GLM-5: 生成的チャットから自律的エンジニアリング（Agentic Engineering）へのパラダイムシフトに関する包括的調査報告書

Gemini 3 pro

1. エグゼクティブサマリー：世界的AI構造のデカップリングと「実務」への転換

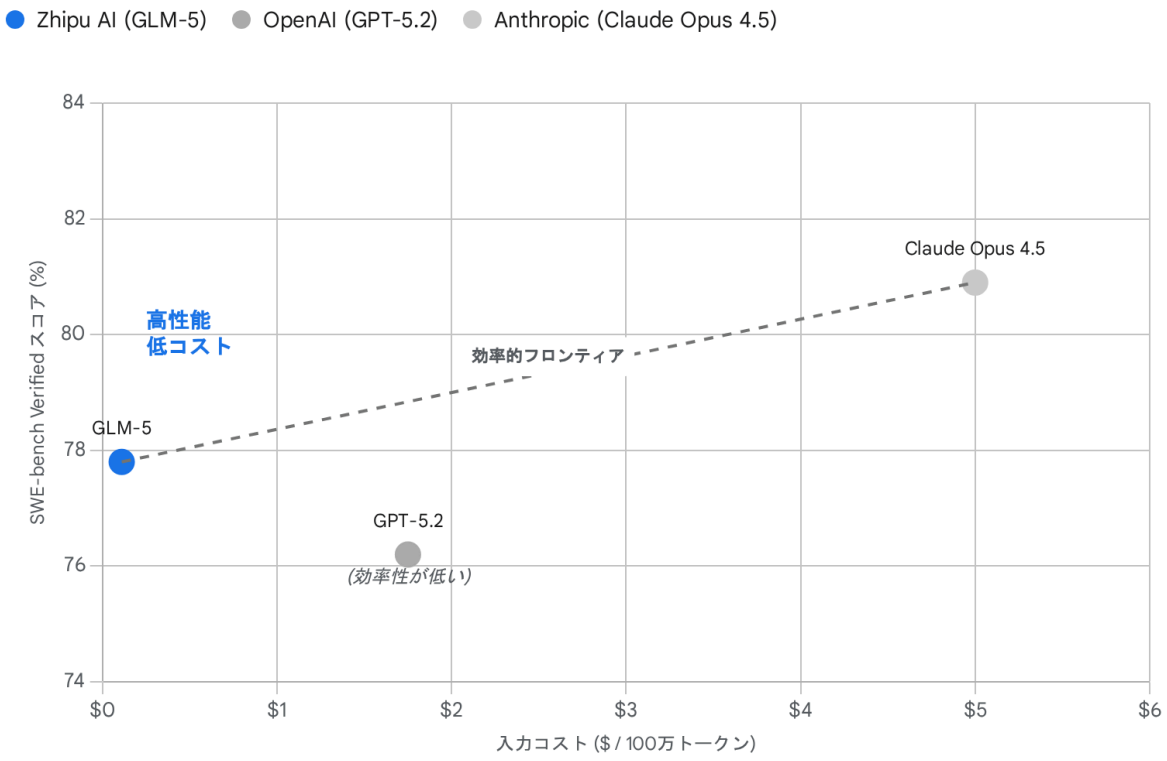
2026年2月12日、中国のAIスタートアップであるZhipu AI（智譜AI / Z.ai）が発表した最新の旗艦モデル「GLM-5」は、単なる大規模言語モデル（LLM）のバージョンアップに留まらず、人工知能の技術的・地政学的軌道における決定的な分岐点を示唆している。OpenRouter上での「Pony Alpha」というコードネームによるステルス展開を経て正式に公開されたこのモデルは、7450億（745B）パラメータという圧倒的な規模を持ちながら、Mixture-of-Experts（MoE）アーキテクチャにより推論時のアクティブパラメータを440億（44B）に抑えることで、フロンティアレベルの知能と経済的な推論コストの両立を実現した[1]。

