Z.ai「GLM-5.4」発表に関する調査報告

事実確認・「ループ型」の技術内容・性能・評価・今後の影響（2026 年 11 月 23 日発表）

調査時点：2026 年 11 月 24 日（発表翌日）

Claude Fable 5.2

---

1. 結論の要約

提示された記述「生成 AI 開発の Z.ai は 11 月 23 日、ループ型を初採用した旗艦モデル『GLM-5.4』を発表した」は、**概ね正確である。ただし「ループ型を初採用」は「Z.ai（GLM シリーズ）として初」の意味に限って正しく、業界初ではない。**

Z.ai は 2026 年 11 月 23 日、公式技術ブログ[1]と開発者ドキュメント[2, 3]で、GLM-5.4 を同社初の Loop Transformer アーキテクチャによる旗艦（flagship）モデルとして公表した[4, 5, 6]。一方、750B 本体の公式ベンチマークは未公表で、API・オープンウェイトも未提供であり[7, 8]、性能評価は現時点で「自己申告＋単一の第三者評価」に依存している[9]。

1.1 要点

- **事実関係**：Z.ai（智譜）が 2026 年 11 月 23 日に GLM-5.4 を公表したことは、公式ドキュメント、公式ブログのメタ情報、同日付の複数報道で確認できる[1, 2, 4]。「旗艦」も公式表現である[2]。「ループ型」は Loop Transformer（重み共有層の反復実行）の訳で、GLM-5 ベースの 28 層（第 27〜54 層）を 2 周させる構造を指す[2, 10]。ただし「初」は同社初にとどまる。Universal Transformer（2018 年）以来の研究系譜があり[11]、Thinking Machines の「Inkling 2 Mini」（2026 年 11 月 20 日、オープンウェイトのループ型）が 3 日先行している[12]。
- **性能**：Z.ai は「2〜3 兆パラメータ級に匹敵」と主張するが[2, 5]、750B 本体のベンチマーク表は未公表で、公開数値は 30B 検証モデルの比較実験のみである[4, 10]。唯一の第三者評価（AICodeKing の私設ベンチ「KingBench 4」）では 33.2/60 で 19 モデル中 7 位、オープン系首位相当だが[8, 13]、所要時間が極端に長い（いわゆる「マラソン思考」）[8, 9]。

- **実務上の示唆**：現時点の提供は GLM Coding Plan（月額 18 ドル〜）、ZCode、Z.ai Chat 経由に限られ、API・重みは安全評価後とされる[6, 14]。智譜は米エンティティリスト掲載企業である[15]。日本企業が研究開発・知財業務で使う場合は、ホスト型サービスへの機密・輸出管理対象技術の入力を避け、重み公開後の閉域運用を前提に評価するのが妥当と考えられる（分析）。

1.2 記述要素ごとの判定

表 1　提示された記述の要素別判定

要素	判定	根拠・留保
① Z.ai が	**確認済み**	開発者ドキュメント（docs.z.ai、中国向け docs.bigmodel.cn）と公式ブログで確認[1, 2, 3]。Z.ai（旧称 Zhipu AI／智譜）は香港上場企業（証券コード 2513）[16, 17]。
② 11 月 23 日に	**確認済み**	同日付の複数報道と公式ブログのメタ情報が 2026 年 11 月 23 日を示す[1, 4, 5]。中国時間の同日午後に Coding Plan での提供開始が報じられ[18, 19]、技術詳細に基づく報道は同日夜に出た[6, 20]。日本時間でも同日（時差 1 時間）。
③ ループ型を初採用した	**一部確認**	「同社初の Loop Transformer 旗艦」は公式表現[2, 21]。ただし業界初ではない（Universal Transformer 2018 年[11]、Inkling 2 Mini 11 月 20 日[12]、GPT-6 Astra に関する報道[22]）。中国メディアの「国内初」は報道側の表現である[23, 24, 25]。
④ 旗艦モデル	**確認済み**	公式ドキュメントが旗艦モデルと明記している[2]。報道側の誇張ではない。
⑤ GLM-5.4 を発表	**確認済み**	正式名「GLM-5.4」、モデル ID は glm-5.4（Claude Code では 1M コンテキスト版の glm-5.4[1m]）[3]。

日本語報道の出典は、Yahoo!ニュース配信のビジネス＋IT 記事と判断される（リード文が提示された記述と一致）[26]。同記事の記載（7500 億パラメータ、28 層を 2 回反復、継続学習による

