調査対象: GLM-5.4 (2026年11月23日発表)
ドキュメント性質: 戦略的インテリジェンス・ブリーフィング
分類: 社外秘 (Confidential)

Z.ai「GLM-5.4」技術評価と実務への影響

ループ型アーキテクチャの全貌と日本企業が直面する知財・輸出管理リスク

Gemini Notebook

エグゼクティブ・サマリー：The Bottom Line



パラメータ増なき「深さ」の獲得

- Z.ai初の「ループ型」旗艦モデル
- 28の層を2回反復通過
- 規模（750B）を維持し実効深さを約36%増加



精度と極端な遅延のトレードオフ

- オープン系首位相当のスコア（KingBench 4: 33.2点）
- 他モデルの約4倍の所要時間（マラソン思考）
- 公式ベンチマークは現在未公表



米国輸出管理規則（EAR）の罠

- 開発元（智譜）は米国エンティティリスト掲載
- ホスト型APIへの機密入力はEAR違反リスクを発火
- 現時点での実務利用は見送るべき

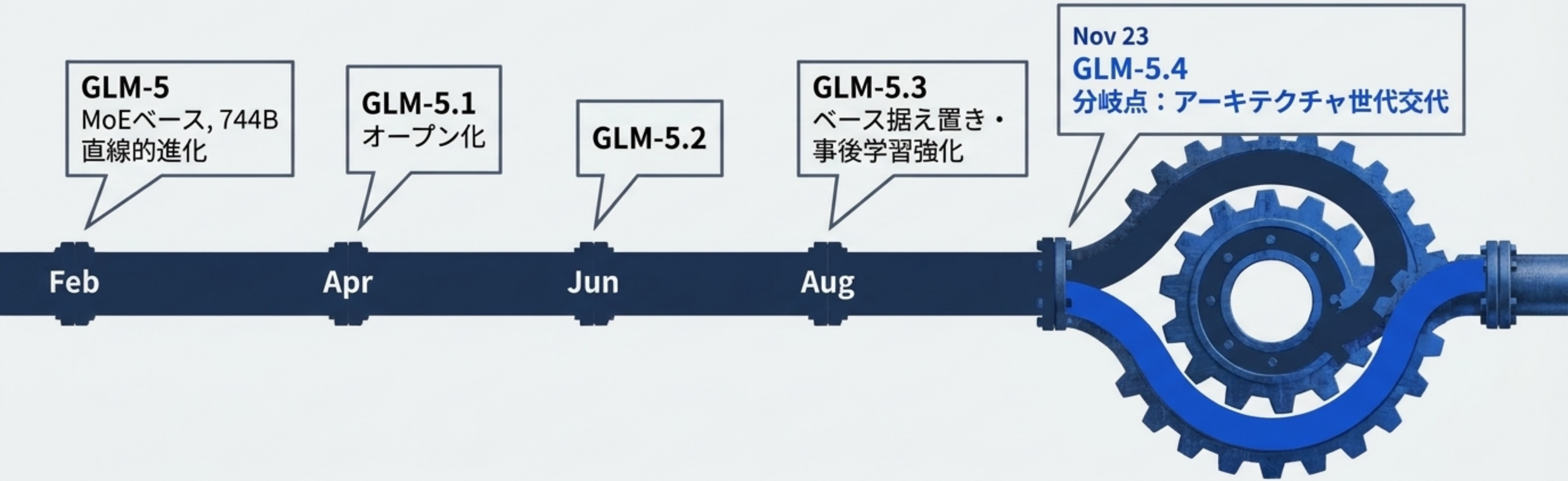
Gemini Notebook

発表内容のファクトチェック（Claim vs. Reality）

Z.aiの主張 / 報道	判定	調査による真実
「11月23日、GLM-5.4をリリース」	[○ 確認済]	公式ブログおよびドキュメントで確認。ただし現在アクセスできるのはCoding Plan等のみ。
「ループ型を初採用」	[△ 一部確認]	Z.aiとしては初だが、業界初ではない。Thinking Machines「Inkling 2 Mini」などが先行。
「2〜3兆パラメータ級に匹敵する知能」	[× 未検証・誇張の可能性]	750B本番モデルの公式ベンチマークは未公表。公開された比較数値は30B検証モデルのものに過ぎない。