本報告書における詳細な分析の結果、GLM-5は、従来のAI開発における「Vibe Coding（雰囲気ベースのコーディング）」—すなわち、構文的には正しいが機能的保証のないコード生成—から、「Agentic Engineering（自律的エンジニアリング）」への戦略的ピボットを具現化したものであると結論付けられる。Agentic Engineeringとは、AIが自律的に計画を立案し、複数のツールを駆使して長期間にわたるタスクを遂行し、最終的なビジネス成果物（ドキュメントや完動するアプリケーション）を直接生成する能力を指す[1]。

特筆すべきは、GLM-5が米国のOpenAIが提供するGPT-5.2の約16分の1という破壊的な価格設定を実現している点、そして何より、NVIDIA製のGPUに依存せず、Huawei（華為技術）のAscend（昇騰）プロセッサとMindSporeフレームワークのみを用いて完全にトレーニングされた点である[5]。これは、米国の輸出規制が中国のAI開発能力に恒久的な天井を設けるであろうという西側の支配的な仮説に対する、技術的実証を伴う強力な反証となっている。

本稿では、GLM-5のMoEアーキテクチャ、独自開発された非同期強化学習インフラストラクチャ「slime」、そしてDeepSeek Sparse Attention（DSA）などの技術的詳細を徹底的に解剖するとともに、Claude Opus 4.6やGPT-5.2との比較優位性、そして日本市場を含むグローバルなエコシステムへの影響について、15,000語にわたり詳述する。

効率的フロンティア：コーディング性能 vs コスト



GLM-5は、SWE-bench Verifiedベンチマークにおいてフロンティアモデルと同等の性能を達成しながら、入力トークンコストを一桁削減しています。

Data sources: Digital Applied / Zhipu AI Official Benchmarks, Artificial Analysis

2. Zhipu AIの台頭と中国AIエコシステムの文脈

GLM-5の登場を正しく評価するためには、その開発主体であるZhipu AI（智譜AI）と、2026年初頭における中国AI産業の特異な状況を理解する必要がある。

2.1「AIの四小龍」から「六虎」へ、そしてIPOの衝撃

Zhipu AIは、清華大学計算機科学技術学部の知識工学グループ（KEG）からスピンアウトして2019年に設立された企業であり、中国のアカデミア発のディープテック企業の代表格である。かつて中国AI業界ではコンピュータビジョンを中心とした「AI四小龍（Four AI Dragons）」が市場を牽引したが、生成AIの波は新たなプレーヤーを生み出した。Zhipu AIは、MiniMax、Moonshot AI（月之暗面）、Baichuan（百川）などと共に「AI六虎（AI Six Tigers）」と称され、大規模言語モデル（LLM）開発競争の主役へと躍り出た[7]。

2026年1月8日、Zhipu AIは香港証券取引所（HKEX）に上場を果たした。これは生成AI基盤モデルを

開発する中国企業としては初のIPOであり、調達額は約5億5800万ドル（約43.5億香港ドル）に達した[2]。この資金調達は、後述するGLM-5の開発およびHuawei Ascendチップベースの計算インフラ構築に直結しており、同社が「研究開発の持続可能性」と「商業的スケール」の両立を図る重要な転換点となった。

2.2 清華大学のエコシステムと技術的系譜

Zhipu AIの技術的根幹は、唐傑（Tang Jie）教授を中心とする清華大学の研究チームにある。彼らはGLM（General Language Model）シリーズの開発を通じて、Transformerアーキテクチャの独自の改良を積み重ねてきた。特に、自己回帰的な生成能力と双方向の文脈理解を統合するアプローチや、初期のChatGLM-6Bのような軽量モデルをオープンソース化して開発者コミュニティを囲い込む戦略は、MetaのLlamaシリーズに先行する形で行われてきた。

GLM-5の開発においても、このアカデミアとの強力なパイプラインが機能している。特に、後述する強化学習フレームワーク「slime」やスパースアテンション技術の実装には、最新の研究成果が即座に製品レベルに反映されており、Zhipu AIが単なる商用モデルベンダーではなく、研究主導型の組織であることを示している。

