ上で完全運用された点である[1]。

外部からの要求トラフィックは、多模態エンコード、プロンプトのプリフィル計算、トークンごとのデコード処理を独立したワーカープールに分離するEPD（Encode-Prefill-Decode）分離アーキテクチャによって受け止められ、自社開発の高帯域インターコネクトを介して10万基規模の国産チップクラスタ上で分散実行される[6]。

3.1 EPD分離アーキテクチャによる負荷分散

国産チップクラスタにおいて、限られたメモリ帯域と容量で100万トークンのマルチモーダル負荷を処理するため、本番環境レベルのEPD分離アーキテクチャが採用された[6]。画像や動画データの多モーダル埋め込みベクトルを生成するEncode Pool、巨大な入力プロンプトのコンテキスト計算を並列実行するPrefill Pool、トークンごとの逐次生成を集中的に行うDecode Poolに処理を分離し、各ステージを個別にスケールアウトさせることで、リソース競合を防ぎクラスタ全体の稼働率を最大化している[6]。

3.2 SGLang推論エンジンとInfra Agentによる自動協調

智譜AIは、オープンソースの高速推論フレームワーク「SGLang」をベースに、国産ハードウェア向けの専用エンジンを構築した[1]。特筆すべき点として、推論スタックのカスタムカーネル開発、ボトルネック診断、オペレータ最適化において、先行フラッグシップであるGLM-5.3駆動のインフラエージェント（Infra Agent）が投入された[27]。

ノード内テンソル並列化、ReplaySSM、W8A8量子化、INT8/FP8/BF16の混合キャッシュ量子化、レイヤースプリット（Layer Split）技術などを組み合わせることで、国産ハードウェアにおける初期ベースラインと比較してエンドツーエンドのサービス性能を3倍向上させ、単一トークンあたりの処理コストを主流のNVIDIA製GPUと同等水準まで引き下げることに成功した[1]。

4. ベンチマーク評価と実務応用性能

GLM-5.3-Flashは、軽量・高効率なFlashクラスのコスト構造でありながら、主要なフロンティアモデルに匹敵する知能スコアを記録している[1]。

4.1 知能指数およびコーディング・エージェント性能

独立系評価機関であるArtificial Analysisの総合知能指数（Intelligence Index v4.1.1）において、

GLM-5.3-Flashは57ポイントを獲得し、Anthropicの最高峰モデル「Claude Opus 4.8」と並び、DeepSeek V4 Pro（53ポイント）を上回った[1]。

評価指標 / ベンチマーク	GLM-5.3-Flash	GLM-5.2	Claude Opus 4.8	評価内容・特性
Artificial Analysis Index	**57** [cite: 1, 2, 28]	< 50（前世代）[1]	57[1]	総合知能指数[1]
GDPval-AA v2	**1773** [cite: 18, 30]	1504[18]	1675[18]	実務作業能力のEloレーティング[18]
DeepSWE v1.1	**63.4** [cite: 18, 32]	46.2[18]	競合水準[18]	GitHub実課題解決ベンチマーク[7]
AutomationBench	**48.8** [cite: 18, 32]	26.2[18]	競合水準[18]	長期エージェント自律タスク[18]
Terminalbench h 2.1	**84.3** [cite: 32]	-	競合水準[32]	CLI環境・ターミナル自律実行[32]
HLE (Humanity's Last Exam)	**55.3** [cite: 32]	-	競合水準[32]	超高難度推論ベンチマーク[32]
Z.ai Code Bench (Max Effort)	**29.0** [cite: 17, 18]	低水準[18]	29.5[17]	Claude Code環境での実コーディング体感[17]
BabyVision	**53.4** [cite: 18, 30]	35.1[18]	46.8（GPT-5.6: 61.6）[18]	（参考：Gemini 3.7 Flashは70.9）[18]
MVBench (動画理解)	**77.8** [cite: 18]	69.4[18]	-	長尺マルチモーダル動画ベンチマーク[18]

