[INDEPENDENT
VERIFICATION REPORT]

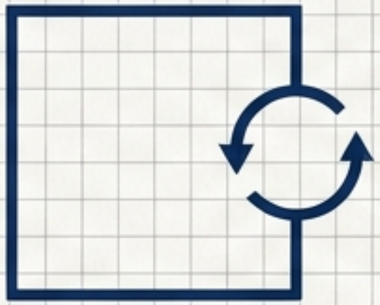
Z.ai「GLM-5.4」検証レポート

最初の旗艦Loop Transformerと「潜在計算（Latent Compute）」へのシフト

レポート分類:	Executive Intelligence Briefing & Technical Teardown
ターゲット:	AIアーキテクト、技術経営層、リサーチャー

Gemini Notebook

Loopアーキテクチャの商用実装



GLM-5.4は実在し、GLM-5（MoE基盤）に中間28層の2回反復を追加した設計を採用。

パラメータを増やさずに計算深度（Effective Depth）を拡張。

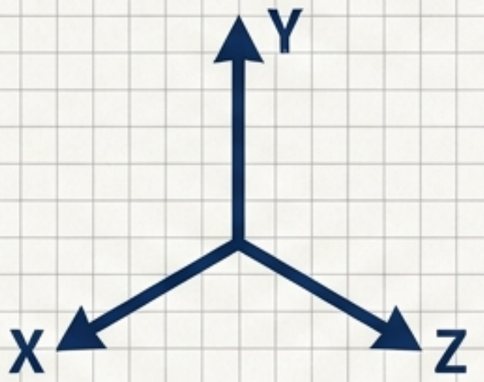
ウェイト未公開によるブラックボックス状態



モデルウェイト、専用config、完全なAPIは未公開（2026年11月24日時点）。

独立環境での再現検証、重み検査、完全なベンチマーク評価は現時点で不可能。

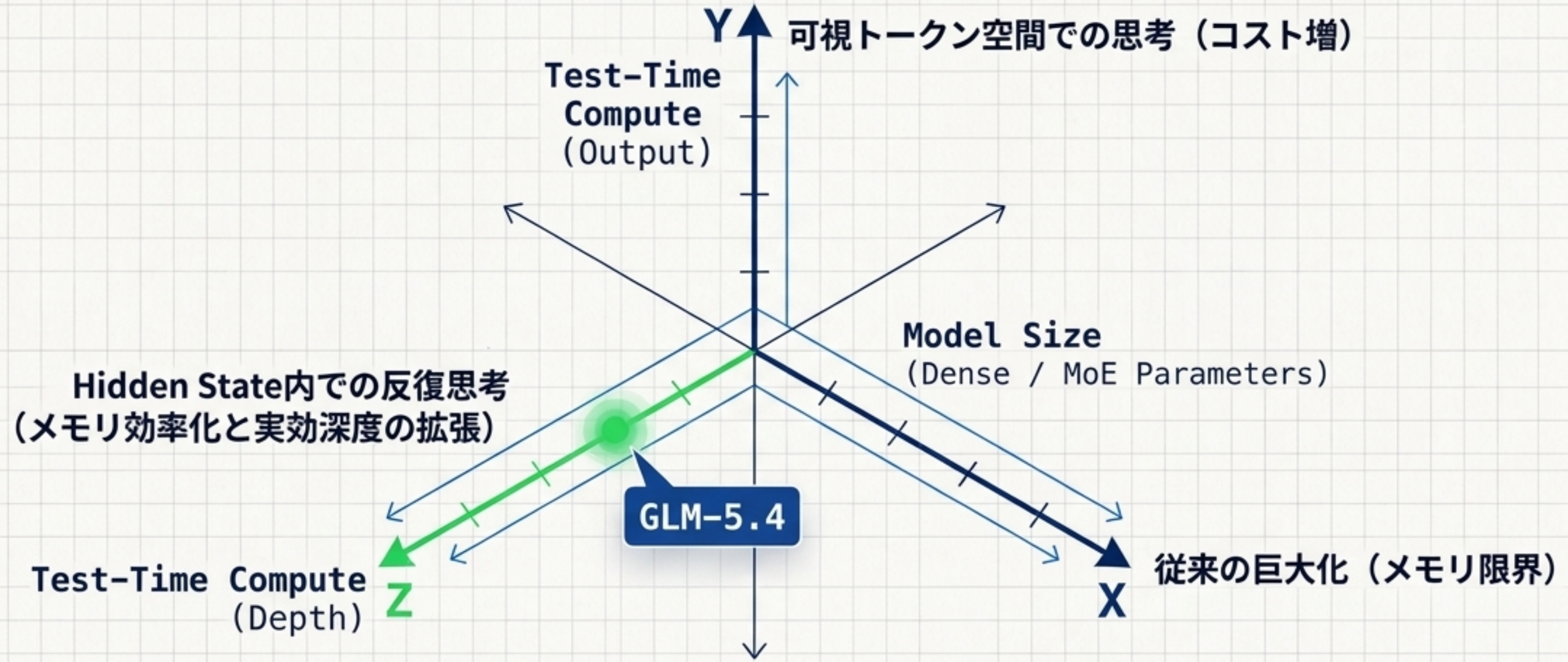
「第3のスケーリング軸」の実証



単なるモデルの巨大化ではなく、限られたメモリ制約下で推論時の内部計算（Latent Compute）を増やすパラダイムへの移行を示す重要な試金石。

Gemini Notebook

インテリジェンス・スケーリングの「第3の軸」



結論インデックス：GLM-5.4の真の意義は、巨大化の代替ではなく、「同じパラメータを複数回通して潜在計算を増やす」 戦略を商用級で実証したことにある。