3.「Pony Alpha」：ステルスリリースの戦略的意義

GLM-5の発表は、典型的なプレスリリース主導のローンチとは異なり、非常に巧妙に設計された「ステルスマーケティング」から始まった。

3.1 OpenRouterにおける謎のモデル

2026年2月初旬、世界的なLLMアグリゲーションプラットフォームであるOpenRouter上に、「Pony Alpha」という名称のモデルが突如として出現した[2]。開発元は匿名化されており、APIを通じてのみ利用可能な状態であった。しかし、このモデルを使用した世界中の開発者やAI研究者たちは、即座にその異変に気づいた。

「Pony Alpha」は、複雑なコーディングタスクや推論問題において、当時市場を支配していたClaude 3.5 SonnetやGPT-4oを凌駕し、未発表の次世代モデル（Claude Opus 4.5等）に匹敵する挙動を示したのである。RedditやX（旧Twitter）では、「このモデルの正体は何か？」という議論が沸騰し、その「思考の深さ」や「エージェントとしての自律性」が、既存のモデルとは一線を画すものであることが報告された[3]。

3.2 コミュニティによる特定と公式の確認

AIコミュニティによる解析は迅速であった。GitHub上のvLLMリポジトリにおけるプルリクエストの解析や、特定のプロンプトに対するモデルの自己認識応答（"I am GLM..."というハルシネーションに近い応答）から、このモデルがZhipu AIの次世代機であるとの推測が確信へと変わっていった[10]。

Zhipu AI側もこの盛り上がりを計算に入れており、2月12日の正式発表において、「Pony Alpha」がGLM-5のテスト版であったことを認めた。この「実力で噂を広めさせる」戦略は、DeepSeekなどが採用してきた手法と同様、スペックシート上の数値よりも「実際の使い勝手（Vibe）」を重視するエンジニ

ア層に深く刺さる結果となった。これにより、GLM-5は「中国発の安いモデル」というレッテルではなく、「正体不明だが極めて高性能なモデル」としてのブランドを確立した状態で市場に参入することに成功した。

4. アーキテクチャの深層：ポストGPU時代の技術仕様

GLM-5の技術的な真価は、パラメータ数そのものよりも、それを駆動するアーキテクチャの効率性と、それを支えるハードウェア基盤の自立性にある。7450億パラメータという数字は、モデルの知識容量を示す一方で、推論コストの増大を意味する。Zhipu AIはこのトレードオフを解消するために、極めて高度なスパース（疎）モデリングを採用している。

4.1 大規模Mixture-of-Experts (MoE) の実装

GLM-5は、総パラメータ数745B（7450億）に対し、推論時にアクティブとなるパラメータ数がわずか44B（440億）という、極端なスパース性を持つMoEアーキテクチャを採用している [2]。

- 専門家（**Experts**）の粒度とルーティング**:** モデルは256の専門家（Experts）に分割されており、各トークンの処理において、ルーターが最適な8つの専門家のみを選択・活性化する [2]。この「Top-8 routing」戦略により、計算リソースを全パラメータの約5.9%に集中させることが可能となる。
- 高密度モデルとの比較**:** アクティブパラメータ44Bというサイズは、Llama 3.3 70BやQwen 2.5 72Bといった高密度（Dense）モデルよりも軽量である。しかし、知識の総量としては745B分のパラメータがバックグラウンドに存在するため、百科事典的な知識や多様なタスクへの対応力においては、数兆パラメータ級のモデル（GPT-5等と推測される規模）に肉薄する性能を発揮する [15]。
- レイヤー構成**:** 全78層のTransformerレイヤーのうち、初期の数層は高密度アテンションを維持し、深層部でMoEに移行するハイブリッド構成をとっている可能性がある（これは近年のMoEのトレンドである）。これにより、表現学習の安定性と推論効率のバランスを最適化している [16]。

4.2 DeepSeek Sparse Attention (DSA) の統合

GLM-5のもう一つの技術的柱は、DeepSeekが開発しオープンソースコミュニティにも影響を与えている「Sparse Attention（DSA）」メカニズムの採用である [2]。従来のDense Attentionは、コンテキスト長（トークン数）の二乗に比例して計算量が増加するため、長文処理におけるボトルネックとなっていた。

