Z.aiの新旗艦モデル「GLM-5.4」徹底解剖：7500億パラメータ・ループ型Transformerの構造革新と実務的評価

Gemini 4 Carbon

生成AI開発のZ.ai（中国・智譜AI／ZAI Co Ltdの国際ブランド、香港証券取引所：02513）は2026年11月23日、中間層の重みを共有して反復実行する「Loop Transformer（ループ型トランスフォーマー）」構造を主力モデルとして初採用した新フラッグシップモデル「GLM-5.4」を発表・提供開始した[1]。同モデルは、2026年2月公開のMixture-of-Experts（MoE）基盤モデル「GLM-5」（7500億パラメータ規模）をベースに、中間28層を2回ループ実行する継続学習によって構築されており、総パラメータ規模を据え置いたまま「2兆～3兆パラメータ（2T～3T）級モデルに匹敵する知能水準」に到達したと同社は主張している[3]。さらに、GLMの旗艦ラインとして初めてネイティブ画像認識（ビジョン）機能を統合し、最大100万（1M）トークンのコンテキストウィンドウと最大12万8000（128K）トークンの出力に対応したほか、コードを汎用インターフェースとして動画制作・3Dモデリング・機械設計（CAD）まで自律処理する「汎用コーディング（Generalized Coding／泛Coding）」の起点と位置づけられている[3]。
提供面では、発表と同時に定額サブスクリプション「GLM Coding Plan」（個人向け月額18米ドル～）、専用開発環境「ZCode」、および「Z.ai Chat」向けに即時開放され、週次利用上限枠（トークン換算の実効枠）は前世代「GLM-5.3」比で1.5倍（50%増）へと拡大された[9]。多段ループ計算を伴う構造でありながら、モデル設計と推論インフラの協調最適化により、オフピーク時で最大200 tokens/s、ピーク負荷時でも70～80 tokens/sの生成速度を確保している[8]。
もっとも、2026年11月23日の発表時点において導入検討企業や開発者が押さえるべき重要な留保条件が4点存在する。第一に、Z.aiは7500億パラメータ本番モデルの公式モデルカードや網羅的な標準ベンチマーク比較表を未公表のまま先行配信しており、現時点の客観的評価は同社の300億（30B）パラメータ対照実験と外部評価者による初期テスト（KingBench 4等）に依存している[8]。第二に、従量課金型の公式Model APIは「セキュリティ評価の完了後」の公開予定とされ、現時点では定額プラン契約者等に対象が限られる[10]。第三に、オープンウェイト（モデル重み）の公開も「必要な安全性強化（safety hardening）の完了後」と予告されるにとどまり、公開時期やライセンス条項は未確定である[8]。第四に、Z.ai自身が公式ブログで認める通り、複雑な課題で過剰に長い自己省察を繰り返す「マラソン思考（marathon thinking）」に陥りやすく、思考強度（Effort）設定を適切に制御しないとタイムアウトやクレジット消費の急増を招く特性がある[3]。

項目	**GLM-5.4** 公表仕様・ステータス（**2026**年**11**月**23**日時点）	前世代旗艦 **GLM-5.3** との比較・備考
開発元・発表日	Z.ai（ZAI Co Ltd／旧・智譜AI、HKEX: 02513）／2026年11月23日[1]	2026年8月14日発表（GLM-5.2と同一ベース重みの事後学習強化版）[17]

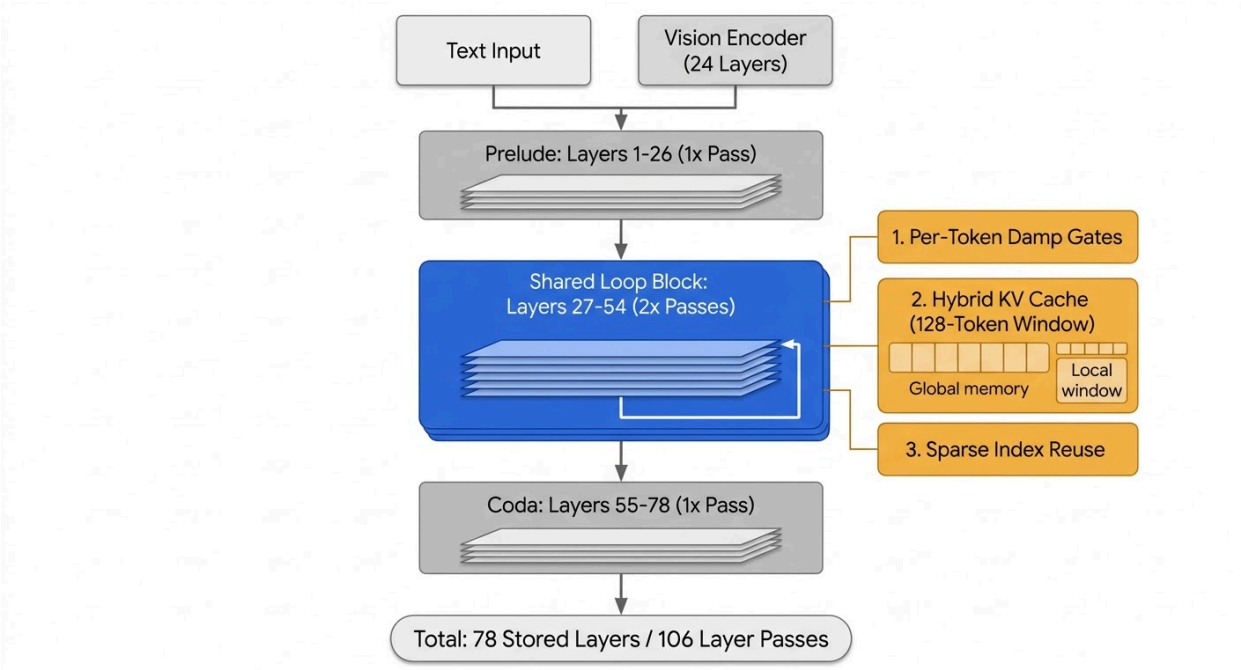
基本アーキテクチャ	Loop Transformer × 疎なMoE（GLM-5ベースモデルの継続学習適応）[3]	標準デコーダ型 疎なMoE（glm_moe_dsa系統、単一パス実行）[18]
パラメータ規模	750Bスケール（ベース744B〜753B＋制御ゲート＋24層視覚エンコーダ、詳細内訳は未公表）[19]	総パラメータ約753B／トークンあたりアクティブ約40B[22]
レイヤー構成・実効深度	物理78層（Prelude 26層＋共有Loop 28層×2周＋Coda 24層）＝計106層パス[7]	各層を1フォワードパスあたり1回のみ直列実行[3]
コンテキスト長／最大出力	最大100万（1M）トークン（コーディング推奨51.2万）／最大128K出力[8]	最大100万（1M）トークン[22]
対応モダリティ	旗艦初のビジョン対応（中国公式仕様は動画・画像・テキスト・ファイル入力／テキスト出力）[9]	テキスト入出力専用（マルチモーダルは総320B／活性18Bの5.3-Flashに分離）[12]
推論スループット	オフピーク最大200 tokens/s／ピーク時 70〜80 tokens/s[10]	ループ型との公式対比速度は非公表
公式ベンチマーク	本番**750B**の公式スコア表・モデルカード未公表（対5.3勝率51.1%の報告と30B検証のみ）[8]	CyberGym 84.5（中国語資料83.5%）、DeepSWE v1.1 66.9等を公表済[26]
外部独立評価（初期）	KingBench 4で33.2/60（55.3%）：全19モデル中7位・オープン系譜1位[14]	KingBench 4で27.5/60（GLM-5.4が+5.7ポイント上回る）[14]
提供チャネル・**API**	GLM Coding Plan、ZCode、Z.ai Chatで即時提供／従量課金**API**は安全評価待ち[10]	定額プランおよび従量課金API（100万トークンあたり入力1.40／出力4.40米ドル）で公開中[14]
オープンウェイト・ライセンス	未公開（「安全性強化の完了後」に公開予定、日程・ライセ	2週間のサイバー安全審査後に独自「GLM-5.3 License