改造、マラソン思考の自認）は公式情報と整合する[2, 5]。ただし同記事の本文は自動取得が拒否されたため、検索結果の抜粋による照合にとどまる。

2. 確認された事実

2.1 発表の経緯・形態

- **発表形態**：①Z.ai 責任者の李子玄（Zixuan Li）氏による「全 GLM Coding Plan 利用者に提供」との告知[19]、②公式 X アカウントによる告知[27]、③技術ブログ「Toward Generalized Coding with Our First Loop Transformer」[1, 28]、④開発者ドキュメント（モデルページ、Coding Plan 概要）[2, 14]。モデルカードや公式ベンチマークは公表されていない[8]。
- **時系列の注意点**：中国メディアは当日午後の時点で「Coding Plan に上線、公式発表・ベンチマークは未公表」と報じており[19, 25]、香港市場の取引時間中は技術詳細が出揃っていなかった。英文メディアは同日、技術ブログに基づいて報じている[4, 5, 6]。
- **系列関係**：GLM-5（2026 年 2 月）→GLM-5.1（3 月下旬提供、4 月 7 日オープン化）→GLM-5.2（6 月）→GLM-5.3（8 月 14 日。ベースは据え置きで事後学習を強化、重みは 8 月 28 日公開）→GLM-5.3-Flash→GLM-5.4（11 月 23 日）[16, 17, 29]。GLM-5 ベースは MoE 構成で総パラメータ約 744B とされるが[30]、出典により 743B〜753B の幅がある[16, 25, 29]。GLM-5.4 は 5.3 と異なり、ベースモデル自体の構造を変えた点で「アーキテクチャ世代交代」に当たる（分析）。

2.2 「ループ型」の正体（アーキテクチャ上の変更）

「ループ型」は英語の Loop Transformer の訳である。中国語の公式ページも英語のまま「Loop Transformer」と表記している[2]。**エージェントの計画・実行・検証ループではなく、モデル内部の計算構造の変更**である[2, 5]。

- **仕組み**：重みを共有した一群の層を 1 トークンの処理中に複数回実行し、表現を磨き上げる。GLM-5.4 は GLM-5 ベースの 28 層を 2 周させる[2, 5]。ループ対象は第 27〜54 層で[10]、保存されている層は計 78 層、1 トークンあたりの層通過は 106 回となる[21]。パラメータを増やさずに実効的な深さが約 36%増える計算である（106÷78 の単純計算）。

- **作り方**：ゼロからの事前学習ではなく、事前学習済みの GLM-5 ベースを継続学習でループ構造に改造した[5, 31]。Z.ai は、ゼロから学習したループモデルは同一計算量では非共有モデルに及ばないと説明している[5]。1B→4B→30B→750B と段階的に検証した[5, 6]。
- **750B 規模での 3 つの工夫**[5, 8]：
  - ゲート付き漸進更新：周回ごと・トークンごとのゲートで残差ストリームへの書き戻し量を制御する。1 周目は控えめに書き、2 周目でゲートを開いて精緻化する。難しいトークンほど 1 周目の書き込みが強くなる。
  - ハイブリッド KV キャッシュ：2 周目は 1 周目の疎な KV を再利用し、自前の KV はローカルのスライディングウィンドウ内だけに持つ。追加ループのキャッシュ増は「全文脈分ではなくウィンドウ 1 つ分」にとどまる。1M コンテキストでも 5.3 比のコスト増は制御可能とされるが、具体的な数値は示されていない[32]。
  - 疎アテンション索引の再利用：2 周目は 1 周目の索引を流用し、索引の追加オーバーヘッドを生じさせない。
  - あわせてカーネル融合により、ループで増えるメモリ読み書きを削減した[5, 33]。
- **位置づけ**：Z.ai は「汎 Coding（Generalized Coding）」を掲げ、ソフトウェアにとどまらず動画・ゲーム・3D・機械設計などへの拡張を打ち出している[2, 34]。調査報告によれば、公式ブログは「記憶集約型タスクにはパラメータ増が、推論集約型タスクには実効計算量増が効く」との考え方を示している[1]。

2.3 仕様（公式に確認できた範囲）

表 2　GLM-5.4 の主な仕様

項目	内容	確度・出典
規模	「750B 規模」（ベースは GLM-5）。アクティブパラメータは非公表	公式（概数）[5, 21]
コンテキスト長	1M（100 万）トークン	公式[2]
最大出力	128K トークン	公式[2]

項目	内容	確度・出典
入出力モダリティ	入力：テキスト・画像・動画・ファイル／出力：テキスト	公式[2]
推論（thinking）	実効 3 段階（low／high／max）。未指定時の既定は max。medium 指定は high に、thinking 無効化は low にマップされる	公式[3]
速度	オフピーク最大約 200 トークン/秒、ピーク時 70〜80 トークン/秒	公式（自己申告）[2]
機能	ストリーミング、Function Calling、コンテキストキャッシュ、構造化出力	公式[2]

情報源間の食い違い

- 規模：中関村在線・快科技は「7000 億（700B）」[23, 24]、智通財経は「745B 基座」[25]とする。公式は「750B 規模」である[21]。700B は発表前の情報に由来する可能性がある（推測）。
- コンテキスト：網易号（鞭牛士）の記事は「200K コンテキスト／最大出力 64K」とするが[18]、公式ドキュメントは 1M／128K であり[2]、公式を採る。公式ページには「30B モデルを 200K 長文 SFT 後に評価」との記述があり[2]、混同の可能性がある（推測）。百度百科にも同様の記載があるとされるが、再確認できなかった[35]。

2.4 性能

(a) 公式（自己申告）

- 「2〜3T 級モデルに匹敵する知能」との主張のみで[2]、**750B 本体のベンチマーク表は未公表**である[4, 6]。
- 公開数値は **30B 検証モデルの同一パラメータ比較**（ループ版の学習計算量は約 10%増）のみである[2, 4]。AIME25 は 19.2→30.4、GPQA-Diamond は 47.6→56.3、AA-LCR（長文脈）は 40.3→47.7、MMLU-Pro と IFBench は+3 点強[4]。平均出力長は多くのベンチマークで 10〜14%短くなった[5, 8]。推論の一部がループ内部に移ったためと説明されている[33]。