Gemini Notebook

GLM-5.4 監査・検証ダッシュボード

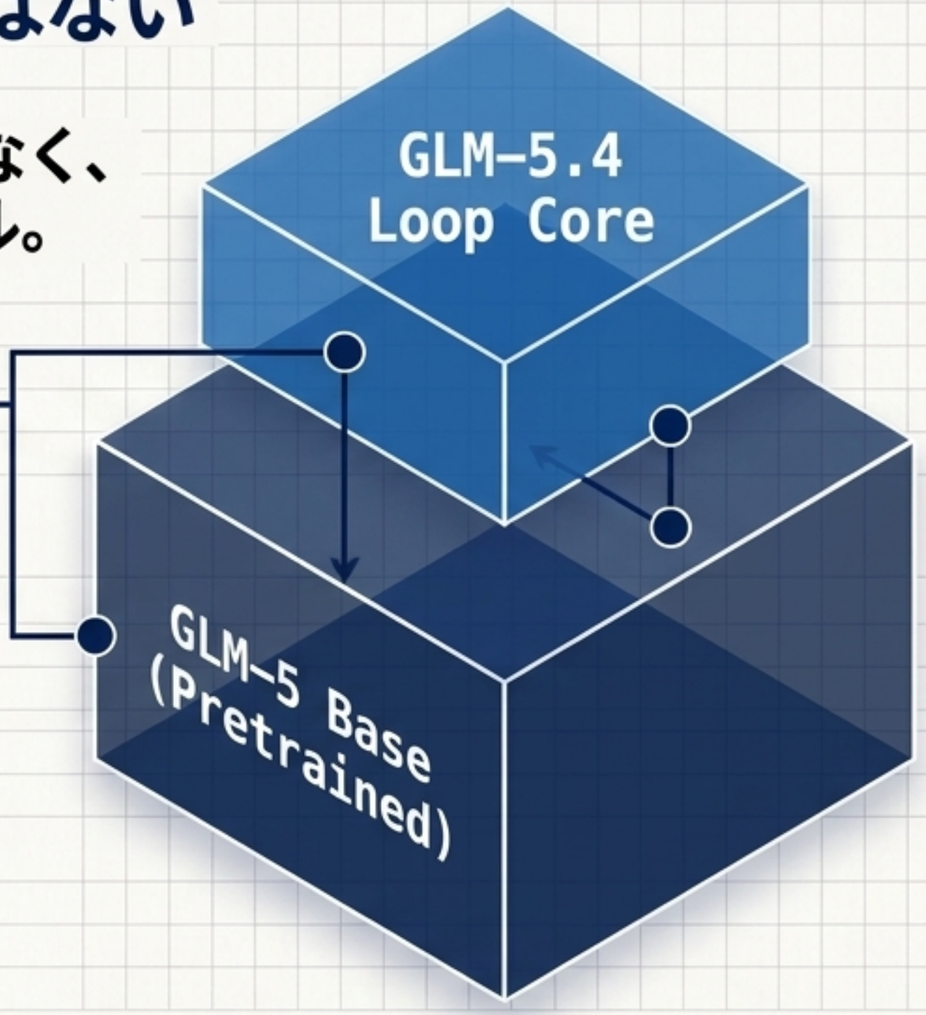
ベンダー主張	ステータス	事実と根拠
旗艦モデルとしてLoop型を初採用	[条件付き]	Z.aiの旗艦としては初だが、AI史では2018年のUniversal Transformer等に先行例あり。
2T–3Tモデル相当のインテリジェンス	[未検証]	ベンダーのマーケティング主張。対応する評価集合や統計手法が一切未公開。
10%〜14%の出力トークン削減	[確認済み]	30Bのiso-parameter実験においてGPQA等で明示的出力の短縮を確認。
オープンウェイト提供	[Pending]	「安全性強化後」の公開予定。現在はZCode経由のブラックボックスAPIのみ。