	ンス未定）[8]	」で重み公開[14]

Loop Transformerアーキテクチャの技術的深層：パラメータ凍結のまま推論深度を拡張する設計

Z.aiは2026年に入り、2月の「GLM-5」を皮切りに、4月の「GLM-5.1」、6月の「GLM-5.2」、8月の「GLM-5.3」と約2カ月周期で高速なモデル刷新を重ねてきた[17]。しかし、前世代のGLM-5.3が新しい事前学習ベースモデルを持たず、環境スケーリングなどの事後学習（post-training）のみでGLM-5.2の重みを引き出すアプローチだったのに対し、GLM-5.4はネットワークの計算構造そのものをループ型へ転換した点で一線を画す[17]。大規模ソフトウェアエンジニアリング、サイバーセキュリティ、3Dモデリング、物理制御といった「超長連鎖推論（ultra-long-chain reasoning）」タスクでは、大域的な論理的一貫性を保ちながら多段階の因果関係を追跡する必要があり、パラメータの単純な積層やデータの暗記だけでは解決できないという技術的判断がその根底にある[3]。Z.aiはゼロからループ型モデルを事前学習するのではなく（同社によれば固定された事前学習予算下ではゼロからのループ学習は標準モデルに遅れをとりやすい）、既存の事前学習済みGLM-5ベースモデル（総パラメータ約744B〜753B、トークンあたりアクティブ約40B、256エキスパート中8つを選択する疎なMoE構造）を継続学習（continued training）によってLoop Transformerへ適応させた[3]。事前に1B、4B、30B、750Bの4段階で段階的なスケーリング検証を経ており、総学習計算量の増加を約10%に抑えながら知能上限の引き上げを図っている[3]。

技術解説メディアによるZ.ai公式発表の分析によれば、GLM-5.4のネットワークは物理的に保存された78層のTransformerレイヤーを4つの実行セクションに分割して処理する[7]。入力段階の「Prelude（第1〜26層）」はトークンごとに1回実行され、テキスト埋め込みと24層ビジョンエンコーダから射影された視覚特徴量を統合して内部表現へ変換する[7]。続く中核部では、第27〜54層の28層ブロックが同一の重みを共有したまま「Loop Pass 1」および「Loop Pass 2」として連続2回反復実行され、1周目の隠れ状態を2周目で精緻化する[3]。最後に「Coda（第55〜78層）」が1回実行され、最終状態を次トークン予測へと変換する[7]。これにより、VRAM上に保持する重みは78層分のままでありながら、1トークンあたり合計106層パス（26＋28×2＋24）の実効深度を獲得する[21]。さらに750B規模かつ100万トークン文脈での実用配備を可能にするため、Z.aiは活性値の発散を防ぐトークン単位の制御ゲート（Damp Gates）、メモリ倍増を回避するハイブリッドKVキャッシュ、および疎アテンションのインデックス再利用という3つの構造改修を組み込んだ[3]。



第一の安定化機構である「トークン単位の門控漸進更新（Gated Progressive Update／Damp Gates）」は、同一重みの層を複数回通過する際に残差ストリームの活性値ノルムが制御不能に膨張する問題に対処するものである[7]。ループ突入前の状態を x_0 、1周目および2周目のループブロック出力をそれぞれ h_1, h_2 、トークンごと・パスごとに算出されるゲート値を b_1, b_2 とすると、各パス後の状態更新は以下の漸進式で表される[21]。

$$x_1 = x_0 + b_1 \times (h_1 - x_0), \quad x_2 = x_1 + b_2 \times (h_2 - x_1)$$

1周目では小さなゲート値 b_1 を用いて値を膨張させずに大域的コンテキストを慎重に構築し、2周目でゲート b_2 を大きく開いて細部の推論結果を書き戻す[8]。同時にこのゲートはトークンの予測難易度に応じてステップ幅を動的に調整しており、予測が難しいトークンに対しては1周目からゲート値を高めて強く介入し、2周目の反復効果を引き出す挙動を示す[7]。

第二・第三の機構は、長文脈推論時のメモリ帯域と計算コストのボトルネックを解消する設計である。素朴なループ実装ではパスごとにKey/Valueキャッシュが倍増してしまうが、GLM-5.4のハイブリッドKVキャッシュは、1周目で選択された長文脈の疎なKVキャッシュを2周目でも大域的に共有しつつ、直近128トークンの局所スライディングウィンドウ内のみ2周目専用のKVキャッシュを独立保持して両者を同時に参照する[8]。これにより、100万トークン文脈下でも追加メモリをスライディングウィンドウ1つ分の微小な増加に抑制した[7]。さらに、ベースモデルが備える疎アテンション（Sparse Attention）においても、2周目はインデクサ（索引器）を再実行せず1周目が生成したインデックスを直接再利用するため、追加の索引計算オーバーヘッドをゼロに保ったまま毎回再計算する場合とほぼ同等の精度を維持している[3]。これに融合GPUカーネル（fused GPU kernels）等のインフラ最適化を組み合わ

せたことが、ループ構造でありながら最大200 tokens/sの高速生成を実現した技術的要因となっている[8]。

性能検証と推論特性：「潜在的深化」と「マラソン思考」のパラドックス

GLM-5.4の性能を客観的に評価する際は、**2026**年**11**月**23**日の発表時点で**750B**本番モデルの公式ベンチマーク表やモデルカードが未公開であるという前提を踏まえる必要がある[8]。技術メディアが引用するZ.aiの報告値として、前世代GLM-5.3との直接比較（head-to-head）で「勝率51.1%、引き分け25.7%、敗北23.1%」という数字が伝えられているものの、SWE-bench等の標準スイートにおける750Bモデルの一次スコアは後日発表とされている[8]。したがって現時点で検証可能な定量データは、(1) Z.aiが公式ブログで開示した30Bスケールでの同規模対照（アブレーション）実験と、(2) 先行アクセスを得た外部評価者AICodeKing氏による6プロジェクト構成の実戦コーディング評価「KingBench 4」の計測データの2つが中核となる[3]。