- Z.ai 自身が弱点として「マラソン思考」（過度に長い思考ループ）を認めている。ただし、追加の再確認が精度を押し上げる面もあると主張する[4, 5]。

(b) 第三者評価（現時点で実質 1 件）

- 開発者系レビュアーAICodeKing の私設ベンチ「KingBench 4」で、GLM-5.4（max）は **33.2/60、19 モデル中 7 位**、オープン系で首位[8, 10, 13]。
- 比較：GLM-5.3 は 27.5[13]、GPT-6.2 Astra は 33.6、Claude Opus 5.5 は 34.7[23, 24]、Gemini 4 Neon は 37.9[9]、首位は Claude Fable 5.2 の 42.0[9, 13]。調査報告は Kimi K3.1 Pro 32.2、DeepSeek V4.1 Flash 31.0、Qwen 4 Max 30.6 も挙げるが、元の投稿を再取得できず再確認していない[36]。
- コスト面：所要時間は Gemini 4 Neon の 109 分に対し **431 分**[8, 9]。ツール呼び出し失敗率は 1.56%（17 モデル中 11 位）だが、失敗の 66%をリトライで回復した[8]。キャッシュヒット率は 98.4%[13]。調査報告は推論ステップ（244 に対し 1,128）、推論トークン（約 7 万に対し約 104 万）も挙げるが、再確認していない[36]。なお daily.dev の要約は、別の厳格スキーマ検査（31 件）でツール呼び出し失敗 0%としており[13]、指標の定義が異なるとみられる。
- 快科技は、ほぼ全ステップで推論を行うためタスク完了が遅くなると伝えている[24]。
- 評価上の留保：単一レビュアーの独自ハーネスによる結果である[9]。Artificial Analysis など主要な第三者指標は現時点で未公表／未確認である[37]。

(c) 比較対象（2026 年 11 月時点の主要モデル）

- 米国：OpenAI GPT-6.2 Astra（11 月 17 日）、Google Gemini 4 Neon（11 月 18 日）[38]、Anthropic Claude Fable 5.2／Mythos 5.2（11 月 9 日）[39]、Claude Opus 5.5（9 月下旬）[38]、xAI Grok 4.7（DataCamp の言及のみ）[8]。
- 中国：Moonshot Kimi K3.1 Pro（10 月 23 日、総パラメータ 2.8T）[40]、Alibaba Qwen3.8-Max（2.4T、オープンウェイト）[41]、Qwen4-Max（11 月 16 日）[42]、DeepSeek V4.1 Flash[43]。
- 位置づけ：GLM-5.4 は、2〜3T 級のモデルを約 1/3〜1/4 の規模で狙うモデルと整理できる（分析）。

(d) 前世代からの改善

- KingBench 4 で+5.7 点（27.5→33.2）[13]。Coding Plan のクレジット消費は 5.3 の約 2/3[7, 14]。速度・コストは公式の定性的主張にとどまる。

2.5 提供形態・価格・ライセンス

- **現在の提供先**：GLM Coding Plan（Lite／Pro／Max）、ZCode、Z.ai Chat[3, 4, 44]。Trae、WorkBuddy、Qoder などのコーディングツールにも接続されていると報じられている[20, 34]。
- **Coding Plan**：月額 18 ドルから[14]（Lite 18 ドル、Pro 80 ドル、Max 168 ドルとの報道がある[7]）。クレジット係数（1 万トークンあたり）は GLM-5.4 が入力 4.6／キャッシュ入力 1.1／出力 16、GLM-5.3 が 6.9／1.7／24 である[14, 45]。同じトークン量なら 5.3 の約 2/3 のクレジットで済み、「週次枠 1.5 倍」[6, 19]はこの係数設定と整合する。オフピークは 50%課金、ピークは平日 14:00〜18:00（UTC+8）[14, 45]。クライアントのコンテキスト設定は 512K が推奨されている[45]。
- **API**：安全評価の完了後に開放とされる[2, 6]。**単価は未公表**[7, 8]。参考として、GLM-5.3 の API 価格は入力 1.40 ドル／出力 4.40 ドル（100 万トークンあたり）である[29]。
- **オープンウェイト**：安全強化（safety hardening）の完了後に公開予定[5, 6]。**時期・ライセンスは未公表**[8]。11 月 24 日時点で Hugging Face の zai-org に GLM-5.4 のリポジトリはなく、最新は GLM-5.3 系である[46]。OpenRouter にも未掲載[8, 37]。先例として、GLM-5.3 は発表 14 日後に重みを公開した（独自ライセンス、Flash 版は MIT）[17, 29]。

