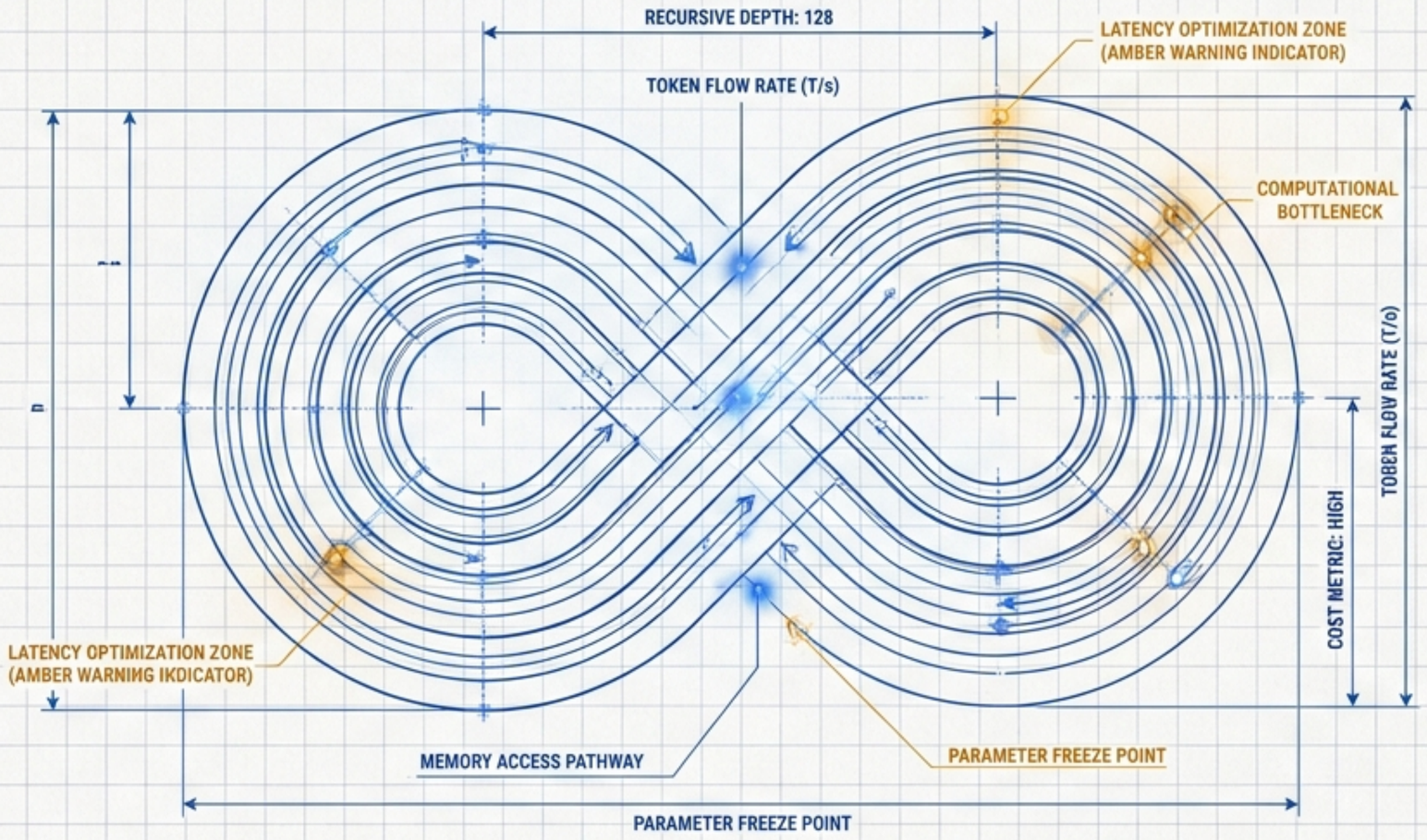
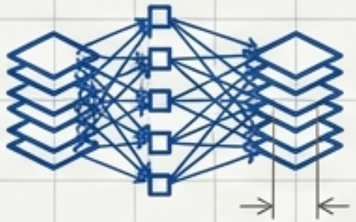
パラメータ凍結のまま
推論深度を拡張する
「The Architectural Blueprint」

