

次世代AI「GLM-5.3-Flash」：フロンティア級の知能を破壊的低コストで

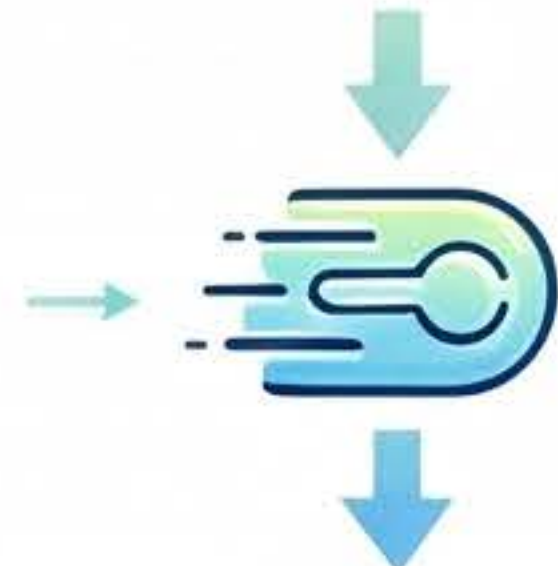
Z.ai (智譜AI) が発表したGLM-5.3-Flashは、320Bという巨大なパラメータを持ちながら、独自のハイブリッド注意機構によりAPI単価を標準版の約9分の1以下に抑制。フロンティア級の知能を「Flash級」の低コストで提供することに特化した、オープンウェイト (MITライセンス) の最新マルチモーダルモデルです。

「Flash」が定義する新たな価値： 知能とコストの再両立

DEFINITION



320Bモデル



推論コスト・API価格
を極限まで削減

「Flash」は小型・最適ではなく「低コスト・高効率」を意味する。320Bの巨大なモデルサイズを維持しつつ、推論コストとAPI価格を極限まで削減。

KEY_FINDING



MITライセンスによる完全オープンな328GBのウェイト。商用利用可能なオープンウェイトとして公認され、企業独自のオンプレミス運用も視野。



STATISTIC



標準版コスト



GLM-5.3-Flashコスト

100万トークンあたり入力50.15/出力50.50という、既存のフロンティアモデルを凌駕する経済性。

主要なフロンティア/Flash系モデルとのAPIコスト比較 (10万入力+1万出力時の試算)

GLM-5.3-Flash
(割引時)

\$0.010

\$0.010

圧倒的な価格効率・マルチモーダル

Gemini 3.7 Flash

\$0.112

\$0.112

Googleの低コスト・マルチモーダル

GPT-5.6 Terra

\$0.160

\$0.160

OpenAIの中位フロンティアモデル

Claude Opus 5

\$0.750

\$0.750

Anthropicの最高峰推論モデル

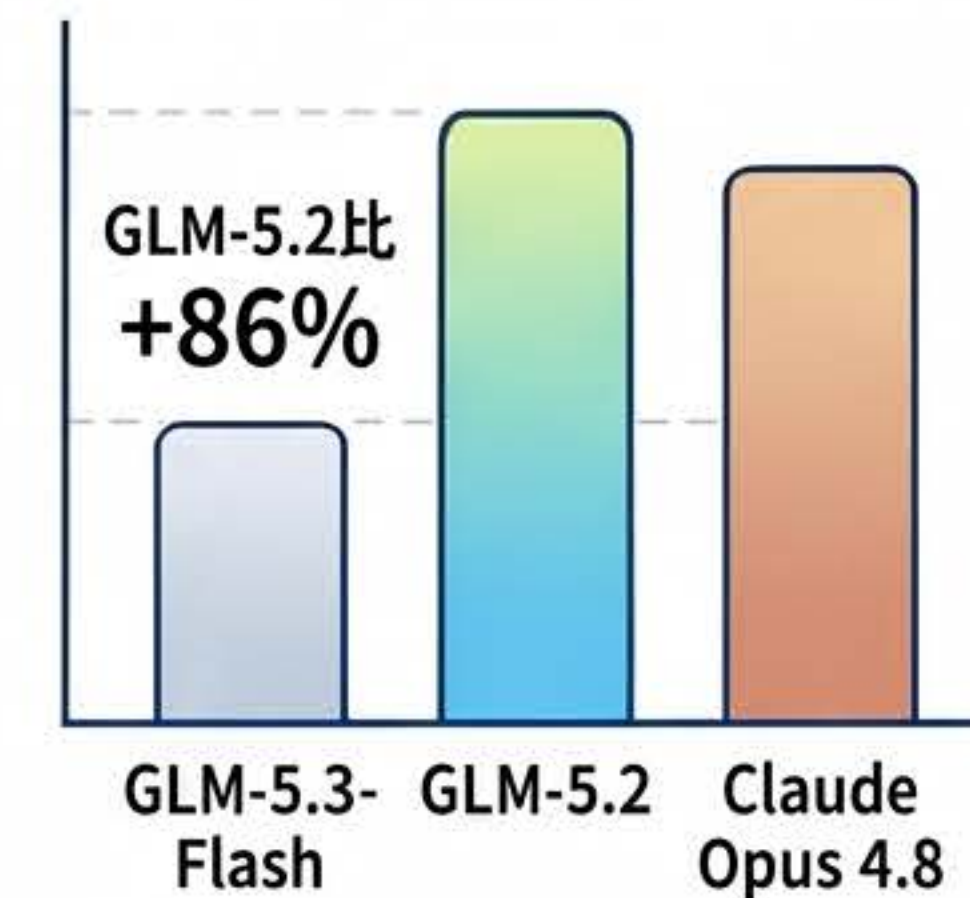
概算費用 (低→高)