4.2 視覚主導型フロントエンド生成とドキュメント・動画解析

GLM-5.3-Flashの実用上の強みは、視覚フィードバックを伴う自律的な開発サイクルにある[9]。Web

デザインのスクリーンショットやUI遷移の画面録画データを入力すると、デザインシステムやコンポーネント構造、アニメーションロジックを解釈した上で、Next.jsおよびTypeScriptを用いたプロジェクトコードを自律生成する[17]。生成後はレンダリング画面をモデル自身がスクリーンショットで再取得し、タイポグラフィや余白、レイアウトの差異を反復修正する自己検証ループが機能する[9]。
さらに、OCR処理を挟むことなく、PDF、複雑なスプレッドシート（XLSX）、プレゼンテーション（PPTX）をダイレクトに画像として解析し、財務分析や文書作成を完遂する[17]。48分間の英語講演動画から57分間で17箇所の重要発言を抽出し、字幕付きの動画クリップとして編集出力した実証例も報告されている[25]。
一方で、幾何認識や一般物体知覚を問う「BabyVision」ベンチマークでは53.4にとどまり、Gemini 3.7 Flash（70.9）に対して明確な差をつけられている[18]。この結果は、GLM-5.3-Flashが一般的な視覚知覚全般よりも、UI/UX設計、ドキュメント解析、ソフトウェア環境の視覚的フィードバック制御に特化して最適化されていることを示している[17]。

5. 価格構造とオープンソースエコシステムへの波及

GLM-5.3-Flashの公開は、性能面のみならず、API価格とライセンス形態の両面で市場に強いインパクトを与えた[1]。

5.1 破壊的価格体系と経済的インパクト

智譜AI公式APIにおける価格設定は、同等の知能水準を持つ商用フラッグシップモデルと比較して極めて低廉に設定されている[1]。

モデル名	入力価格 (100万トークン)	キャッシュ入力価格	出力価格 (100万トークン)	価格比率（対 Opus 4.8）
Claude Opus 4.8	$5.00（約800円）[29]	提供形態による	$25.00（約4,000円）[29]	基準 **(1/1)** [cite: 1, 29]
GLM-5.3 (Text Only)	$1.50 / 8.0元[28]	2.3元[28]	$2.50 / 28.0元[28]	約 1/4 ～ 1/10[1]
GLM-5.3-Flash (通常価格)	**$0.15 / 0.8元** [cite: 3, 28, 33]	**$0.03 / 0.23元** [cite: 3, 28]	**$0.25 / 2.8元** [cite: 3, 28, 33]	約 **1/40 (2.5%)** [cite: 1, 5, 29]
GLM-5.3-Flash (公開記念半額)	**$0.075 / 0.4元** [cite: 26, 33, 34]	**0.115元** [cite: 33]	**$0.125 / 1.4元** [cite: 33]	約 **1/80 (1.25%)** [cite: 5, 29, 35]

公開から2週間のプロモーション期間中は半額が適用され、入力100万トークンあたり0.4元（約0.075）、出力1.4元（約0.125）で提供された[3]。これはClaude Opus 4.8の40分の1から80分の1の価格であり、DeepSeekなどの競合企業が牽引してきた価格競争をさらに一段階押し進める結果となった[5]。

5.2 重み公開とローカル推論環境の要件

モデルの重みはHugging FaceやOllama、Unsloth等でMITライセンスの下で無償公開され、商用利用を含めた自由な改変・再配布が認められている[1]。ネイティブFP8形式の重み（約328GB）に加え、コミュニティ主導でUnsloth Dynamic GGUF（UD-Q4_K_XL、UD-Q2_K_XLなど）が提供されている[21]。フル精度の運用には大規模クラスタを要するものの、4-bitや2-bit量子化を適用することで、ワークステーション規模（NVIDIA Sparkクラスタ やMac Studio等の統合メモリ環境）でのローカル推論・エージェント運用の道が開かれている[37]。

6. 市場の評価と戦略的・地政学的インプリケーション

6.1 開発者コミュニティの反響と技術的論点

開発者コミュニティにおける評価は、コーディングおよびエージェントタスクにおける実用性の高さに集中している[7]。デバッグ精度の高さ、簡潔なコード生成、Claude CodeやClineなどのエージェントツールとの親和性が称賛されている[7]。また、320Bという総パラメータ規模を持ちながら、18Bのアクティブパラメータとハイブリッドアテンションによって実現された高速なトークン生成速度（TTFTおよびスループット）が高く評価されている[8]。

一方で、約328GBに達するモデルサイズは単一のコンシューマ向けGPU（RTX 4090等）でのローカル展開を困難にしており、さらなる軽量量子化手法や投機的サンプリング（Speculative Decoding）の整備を求める議論も活発に行われている[37]。

6.2 競合環境におけるポジショニングと産業構造への示唆

GLM-5.3-Flashは、AI市場におけるポジショニングとして、Claude Opus 4.8などの最高峰クローズドモデルに匹敵する知能指数を、極限まで抑えられたFlashモデルの価格帯とオープンウェイト形式で提供するという独自のポジションを確立している[1]。DeepSeekがV4シリーズの価格改定を実施した局面において、同等クラスのパラメータ規模（320B-A18B）を持ちながらマルチモーダル入力をネイティブ統合したことで、オープンソース陣営における有力な選択肢として認知された[14]。