Gemini Notebook

モデル進化の系譜：スケールから「構造」への転換

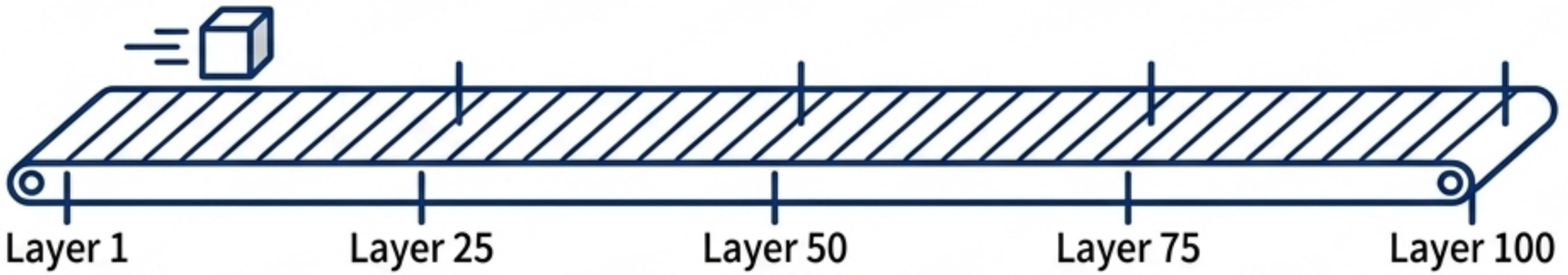


「パラメータの巨大化（横の拡大）」から
「継続学習によるループ化（縦の反復）」へ。
ゼロからの学習ではなく、GLM-5ベースを改造して構築。

Gemini Notebook

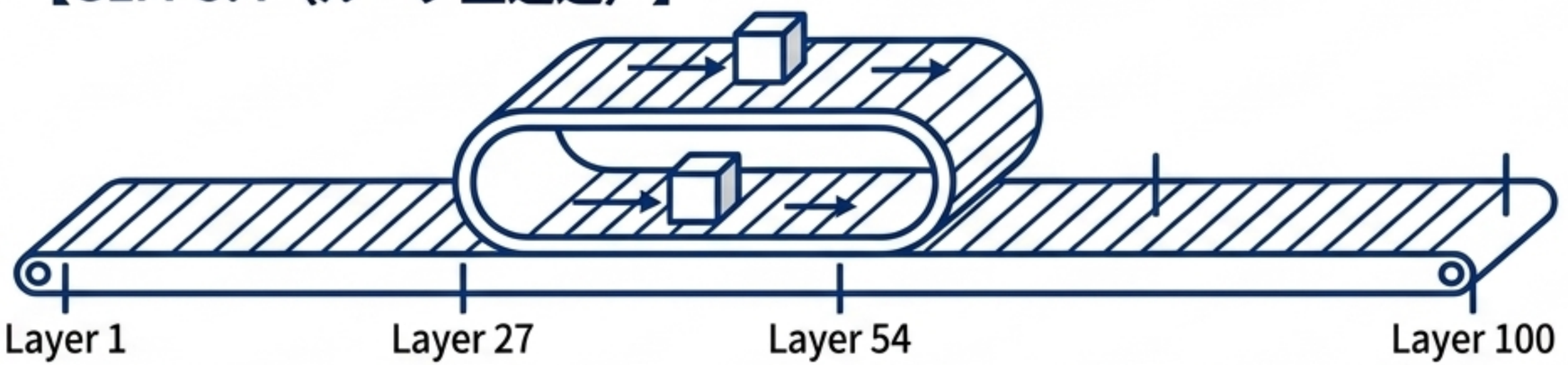
視覚解剖：ループ・トランスフォーマー（Loop Transformer）の仕組み