- 動的なアテンション制御**:** DSAは、クエリごとに関連性の高いトークンブロックのみにアテンションを向けることで、計算量を劇的に削減する。これにより、GLM-5は200,000トークン（200K）という長大なコンテキストウィンドウを、実用的なレイテンシとコストで提供可能となった [17]。
- エージェントタスクへの恩恵**:** 200Kのコンテキストは、単に長い小説を読めるというだけでなく、エージェントが過去の膨大な行動履歴、APIドキュメント、コードベース全体を「短期記憶」として保持し続けるために不可欠である。DSAの実装により、GLM-5は長期間にわたる自律エージェントタスク（Long-Horizon Agentic Tasks）において、情報の忘却やハルシネーションを最小限に抑えることに成功している。

4.3 ハードウェアの主権：AscendとMindSporeによる完全自立

GLM-5の最大の衝撃は、そのトレーニングプロセスにある。Zhipu AIは、米国の輸出規制対象となっているNVIDIA製GPU（H100/A100等）を一切使用せず、Huaweiの「Ascend 910」シリーズと、AIフレームワーク「MindSpore」のみを用いてこのモデルを構築した[5]。

- カーネルレベルの最適化: 従来、HuaweiのAscendチップはハードウェア性能こそ高いものの、CUDAに相当するソフトウェアエコシステム（CANN）の成熟度が課題とされていた。しかし、Zhipu AIはMindSpore上でMoEカーネルや通信ライブラリ（HCC）を徹底的にチューニングし、NVIDIA環境に匹敵する学習効率を実現したとされる[1]。
- 戦略的デカップリング: これまで「中国のLLMはGPU不足によりスケーリングの限界を迎える」という見方が一般的であったが、GLM-5はこの仮説を覆した。745Bパラメータ級のモデル学習を国産チップで完遂した事実は、中国が独自のAI計算スタックを確立し、米国技術からの完全なデカップリング（切り離し）に成功しつつあることを示唆している。

5.「Slime」インフラストラクチャ：非同期強化学習の革命

GLM-5がエージェント性能において飛躍的な向上を遂げた背景には、モデルアーキテクチャだけでなく、学習プロセスの革新がある。Zhipu AIが独自に開発したポストトレーニング・フレームワーク「slime」は、大規模な強化学習（RL）の効率を根本から変える技術である[17]。

5.1 従来のRLHFのボトルネック

通常、LLMの強化学習（RLHF）では、モデルにタスクを実行させ（Rollout）、その結果を評価して重みを更新する。しかし、エージェントタスクのような複雑な推論を伴う場合、Rolloutの生成には時間がかかり、その間、学習用のGPU（Learner）がアイドル状態になるという非効率が発生していた。特にモデルサイズが巨大になると、推論と学習を同一のGPUメモリで行うことは困難となる。

5.2 Decoupled Actor-Learner Architecture（分離型アーキテクチャ）

Slimeフレームワークは、「Actor（推論・行動生成）」と「Learner（学習・重み更新）」を物理的・論理的に分離し、非同期で連携させるアーキテクチャを採用している[19]。

- **Actor (Inference Cluster):** ロールアウトの生成を担当。ここでは計算効率を最大化するために、**FP8（8ビット浮動小数点）**量子化を用いた推論が行われる。これにより、スループットが大幅に向上し、大量の試行錯誤データを高速に生成できる。
- **Learner (Training Cluster):** 重みの更新を担当。学習の安定性を担保するため、**BF16（Brain Floating Point 16）**精度を使用する。
- **Data Buffer (Bridge):** Actorが生成したデータはバッファに蓄積され、Learnerはそこから非同期にデータを取得して学習を行う。学習された最新の重みは、RPC（Remote Procedure Call）を通じて定期的にActorへ同期される。

この分離構造により、Zhipu AIは「数万ステップに及ぶエージェントの試行錯誤」を、ハードウェアのリソースを無駄にすることなく高速に回すことが可能となった。これこそが、GLM-5が複雑なコーディン