2.6 評価・評判

- **肯定的評価**：中国メディアは「開源最強」「Opus 5.5 との差 5%以内」[24]、「200 の実戦シナリオで Coding 体感最強の国産モデル」[20]と報じた。調査報告によれば、AICodeKing は「現時点で最良のオープンモデル」と評している[36]。NVIDIA 開発者フォーラムでは、メモリ帯域が制約される環境（DGX Spark 等）に有利との議論がある[43]。
- **懐疑・批判**：Hacker News では、ベンチマークがないのは結果が芳しくないからではないかとの初期反応があった[28]。Wccftech は見出しでベンチマークの省略を指摘し[33]、The Decoder は 750B 本体のスコアも独立評価もないと明記した[4]。光子星球も、公開データは

主に社内評価で第三者検証が不足していると指摘する[32]。調査報告は、X 上に「Opus 5.5 から蒸留された文体」とする投稿があるとするが、根拠は文体比較のみの未検証の主張であり、本文書では該当投稿を再取得できていない。

- **資本市場**：智譜（02513.HK）は 11 月 23 日、場中に一時約 4%上昇した（346.8 香港ドル）[25]。調査報告は同日終値を 340.20 香港ドル（+1.86%）とするが、再確認できていない。52 週高値は 2,980 香港ドルで[47]、足元はそこから 9 割近く下落した水準にある。HSBC は 11 月中旬、目標株価を 1,375 香港ドルから大幅に引き下げた（調査報告では 500 香港ドル）[48]。恒生中国企業指数への採用（12 月 7 日発効）も材料である[25]。経営陣が 12 月に約 3 兆パラメータの「GLM-5.5」を投入予定と新浪科技・智通財経が報じているが[19, 25]、公式には未確認である。

3. 分析・推測（今後の影響）

以下は確認済みの事実に基づく分析であり、断定ではない。

1. **米中競争への含意**：智譜はエンティティリスト掲載により米国由来の先端技術の調達が制約されており[15]、「パラメータを増やさず計算の深さを増やす」ループ型は合理的な適応策と考えられる。GLM-5.4 はその最初の大規模実証と言える。第三者評価 1 件では GPT-6.2 Astra に肉薄するが、首位の Claude Fable 5.2 とは約 9 点差がある[13, 23]。「中国は数カ月遅れ」という構図を覆す証拠はまだない。むしろ、同等の結果を得るのに数倍の時間を要する点が実効コスト上の弱点である[8]。
2. **オープンウェイト・価格競争**：重みが公開されれば、Kimi K3.1 Pro（2.8T）[40]や Qwen3.8-Max（2.4T）[41]よりはるかに小さい 750B 級のオープンモデルとなり、自社ホスティングの経済性で優位に立つ可能性がある。ただし 5.3 と同様に公開が遅れ、独自ライセンスとなる可能性も考慮すべきである[29]。米 NIST の CAISI は 2026 年 9 月 17 日の評価で、GLM-5.3 を「これまでに公開された中で最もサイバー能力の高いオープンウェイトモデル」と位置づけた（ただし米国のフロンティアモデルより大幅に低く、約 4 カ月遅れとする）[17]。Anthropic も、GLM-5.3 のエクスプロイト構築能力が Claude Mythos Preview に迫ると分析している[49]。このため、公開判断はより慎重になる可能性がある。

3. **設計潮流**：11 月だけでも Thinking Machines（Inkling 2 Mini）[12]と Z.ai が相次いでループ型を出した。OpenAI の GPT-6 Astra もループ型（recurrent depth）と報じられたが、The Information の報道に基づくもので、OpenAI は確認していない[22, 50]。「パラメータのスケーリング」に加えて「深さ（ループ回数）のスケーリング」が新たな軸になりつつある。調査報告によれば、Z.ai 自身もトークン難度に応じてループ回数を変える適応計算を次の課題に挙げている[1]。一方、推論過程が出力トークンに現れにくくなりうるため、思考過程の監視可能性（CoT monitorability）をめぐる議論が強まると見込まれる[22]。もっとも、ループ化が監視可能性低下の直接の原因とは言い切れないとの指摘もある[50]。
4. **特許・知財**：ループ型アーキテクチャに関する智譜・他社の特許出願動向は**未確認**である。基本概念は 2018 年の Universal Transformer 論文で公知であり[11]、権利化の焦点は「事前学習済みモデルの継続学習によるループ化」「ゲート付き書き戻し」「ループ間 KV 共有＋スライディングウィンドウ」「疎アテンション索引の再利用」といった実装レベルの工夫に移ると考えられる。中国・米国の公開公報の監視を推奨する。

4. 推奨事項（日本企業・知財実務向け）

- **現時点では本番採用を見送り、評価対象リストに入れる**：公式ベンチマーク・API・重みがいずれも未提供で、判断材料が不足している[4, 8]。公式スコア、独立評価、重みとライセンスの公開を待って再評価するのが妥当である。
- **ホスト型（Coding Plan／Z.ai Chat）に機密情報を入れない**：智譜は 2025 年 1 月 16 日付の米連邦官報（90 FR 4617）でエンティティリストに追加された。EAR 対象の全品目について許可が必要で、審査方針は「原則不許可（presumption of denial）」である[15]。米法律事務所 Sheppard Mullin は 2026 年 7 月のブログで、輸出管理リスクを負う企業にとっては中国製 AI モデル（GLM 等）との関わり方自体が違反リスクになりうると指摘している[51]。調査報告によれば、同ブログは「公開重みのダウンロード自体は対 Zhipu 輸出に当たらないが、EAR 対象技術を Zhipu のホスト型 API に送ることは許可を要する輸出に当たる」と整理している（該当箇所は再取得できていない）。ホスト型 API のプロンプトは中国内の智譜イン