【従来型（直線通過）】



パラメータ増＝層の追加が必要

【GLM-5.4（ループ型通過）】



パラメータ（重み）のメモリを増やさずに、実効的な深さを増加

保存層数: 計78層

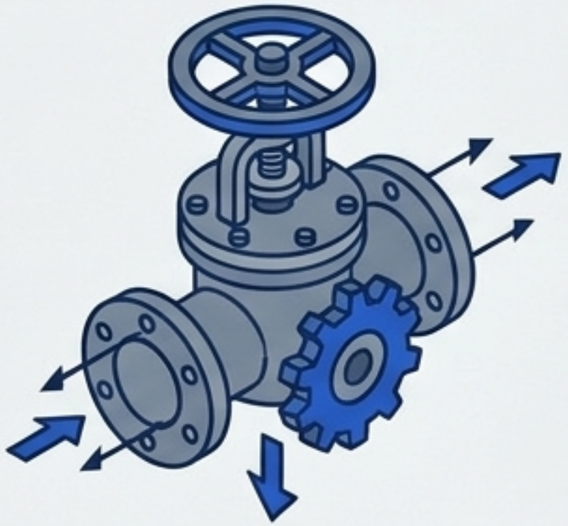
ループ対象:
第27～54層
（28層分）を2周

実効通過回数:
1トークンあたり
106回

実効深さ増加:
約36%

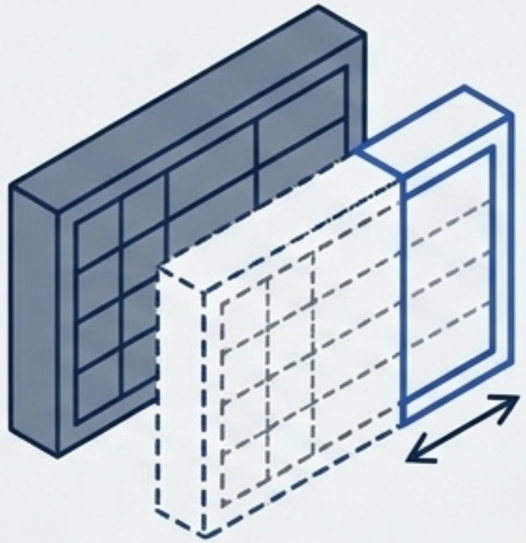
Gemini Notebook

750B規模を実現した3つの技術的ブレイクスルー



ゲート付き漸進更新

周回・トークンごとに情報の書き戻し量を制御。1周目は控えめ、2周目でゲートを開き精緻化。難しいトークンほど適応的に反応する。



ハイブリッドKVキャッシュ

2周目は1周目の疎なキャッシュを再利用し、自前はスライディングウィンドウ内のみ保持。1Mコンテキストの超長文でもメモリ爆発を防ぐ。



疎アテンション索引の再利用

1周目で作成した検索インデックス（索引）を2周目で流用。追加の計算オーバーヘッドを発生させない設計。

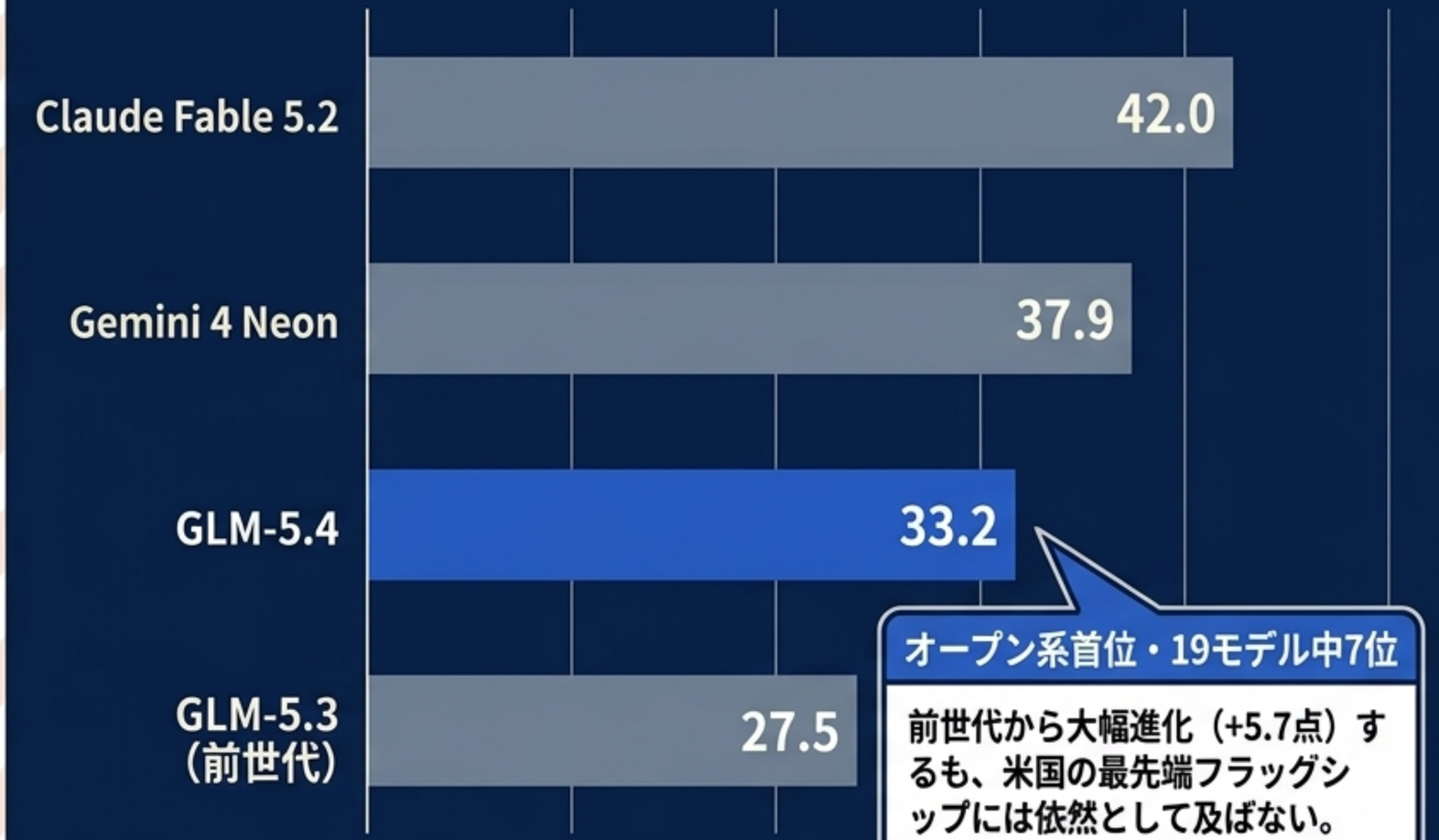
Gemini Notebook

性能評価の現実（Performance Reality Check）

公式スコア未公表