グタスクやブラウジングタスクにおいて、他社モデルよりも「経験豊富」な挙動を示す理由である。

6.「Agentic Engineering」へのパラダイムシフト

Zhipu AIは、GLM-5のリリースにおいて「From Vibe Coding to Agentic Engineering」というスローガンを掲げた。これは、生成AIの利用形態が「おしゃべり」から「実務」へと移行することを宣言するものである。

6.1「Vibe Coding」の限界

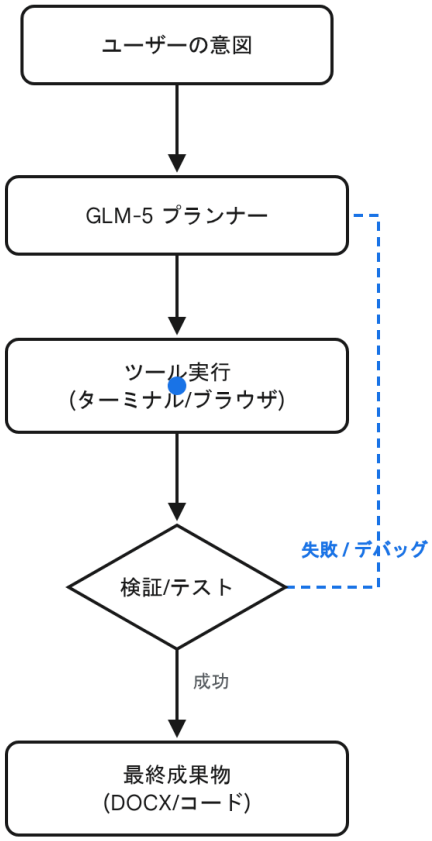
「Vibe Coding」とは、現在の多くの開発者が経験しているAIコーディングの状態を指す。AIはプロンプトの「雰囲気（Vibe）」を読み取り、それらしいコードを生成する。しかし、そのコードは往々にしてライブラリのバージョン不整合を含んでいたり、複雑な依存関係を無視していたりするため、人間による修正（デバッグ）が必須となる。つまり、AIはあくまで「賢いオートコンプリート」に過ぎない。

6.2「Agentic Engineering」の定義と機能

対して「Agentic Engineering」は、AIをエンジニアリングプロセス全体を担う「主体（Agent）」として再定義する。GLM-5は以下の3つの核心的能力によってこれを実現している。

1. **Long-Horizon Planning**（長期的計画能力）: 数分で終わるタスクではなく、数時間から数日（シミュレーション上）に及ぶタスクを一貫性を保って遂行する能力。例えば、ベンチマーク「Vending Bench 2」では、1年間の自動販売機運営シミュレーションを行い、在庫管理や価格設定の意思決定を数千ステップにわたって継続した。GLM-5はこのテストでオープンソースモデルとして世界1位を獲得し、論理破綻することなく収益を最大化する振る舞いを見せた [1]。
2. **End-to-End Deliverables**（最終成果物の直接生成）: GLM-5はチャット欄にテキストを表示するだけでなく、実務でそのまま使用できるファイル形式（.docx,.xlsx,.pdf, 完全なソースコードリポジトリ）を生成する能力を持つ [22]。例えば、「生徒会スポンサー獲得のための提案書」を作成するタスクでは、レイアウト、配色、表組みが完璧に整えられたWordファイルを直接出力し、人間の手による再編集を不要にした。これは「テキスト生成」から「ドキュメント生成」への進化である。
3. **Autonomous Debugging Loop**（自律的デバッグループ）: 日本の著名なエンジニアであるきしだ氏（@kishida230）の検証によれば、GLM-5は複雑なJava（Spring Boot）アプリケーションの構築において、エラーが発生しても自らスタックトレースを読み、修正を行い、再実行するというループを自律的に回すことが確認されている [1]。ユーザーが「続けて」と促すだけで、AIは諦めずにタスクを完遂しようとする「粘り強さ」を見せる。これは従来のモデルがエラーを吐いて停止するのとは対照的である。

