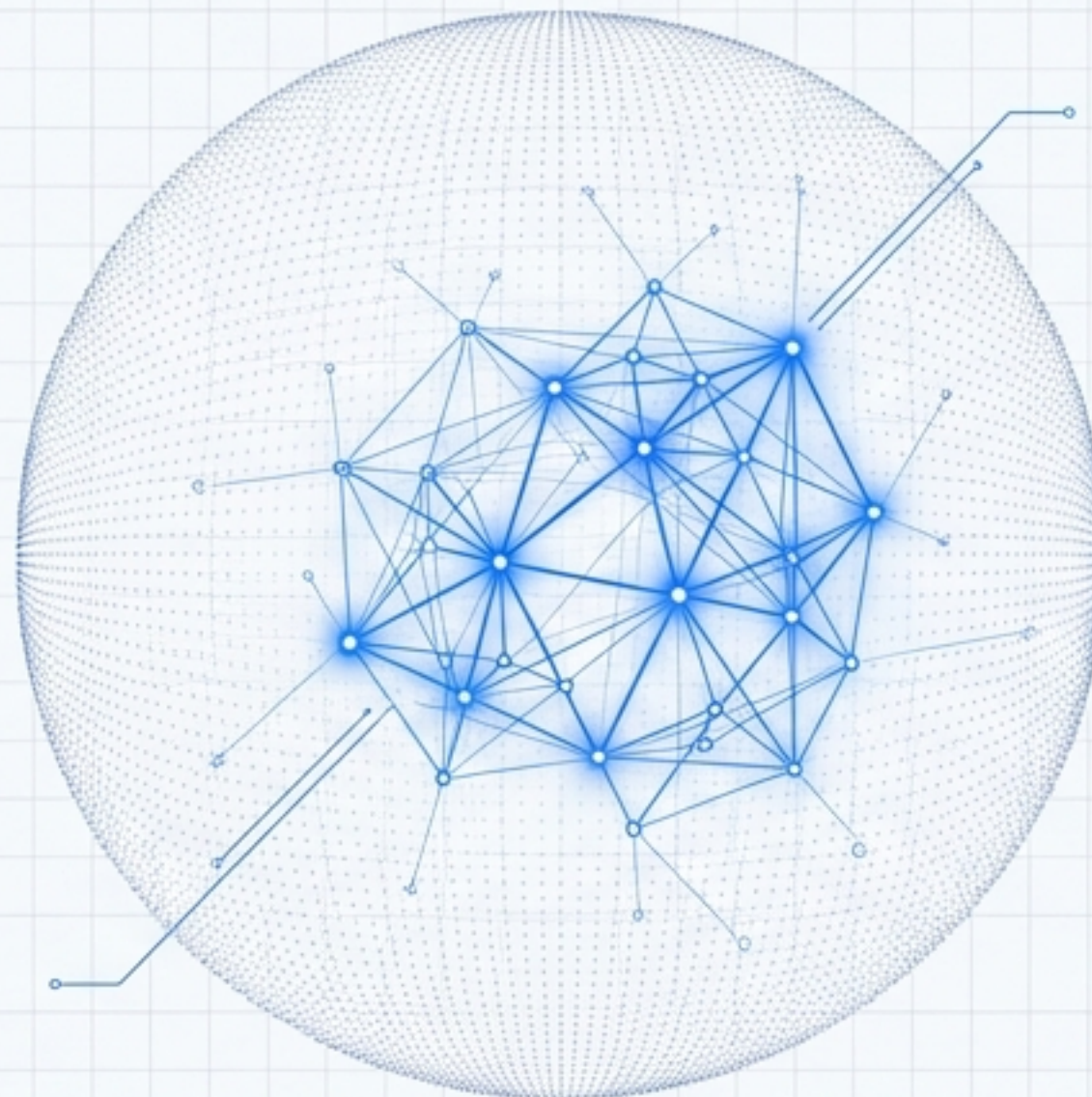




GLM-5 Strategic Intelligence Briefing

Zhipu AIフラッグシップモデルの技術的解剖と戦略的示唆

Primary Source Analysis & Technical Due Diligence (一次情報に基づく厳密分析)

本レポートは、Zhipu AIの公式ドキュメントおよびGitHubリポジトリ(2026-02-11更新)の一次情報を基盤とし、市場の噂と事実の乖離を検証したものである。

エグゼクティブサマリー：効率性を極めた「実用」重視の巨人

CORE SPECS (仕様)

744B

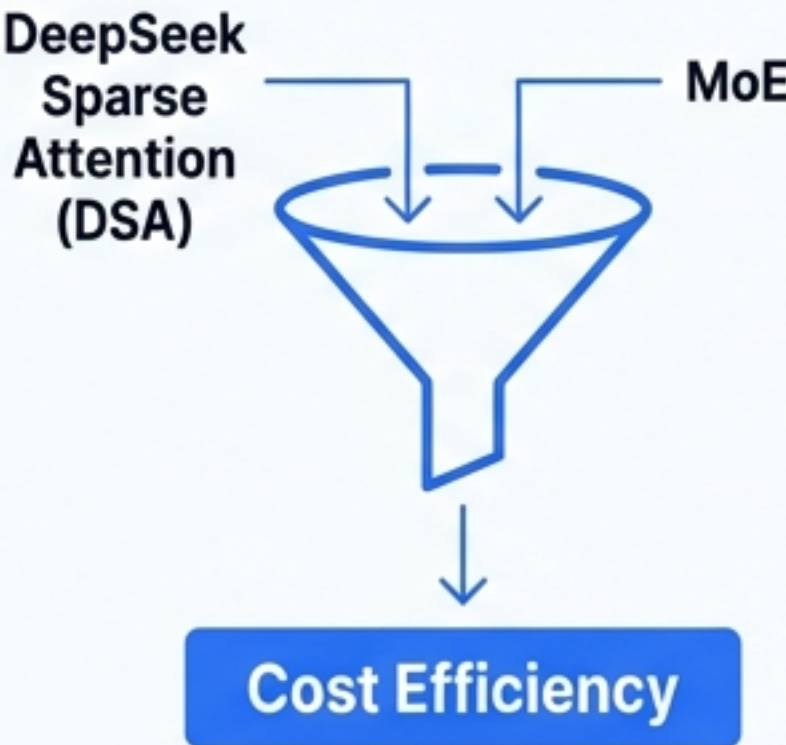
Total Parameters
Inter/Noto Sans JP Regular