750B本番モデルの公式ベンチマークは未公表。Artificial Analysis等の主要な第三者評価も現時の点で不在。自己申告に依存するリスクに留意。

KingBench 4 スコア比較（第三者評価）

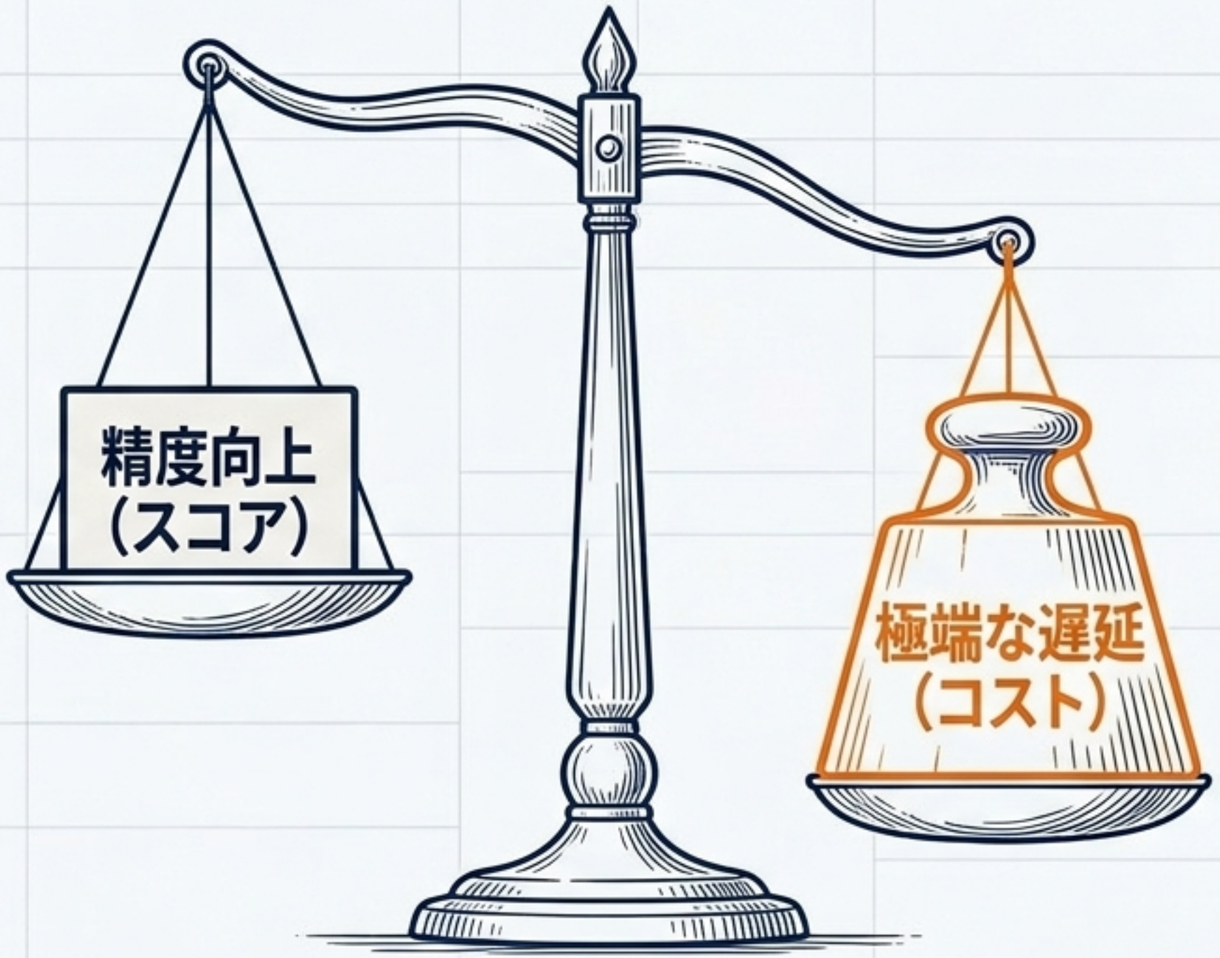


Gemini Notebook

トレードオフ：「マラソン思考」のジレンマ

精度向上（スコア）

- ツール呼び出し失敗の66%を「再確認（リトライ）」で回復。
- ほぼ全ステップでの反復推論による深い思考。



極端な遅延（コスト）

- 所要時間：431分（Gemini 4 Neonの109分に対し、約4倍の遅延）
- 平均出力長が10〜14%減少（推論が内部ループに隠蔽されるため）。

ループ構造は思考の深さをもたらすが、実効コスト（時間）とのトレードオフが極めて大きい。

Gemini Notebook

グローバル・コンペティティブ・マトリクス（2026年11月版）

比較項目	GLM-5.4（中国）	Claude Fable 5.2（米）	Gemini 4 Neon（米）	GPT-6.2 Astra（米）	Kimi K3.1 Pro（中国）
パラメータ規模	～750B	非公開	非公開	非公開	2.8T
アーキテクチャ	ループ型	従来型	従来型	ループ型（報道）	従来型
提供形態	ホスト型APIのみ（重み未公開）	API	API	API	オープンウェイト