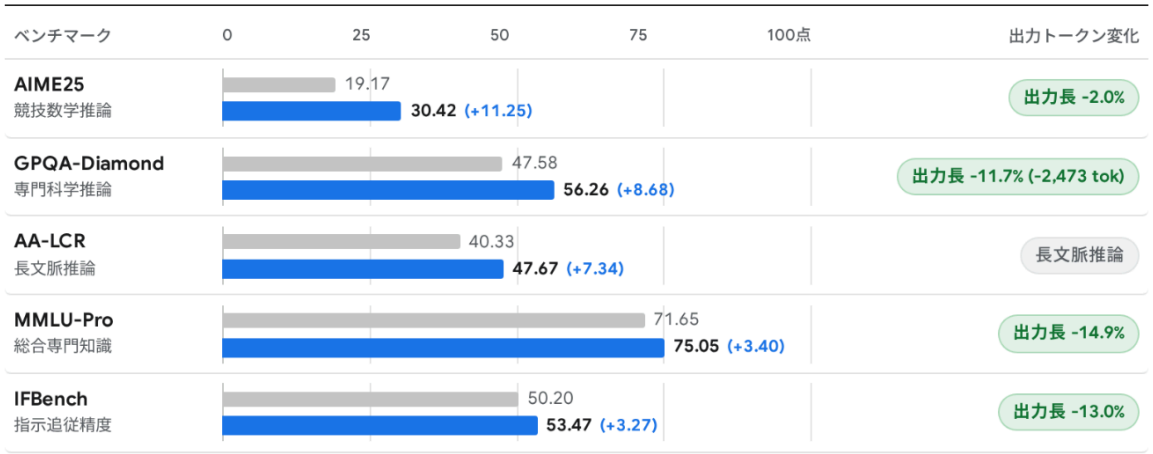
単発推論での出力トークン短縮と実戦エージェントでの超多段ステップの二面性

PANEL 1 · 基礎アーキテクチャ検証（Z.ai 内部対照実験）

検証1：30B同規模モデル対照実験（非ループ vs 2周ループ構造）

20万トークン長文脈SFT後・思考モード有効（学習計算量 +約10%）／1問あたりの明示的CoT出力長が短縮

非ループ 30B（Base）　2周ループ 30B（Looped）

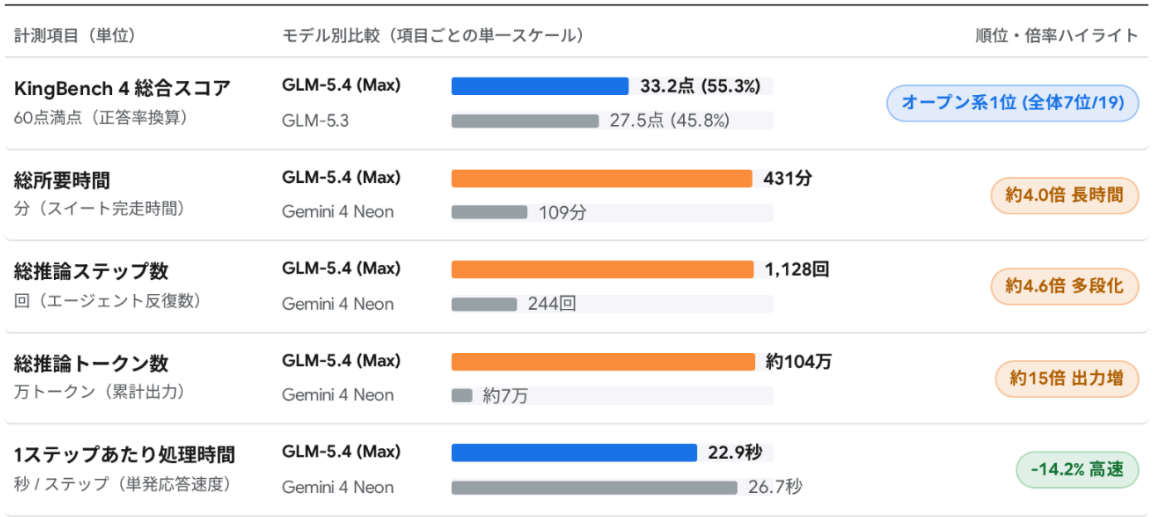


PANEL 2 · 本番750Bフラッグシップ外部計測（AICodeKing / KingBench 4）

検証2：本番GLM-5.4 エージェント実戦計測（KingBench 4・Max設定）

各ステップは高速ながら細かな検証を1,128回反復し、総推論トークンと所要時間が拡大（項目別独立スケール）

GLM-5.4（スコア・速度）　GLM-5.4（長考コスト）　比較対象モデル



思考ほぼゼロのステップ率：**1%未満**（GPT系は14〜39%）　ツール失敗からのリトライ回復率：**66%**（失敗率 1.56%）

キャッシュヒット率：**96〜99%**（初回トークン中央値 4.1秒）

上段はZ.aiによる300億パラメータ同規模対照実験（20万トークン長文脈SFT後・思考モード有効、学習計算量+約10%）の結果であり、数学・科学・長文脈推論で大幅なスコア向上を果たすと同時に、推論の一部がループの潜在深度へ移行することで出力トークン長が最大14.9%短縮された。一方、下段の外部評価（AICodeKingによるKingBench 4、Max設定）では、本番GLM-5.4がオープン系モデル1位（全体7位）を記録したものの、1ステップあたり22.9秒と高速でありながら計1,128ステップの細かな検証を積み重ねた結果、総推論トークン数と所要時間が大幅に増加している。

データ出典: Z.ai Research Blog, DEV Community, THE DECODER, DataCamp, OneAI API Key, Fello AI

これら2つのデータは、単発の知識・論理推論と長期エージェント実行における興味深い二面性を示している。Z.aiの30Bモデル対照実験（20万トークン長文脈SFT後・思考モード有効、学習計算量+約10%）では、数学「AIME25」が19.17から30.42（+11.25ポイント）、科学推論「GPQA-Diamond」が47.58から56.26（+8.68ポイント）、長文脈推論「AA-LCR」が40.33から47.67（+7.34ポイント）、さらに「MMLU-Pro」で+3.40、「IFBench」で+3.27と軒並み向上した[3]。同時に、GPQA-Diamondで平均2,473トークン（-11.7%）、MMLU-Proで-14.9%、IFBenchで-13.0%と、明示的な思考連鎖（**CoT**）の出力トークン長が平均**10〜14%**短縮された[3]。これは、従来テキストトークンとして外部出力していた推論の一部が、ループ構造の隠れ状態反復（潜在的深度：implicit depth）へと移行したことを裏付ける実験結果となっている[3]。