Gemini Notebook

基盤アーキテクチャ：ゼロからの構築ではない

GLM-5.4は「MoEを捨ててループ型にした」のではなく、既存の巨大MoEにrecurrent depthを追加したモデル。

GLM-5 Base Spec（Data Panel）
総パラメータ数: 744B
Activeパラメータ/token: 約40B
層構造: 3 Dense Layers + 75 MoE Layers（計78層）
アテンション: DeepSeek Sparse Attentionを採用

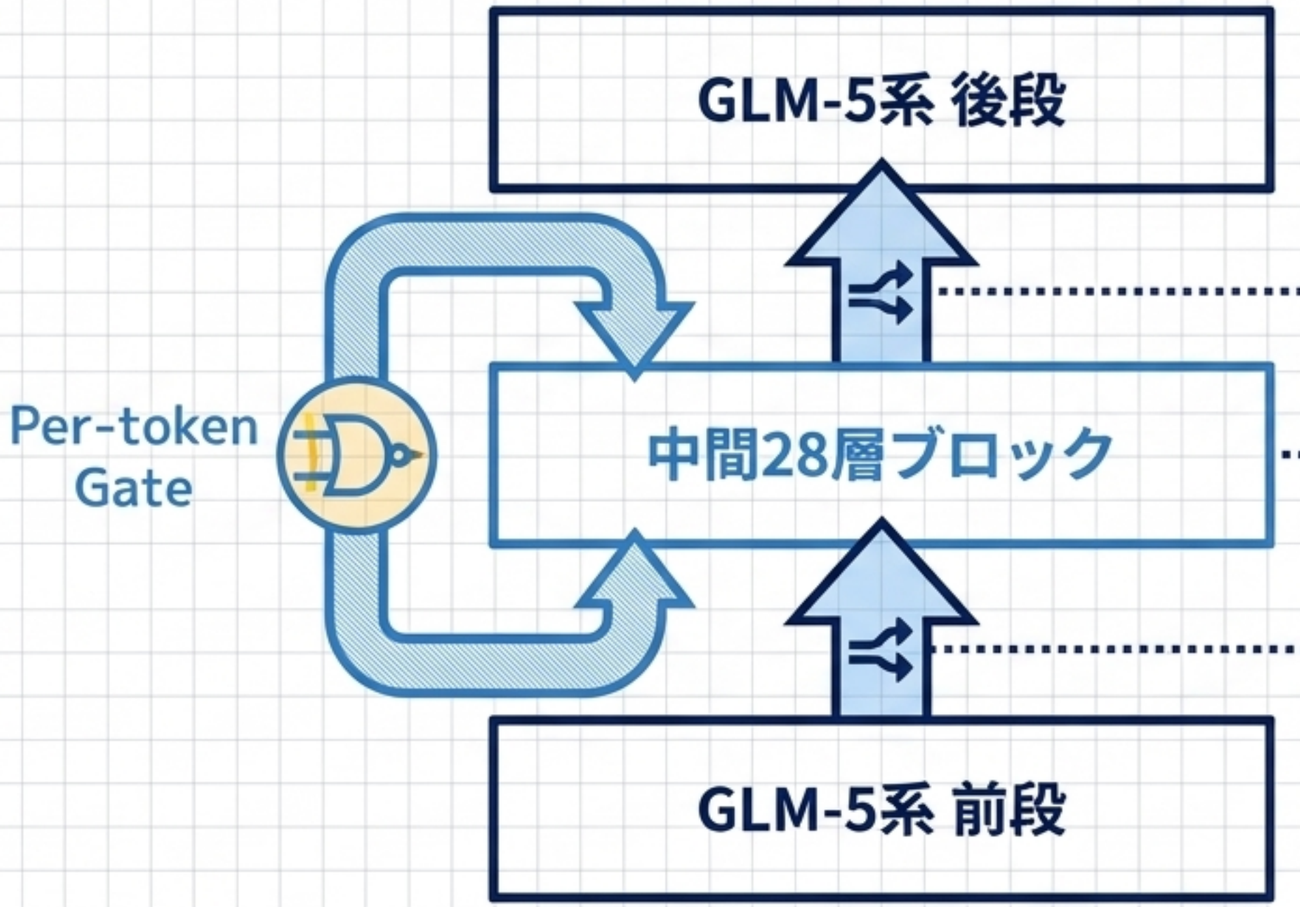


Insight

Insight: Scratchからのweight-sharing学習は非共有モデルに劣るため、既存の強力なMoE基盤を流用。

Gemini Notebook

Partial Recurrence Core（部分ループ回路）の解剖



【Pass 1】
通常の順伝播。First-pass Sparse IndexとGlobal KVを生成。

【Per-token Gate】
トークンごとに残差ストリームの更新を制御するゲート機構。

【Pass 2】
同じ28層の重み（同一パラメータ）を再実行。独立したSliding-window KVを保持。

技術的意義: 78層すべてではなく中間サブスタックのみを固定2周させることで、保存パラメータを増やさずに実効深度（Effective Depth）を倍増。

Gemini Notebook

継続学習パイプライン（Post-training Recurrent Depth）

事前学習済み GLM-5



（既存の巨大MoE モデルをロード）

Loop Core化



（中間28層を抽出して反復構造へ変換）

ゲート追加 & Hybrid KV対応



（Iteration別Gateと Sparse Index Reuse メカニズムの統合）

Continued Training



（Long-context SFTを含む継続学習の実行）

結論: Pretrained modelを後からRecurrent Depth化する手法を採用。
同じ計算量ならscratchからの学習よりも有利。

Gemini Notebook

メモリの壁の突破：Hybrid KV Cache設計

課題: 単純なループ / Naive Loop



単純なループでは、反復に比例して長文推論時のKVメモリが線形に爆発する。

解決策: GLM-5.4 Hybrid KV



第1周目のSparse Attention Indexを第2周目でもそのまま再利用し、メモリアクセスと計算オーバーヘッドを最小化。

Gemini Notebook

アーキテクチャ進化と「思考」の局在化

方式	パラメータ量	追加計算コスト	思考が発生する場所
通常Transformer	深さに比例	深さに比例	Hidden State（内部）
MoE（GLM-5）	総容量大（Active小）	Routing + Active Experts	Hidden State（内部）