また、本モデルの成功は、半導体供給制約下におけるAIインフラのあり方に重要な示唆を与えている[6]。ハードウェア単体のメモリ帯域や処理能力の制約を、ハイブリッドアテンション（IndexPool）、mHC、EPD分離アーキテクチャ、推論エージェントによるカーネル最適化といったソフトウェア・アルゴリズム工学の統合によって相殺できることを実証した[1]。23兆トークンを超えるトラフィックを全量国産チップで処理し、NVIDIA製GPUと同等の単一トークンコストを達成した実績は、大規模AIサービスの基盤が特定ベンダーのハードウェアのみに依存しない新たなフェーズへ移行しつつあることを示している[1]。

7. 結論

智譜AIによるGLM-5.3-Flash（320B-A18B）の公開は、オープンソースAIの新たな進化段階を明確に示した[1]。

本モデルは、「Ox Alpha」としての匿名テストを通じて実証された高度なコーディング・エージェント実行能力を備え、Claude Opus 4.8級の知能を40分の1の価格水準で提供する[1]。技術的には、スパース＆リニアアテンションのハイブリッド統合、IndexPoolによるKVキャッシュの4.44倍圧縮、mHCによるスケーリング安定化など、極限の推論効率を追求したアーキテクチャが結実している[1]。さらに、これらの処理が中国国産アクセラレータクラスタ上で完全に実証されたことは、グローバルなAIインフ

ラ競争および半導体エコシステム全体に対しても極めて深い地政学的・技術的示唆を与えている[6]。

引用文献

1. 智譜（Zhipu）がGLM-5.3-Flashをオープンソース化：中国製チップクラスターで稼働、価格はClaude Opus 4.8の40分の1, https://finance.biggo.jp/news/8ac68e37-02c9-4e8e-bc5a-a6f9e696b5e3
2. GLM-5.3-Flash: Frontier Intelligence, Flash Cost - Z.ai, https://z.ai/blog/glm-5.3-flash
3. 謎の高性能AI「Ox Alpha」の正体は「GLM-5.3-Flash」だったことが判明、Claude Opus 4.8と同等性能のオープンモデル＆中国製チップで運用可能, https://gigazine.net/news/20260827-glm-5-3-flash/
4. Ox Alpha现身：智谱开源GLM-5.3-Flash原生多模态模型, https://h5.ifeng.com/c/vivoArticle/v0020X1h7CS--v9TE88fG3hjOLBVvcdYddfSc0ZiCy--6n--Zc__?isNews=1&showComments=0
5. 智谱上线并开源GLM-5.3-Flash，系此前Ox-Alpha（中文社区称作牛来）模型, https://finance.eastmoney.com/a/202608263855190478.html
6. 智谱：GLM-5.3-Flash跑在国产芯片上, https://finance.ifeng.com/c/8vvl5CyfFKU
7. Mystery Ox Alpha Debuts on OpenRouter — Free, Multimodal LLM, https://app.chaingpt.org/ai-web3-news/50042/mystery-ox-alpha-debuts-on-openrouter-free-multimodal-llm-outperforms-top-coding-models
8. Opus 4.8級を手元に！320Bのモデル「GLM-5.3-Flash」無償公開, https://pc.watch.impress.co.jp/docs/news/2136012.html
9. Ox Alpha揭晓：智谱开源GLM-5.3-Flash，跑在国产芯片上, https://www.163.com/tech/article/L5B02LT100097U7T.html
10. Ox Alpha mystery solved, powerful AI model on OpenRouter is made by this Chinese company, https://www.indiatoday.in/technology/news/story/ox-alpha-mystery-solved-powerful-ai-model-on-openrouter-is-made-by-this-chinese-company-2980398-2026-08-26
11. How an Anonymous AI Model Became Z.AI's GLM Reveal - Atoms, https://atoms.dev/blog/ox-alpha-anonymous-ai-model-opencode-benchmark-zai-glm
12. Ox Alpha (GLM-5.3-Flash): Z.AI, GLM & Benchmarks | Atoms, https://atoms.dev/models/ox-alpha
13. Zhipu (Z.ai) Unmasks The Mystery Ox Alpha Model as GLM-5.3-Flash, Revealing That It Was Run Entirely On Chinese GPUs While Serving 100 Trillion Tokens/Day, https://wccftech.com/zhipu-z-ai-unmasks-the-mystery-ox-alpha-model-as-glm-5-3-flash-revealing-that-it-was-run-entirely-on-chinese-gpus-while-serving-100-trillion-tokens-day/
14. 刚刚！智谱发布GLM-5.3模型、“为编程而生”，性能比肩Fable 5 - 网易, https://www.163.com/dy/article/L4A2D28D05198NMR.html
15. GLM-5.3：智谱开放权重编程与安全模型，评测排名与规格, https://www.datalearner.com/ai-models/pretrained-models/glm-5-3
16. 智谱が「牛来」モデル GLM-5.3-Flashをオンラインで公開し,