一方で、本番750Bモデルを評価したAICodeKing氏の「KingBench 4」独立テスト（Max設定）では、GLM-5.4は60点満点中33.2点（55.3%）を記録し、GLM-5.3（27.5点）を5.7ポイント上回って全19モデル中7位、オープン系譜モデル群（Qwen4 Max、DeepSeek-V4.1-Flash、Kimi K3.1 Pro、Mistral Large 4等）の中で1位に立った（ただし、GLM-5.4自体の重みは未公開である点に注意が必要である）[8]。なお、クローズド最上位群を比較する別指標「Artificial Analysis Intelligence Index v5.0.1」ではGPT-6.2 Astra Maxが62、Gemini 4 Neon Xhighが61、Claude Opus 5.5 Maxが59であるのに対し前世代GLM-5.3 Maxは32に位置しており、API未開放のため同指標に未収載のGLM-5.4がクローズド最上位層との差をどこまで詰めたかは今後の全体検証を待つ段階にある[12]。KingBench 4の詳細テレメトリを見ると、比較対象のGemini 4 Neonが244ステップ・約7万推論トークン・109分で完走したのに対し、GLM-5.4は1,128ステップ（約4.6倍）・約104万推論トークン（約15倍）・431分（約4倍）を要した[8]。1ステップあたりの処理時間は22.9秒（初回トークン到達時間の中央値4.1秒、全体時間の75%が生成・25%がビルドとテスト）とNeonの26.7秒より高速であるにもかかわらず総所要時間が4倍に膨らんだ主因は、GLM-5.4が細かなステップごとに必ず推論を挟む点にある[12]。推論をほぼ省略して行動するステップの比率はGPT系モデルの14〜39%に対しGLM-5.4では1%未満にとどまっており、反復構造と強化学習による自己検証傾向が重なることで長期の「マラソン思考」が生じる構造になっている[3]。

さらに運用上の注意点として、Z.aiのCoding Plan環境では思考強度（Effort）が未指定の場合にデフォルトで最大強度「max」が適用される（内部マッピングでは minimal / light / low が low へ、medium / high が high へ、xhigh / max / ultra および未指定が max へ変換される）ほか、APIの thinking.clear_thinking（過去ターンの推論内容を消去する設定）の既定値も従来モデルの true からGLM-5.4以降では false に変更されている[8]。独立テストのセキュリティ課題では、上位設定（high / max）でモデルが思考を続け180秒のタイムアウトに達して失敗した一方、1段階下げた設定（medium リクエスト）では95秒で正常にクリアする逆転現象が報告されており、評価者らはデフォルトの max 放置を避けて high 相当を基準に運用することを推奨している[8]。

初の旗艦ビジョン統合と「汎用コーディング（Generalized Coding）」の実務適性

プロダクト機能における大きな転換点が、旗艦モデルへのネイティブ視覚（ビジョン）統合と「汎用コーディング（Generalized Coding／泛Coding）」の打ち出しである[3]。前世代では旗艦のGLM-5.3がテキスト専用であり、画像等のマルチモーダル入力は軽量版の「GLM-5.3-Flash」（総320B／活性18B）が担う分業体制だったため、ZCode上でもスクリーンショット理解などは5.3-Flashが併用されていた[12]。GLM-5.4がビジョンを本体に統合したことで、100万トークンの大規模コードベースとUIモックアッ

プ、エラーダイアログのスクリーンショット、設計ホワイトボード画像を単一プロンプトで同時に渡し、画面崩れの視覚検証やコンポーネント実装を行うワークフローが一本化された[12]。なお、中国国内のBigModel公式ドキュメントでは入力モダリティとして「動画・画像・テキスト・ファイル」の4種が明記されている一方、国際版Coding Planの初期案内や外部検証ではテキスト・画像入力を中心に取り扱っており、対応画像サイズ上限や視覚ベンチマーク単体の公式スコアは発表時点では開示されていない[10]。

この視覚認識と長連鎖コーディング能力を掛け合わせてZ.aiが提唱する「汎用コーディング」は、プログラムコードをソフトウェア開発にとどまらず、外部の専門ソフトや物理設計を制御する汎用プロトコルとして扱う構想である[3]。Z.aiは発表に際し、オープンソースCAD「FreeCAD」のPython APIを制御して工学基準に沿った6軸ロボットアームをモデリングしたうえで「Blender」で動作レンダリングと音声合成まで行いシミュレーション動画を出力した事例をはじめ、純粋な数学的幾何学とカスタムGLSLシェーダーによるインタラクティブな3D野鳥観察シーンの構築、企画・編曲から描画までを一気通貫処理した自己紹介ミュージックビデオ『Hello World』の制作、物理演算や戦闘システムを備えた完全なゲームの直接構築という4つのデモンストレーションを披露した[3]。（開発者コミュニティの一部からは、推論や設計状態をコードとして外部化することで長文脈の劣化やコンテキスト圧縮を乗り切る実用的なアプローチだとの評価と批評の双方が寄せられている[37]。）

しかし、これらの公式デモと第三者による実務テストの評価には明確な適性差が表れている。DataCampやAICodeKingの独立検証によれば、テキスト主導で段階的に論理を積み上げるGLM-5.4の推論スタイルは、素早い視覚的試行錯誤や空間的調整を要するフロントエンド開発や3D開発には不向きと結論づけられており、大量のコードを一括執筆するエンドツーエンド生成役としても時間・トークン効率が悪い[8]。対照的に独立評価で高く評価されたのが、セキュリティ脆弱性監査・深層バグ検出・事前アーキテクチャ設計の役割である[8]。ほぼ全ステップで推論を行う徹底した慎重さにより、Anthropicの「Claude Fable 5.2」が見逃した微細なアプリケーションバグや脆弱性を検出したほか、ツール呼び出しでは初回失敗率こそ1.56%（17環境中11位）だったものの、失敗後のリトライによる自己復旧率が66%とテスト対象の全Claude系・Codex系モデルを上回る粘り強さを記録した[8]。

想定ユースケース	外部独立検証における実務適性	推奨**Effort**設定・運用上の勘所
セキュリティ監査・脆弱性診断	最適：競合上位モデル（Claude Fable 5.2等）が見落とす深層バグや脆弱性を検出しやすい[8]	High（または Medium 指定）：Max でのタイムアウトを回避しつつ深い検証を維持[8]
リリース前コードレビュー・設計計画	最適：100万トークン文脈を活かした全体構造の分解や、他モデル実装に対する監査役として有効[8]	High：コード執筆者ではなく「セカンドレビュアー／プランナー」として併用[8]
長時間の自律エージェント・	適：初回ツール失敗率1.56%	High＋長めのタイムアウト設