Z.ai「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 Notebook



【アーキテクチャ】

750Bで2T〜3T級の知能へ。総パラメータを据え置き、「Loop Transformer」構造により推論深度を拡張。



【モダリティ】

旗艦初のビジョン統合。コードを汎用インターフェースとし、CADから3D空間まで自律処理する「汎用コーディング」。



【コスト構造】

日本企業に有利なエコノミクス。
実質1.5倍のクレジット枠拡張と、JSTに最適化されたオフピーク割引の恩恵。



【推論特性】

「マラソン思考」の二面性。
思考率99%超の超多段推論による驚異的精度と、時間・コスト管理の重要性。⚠

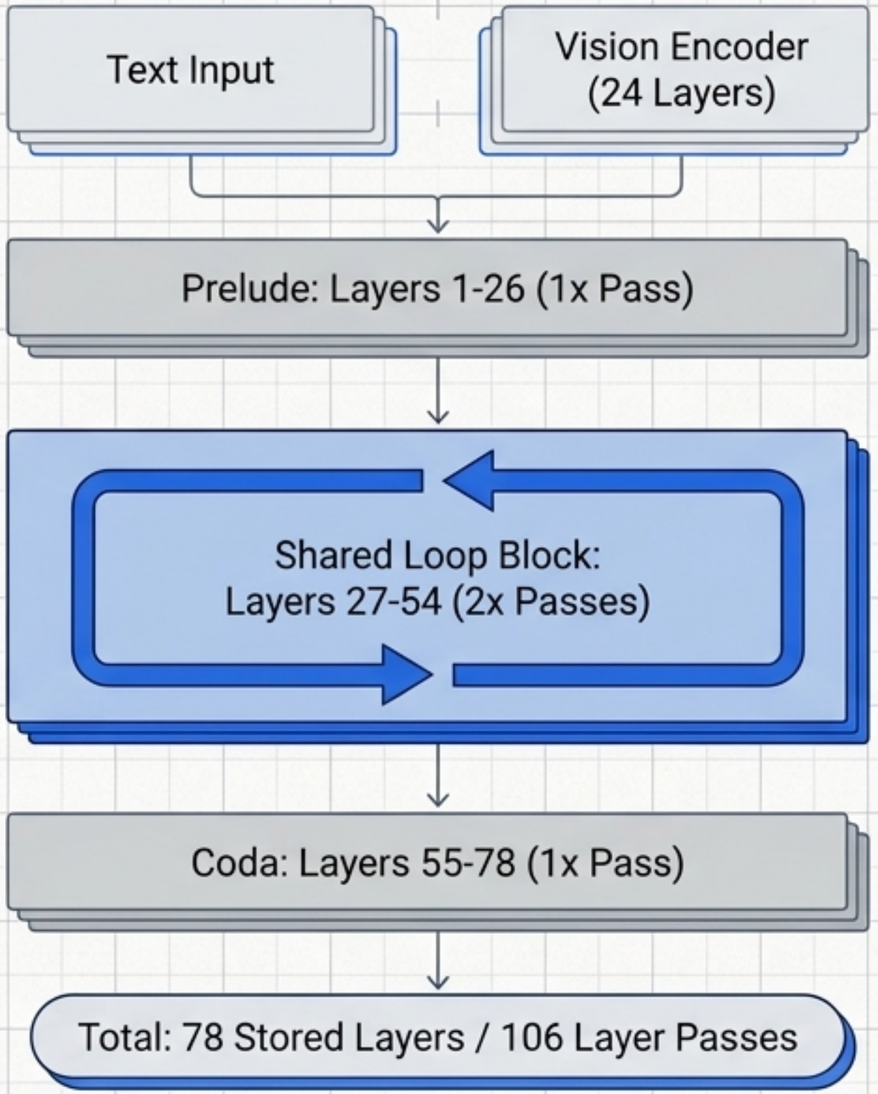
Gemini Notebook

Evolution Matrix: GLM-5.3 vs GLM-5.4

比較項目	前世代 GLM-5.3	新旗艦 GLM-5.4
基本アーキテクチャ	単一パス実行の疎なMoE	Loop Transformer × 疎なMoE
レイヤー構成・実効深度	各層1回直列実行	物理78層 / 共有ループ2回による計106層パス
コンテキスト・出力長	最大1M入力	最大1M入力 / 最大128K出力
対応モダリティ	テキスト専用 （マルチモーダルはFlash版）	ビジョン統合 (動画・画像・テキスト・ファイル入力)
推論スループット	（比較非公表）	オフピーク最大200 tokens/s、 ピーク時70〜80 tokens/s