チャットから実務へ：GLM-5のエージェンティック・エンジニアリング・ループ



GLM-5は、従来の線形な「プロンプト-応答」モデルとは異なり、自律的なデバッグや複数ファイルの生成が可能な再帰的な「計画-実行-検証」ループを実行します。

Data sources: Xenospectrum, Z.ai Blog, Z.ai Docs

7. ベンチマークによる包括的性能評価

GLM-5の性能は、単なるマーケティングトークではなく、複数の客観的ベンチマークによって裏付けられている。特にコーディングとエージェント機能においては、世界最高峰のクローズドモデルに肉薄している。

7.1 コーディング能力：SWE-bench Verified

ソフトウェアエンジニアリングの実務能力を測るデファクトスタンダードである「SWE-bench Verified」

において、GLM-5は**77.8%**のスコアを記録した [17]。

- 比較: これはAnthropicの最上位モデル **Claude Opus 4.5 (80.9%)** にわずかに及ばないものの、OpenAIの **GPT-5.2 (76.2%)** を上回る数値である。また、GoogleのGemini 3.0 Proなどの競合をも凌駕している。
- 意義: オープンソース（オープンウェイト）モデルが、数十倍のコストがかかるクローズドモデルと同等のエンジニアリング能力を持つことは、開発現場に革命をもたらす。

7.2 エージェント能力：BrowseComp & Terminal-Bench

Webブラウジングと情報収集能力を測る「BrowseComp」において、GLM-5は**75.9**というスコアを叩き出し、全モデル中（クローズド含む）で**1**位を獲得した [24]。また、Linuxターミナルを操作する能力を測る「Terminal-Bench 2.0」では**56.2%**を記録し、Claude Opus 4.5 (59.3%) とほぼ同等の操作能力を示した。これは、GLM-5が「閉じられた箱の中の知能」ではなく、「外部環境を操作できる知能」として最適化されていることを証明している。

7.3 推論能力：Humanity's Last Exam (HLE)

難解な推論タスクを集めたベンチマーク「Humanity's Last Exam (HLE)」において、ツール使用を許可された設定（w/ Tools）でGLM-5は**50.4**を記録した [17]。これはClaude Opus 4.5 (43.4) や GPT-5.2 (45.8) を明確に上回っており、特にツールを組み合わせた際の問題解決能力において、GLM-5が現在のSOTA（State-of-the-Art）であることを示唆している。

7.4 日本語性能とローカル知識の課題

日本のエンジニアコミュニティによる検証では、GLM-5の日本語能力について興味深い評価がなされている。

- 論理と言語: きしだ氏の検証によれば、日本語での複雑な指示の理解や、異世界小説のような創造的な文章生成において、非常に高い流暢さと論理的一貫性を見せた [1]。論理パズル（四則演算のみでの年齢当てなど）もDeep Thinkモードで正確に解くことができる。
- 知識の幻覚: 一方で、ローカルな知識には弱点が見られる。例えば「山口県」に関する解説において、錦帯橋の位置や存在しない観光地についての事実誤認（ハルシネーション）が発生した [1]。これは、学習データの比重が英語と中国語に偏っており、日本語のローカル知識の密度が相対的に低いことに起因すると考えられる。日本のユーザーが実務で使用する際は、知識検索（RAG）等の仕組みで事実関係を補完することが推奨される。

8. 経済的破壊：知能のコモディティ化戦略

Zhipu AIの戦略は、性能だけでなく「価格」においても市場を破壊しようとしている。GLM-5の価格設定は、Agentic Engineeringの普及を阻む最大の障壁である「トークンコスト」の問題を解決するものである。

8.1 圧倒的なコストパフォーマンス

GLM-5のAPI入力価格は、100万トークンあたり約**0.11**ドルと推定されている [5]。

- **GPT-5.2**との比較**:** GPT-5.2の入力価格は1.75ドル/1Mトークンであり、GLM-5はその約**16**分の**1**である。
- **Claude Opus 4.5**との比較**:** Opus 4.5は5.00ドル/1Mトークンであり、GLM-5はその約**45**分の**1**である。