フラを経由するとの解説もある[52]。未公開の発明内容、技術ノウハウ、出願前の明細書草稿、米国由来の技術データは入力しないこと。

- **重み公開後はオンプレ／国内クラウドでの閉域運用を前提に検証する**：先行技術調査の補助、コードやアルゴリズムのレビュー、脆弱性指摘など、慎重さが評価されている用途での比較検証が有望である。長考による遅延を考慮し、エフォートは明示的に high から試すのがよい[8]。
- **社内ルールの整備**：モデル選定基準に「提供主体の所在国・規制リスト該当性」「データの処理地（中国本土か、シンガポール法人経由の国際サービスか）」[52]「ライセンス条件（MIT か独自か）」を明記する。
- **出願・技術動向の監視**：ループ型実装に関する特許公開公報（CN／US／WO）を定期的にウォッチし、自社の生成 AI 関連出願の新規性・進歩性評価に反映する。

5. 留保事項

- 発表翌日の調査であり、公式ベンチマーク、API 価格、重み公開時期、ライセンス、アクティブパラメータ数は**現時点で未公表**である。
- 公式技術ブログは本文を直接取得できず[1]、内容は同ブログに基づく複数の報道と、取得できた公式ドキュメントで照合した。英語版モデルページは取得エラーであった[53]。
- 性能の第三者データは単一レビュアーのものに限られる。中国メディアの「開源最強」「国内初」等の表現は宣伝的色彩が強い。
- GPT-6 Astra がループ型であるという点は報道に基づくもので、OpenAI は公式に確認していない。
- 日本語報道（ビジネス＋IT／Yahoo!ニュース）は本文全文を取得できず、検索抜粋で照合した。

6. 出典の再確認結果（本文書作成時）

本文書の作成にあたり、調査報告の出典を 2026 年 11 月 24 日に再取得し、参考文献リストに確認状況を付した。調査報告から修正・補足した点は次のとおりである。

- 米連邦官報の引用を、調査報告の「90 FR 4619」から「90 FR 4617」（4617〜4620 頁。智譜関連の掲載は 4620 頁）に改めた[15]。
- 実効深さ「約 36%増」は、調査報告では FourWeekMBA の試算としていたが、該当記事を特定できなかった。このため、層数（78 層、106 回通過）[21]からの単純計算として記載した。
- GLM-5.5（12 月、約 3 兆パラメータ）の報道元は、再確認できた新浪科技・智通財経とした[19, 25]（調査報告は第一財経・新浪）。
- ベースモデルの総パラメータ数は、出典間で 743B〜753B の幅があることを明記した。調査報告にあった「アクティブ 40B」「DeepSeek Sparse Attention 採用」は再確認できなかったため本文から外した。
- Claude Opus 5.5 の公表日（調査報告は 9 月 22 日）は、「9 月下旬」までの確認にとどまる[38]。
- 次の数値・記述は再確認できなかった：30B 検証モデルの平均出力長短縮の個別値（MMLU-Pro −14.9%、GPQA-Diamond −11.7%、AIME25 −2%）、KingBench 4 の一部数値（推論ステップ、推論トークン、Kimi K3.1 Pro 等のスコア）、11 月 23 日終値、Inkling 2 Mini の詳細仕様（総パラメータ 26.4B／アクティブ 3.4B、中間 12 層を 3 回ループ、Apache 2.0）、公式図の層構成（Prelude 26 層・Coda 24 層）、X 上の個別投稿の内容。
- 調査報告が言及した IT 之家・第一財経・鳳凰網の記事は出典 URL を特定できなかった（鳳凰網科技の内容は搜狐の転載で確認[20]）。

参考文献

確認状況の凡例：【確認済】本文書作成時に内容を取得して確認／【一部確認】書誌情報・メタ情報または検索抜粋のみ確認／【未確認】取得できず、調査報告の記載による。URL の閲覧日はいずれも 2026 年 11 月 24 日。

[1] Z.ai, “GLM-5.4: Toward Generalized Coding with Our First Loop Transformer”（公式技術ブログ）, 2026 年 11 月 23 日. https://z.ai/blog/glm-5.4 **【一部確認】**タイトル等のメタ情報のみ取得。本文は直接取得できず、同ブログに基づく報道で照合

[2] 智譜 AI 開放平台ドキュメント，「GLM-5.4」（モデル紹介ページ），日付表示なし．https://docs.bigmodel.cn/cn/guide/models/glm-5.4 **【確認済】**

[3] Z.ai Developer Docs, “How to Switch Models”, 日付表示なし. https://docs.z.ai/devpack/latest-model **【確認済】**

[4] The Decoder（Jonathan Kemper），“Z.ai follows OpenAI into loop transformer territory with its new 750B flagship model”, 2026 年 11 月 23 日. https://the-decoder.com/z-ai-follows-openai-into-loop-transformer-territory-with-its-new-750b-flagship-model/ **【確認済】**