ツール連携	に対しリトライ復旧率66%と高い自己修復力を発揮[8]	定＋リトライ有効化（Maxは所要時間が約4倍に長期化）[8]
フロントエンド**UI・3D**開発・大量一括実装	非推奨：GLSL等の公式デモに反し、高速な視覚的反復や大量コードの一括生成には不向き[8]	他の高速・視覚特化モデルを主軸に採用することを推奨[8]

提供形態・料金エコノミクス・オープンウェイト戦略の実務的影響

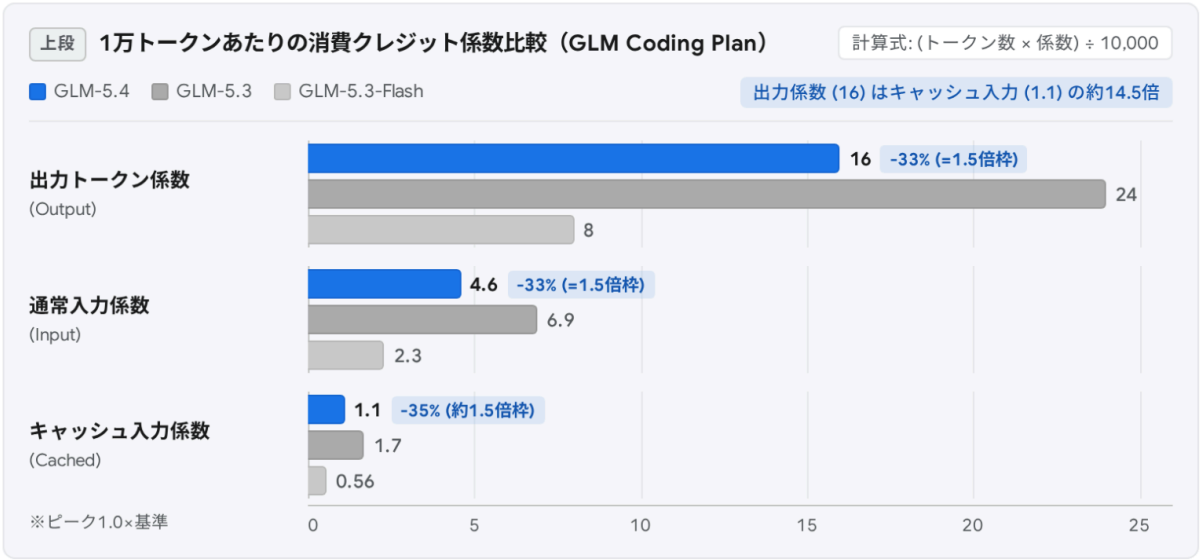
2026年11月23日時点でGLM-5.4を利用できる公式チャネルは、「GLM Coding Plan」「ZCode」「Z.ai Chat」（中国本土はBigModelアカウントと自動連携）の3つである[14]。GLM Coding Planは個人向けにLite（月額18米ドル／年払い換算月額12.60米ドル、5時間枠2,000クレジット・週枠10,000クレジット）、Pro（月額80米ドル／年払い月額56米ドル、5時間枠12,000・週枠60,000クレジット）、Max（月額168米ドル／年払い月額117.60米ドル、5時間枠28,000・週枠140,000クレジット）の3プランと法人向けTeamプラン（最低2シート～）を提供し、ZCodeやClaude Code、OpenCode、Cline、Codexなど20以上のエージェントツールで利用できる[14]。外部APIアグリゲータでは、AIHubMixが同日中にモデルID coding-glm-5.4 として100万トークンあたり未キャッシュ入力0.06米ドル／キャッシュ入力0.015米ドル／出力0.22米ドルで提供を開始した一方、OpenRouterでは同日時点で z-ai/glm-5.4 が未収載（404）であり、Z.aiおよびBigModel公式の従量課金API単価表も安全審査完了まで未公表となっている（参考として前世代GLM-5.3の公式API単価は100万トークンあたり入力1.40米ドル／キャッシュ0.26米ドル／出力4.40米ドル、中国BigModelでは入力8元／キャッシュ2元／出力28元）[10]。

注目された「GLM-5.3比で週次利用枠が1.5倍（50%増）」という施策の仕組みは、付与クレジット総数の変更ではなく、**GLM-5.4**のトークンあたりクレジット消費係数を引き下げたことによる実質ディスカウントである[9]。2026年7月30日導入のクレジット制プランにおけるモデル消費計算式は以下の通り定義されている[14]。

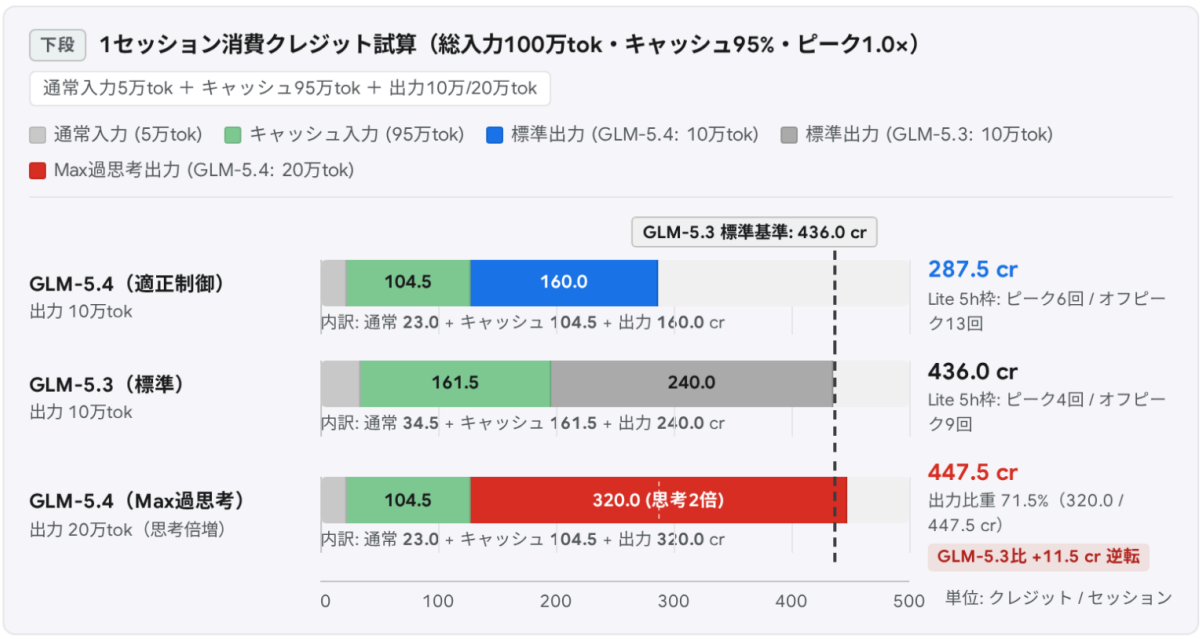
$$\text{消費クレジット} = \frac{(\text{未キャッシュ入力} \times M_{\text{in}}) + (\text{キャッシュ入力}}{10,000}$$

