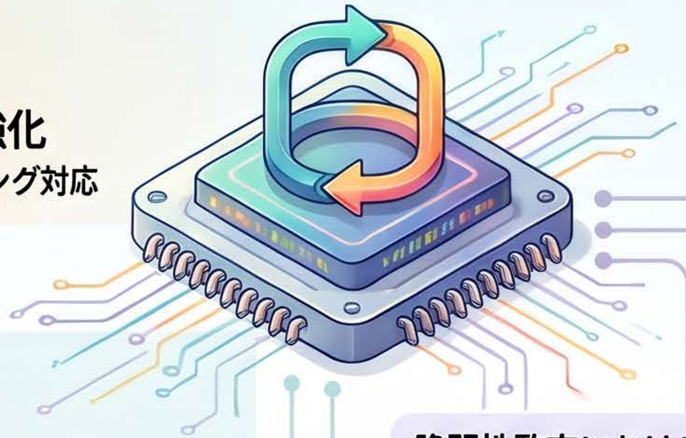
Z.ai「GLM-5.4」徹底解剖：Loop Transformer構造と実務活用ガイド

Loop Transformer構造

中間層をループ実行し、長文脈・複雑推論を強化

バラメータ規模7,500億、ネイティブ画像認識・汎用コーディング対応



Loop Transformer構造の技術的革新

物理78層で106層相当の実行深度



1-25 → 26-53 → 54-78

中間28層を2周反復
VRAM消費を抑えたまま推論深度を拡張

制御ゲートによる活性値発散の防止

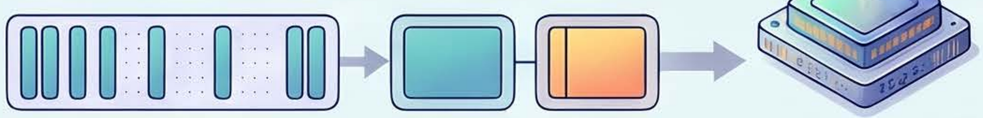






トークン単位の漸進的更新により、反復処理時の数値膨張を防ぎ動的制御を実現

ハイブリッドKVキャッシュでのメモリ削減



共有長文脈 (疎) → 局所ウィンドウ (保持) → データアニップ処理

長文脈の疎なKVを共有しつつ局所ウィンドウのみ保持し、100万トークン長文脈に対応

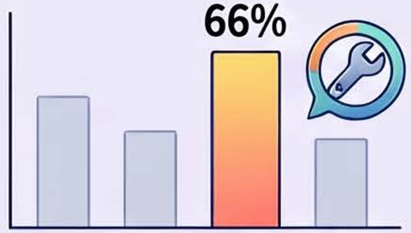
実務評価と運用最適化のポイント

思考強度の明示的調整

High | Medium | ⚠ Max (非推奨)

デフォルトのMaxを避けHighやMediumに指定して
過剰思考によるタイムアウトを回避

脆弱性監査における高い自律修復力



GLM-5.4

ツール連携エラー後のリトライ修復率は
66%を記録し、他モデルを凌駕

日本時間 (JST) のオフピーク割引活用



平日15〜19時以外は消費クレジット半額となり、夜間や週末の低コスト運用が可能

GLM-5.4と前世代GLM-5.3の主要スペックおよび実戦評価比較

比較項目	GLM-5.4 (新モデル)	前世代 GLM-5.3
アーキテクチャ	Loop Transformer (中間28層×2周)	標準デコーダ型 MoE (単一パス)
入力モダリティ	テキスト・画像 (ビジョン統合)	テキスト専用
KingBench 4スコア	33.2点 (オープン系1位)	27.5点

Gemini Notebook