

GLM-5 : Zhipu AIの次世代巨大MoEモデル徹底解剖

GLM-5はZhipu AIのフラッグシップ基盤モデル。MoE構造で推論時のみアクティブ化。事前学習し、コンテキスト長をサポート。高効率と長文処理能力を両立。

基本仕様とスケール



28.5T

事前学習トークン
(23T から 28.5T増加)



200K

コンテキスト長
(最大202,752トークン処理)



商用API提供
(Z.ai)



オープンウェイト
(MITライセンス,
Hugging Face等)

GLM-5 MODEL CORE (MoE)

総パラメータ: 744B

推論時アクティブ: 40B
(Top-k=8 Experts)

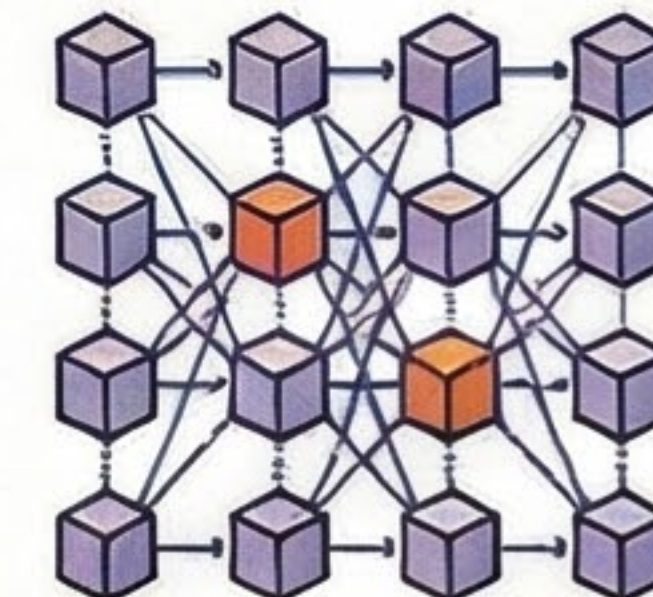
技術的イノベーション



DeepSeek Sparse
Attention (DSA) 統合
計算複雑性を低減、
トークン効率制的改善



非間隔RL基盤「slime」採用
Megatronと5GLangを接続、
高スループット強化学習



MoE構成: 256エキスパート
(Top-k=8 アクティブ化)

ベンチマーク性能 (推論・コード・エージェント)

SWE-bench Verified: 77.8 (コーディング能力)

GPQA-Diamond: 86.0 (専門知識推論)

HLE (Humanity's Last Exam): 30.5 / 50.4 (w/ Tools)
(議解試験正答率)

コスト効率と環境負荷



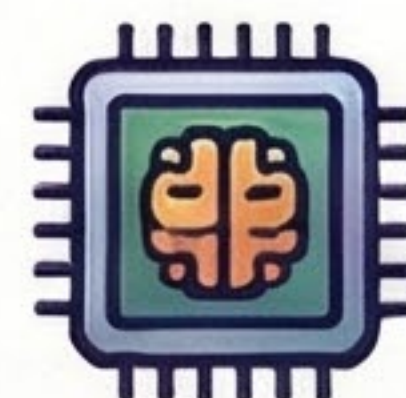
出力単価:
GLM-5 (\$3.20/1M) vs GPT-5 (\$10.00/1M)

主要モデル比較 (API \$/1M pts):

- GLM-5 (200K, 入力\$1.00, 出力\$3.20)
- GPT-5 (400K, 入力\$1.25, 出力\$10.00)
- GPT-4o (未指定, 入力\$2.50, 出力\$10.00)



推定CO2排出量:
約1,130~1,430トン
(学習総電力量 約2.30GWh)



推定計算量
 6.84×10^{24} FLOPs
(H100 GPU約11.4万日分相当)

留意点と矛盾点



公開情報の矛盾:
公式は「744B」を正とすべき
(一部メディア「745億」表記)



透明性と安全性の課題:
学習データ詳細・安全性評価文書は
「Coming Soon」