[5] LavX News（Elena Varga），“Z.ai Introduces GLM-5.4, Its First Loop Transformer Model”, 2026 年 11 月 23 日. https://news.lavx.hu/article/z-ai-introduces-glm-5-4-its-first-loop-transformer-model **【確認済】**

[6] The Next Web（Ana Maria Constantin），“Z.ai’s GLM-5.4 runs its layers twice to rival 2T-parameter models”, 2026 年 11 月 23 日. https://thenextweb.com/news/zai-glm-5-4-loop-transformer-coding-model **【確認済】**

[7] Atomic Agent（Nadya Dudka, Andrew Dyuzhov），“GLM 5.4 vs GLM 5.3: What’s New, Price, Credits, and How to Use It”, 2026 年 11 月 23 日. https://atomicagent.io/blog/glm-5-4/ **【確認済】**

[8] DataCamp（Matt Crabtree），“GLM-5.4: Z.ai’s First Loop Transformer and the Push Toward Generalized Coding”, 2026 年 11 月 23 日. https://www.datacamp.com/blog/glm-5-4 **【確認済】**

[9] Fello AI（Lenka Vojtechova），“GLM 5.4: Z.ai’s First Loop Transformer, Where to Use It and What It Costs”, 2026 年 11 月 23 日更新. https://felloai.com/glm-5-4/ **【確認済】**

[10] BeatAPI, “What Is GLM-5.4? Z.ai’s Loop Transformer, Early Tests, and API Access”, 2026 年 11 月 23 日. https://beatapi.io/blog/what-is-glm-5-4 **【確認済】**発行元は競合モデルの API を販売する事業者

[11] arXiv:1807.03819（ICLR 2019）（M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, Ł. Kaiser），“Universal Transformers”, 2018 年 7 月 10 日投稿. https://arxiv.org/abs/1807.03819 **【確認済】**

[12] Artificial Analysis, “Thinking Machines has released Inkling 2 Mini, an open weights looped transformer”, 2026 年 11 月 20 日. https://artificialanalysis.ai/articles/thinking-machines-has-released-inkling-the-new-leading-u-s-open-weights-model **【一部確認】**記事一覧の見出し・日付のみ確認。URL は見出しが掲載されていた同サイトの別記事ページ

[13] daily.dev, “GLM-5.4 scores 33.2 on KingBench 4, the top open-weight coding result so far”, 2026 年 11 月 23 日. https://daily.dev/posts/glm-5-4-scores-33-2-on-kingbench-4-the-top-open-weight-coding-result-so-far-fahtrnytv **【確認済】**

[14] Z.ai Developer Docs, “Overview”（GLM Coding Plan），日付表示なし．https://docs.z.ai/devpack/overview **【確認済】**

[15] 米商務省産業安全保障局（BIS）／Federal Register, “Addition of Entities to and Revision of Entry on the Entity List”, 90 FR 4617（FR Doc. 2025-00704）, 2025 年 1 月 16 日. https://www.govinfo.gov/content/pkg/FR-2025-01-16/html/2025-00704.htm **【確認済】**

[16] Wikipedia（英語版）, “Z.ai”, 2026 年 11 月 24 日閲覧. https://en.wikipedia.org/wiki/Z.ai **【確認済】**

[17] NIST CAISI, “CAISI’s Assessment of Z.ai’s GLM-5.3 Cyber Capabilities”, 2026 年 9 月 17 日. https://www.nist.gov/news-events/news/2026/09/caisis-assessment-zais-glm-53-cyber-capabilities **【確認済】**検索結果で要旨を確認

[18] 網易号（鞭牛士）,「智谱 GLM-5.4 正式上线 Coding Plan」, 2026 年 11 月 23 日 13:42. https://c.m.163.com/news/a/LCE36DBO0511A6N9.html **【確認済】**

[19] 新浪財経（新浪科技）,「智谱 GLM-5.4 官宣上线：新增视觉能力，编码周额度提升 50%」, 2026 年 11 月 23 日 15:44. https://finance.sina.com.cn/jjxw/2026-11-23/doc-iniztfnh0627708.shtml **【確認済】**

[20] 搜狐（来源：和訊網／鳳凰網科技）,「智谱发布 GLM-5.4：用 Loop Transformer 循环 28 层，主攻泛 Coding」, 2026 年 11 月 23 日 19:33. https://www.sohu.com/a/1105367435_639898 **【確認済】**

[21] DEV Community（jamilxt）, “GLM-5.4 Explained: How Z.ai Turned a 750B Model Into a Loop Transformer”, 2026 年 11 月 23 日. https://dev.to/jamilxt/glm-54-explained-how-zai-turned-a-750b-model-into-a-loop-transformer-53dl **【確認済】**

[22] Le Grand Continent, GPT-6 Astra のアーキテクチャと監視可能性に関する記事（仏語）, 2026 年 9 月 4 日. https://legrandcontinent.eu/fr/2026/09/04/avec-astra-la-boite-noire-dopenai-devient-encore-plus-noire/ **【一部確認】**検索抜粋のみ確認。正式タイトル未確認

[23] 中関村在線,「智谱发布 7000 亿参数 GLM-5.4 大模型，Loop 架构引领轻量高效新范式」, 2026 年 11 月 23 日. https://ai.zol.com.cn/1276/12763391.html **【確認済】**