技術的優位性と独立評価の実態



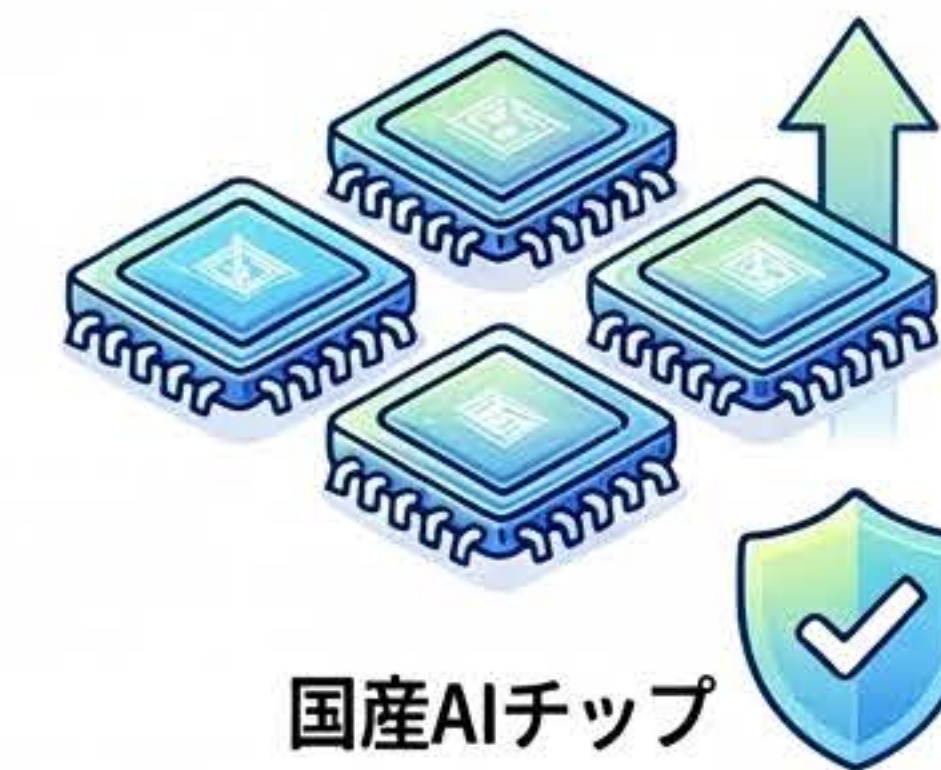
PROCESS_STEP

線形+疎注意のハイブリッド構造による効率化。標準版比で計算量を約3倍、KVキャッシュを4.4倍削減し、1Mコンテキストに対応。



COMPARISON

旧世代を圧倒し、最新競合モデルに迫るエージェント能力。AutomationBench等の実務タスクでGLM-5.2を86%上回り、Claude Opus 4.8に肉薄。



国産AIチップへの高度な最適化

中国製AIチップへの高度な最適化。数億個の国産チップでの通用を目標に、独自のサービングスタックで性能を3倍に改善。