Chain-of-Thought	不変	出力トークン数に比例	可視トークン空間（外部）
Agent Loop	不変	API呼び出し回数に比例	外部メッセージ列（外部）
GLM-5.4（Loop）	追加なし（重み共有）	1周より増大（反復回数）	Hidden State（内部反復）

キーインサイト: CoTが「外部」で可視トークンを消費して思考するのに対し、GLM-5.4は1回のForward passの「内部」で計算深度を深める。

Gemini Notebook

パフォーマンス評価：主張とブラックボックスのギャップ

確認済みのベンダー主張（30Bアブレーション）

+10%

学習計算量

⬇ -10〜14%

出力トークン削減

GPQA-Diamond / MMLU-Pro等で平均10〜14%減。
明示的なCoTを内部計算にオフロードできた証左。

欠落している独立評価（2026-11-24時点）



⚠ 【未確認】
GPQA-Diamond / MMLU-Pro のGLM-5.4本体確定値

⚠ 【未確認】
SWE-bench (Verified/Pro) の確定値

⚠ 【未確認】
LiveCodeBench 等の主要リーダーボードへの掲載

総評: 前世代（GLM-5.1）で公開されていた包括的スコア表が欠落しており、「2T-3Tモデル相当」の主張は現時点で独立検証不可。

Gemini Notebook

アクセス状況と商用利用条件

ZCode Coding Plan	[利用可能]	モデルID `glm-5.4`。1Mコンテキスト(約96.6万で自動圧縮)。 推論強度はlow/high/maxの3段階。Quota優遇措置あり。
従量課金 Model API	[Coming Soon]	公式ブログで近日公開予定。 固有のトークン単価は未確認。
Open Weights	[未公開]	「Safety hardening完了後に公開予定」。 ローカル推論不可。
商用ライセンス	[未確認]	GLM-5のMITライセンスを自動適用してはならない。

Gemini Notebook

研究者のための独立検証ロードマップ

- [ ] 1 モデル同一性の確認: Coding APIでの `glm-5.4` と直前世代 `GLM-5.3` の同一プロンプトA/Bテスト。
- [ ] 2 トークン効率の検証: 推論難易度を固定し、GLM-5比で出力生成トークンが実際に10–14%削減されるかの計測。
- [ ] 3 Long-contextストレステスト: Needle-in-a-Haystackだけでなく、RULER等で900K付近の劣化曲線を測定（クライアント側圧縮との分離）。
- [ ] 4 実効速度の測定: TTFTおよびDecode tokens/sの計測。公式の「最大200 t/s」の環境条件の特定。
- [ ] 5 ループ固有性の直接検証（ウェイト公開後）: Forward hookを用いて同一パラメーテンソルが本当に2回呼ばれるかをプロファイリング。