コンテキスト長	1M	非公開	非公開	非公開	1M

GLM-5.4は、Kimi等の2～3T級巨大モデルを「3分の1の規模（750B）+ループ構造」で追撃する設計思想。

Gemini Notebook

開発背景のシンセシス：なぜ「ループ型」への移行が必要だったのか？

地政学的制約

- Z.ai（智譜）の米国エンティティリスト掲載
- 最先端GPU（演算リソース）の調達経路の分断

スケーリングの壁

- 従来のアプローチ（パラメータ数の単純拡大）は、メモリと計算資源の制約で維持困難に

適応進化としてのループ型

- 「横への拡大（巨大化）」から「縦への反復（計算の深さ増）」へのピボット

結論：ループ・トランスフォーマーは純粋な技術的探求というより、制約下の演算リソースから最大性能を絞り出すための「生存戦略的アーキテクチャ」である。

Gemini Notebook

現在の提供形態とコスト構造

利用可能なプラットフォーム

- GLM Coding Plan（月額18ドル〜）
- ZCode / Z.ai Chat / Trae等のコーディングツール

※API開放およびオープンウェイト公開は「安全評価完了後」であり、時期未定・ライセンス不明。

コスト構造の改善（クレジット係数比較）

入力トークン係数:



GLM-5.4: 4.6

GLM-5.3: 6.9

出力トークン係数:



