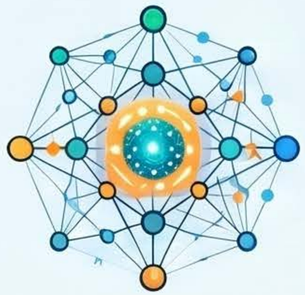
次世代AI「GLM-5.2」：100万トークンが変える開発と知財の未来

技術的ブレークスルー：極大コンテキストと自律性

実用的な「100万トークン」の壁を突破

DeepSeek Sparse Attention (DSA) メカニズムにより、長文処理の計算負荷を劇的に削減。大規模なコードベースや数百ページの特許文書を一度に処理可能です。



744Bパラメータの MoEアーキテクチャ

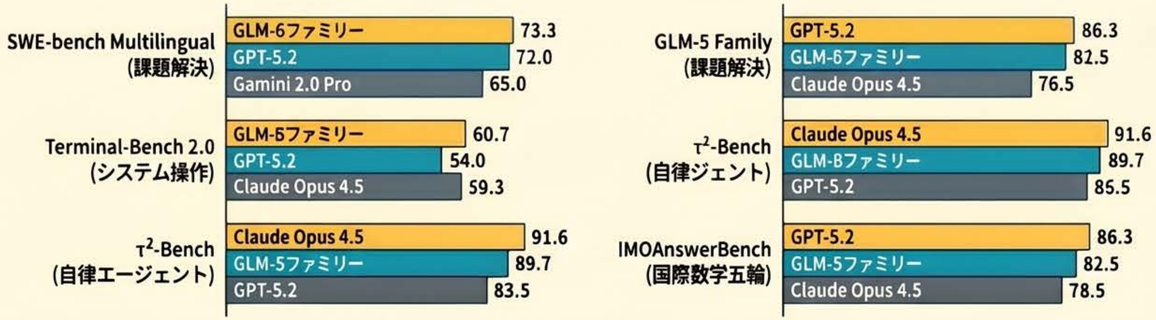
総パラメータ数7,440億のうち、推論時にアクティブになるのは約400億に抑制し、高い表現力と実用的な推論コストを両立しています。

非同期強化学習インフラ「Slime」

データ生成、採点、検証を同一データバスでシームレスに結合し、長期的な計画能力と「思考の連鎖(CoT)」をモデル内部にネイティブ構築しました。

パフォーマンス比較：商用トップモデルへの肉薄

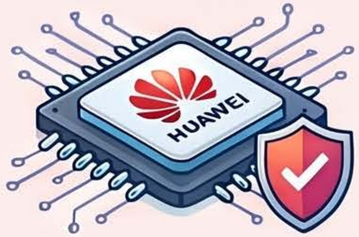
GLM-5ファミリーと主要な競合モデルのベンチマーク比較



エンジニアリング領域でGPT-5.2を凌駕:
多言語コーディングやシステムレベルのターミナル操作において、商用の最先端モデルを上回るベンチマークスコアを記録しています。

圧倒的なコスト競争力:
月額サブスクリプション (約18ドル～) を通じて、Claude Opusクラスのモデルを他社比数倍のトークン効率で利用可能です。

地政学的・戦略的意義



NVIDIA非依存の「独立性」を証明
米国の輸出規制対象外であるHuaweiの「Ascend」チップ上で完全学習。特定国のハードウェアに縛られない生存戦略の中核です。

完全MITライセンスによる「民主化」
API遮断リスクを回避するため、ローカルホスト可能なオープンウェイトとして公開。機密性の高い金融・政府機関での利用を可能にします。

知財（IP）実務における変革

特許解析の「文脈の非断片化」
請求項と実施形態、審査履歴を分割せずに一括解析。法的な依存関係を損なう「チャンキング」の問題を解消します。

弁理士コストの劇的削減
数百ドルの人件費がかかる特許マッピングをわずか数セントで実行。1移分の弁理士コストで1件の解析が完了する計算です。

ハイブリッド戦略：長文脈 × RAG
1Mトークンで全体を俯瞰しつつ、特定した重要箇所を精密に再解析することで、捏造を抑えた厳格な証拠性を確保します。