Gemini Notebook

The Loop Architecture：再帰的深度の獲得



内部表現への変換

同一重みで2周旋回。
1周目の隠れ状態を2周目で精緻化。

最終状態から次トークン予測へ

Gemini Notebook

破綻を防ぐ3つの安定化機構

1

Damp Gates (トークン単位の門控漸進更新)

活性値ノルムの発散を制御。1周目は小さくゲートを開き文脈を構築し、2周目で大きく開き細部を書き戻す。

$$\Delta\theta_t = \eta_t \cdot f(x_t, g_t)$$

Dynamic step adjustment

$$\Delta\theta_t = \eta_t \cdot f(x_t, g_t),\ g_t \in [0,1]$$

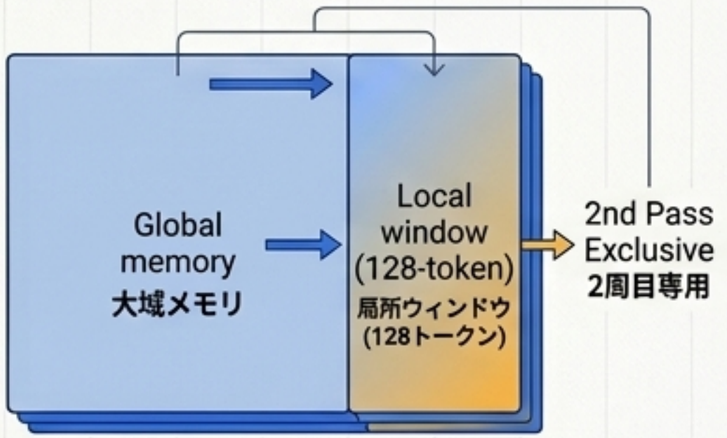
$$= \sigma_t [f(x_t, g)]$$

$$= \sigma^{+v} \in (x_t, g_t))$$

2

ハイブリッドKVキャッシュ

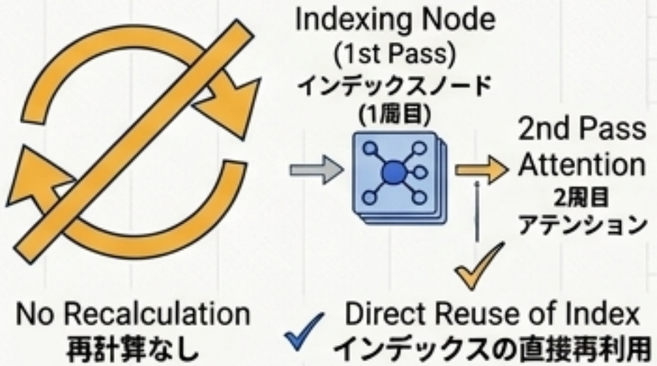
メモリ倍増の回避。1周目の大域的KVキャッシュを共有しつつ、直近128トークンの局所スライディングウィンドウのみ2周目専用として独立保持。