https://www.chaincatcher.com/ja/article/2285696
17. GLM-5.3-Flash - Overview - Z.AI DEVELOPER DOCUMENT, https://docs.z.ai/guides/vlm/glm-5.3-flash
18. glm-5.3-flash - Ollama, https://ollama.com/library/glm-5.3-flash
19. Zhipu AI Launches and Open-Sources GLM-5.3-Flash, Its First Native Multimodal Model, https://www.binance.com/en/square/post/359886970606715
20. Z.ai、Ox AlphaをGLM-5.3-Flashとして公開：中国製AIチップ数万基で運用, https://xenospectrum.com/z-ai-ox-alpha-reveal/
21. zai-org/GLM-5.3-Flash - vLLM Recipes, https://recipes.vllm.ai/zai-org/GLM-5.3-Flash
22. The Computing Power Secret Behind Zhipu AI's "Niu Lai" Model - 36氪, https://eu.36kr.com/en/p/3956946155355267
23. Z.ai Reveals Mystery Model 'Ox Alpha' Was GLM-5.3-Flash, Run on Chinese AI Chips, https://xenospectrum.com/en/z-ai-ox-alpha-reveal/
24. GLM-5.3 - 智谱AI开放文档, https://docs.bigmodel.cn/cn/guide/models/text/glm-5.3
25. 连夜实测智谱GLM-5.3-Flash：剪视频、复刻《深海迷航》，价格降低97.5%, https://zhidx.com/p/588399.html
26. Models | ZenMux AI Model Routing, https://zenmux.ai/models
27. 智谱(02513)GLM-5.3-Flash：前沿智能进入普惠时代, https://finance.sina.com.cn/stock/t/2026-08-26/doc-inipsmip5869682.shtml
28. 智谱认领神秘模型“牛来”跑在10万国产卡上，SemiAnalysis评价“英伟达护城河再受考验”, https://news.futunn.com/post/78319729/zhipu-ai-claims-its-mysterious-model-niu-lai-runs-on
29. Zhipu Open-Sources GLM-5.3-Flash: Running on Chinese-Made Chip Clusters, Priced at 1/40th of Claude Opus 4.8, https://finance.biggo.com/news/8ac68e37-02c9-4e8e-bc5a-a6f9e696b5e3
30. Z.aiがGLM-5.3-Flashを公開 Opus級に迫って価格は10分の1、覆面, https://www.ai-jitan-hub.com/news/glm-5-3-flash
31. GDPval-AA v2 Leaderboard - Artificial Analysis, https://artificialanalysis.ai/evaluations/gdpval-aa
32. zai-org/GLM-5.3-Flash - Hugging Face, https://huggingface.co/zai-org/GLM-5.3-Flash
33. 这一次，智谱用GLM-5.3-Flash实现了真正的智能平权 - 51CTO, https://www.51cto.com/article/854129.html
34. GLM 5.3 Flash - API Pricing & Benchmarks - OpenRouter, https://openrouter.ai/z-ai/glm-5.3-flash
35. 智谱认领“牛来”并宣布上线、开源GLM-5.3-Flash, https://wap.eastmoney.com/a/202608263855217025.html
36. The mysterious high-performance AI 'Ox Alpha' has been identified as 'GLM-5.3-Flash,' an open model with performance equivalent to Claude Opus 4.8, and capable of running on a Chinese-made chip., https://gigazine.net/gsc_news/en/20260827-glm-5-3-flash/
37. GLM-5.3-Flash weights released (Ox Alpha),

https://forums.developer.nvidia.com/t/glm-5-3-flash-weights-released-ox-alpha/381345
38. unsloth/GLM-5.3-Flash-GGUF - Hugging Face, https://huggingface.co/unsloth/GLM-5.3-Flash-GGUF
39. Kimi K2.6 - How to Run Locally | Unsloth Documentation, https://unsloth.ai/docs/models/kimi-k2.6
40. Kimi K2.7 Code - How to Run Locally | Unsloth Documentation, https://unsloth.ai/docs/models/kimi-k2.7-code
41. GLM-5.2 - How to Run Locally | Unsloth Documentation, https://unsloth.ai/docs/models/glm-5.2
42. stealth/ox-alpha : r/LocalLLaMA - Reddit, https://www.reddit.com/r/LocalLLaMA/comments/1vwk6bh/stealthoxalpha/