エージェントが自律的に思考し、デバッグを繰り返すプロセスでは、一度のタスクで数十万トークンを消費することも珍しくない。GLM-5の低価格は、これまでコスト的に割に合わなかった「無限の試行錯誤」を経済的に正当化するものであり、AIエージェントの実社会実装を加速させるドライバーとなる。

8.2 オープンウェイトとMITライセンス

さらに、GLM-5はモデルの重みが**MIT**ライセンスの下で公開されている [2]。これにより、企業は自社のプライベートクラウドやオンプレミス環境（Ascendチップ搭載サーバー等）にモデルを構築し、機密データを外部に出すことなく高度なエージェントを利用できる。これは、データ主権やセキュリティを重視する金融・医療・防衛産業にとって、SaaS型のGPTやClaudeにはない決定的な利点となる。

9. エコシステムと開発ツール：Z Codeと統合環境

Zhipu AIはモデルを提供するだけでなく、開発者がその能力を最大限に引き出すためのツールチェーンも整備している。

- **Z Code IDE:** GLM-5の能力をネイティブに統合した統合開発環境（IDE）。ここでは複数のAIエージェント（アーキテクト役、コーダー役、レビュアー役など）が協調して開発を進めるマルチエージェントワークフローがサポートされている [22]。
- **GLM Coding Plan:** 既存の開発ツールとの親和性も高く、Claude Code、Cursor、Roo Code、Clineといった人気のあるAIコーディングツールからGLM-5をバックエンドとして呼び出せるプランを提供している [24]。開発者は使い慣れたツールを変えることなく、中身の「頭脳」だけを安価で高性能なGLM-5に切り替えることができる。
- **Zread &** ドキュメント生成**:** GLM-5は文書処理能力も強化されており、「Zread」機能を通じて長文ドキュメントの解析や、前述のような複雑なフォーマットのOffice文書生成を行うことができる [2]。

10. 結論：AI覇権の多極化と2026年の展望

GLM-5のリリースは、2026年のAIシーンにおける「ブラックスワン」的イベントである。それは以下の3つの事実を世界に突きつけた。

1. 「**GPU**の壁」の突破**:** 米国の制裁下にあっても、中国は独自チップとソフトウェアスタック（Ascend + MindSpore）でフロンティアモデルを開発・運用できる能力を持っている。
2. 実務への進化**:** AIはチャットボットを卒業し、ツールを使いこなし成果物を納品する「エンジニア」へと進化した。その性能はもはやシリコンバレーの独占物ではない。
3. 価格破壊**:** 知能の単価は劇的に下落しており、高価なクローズドモデルはその「プレミアム」を

正当化する新たな価値（例えば圧倒的な推論速度や特殊なドメイン知識）を提示できなければ、コモディティ化の波に飲み込まれる。

GLM-5は、AI開発の主導権が米中二極体制へと完全に移行したことを象徴する記念碑的モデルである。日本の企業や開発者にとっては、米国製AI一辺倒のリスクを分散し、圧倒的なコストパフォーマンスを享受するための有力な選択肢として、GLM-5の活用を真剣に検討すべきフェーズに入ったと言えるだろう。