3

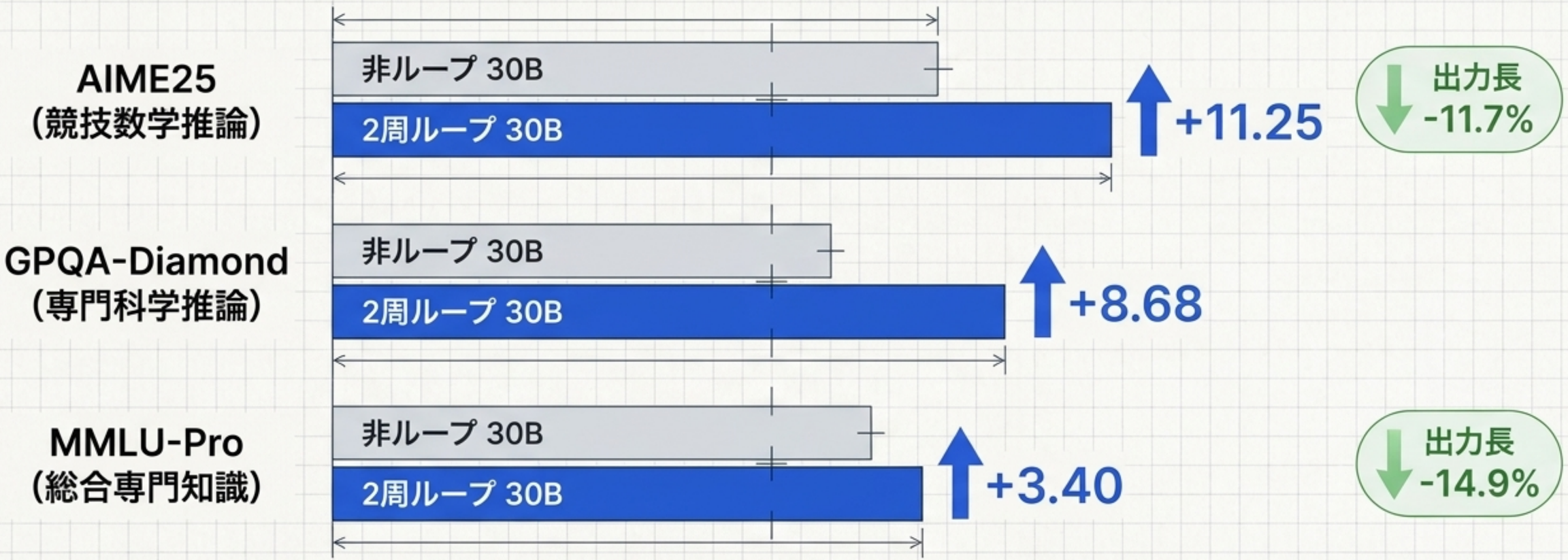
疎アテンションのインデックス再利用

計算コストゼロへの挑戦。2周目はインデクサを再実行せず、1周目のインデックスを直接再利用し計算オーバーヘッドを抑制。



Gemini Notebook

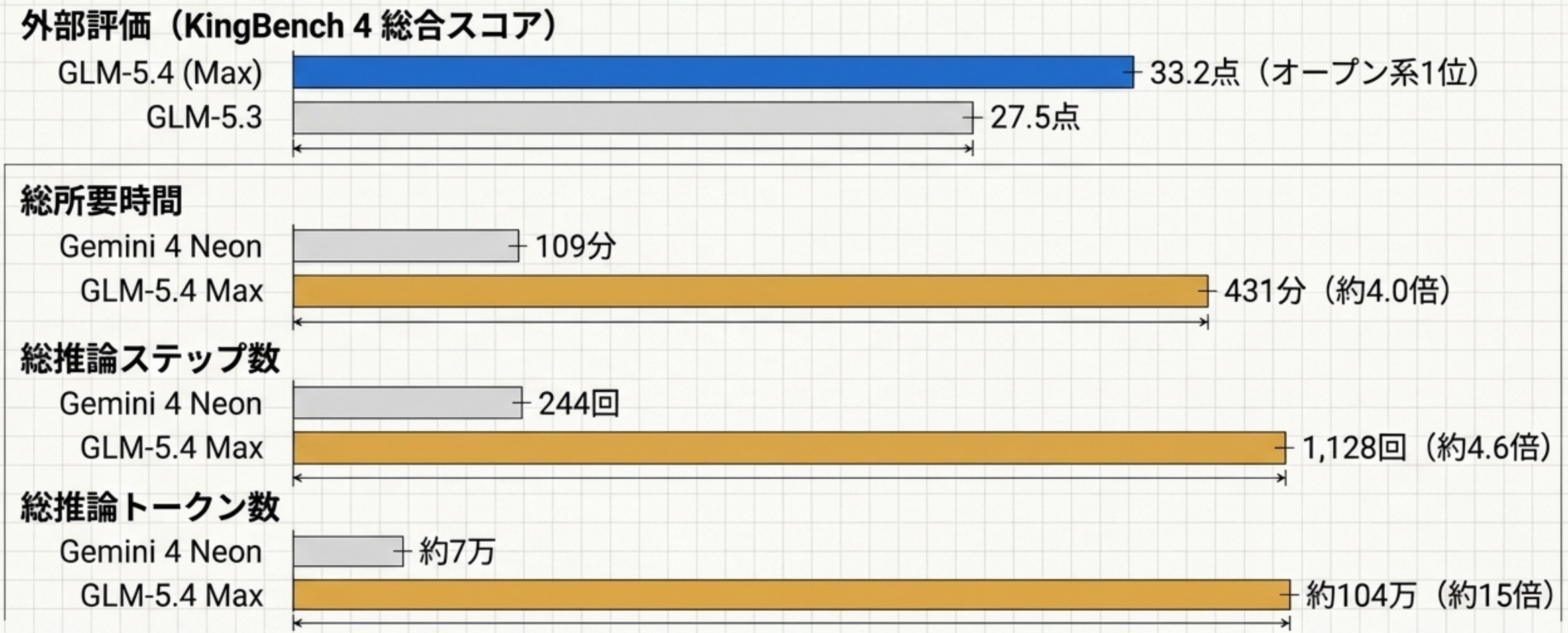
The “Latent Depth” Paradox（潜在的深化の証明）



テキストとして外部出力していた明示的な推論（CoT）の一部が、
ループ内の『隠れた思考（暗黙の深度）』へと移行したことを証明。

Gemini Notebook

Marathon Thinking（実戦における超多段ステップのジレンマ）



原因：1ステップあたりの処理は高速だが、細かなステップごとに必ず自己検証を挟む特異な挙動（思考を省略するステップ率が1%未満）。これが『マラソン思考』を誘発。

Gemini Notebook

「汎用コーディング」の幕開け（Generalized Coding）



概念：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」として扱う新構想。
100万トークンのコードとUI画像を単一プロンプトで同時解釈。

Gemini Notebook

Use-Case Diagnostic Matrix（実務適性マトリクス）

ステータス	ユースケース	診断理由	推奨設定
 【最適】	セキュリティ監査・脆弱性診断	特異な粘り強さで、上位モデルが見落とす深層バグを検出可能。	推奨 Effort: High
 【最適】	リリース前コードレビュー・設計	100万文脈を活かし、全体構造の分解・監査役として極めて有効。	推奨 Effort: High
 【適応】	長時間の自律エージェント	初回ツール失敗率はあるが、リトライによる自己修復率が66%と高い。	長めのタイムアウト設定
 【非推奨】	フロントエンドUI/3D一括生成	素早い視覚的試行錯誤や大量コードの一括執筆には、時間・トークン効率が悪化。	他モデルの主軸採用を推奨