[24] 快科技,「智谱 GLM-5.4 大模型发布：仍是开源最强、比 Opus 5.5 差距 5%内」, 2026 年 11 月 23 日. https://news.mydrivers.com/1/1167/1167847.htm **【確認済】**

[25] Investing.com 香港（智通財經）,「港股異動｜智譜(02513)盤中漲超 4% GLM-5.4 模型上線 Coding Plan 訂閱體系」, 2026 年 11 月 23 日 15:10. https://hk.investing.com/news/stock-market-news/article-1771708 **【確認済】**

[26] Yahoo!ニュース（ビジネス＋IT）,「7500 億が 2 兆〜3 兆級に化ける？ Z.ai、ループ型を初採用した新旗艦 GLM-5.4 を発表」, 2026 年 11 月 23 日頃. https://news.yahoo.co.jp/articles/617595132c087d9debcc0fa372d9d2222859ba37 **【未確認】**取得拒否。タイトル・内容は調査報告（検索抜粋）による

[27] X, GLM-5.4 関連のトレンド要約ページ（公式アカウントの告知等）, 2026 年 11 月 23 日頃. https://x.com/i/trending/2124722750279348595 **【未確認】**取得不可。内容は調査報告による

[28] Hacker News, “GLM-5.4: Toward Generalized Coding with Our First Loop Transformer”（スレッド）, 2026 年 11 月 23 日頃. https://news.ycombinator.com/item?id=50609685 **【確認済】**

[29] Codersera, “GLM-5.3 Weights Are Out: Full vs Flash (2026 Guide)”, 2026 年 8 月 14 日（10 月 5 日更新）. https://codersera.com/blog/glm-5-3-launch-guide-2026/ **【確認済】**

[30] TechTimes, Zhipu AI 株急騰と GLM-5.2 公表に関する記事, 2026 年 6 月 15 日. https://www.techtimes.com/articles/318398/20260615/zhipu-ai-stock-surges-33-percent-fable-5-export-ban-resets-ai-vendor-race.htm **【一部確認】**検索抜粋のみ確認。正式タイトル未確認

[31] Unite.AI, “Z.ai’s GLM-5.4 Uses Iterative Depth Without Adding Parameters”, 2026 年 11 月 23 日. https://www.unite.ai/z-ais-glm-5-4-uses-iterative-depth-without-adding-parameters/ **【確認済】**AI 生成記事と表示

[32] 網易号（光子星球）（黄梦宇）,「架构、RSI、商业化：读懂智谱 GLM-5.4 的三条线」, 2026 年 11 月 24 日. https://www.163.com/dy/article/LCFAJ8E00552GIEM.html **【確認済】**

[33] Wccftech（Rohail Saleem）, “Z.ai Debuts GLM-5.4 With a Novel Looped Transformer Architecture, Skipping AI Benchmarks Entirely”, 2026 年 11 月 23 日. https://wccftech.com/z-ai-debuts-glm-5-4-with-a-novel-looped-transformer-architecture-skipping-ai-benchmarks-entirely/ **【確認済】**

[34] OODAloop, “Z.ai Unveils GLM-5.4 Built on Loop Transformer Architecture to Advance ‘Generalized Coding’”, 2026 年 11 月 23 日. https://oodaloop.com/briefs/technology/z-ai-unveils-glm-5-4-built-on-loop-transformer-architecture-to-advance-generalized-coding/ **【確認済】**

[35] 百度百科,「GLM-5.4」, 日付未確認. https://baike.baidu.com/item/GLM-5.4/69833549 **【未確認】**取得拒否。内容は調査報告による

[36] AICodeKing（X 投稿）, GLM-5.4 の KingBench 4 評価に関する投稿, 2026 年 11 月 23 日頃. https://x.com/aicodeking/status/2124731584490111224 **【未確認】**取得不可。内容は調査報告による

[37] Kingy AI（Curtis Pyke）, “GLM-5.4 launch: API, pricing, benchmarks and weights”, 2026 年 11 月 22 日（米太平洋時間）. https://kingy.ai/blog/glm-5-4-launch/ **【確認済】**

[38] Respan（Dylan Cable）, “Gemini 4 Argon vs GPT-6.2 Astra vs Opus 5.5: Comparison (2026)”, 2026 年 10 月 1 日（11 月 20 日更新）. https://www.respan.ai/articles/gemini-4-argon-vs-gpt-6-astra-vs-claude-opus-5-5 **【確認済】**

[39] Constellation Research, “Anthropic releases Mythos 5.2, Fable 5.2”, 2026 年 11 月 9 日. https://www.constellationr.com/insights/news/anthropic-releases-mythos-52-fable-52 **【一部確認】**検索抜粋のみ確認

[40] MarkTechPost, “Moonshot AI Releases Kimi K3.1 Pro: A 2.8T Total/104B Active Params (MoE) Model with 1M Context”, 2026 年 10 月 23 日. https://www.marktechpost.com/2026/10/23/moonshot-ai-kimi-k3-1-pro-open-weight-model/ **【一部確認】**検索抜粋のみ確認