GLM-5.4: 16

GLM-5.3: 24

効果：前世代比でクレジット消費を約2/3に抑制。「週次枠1.5倍」のマーケティング文句と整合する実質的なコストダウン。

Gemini Notebook

コンプライアンスの罠：米国輸出管理規則（EAR）の波及リスク

法的ファクト

- 開発元「智譜（Z.ai）」は、米国エンティティリスト（90 FR 4617）に追加済み。
- 審査方針は「原則不許可（presumption of denial）」。



日本企業への波及リスク

- 公開された重み（オープンウェイト）をダウンロードする行為自体は輸出に当たらない。
- しかし、米国のEAR対象技術（技術データ等）をZ.aiのホスト型APIに送信する行為は、許可を要する「輸出」に該当する恐れが。

Gemini Notebook

データフロー・リスクマップ（想定利用シナリオと判定）



【OUT】EAR違反発火リスク・知財漏洩リスク



【SAFE】コンプライアンス準拠・安全性担保

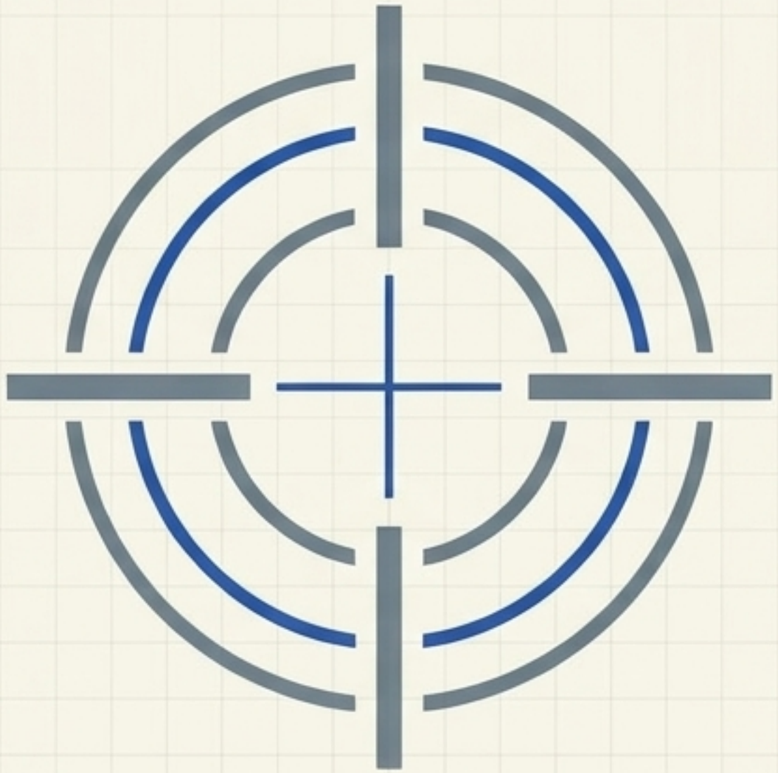
Gemini Notebook

知財動向と特許監視（IP Watchlist）

背景：「ループ型」の基本概念はすでに公知（Universal Transformer 2018等）。
今後の知財競争の焦点は「実装の工夫」に移行する。

論点1
事前学習済みモデルの「継続学習」によるループ化手法

論点2
周回ごとの「ゲート付き書き戻し (Gated progressive updates)) 」回路



論点3
ループ間でのKV共有と「スライディングウィンドウ」のハイブリッド管理

論点4
2周目以降の「疎アテンション索引の再利用」システム

Gemini Notebook

日本企業向け：戦略的アクションプラン

フェーズ1：現時点

【採用見送り・リスト入り】

- 指示: 本番環境への採用は見送り、「要評価リスト」にとどめる。
- 理由: 公式スコア、独立評価、重みデータが未提供であり、判断材料が決定的に不足しているため。

フェーズ2：厳守事項

【APIへの機密入力禁止】

- 指示: Z.aiが提供するホスト型サービス（Coding Plan等）への機密情報・EAR対象技術の入力を社内ルールで全面禁止する。

フェーズ3：次の一手

【重み公開後のローカル検証】

- 指示: オープンウェイトとライセンスが公開された後、自社オンプレミス環境にて「エフォート：High」を指定し、コードレビュー等の特定用途で遅延と精度の検証を行う。

Gemini Notebook