Gemini Notebook

The Credit Illusion & Cost Inversion（コスト逆転の罠）

The Illusion

通常入力係数: 4.6
（実質1.5倍枠）
→ from GLM-5.3

キャッシュ入力係数: 1.1
（実質1.5倍枠）

出力係数: 16
（※キャッシュ入力の約14.5倍と極めて重い）

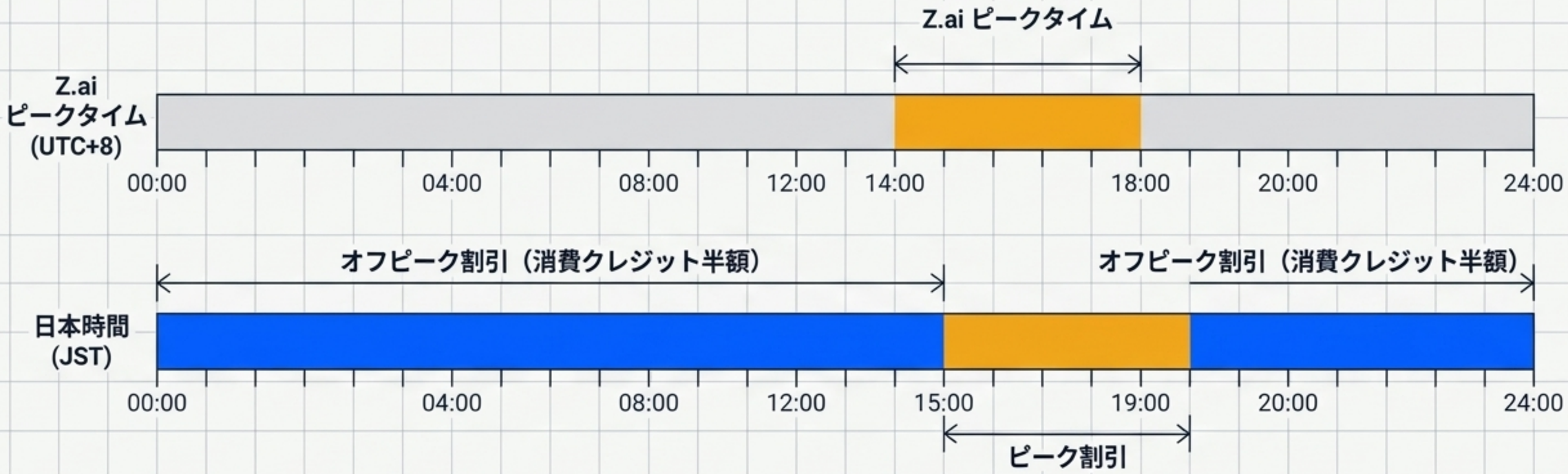
The Inversion



デフォルトの「Effort: Max」設定によりマラソン思考が発生し、推論出力トークンが倍増すると、出力係数 (16) の重さによりディスカウントが相殺され赤字に転落する。

Gemini Notebook

The "Japan Advantage"（日本時間運用の圧倒的優位性）



- **ピークタイムのズレ**：日本のピーク帯は平日15:00〜19:00のわずか4時間のみ。
- **大幅なコスト削減**：日本の午前中、19時以降の夜間、および週末終日はすべてオフピーク割引が適用。
- **戦略的活用**：Zcodeの無料キュー実行機能と組み合わせることで、将来的なAPIコストを最大92%節約可能。

Gemini Notebook

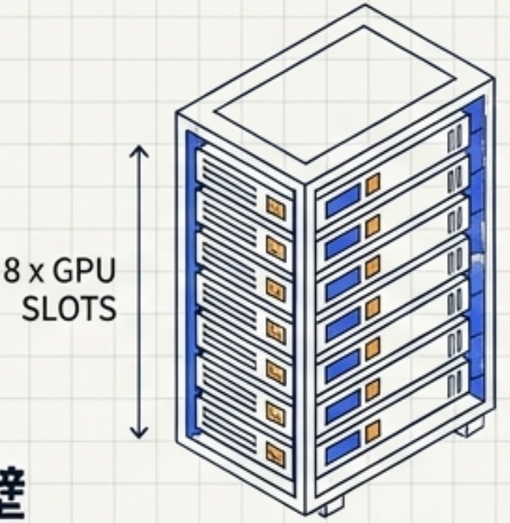
Security & The Open-Weight Dilemma（セキュリティと重み公開の行方）



公開保留の背景

Anthropicによる調査レポートで、前世代GLM-5.3が『自律的エクスプロイト開発能力において脅威水準を引き上げる』と警告。

脆弱性検出に優れた5.4の重み・API公開は『安全性強化の完了後』へと先送りされている。



セルフホストの巨大な壁

オンプレミス運用におけるハードウェア要件の実態：

- FP8精度推論でも約744GBのVRAMが必須。
- NVIDIA H200×8基 または B200×8基 クラスの巨大クラスタが必要。
- 推論エンジンの課題：現行のvLLM等は単一パス前提。多段ループ・トークン単位ゲートの効率処理はエンジン側の対応待ち。

Gemini Notebook

4-Step Implementation Guide（実務検証のための4ステップ）

[Step 1] 専用APIキーとエンドポイント設定
定額プラン用キーを発行し、互換ツールに応じ専用エンドポイントを指定。
`endpoint: "https://api.z.ai/api/coding/paas/v4"`