[41] Wikipedia（英語版）, “Qwen”, 2026 年 11 月 24 日閲覧. https://en.wikipedia.org/wiki/Qwen **【確認済】**

[42] Alibaba Cloud, “Alibaba Unveils Qwen4-Max: New AI Teammate for Work, Creativity, and Discovery”, 2026 年 11 月 16 日. https://www.alibabacloud.com/press-room/alibaba-unveils-qwen4-max **【一部確認】**検索抜粋のみ確認

[43] NVIDIA Developer Forums, “GLM-5.4 and loop transformers to the rescue of bandwidth constrained environments”, 2026 年 11 月 23 日. https://forums.developer.nvidia.com/t/glm-5-4-and-loop-transformers-to-the-rescue-of-bandwidth-constrained-environments/389904 **【確認済】**

[44] TestingCatalog, “Z.ai launches GLM-5.4, its first Loop Transformer model”, 2026 年 11 月 23 日. https://www.testingcatalog.com/z-ai-launches-glm-5-4-its-first-loop-transformer-model/ **【確認済】**

[45] 智譜 AI 開放平台ドキュメント,「套餐概览」（GLM Coding Plan）, 日付表示なし. https://docs.bigmodel.cn/cn/coding-plan/overview **【確認済】**

[46] Hugging Face, “zai-org”（組織ページ）, 2026 年 11 月 24 日時点. https://huggingface.co/zai-org **【確認済】**

[47] 英為財情 Investing.com,「智谱(2513)」株価ページ, 2026 年 11 月 24 日閲覧. https://cn.investing.com/equities/knowledge-atlas-tech-joint-stock **【一部確認】**取得時の表示は 10 月 7 日時点の株価。52 週高値のみ確認

[48] 同花順財経, 智譜の自社株買い計画と株価に関する記事, 2026 年 11 月 12 日. https://stock.10jqka.com.cn/20261112/c681749243.shtml **【一部確認】**検索抜粋のみ確認。正式タイトル未確認

[49] The Next Web, “Anthropic says China’s GLM-5.3 nearly matches Mythos at cyber exploits”, 2026 年 9 月下旬. https://thenextweb.com/news/glm-5-3-cyber-exploits-mythos-anthropic-safeguards **【一部確認】**検索抜粋のみ確認

[50] daily.dev（原記事：Sebastian Raschka）, “GPT-6 Astra, Looped Transformers, and Hidden Reasoning”, 2026 年 9 月 9 日. https://daily.dev/posts/gpt-6-astra-looped-transformers-and-hidden-reasoning-eomtvqqoa **【確認済】**

[51] Sheppard Mullin（Jordan Mallory, Reid Whitten, Julien Blanquart）, “Choosing Between U.S. and Chinese AI Models: The Export Control Risks on Both Sides”, 2026 年 7 月 23 日. https://www.sheppard.com/insights/blogs/us-vs-chinese-ai-models-export-control-risks **【一部確認】**書誌情報・概要と検索抜粋のみ確認。本文は取得できず

[52] Layer3Labs, “Z.AI Explained: Zhipu, GLM Models, Safety & Pricing”, 2026 年 11 月 9 日更新と表示. https://www.layer3labs.io/guides/z-ai-explained **【確認済】**ページ内の日付表示に不整合あり

[53] Z.ai Developer Docs, “GLM-5.4”（英語版モデルページ）, 日付未確認. https://docs.z.ai/guides/llm/glm-5.4 **【未確認】**取得エラー

その他の参照資料

調査報告の出典一覧に含まれるが、本文書では個別に引用していない資料。括弧内は再取得の結果。

- mulianju.com「GLM-5.4」解説記事. https://www.mulianju.com/glm-5-4-release/（取得不可）
- pasqualepillitteri.it “Z.ai launches GLM-5.4, its first loop…”. https://pasqualepillitteri.it/en/news/36238/z-ai-launches-glm-5-4-its-first-loop（ボット認証画面のみ）
- MLQ.ai “Z.ai launches GLM-5.4…”. https://mlq.ai/news/zai-launches-glm-54-a-750-billion-parameter-model-built-on-a-looped-transformer/（取得エラー）
- Medium（Data Science in Your Pocket）“GLM-5.4 is here”. https://medium.com/data-science-in-your-pocket/glm-5-4-is-here-8c3e588d3e79（取得エラー）
- Medium “Does Gemini 4 Neon actually beat Opus 5.5 on every benchmark?”. https://medium.com/@amitsrivastava115/does-gemini-4-neon-actually-beat-opus-5-5-on-every-benchmark-0f86926999de（取得エラー）
- Medium “The GPT-6 Astra Architecture Puzzle”. https://medium.com/@lyx_62906/the-gpt-6-astra-architecture-puzzle-05fd2ff781af（取得エラー）
- Wikipedia（英語版）“Claude Mythos”. https://en.wikipedia.org/wiki/Claude_Mythos（取得済みだが5.2 系の記載なし）
- 雪球「智谱(02513)」. https://xueqiu.com/S/02513/hots?page=29（未取得）
- X 投稿（@pupposandro）. https://x.com/pupposandro/status/2124766042383609956（取得不可）
- X アカウント（@0xRemakw）. https://x.com/0xRemakw（取得不可）