Z.ai「GLM-5.4」検証レポート：ループ型Transformerの実効性と課題

入力

GLM-5.4 モデル

初期・最終層

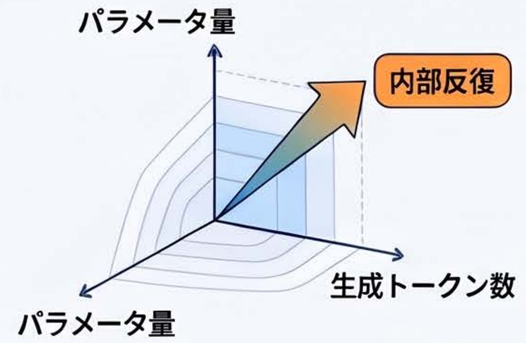
中間28層（ループ処理領域）

2周再実行

MoEとHybrid KVの融合

744B MoE構造を維持し、2周目のKV再利用でメモリ増強を抑制。

内部反復という新スケーリング



パラメータ量・生成トークン数に続く「第3のスケーリング軸」を実験。

検証状況と実務的インプリケーション

性能と効率の主張

推論時トークン数を
10–14%削減

30B比較アプレーションにおいて内部反復により出力長を効率化。

ベンダー主張の性能



「2T–3T相当」はベンダー主張

独立した評価環境や詳細なスコア表は未公開であり検証不可。

提供形態と再現性

ウェイト未公開と提供形態

ZCode等で利用可能だがウェイト未公開のため完全再現は不可。

GLM-5.4発表内容の確認状況

項目	確認状況	技術的留意事項
ループ型初採用	確認済み (Z.al揃初)	AI全体の初ではなく Universal Transformer 等の先行研究あり
性能ベンチマーク	自社主張のみ	第三者による独立測定結果は未反映
オープンウェイト	未公開	安全性検証後に公開予定、現状はローカル構築不可

Gemini Notebook