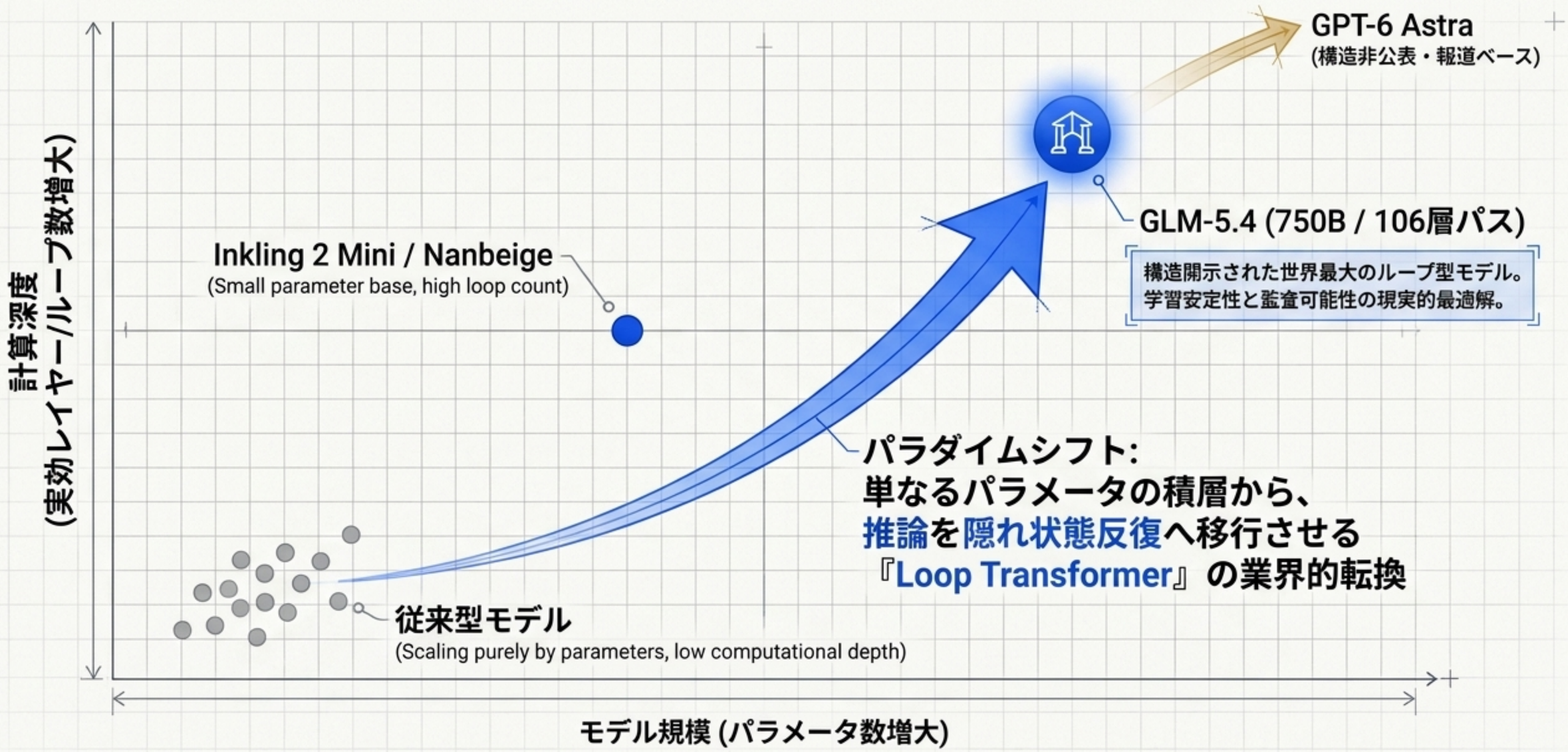
[Step 2] コンテキスト圧縮の制限
自動圧縮閾値を51.2万トークンに設定し、実用コーディング精度を維持。
`AUTO_COMPACT_WINDOW = 512000`

[Step 3] Effort（思考強度）の明示的指定
デフォルトのMaxによる課金爆発を防ぐため、設定内で強度を明示。
`/effort high` or `/effort medium`

[Step 4] APIタイムアウト値の拡張
大規模監査時の1000ステップ超の検証に備え、クライアント側の待機時間を延長。
`API_TIMEOUT_MS = 3000000`