Gemini Notebook

リスクと限界：「マラソン・シンキング」の罠

固有リスク: 過剰反復による自己強化

内部反復と長い外部CoTが結びつくと、誤った仮説を何度も自己強化するハルシネーション・ループに陥る危険性。

知識追加の限界

計算深度（Effective Depth）は増えるが、新しい知識を追加するわけではない。
「深く考える＝事実を知っている」の非等価性。

安全性情報の欠落

ハルシネーション率、脱獄耐性、安全拒否率を網羅したSystem Cardが未公開。
Looped modelの堅牢性検証の不足。



Gemini Notebook

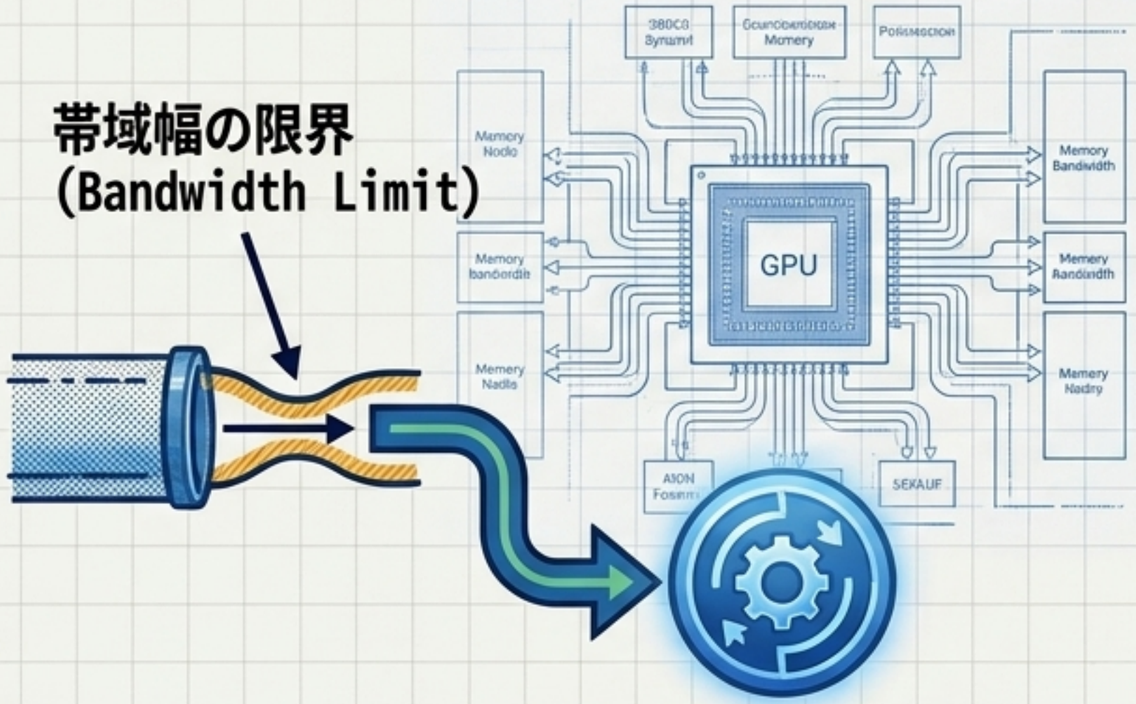
業界への戦略的含意：ハードウェア制約下の最適解

コンテキスト

HBM（メモリ）と通信帯域が極度に制約されるGPU環境。

パラダイムシフト

- 旧来：パラメータ競争（重みの保存量を増やす＝メモリ制約に直面）
- 今後：有効深度（Effective Depth）とメモリ効率化の共同最適化



インサイト： GLM-5.4は魔法ではない。保存する重みを増やす代わりに、同じ重みを賢く再利用し、モデル配布サイズやメモリ容量を「推論時の計算量」とトレードオフする経済的な生存戦略である。

Gemini Notebook

最終判定（Final Verdict）

実在性：[高確度]

Loop型採用とMoE基盤の実装は事実。

新規性：[高]

学術概念にとどまっていたLoop型を商用級の大規模MoEに実装した工学的価値は大きい。

性能優位：[中・要検証]

性能向上はベンダーの自己申告に依存。

再現性：[低]

ウェイト・Config未公開により独立検証不可。

Next Steps（注視すべきトリガー）：

1. 公式Technical Reportの発表
2. Open Weightsおよび専用Configの公開
3. SWE-bench / GPQAの第三者リーダーボードへの登場

結語：現段階での「2T-3Tモデル相当」という表現は、興味深いが独立検証されていないマーケティング主張として扱うべきである。

Gemini Notebook