GLM-5.3の係数（入力6.9／キャッシュ1.7／出力24）に対し、GLM-5.4は入力4.6／キャッシュ1.1／出力16へと約3分の2に引き下げられており、同一トークン構成であれば1クレジットあたり1.5倍のトークンを処理できる[14]。しかし、出力トークンの係数（16）がキャッシュ入力係数（1.1）の約14.5倍と極めて重いため、デフォルトの max 設定でマラソン思考が発生して出力（推論）トークンが倍増すると、単価割引が相殺されてGLM-5.3よりむしろ高コストになる逆転現象が生じる[14]。

クレジット係数引き下げによる実質1.5倍枠と「過思考」が招くコスト逆転



▼ 単価係数は一律約3分の2（実質1.5倍枠）だが、デフォルトのMax effortで出力トークンが倍増すると出力係数（16）の重みが直撃



上段はGLM Coding Plan（クレジット制）における1万トークンあたりの消費係数比較であり、GLM-5.4はすべての区分でGLM-5.3比約3分の2に設定されているため、同一出力長であれば1.5倍のトークンを消費できる。しかし下段の1セッション試算（総入力100万トークン［うちキャッシュヒット95%＝95万トークン］・ピーク時間帯1.0倍レート適用）が示す通り、デフォルトのMax設定で思考連鎖が長期化して出力トークンが10万から20万へ倍増した場合、出力係数（16）の重みによって総消費量は447.5クレジットに跳ね上がり、GLM-5.3の標準コスト（436.0クレジット）を逆転する。（なお、平日15:00〜19:00 JST以外のオフピークおよび週末終日は上記消費クレジットが一律半額となる。）

データ出典: Z.ai Developer Docs (GLM Coding Plan Overview)、 Z.ai Developer Docs (Usage Revision)、 Fello AI

Effort設定を適正化して過思考を防げば、キャッシュヒット率95%想定でのGLM-5.4の週間推論可能量は、Liteプランで7,200万〜1億4,600万トークン、Proプランで4億4,000万〜8億7,000万トークン、Maxプランで10億1,400万〜20億2,800万トークンに達する（旧Legacy V2プランでもGLM-5.3のオフ

ピーク1倍／ピーク3倍に対し、GLM-5.4はオフピーク0.67倍／ピーク2倍へ優遇されている）[14]。さらに日本企業にとって有利な条件として、Z.aiのピーク時間帯は「月曜～金曜の14:00～18:00（UTC+8＝シンガポール標準時）」、すなわち日本時間（**JST**）の平日**15:00～19:00**の**4**時間のみに設定されており、日本の午前中・19時以降の夜間・週末終日はすべてオフピーク割引（消費クレジット半額、公式ドキュメント試算では将来のGLM-5.4標準API従量課金比で最大92%節約）が適用されるほか、ZCodeの「Idle-time tasks（空き容量時間帯の無料キュー実行）」も活用できる[14]。

他方、厳格なデータガバナンスからオンプレミス運用や専有クラウドでのセルフホストを求める企業にとっては、オープンウェイト未公開とライセンス方針の変化が最大の障壁となる[8]。Z.aiはGLM-5からGLM-5.2までApache-2.0やMITといった寛容なライセンスで重みを即時公開してきたが、前世代GLM-5.3では2週間のサイバー安全審査を経て重みを公開し、100億米ドル規模の商用モデルホスティング閾値でZ.aiの事前承認を求める独自「GLM-5.3 License」へと制限を強めた[14]。その背景には、GLM-5.3がサイバー評価CyberGymで84.5（中国語資料では83.5%）を記録し、2026年9月29日付の米Anthropicによる調査レポート『GLM-5.3 and the spread of advanced cyber capabilities』において、同社の「Claude Mythos Preview」に迫る自律的エクスプロイト開発能力が攻撃者の脅威水準を一段引き上げると警告された経緯がある[7]。脆弱性検出能力がさらに向上したGLM-5.4において重み公開が「安全性強化の完了後」とされているのもこの文脈にあり、視覚エンコーダを含む完全な重みが公開されるかどうかも現時点では未確定である[8]。

将来重みが公開された場合でも、同規模（約753B MoE、アクティブ約40B）の前世代GLM-5.3の配備実績から推計すると、本番推論標準のFP8精度で約744GBのVRAMが必要となるため、単一ノード運用にはNVIDIA H200（141GB）×8基（計1,128GB VRAM）またはB200×8基（計1,536GB VRAM）クラスのGPUクラスタが求められ、H100（80GB）×8基（計640GB）ではINT4量子化（約372GB）に落とす必要がある（BF16フル精度は約1.5TBを要しB300×8基クラスに限定される）[22]。さらにGLM-5シリーズはvLLMやSGLang（NVIDIAおよびHuawei Ascend環境）に対応しているものの、既存のglm_moe_dsa 実装は単一パス前提であるため、28層2周ループ・トークン単位ゲート・128トークン局所ウィンドウ併用KV共有を効率処理する推論エンジン側の対応完了を待つ必要がある（現時点で直ちにMITライセンス下でセルフホスト可能な同世代代替としては「DeepSeek-V4.1」系が挙げられる）[8]。

日本企業が日本語環境でテスト導入するにあたっては、公式ドキュメント上の対応言語表記が「英語・中国語（English & Chinese）」であり、日本語性能の公式ベンチマークが未公表である点にも留意したい[1]。ZCodeの最新更新では日本語・中国語混在時のフォント描画不具合が修正されており、日本語での指示自体は可能だが、前世代でも日本語特有の言葉遊び等には弱さが指摘されていたため、まずはコード監査や設計レビューを中心に実務検証を行うのが堅実である[1]。Coding Plan経由で検証環境を構築する際の具体的な推奨設定は以下の4ステップに集約される。

1. **Coding Plan**専用**API**キーの発行とエンドポイント設定: 2026年10月16日以降、定額プラン用キーは従量課金キーと分離されているため、「Personal / Team Coding Plan」画面で専用キーを発行する。接続先はClaude Code（Anthropic互換）では https://api.z.ai/api/anthropic とモデル名 glm-5.4[1m]、Cline等のOpenAI互換ツールでは https://api.z.ai/api/coding/paas/v4 と glm-5.4（画像サポート有効化）、OpenCodeでは zai-coding-plan/glm-5.4 を指定する[8]。
2. クライアント側コンテキスト圧縮閾値の制限: モデルの最大文脈は100万トークンだが、実用上のコーディング精度を維持するため、クライアント側の自動圧縮設定（例：CLAUDE_CODE_AUTO_COMPACT_WINDOW）は 512000（51.2万トークン）に設定することが中国語公式ガイドおよび外部検証で推奨されている[8]。