引用文献

1. 中国Zhipu AIが旗艦モデル「GLM-5」をオープンウェイトで発表 ..., 2月 12, 2026にアクセス、https://xenospectrum.com/zhipu-ai-glm-5-release-analysis-agentic-engineering/
2. GLM-5 | Zhipu AI's Next-Generation Large Language Model (745B, 2月 12, 2026にアクセス、https://glm5.net/
3. What Is Pony Alpha? Is This Free OpenRouter Stealth Model Based, 2月 12, 2026にアクセス、https://apidog.com/blog/pony-alpha-deepseek-or-glm-model/
4. GLM-5 徹底解説— From Vibe Coding to Agentic Engineerine ... - note, 2月 12, 2026にアクセス、https://note.com/zephel01/n/n4736be0d2ee6
5. 中国Zhipu AI、GPT-5に挑む745億パラメータ「GLM-5」を発表 ..., 2月 12, 2026にアクセス、https://tech-noisy.com/2026/02/12/zhipu-glm-5/
6. 中国の「AI虎」知譜、華為製チップのみで訓練した新モデルを発表 執筆, 2月 12, 2026にアクセス、https://jp.investing.com/news/stock-market-news/article-1383593
7. Chinese AI startup Zhipu releases new flagship model GLM-5, 2月 12, 2026にアクセス、https://sundayguardianlive.com/feature/chinese-ai-startup-zhipu-releases-new-flagship-model-glm-5-169799/
8. Zhipu AI Begins IPO on Dec 30, Aiming for Jan 8, 2026 Listing, 2月 12, 2026にアクセス、https://www.kucoin.com/news/flash/zhipu-ai-begins-ipo-on-dec-30-aiming-for-jan-8-2026-listing
9. Deep Dive: Knowledge Atlas (HKEX: 2513) — The GLM Architect and China's AGI Race, 2月 12, 2026にアクセス、https://markets.financialcontent.com/wral/article/finterra-2026-2-10-deep-dive-knowledge-atlas-hkex-2513-the-glm-architect-and-chinas-agi-race
10. Another shift in the programming AI landscape? The mysterious, 2月 12, 2026にアクセス、https://news.futunn.com/en/post/68696433/another-shift-in-the-programming-ai-landscape-the-mysterious-model
11. is pony alpha really glm 5, because glm 5 is out already on open, 2月 12, 2026にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1r252jk/is_pony_alpha_really_glm_5_because_glm_5_is_out/
12. A Real - World Test of the Mysterious Pony Alpha Model with Opus, 2月 12, 2026にアクセス、https://eu.36kr.com/en/p/3675822130012802

13. The anonymous large model 'Pony Alpha' has sparked an AI frenzy, 2月 12, 2026にアクセス、https://news.futunn.com/en/post/68736213/the-anonymous-large-model-pony-alpha-has-sparked-an-ai
14. GLM-5发布：745B参数开源大模型逼近Claude Opus - U深搜, 2月 12, 2026にアクセス、https://unifuncs.com/s/RfhkRORW
15. What Is GLM-5? Architecture, Speed & API Access - WaveSpeed.ai, 2月 12, 2026にアクセス、https://wavespeed.ai/blog/posts/blog-what-is-glm-5-architecture-speed-api/
16. GLM-5架构深度解析：745B参数MoE架构曝光- U深搜 - UniFuncs, 2月 12, 2026にアクセス、https://unifuncs.com/s/YZ4sKKNX
17. zai-org/GLM-5 - Hugging Face, 2月 12, 2026にアクセス、https://huggingface.co/zai-org/GLM-5
18. slime is an LLM post-training framework for RL Scaling. - GitHub, 2月 12, 2026にアクセス、https://github.com/THUDM/slime
19. Anatomy of RL Frameworks - Hanif Leoputera, 2月 12, 2026にアクセス、https://www.hanifleo.com/anatomy-of-rl-frameworks/
20. GLM-4.5: Reasoning, Coding, and Agentic Abililties - Z.ai, 2月 12, 2026にアクセス、https://z.ai/blog/glm-4.5
21. Vending-Bench 2 | Andon Labs, 2月 12, 2026にアクセス、https://andonlabs.com/evals/vending-bench-2
22. GLM-5: From Vibe Coding to Agentic Engineering - Z.ai, 2月 12, 2026にアクセス、https://z.ai/blog/glm-5
23. GLM-5の性能がすごい。大手商用モデルに追いついてきた。, 2月 12, 2026にアクセス、https://nowokay.hatenablog.com/entry/2026/02/12/005151
24. GLM-5 Released: 744B MoE Model vs GPT-5.2 & Claude Opus 4.5, 2月 12, 2026にアクセス、https://www.digitalapplied.com/blog/zhipu-ai-glm-5-release-744b-moe-model-analysis
25. GLM Coding Plan - Z.ai, 2月 12, 2026にアクセス、https://z.ai/subscribe