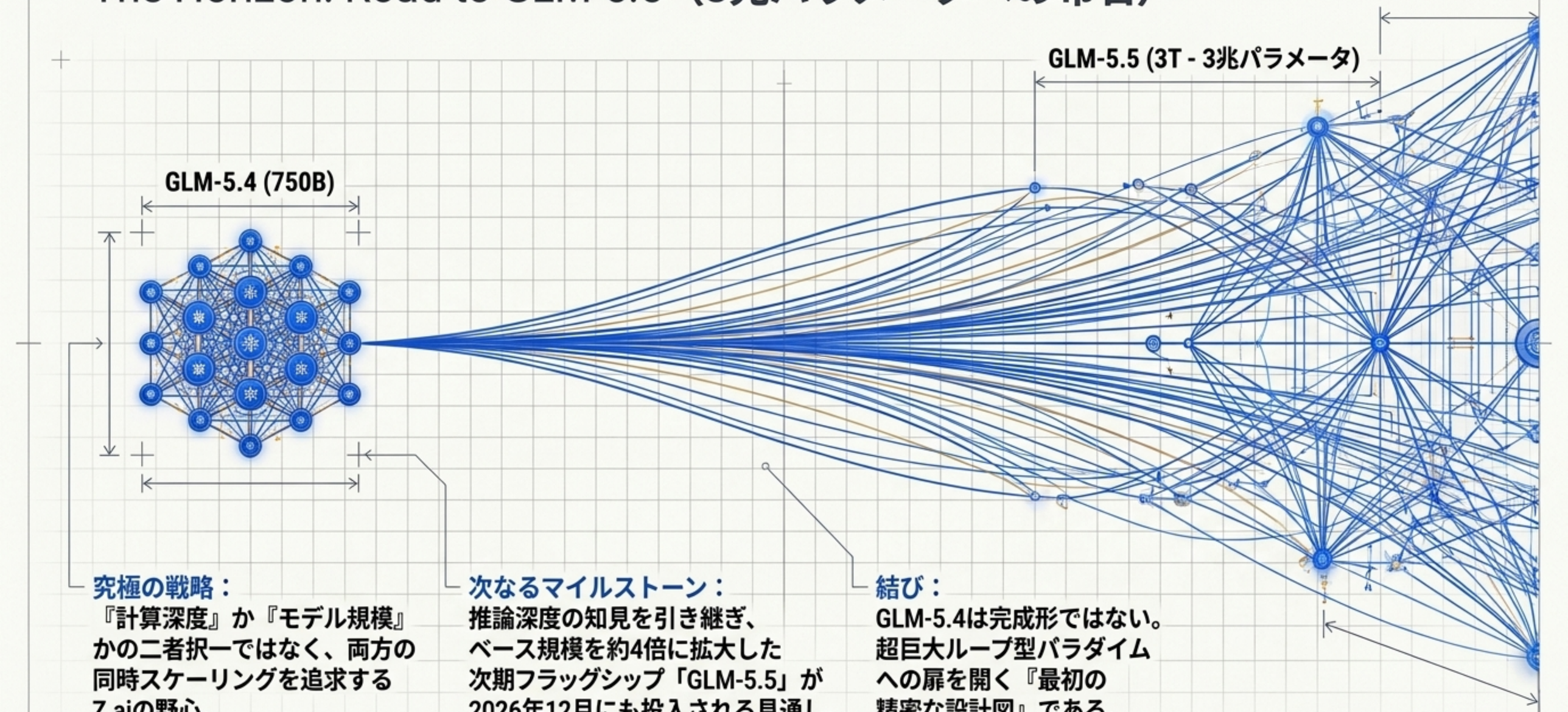
Gemini Notebook

Industry Synthesis: The Loop Paradigm（再帰的深度スケーリングの潮流）



Gemini Notebook

The Horizon: Road to GLM-5.5（3兆パラメータへの布石）



究極の戦略：
『計算深度』か『モデル規模』かの二者択一ではなく、両方の同時スケーリングを追求するZ.aiの野心。

次なるマイルストーン：
推論深度の知見を引き継ぎ、ベース規模を約4倍に拡大した次期フラッグシップ「GLM-5.5」が2026年12月にも投入される見通し。

結び：
GLM-5.4は完成形ではない。超巨大ループ型パラダイムへの扉を開く『最初の精密な設計図』である。

Gemini Notebook