3. 思考強度（**Effort**）の明示的指定: デフォルトの max によるタイムアウトと出力クレジット急増を防ぐため、Claude Code等のセッション内または設定ファイルで /effort high（定型実装では medium や low）を明示指定する[8]。
4. 長時間推論に備えたタイムアウト値の拡張: 大規模リポジトリの監査では1,000ステップ超の反復検証を行う場合があるため、クライアント側のAPIタイムアウト（例：API_TIMEOUT_MS: 3000000）を標準値より十分に長く設定する[8]。

業界潮流における位置づけと「3兆パラメータ級GLM-5.5」への布石

技術史の文脈で見ると、GLM-5.4の発表は2026年後半に世界のフロンティアAI開発で急速に台頭している「Loop Transformer（再帰的深度スケーリング）」潮流の象徴的なマイルストーンである[12]。重み共有層を反復適用する着想は2018年の「Universal Transformer」（Dehghani et al.）や2019年の「ALBERT」に起源を持つが、2025年にSaunshi et al.が「k層のモデルをL回ループさせるとk×L層の標準モデルに匹敵する推論性能を示し、T回のループが潜在空間内でTステップの明示的CoTをシミュレートできる」ことを理論・実験で示し、Geiping et al.（Huginn-0125／Recurrent-Depth）、Zhu et al.（Ouro／LoopLM）、Bae et al.（Mixture-of-Recursions）らの研究が続いたことで実用段階に入った[19]。2026年秋には、OpenAIの「GPT-6 Astra」（*The Information* や *The Decoder* 等でループ構造の採用が報じられているがOpenAI公式では構造非公表）をはじめ、Thinking Machinesのローカル向けMoEモデル「Inkling 2 Mini」、Nanbeige LLM Labの「Nanbeige 4.2-3B」、SenseTime系の「SenseNova-Loop-9B」が相次いで登場しており、その中でGLM-5.4は構造が公式開示されたループ型モデルとして世界最大の750B規模となる[7]。

モデル名（開発元）	発表・確認時期	パラメータ構成	ループ構成・実効深度	学習手法・KVキャッシュ設計の特徴
GLM-5.4（Z.ai／智譜AI）	2026年11月23日[3]	750Bスケール MoE（ベース約40Bアクティブ）[8]	物理78層中、中間28層を2周＝**106**層パス[21]	事前学習済みモデルの継続学習適応（+約10%計算量）／大域共有＋128トークン局所独立KV[21]
Inkling 2 Mini（Thinking Machines）	2026年11月[19]	保存26.4B MoE／1パス3.4B／実効5.4Bアクティブ[19]	物理24層中、中間12層を3周＝**48**層パス[19]	32GB以上RAMのMac等でローカル動作／llama.cppではパス別KV確保で大域キャッシュ倍増[7]

Nanbeige 4.2-3B（Nanbeige LLM Lab）	2026年[54]	非埋め込み 約3B（Dense）[54]	物理22層全体を2周＝**44**層パス[54]	28兆トークンでゼロから事前学習（4周は不安定化と報告）／パスごとに独立KV保持[54]
GPT-6 Astra（OpenAI）	2026年9月以降[7]	非公表（クローズド旗艦）[12]	ループ構造・再帰深度の採用が一部メディアで報道（公式未確認）[7]	サイバーCritical閾値到達に伴いCoT監視モニタを導入・潜在推論の可監査性が議論に[37]

このループ型アーキテクチャの普及は、同時にAI安全性と解釈可能性（Interpretability）をめぐる議論を呼んでいる。明示的なテキストCoTであれば人間や安全監視モデルが推論過程を監査できる（OpenAIもAstraのサイバー対策としてCoTモニタへの投資を公表している）のに対し、内部ループ回数を過度に増やして推論を連続的な隠れ状態（「neuralese」）へ移行させると、思考過程の監査が困難になるという懸念が研究者コミュニティで指摘されている[37]。またNanbeige 4.2-3Bの実験でも、2周ループが高い効率を示した一方、4周への増加は学習の減速や最適化の不安定化を招いたと報告されており、GLM-5.4がループを「固定2周」にとどめつつ明示的CoT出力も併用している設計は、学習安定性・推論コスト・監査可能性の現実的な均衡点といえる[37]。

さらに中国のフロンティアAI競争という観点では、Moonshot AIの「Kimi K3.1 Pro」（2.8兆パラメータ）やAlibabaの「Qwen3.8-Max」（2.4兆パラメータ）および「Qwen4 Max」など、総パラメータ2兆〜3兆（2T〜3T）級への巨大化が進む中、先端GPUやHBMメモリ帯域の制約下で3分の1以下のメモリフットプリントから2T〜3T級に迫る推論深度を引き出すLoop Transformerの戦略的価値は大きい[7]。しかもZ.aiは公式ブログの結びにおいて、「計算深度のスケーリングとモデル規模のスケーリングのどちらか一方を選ぶのではなく、両方を同時に追求している。GLM-5.4は計算深度を実証したものであり、ここで得たすべての知見を引き継ぐ、大幅に巨大な次期モデルがまもなく完成する」と明言した[3]。複数の中国テクノロジー・金融メディアの報道や経営陣の発言によれば、Z.aiは現行ベースの約4倍にあたる約**3**兆（**3T**）パラメータ規模の次期フラッグシップ「**GLM-5.5**」を**2026**年**12**月にも投入する見通しと伝えられており、GLM-5.4はその超巨大ループ型モデルへ向けた重要な技術的布石と位置づけられる[4]。

引用文献

1. GLM-5.4とは？Z.ai初のLoop Transformerモデルの特徴・料金, https://ai-revolution.co.jp/media/what-is-glm-5-4/
2. 脆弱性の監視サービス | 公開情報の確認とお知らせ・AIセキュリティ, https://crystal-method.com/vulnerability-watch/
3. GLM-5.4: Toward Generalized Coding with Our First Loop Transformer, https://z.ai/blog/glm-5.4

