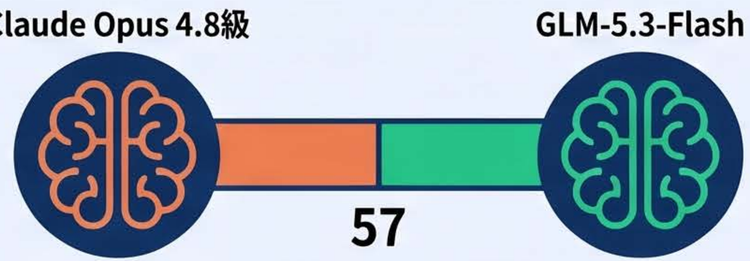
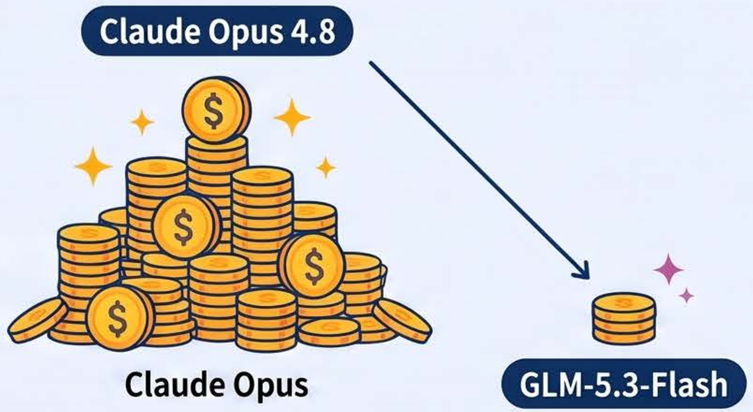
GLM-5.3-Flash：最高峰の知能を「Flash」の価格で提供するAIの新基準

圧倒的なパフォーマンスと破壊的経済性



Claude Opus 4.8級の知能を1/40のコストで

総合知能指数で57を記録し、最高峰モデルと同等の実力を圧倒的な安価で提供。



実務・コーディングにおける高い課題解決力

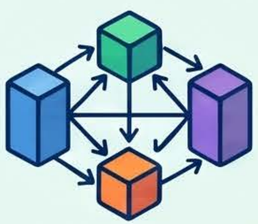
DeepSWEベンチマークで63.4を記録し、既存の主要フロンティアモデルを凌駕。

ネイティブ・マルチモーダル＆常時思考モード

画像・動画を直接理解し、タスクの複雑さに応じて思考深度を動的に切り替え可能。

革新的なアーキテクチャと国産インフラの最適化

A18B
アクティブパラメータ
(MoE & Hybrid Attention)



KVキャッシュ

4.44倍圧縮

推論を高速化

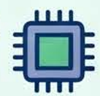
疎なMoE構造とIndexPool技術により、KVキャッシュを4.44倍圧縮し推論を高速化。

圧倒的な価格優位性

(100万トークンあたり)

モデル名	入力価格 (USD)	出力価格 (USD)	対Opus 価格比
Claude Opus 4.8	$5.00	$25.00	100% (基準)
GLM-5.3-Flash	$0.15	$0.25	約 2.5%
期間限定セール	$0.075	$0.125	約 1.25%

10万基規模の国産チップクラスタで完全運用



特定ベンダーのハードウェアに依存せず、NVIDIA製GPUと同等のコスト効率を達成。

EPD分離アーキテクチャによる負荷分散

エンコード (E) | 計算 (P) | デコード (D)

100万トークンの長文脈処理を安定化

エンコード・計算・デコードを分離し、100万トークンの長文脈処理を安定化。