4. ZhipuのGLM-5.4がCoding Planで提供開始、視覚理解を追加し週次枠を5割拡大, https://finance.biggo.jp/news/49b1fc82-8bc5-4894-8a94-b660e0635458
5. ZAI Co Ltd、Loop Transformerアーキテクチャ採用の新モデル「GLM-5.4」を発表, https://jp.investing.com/news/stock-market-news/article-93CH-1767829
6. Z.aiのGLM-5.4、パラメータを増やさず反復処理で推論を深める, https://www.unite.ai/ja/z-ais-glm-5-4-uses-iterative-depth-without-adding-parameters/
7. https://the-decoder.com/z-ai-follows-openai-into-loop-transformer-territory-with-its-new-750b-flagship-model/
8. GLM-5.4: Z.ai's First Loop Transformer and the Push ... - DataCamp, https://www.datacamp.com/blog/glm-5-4
9. Z.ai「GLM-5.4」を予告なしに公開、「しれっと」の声も——画像認識を初搭載、週次クォータ5割増に, https://hashout.jp/ai/14787/
10. GLM-5.4 - 智谱AI开放文档, https://docs.bigmodel.cn/cn/guide/models/glm-5.4
11. 刚刚，GLM-5.4上线！全新的Loop Transformer架构亮相, https://finance.sina.com.cn/tech/roll/2026-11-23/doc-iniztmua8293720.shtml
12. https://oneaiapikey.com/resources/blog/glm-5-4-kingbench
13. Overview - Z.AI DEVELOPER DOCUMENT, https://docs.z.ai/guides/overview/overview
14. GLM 5.4: Z.ai's Loop Transformer, Access and Real Cost - Fello AI, https://felloai.com/glm-5-4/
15. Z.ai launches GLM-5.4 and makes 28 layers pull a double shift, https://digg.com/tech/bo3tksri
16. GLM-5.4: Toward Generalized Coding with Our First Loop Transformer, https://www.reddit.com/r/LocalLLaMA/comments/1y3enps/glm54_toward_generalized_coding_with_our_first/
17. GLM-5.3とは？性能や料金、GPT-6・Claude Fable 5.1との違いを解説, https://www.ai-souken.com/article/what-is-glm-5-3
18. GLM-5.2とは？Z.aiの100万トークンコンテキスト・オープン, https://www.orcarouter.ai/ja/blog/what-is-glm-5-2
19. https://vdf.ai/resources/loop-transformer/
20. GLM-5、GLM-5.1、GLM-5.2、GLM-5.3 - Hugging Face, https://huggingface.co/docs/transformers/ja/model_doc/glm_moe_dsa
21. GLM-5.4 Explained: How Z.ai Turned a 750B Model Into a Loop, https://dev.to/jamilxt/glm-54-explained-how-zai-turned-a-750b-model-into-a-loop-transformer-53dl
22. GLM 5.3 Hardware Requirements: GPU, Memory, and Deployment, https://www.yottalabs.ai/post/glm-5-3-hardware-requirements-gpu-memory-2026
23. https://aihubmix.com/blog/ja/glm-5-4-api-access-should-you-wait-for-openrouter
24. GLM-5.3-Flashとは？料金・無料枠・使い方と性能を解説【2026年, https://ai-revolution.co.jp/media/what-is-glm-5-3-flash/
25. GLM-5.4 launches on Z.ai's Coding Plan with vision, https://pub1j5sn0fits0.mybusinesses.blog/nhl/teams/NYR/new-york-rangers/

26. GLM-5.3 のサイバーセキュリティ能力：ベンチマークの主張を読む，https://evolink.ai/ja/blog/glm-5-3-cybersecurity
27. GLM 5.3 の実力と使い方：ベンチマーク、API提供状況、重み公開の，https://ofox.ai/ja/blog/glm-5-3-benchmarks-access-2026/
28. GLM-5.4 scores 33.2 on KingBench 4, the top open-weight, https://daily.dev/posts/glm-5-4-scores-33-2-on-kingbench-4-the-top-open-weight-coding-result-so-far-fahtrnytv
29. GLM 5.4 is here - Medium, https://medium.com/data-science-in-your-pocket/glm-5-4-is-here-8c3e588d3e79
30. Z.ai launches GLM-5.4, a 750 billion-parameter model built ... - MLQ.ai, https://mlq.ai/news/zai-launches-glm-54-a-750-billion-parameter-model-built-on-a-looped-transformer/
31. GLM-5.3: Specs, Pricing, API & Free Chat, https://glm-ai.chat/models/glm-5-3/
32. GLM 5.3 Parameters: 753B, 1M Context Window - Layer3Labs, https://www.layer3labs.io/guides/glm-5-3-explained
33. Z.ai Opens GLM-5.3 Weights With a $10 Billion Security Gate, https://magica.com/news/glm-5-3-open-weights-commercial-use-safety-review
34. New Released - Overview - Z.AI DEVELOPER DOCUMENT, https://docs.z.ai/release-notes/new-released
35. GLM-5.3を徹底解説:ベースモデルは同じなのに性能50%増 - note, https://note.com/ai_driven/n/ndbf047521ff4
36. Z.aiがGLM-5.4を発表、7500億パラメータ初のループ型モデル，https://pasqualepillitteri.it/ja/news/36246/glm-5-4-z-ai-loop-transformer-ja
37. https://news.ycombinator.com/item?id=50609685
38. Chat Completion - Overview - Z.AI DEVELOPER DOCUMENT, https://docs.z.ai/api-reference/llm/chat-completion
39. Z.ai's Official Agent Harness for GLM-5.4 - ZCode, https://zcode.z.ai/en
40. ZCode Docs | GLM-5.4 Agentic Coding Guide - Z.ai, https://zcode.z.ai/en/docs/welcome
41. Overview - Z.AI DEVELOPER DOCUMENT, https://docs.z.ai/devpack/overview
42. Team Plan Benefits - Overview - Z.AI DEVELOPER DOCUMENT, https://docs.z.ai/devpack/teamplan
43. Plan Update Announcement - Overview - Z.ai docs, https://docs.z.ai/devpack/notice/usage-revision
44. zai-org/GLM-5 - From Vibe Coding to Agentic Engineering - GitHub, https://github.com/zai-org/GLM-5
45. Anthropic Warns GLM-5.3 Builds Exploits Like Mythos, https://www.trendingtopics.eu/anthropic-glm-5-3-cyber-warning/
46. Reading Anthropic's GLM-5.3 Report Alongside Opus, GPT, Fable, https://note.com/zephel01/n/ncd4d8755f15e?hl=en
47. GLM-5.4 and loop transformers to the rescue of bandwidth, https://forums.developer.nvidia.com/t/glm-5-4-and-loop-transformers-to-the-rescue-of-bandwidth-constrained-environments/389904
48. ZCode Releases & Updates | GLM-5.4 Coding Agent,

https://zcode.z.ai/en/changelog
49. GLM-5.3が公開！！日本語の言葉遊びはイマイチだった - note, https://note.com/currypurin/n/n66c0bc334e54
50. Tool Integration - Overview - Z.AI DEVELOPER DOCUMENT, https://docs.z.ai/devpack/tool/others
51. 套餐概览 - 智谱AI开放文档, https://docs.bigmodel.cn/cn/coding-plan/overview
52. Claude Code - Overview - Z.AI DEVELOPER DOCUMENT, https://docs.z.ai/devpack/tool/claude
53. SpiralFormer: Looped Transformers Can Learn Hierarchical, https://aclanthology.org/2026.emnlp-main.1602.pdf
54. https://blogs.gruheshkurra.com/blog/looped-transformers-explained/
55. GLM-5.4: Z.ai's Loop Transformer, Access and Limits - i10X News, https://i10x.ai/news/glm-5-4-zai-loop-transformer-coding-plan
56. 智谱最新模型新增视觉能力，小参数实现高智能水平 - 网易, https://www.163.com/dy/article/LCESFV0A05129QAF.html
57. 智谱最新模型新增视觉能力，小参数实现高智能水平 - 搜狐, https://www.sohu.com/a/1105385965_161795