40B

Active (Inference)

- Context: **200K Tokens**
- Training Data: **28.5T Tokens**

KEY DIFFERENTIATORS (差別化要因)

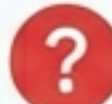


- 長文性能と推論コスト低減の両立
(Balancing Long-Context & Low Cost)
- 非同期RL基盤「slime」による
エージェント能力の強化

MARKET POSITION (市場戦略)

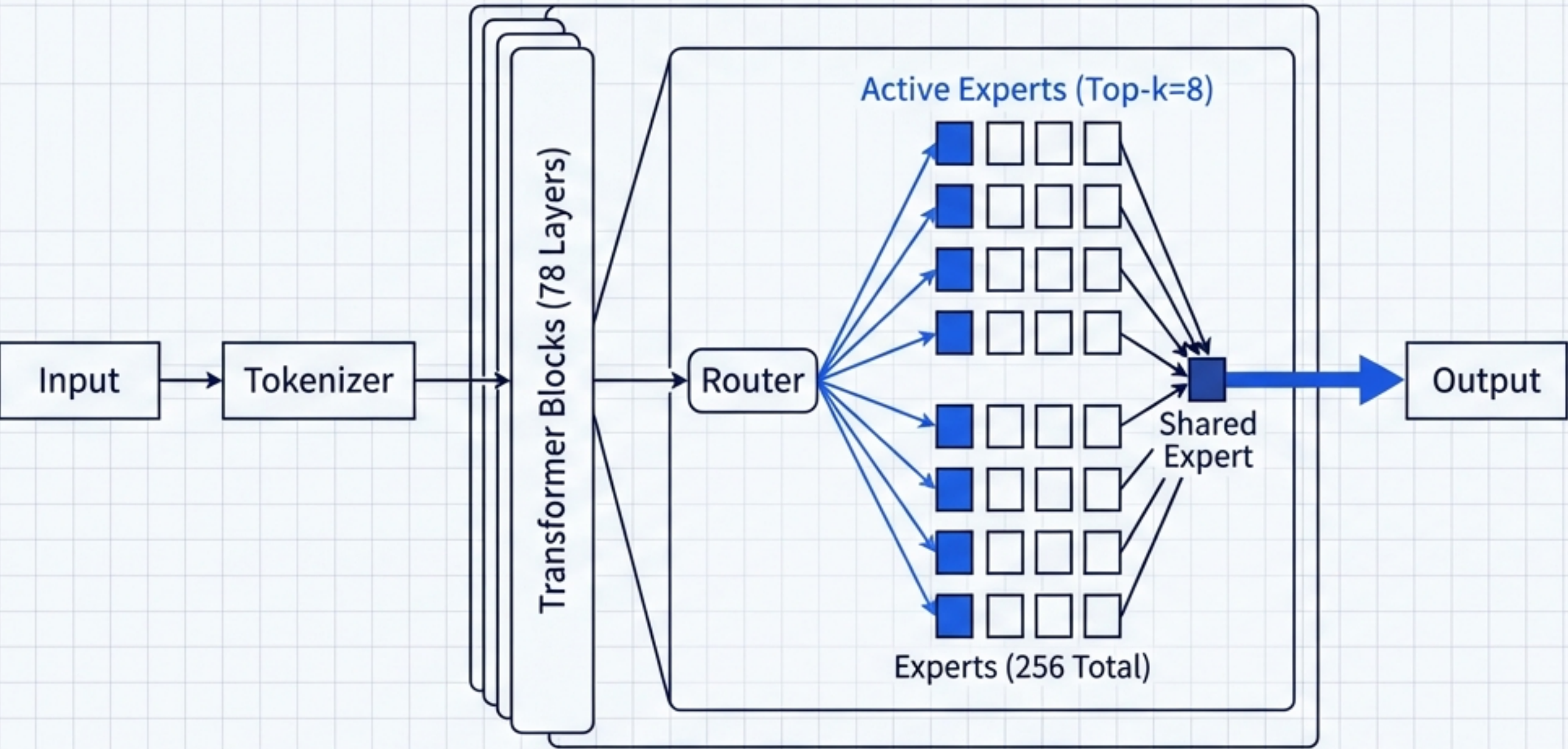
- Status: 商用API + オープンウェイト (MIT License) 
- Focus: **Coding & Agents**
(SWE-bench Verified **77.8**)
- Economics: 出力コストがGPT-5の約1/3 

情報の厳密性検証：市場の噂 vs 一次情報

MEDIA RUMORS (報道・噂)		OFFICIAL SOURCES (公式発表 - GitHub/Docs)	
Inter/Noto Sans JP Regular		Inter/Noto Sans JP Regular	
Parameters	745億 (74.5B) / 44億	Parameters	744B (7440億) / 40B Active 
Hardware	Huawei Ascend + MindSporeで学習  Unverified	Hardware	未指定 (Unspecified) (ただしAscend推論対応は確認済)
Tech Noisy記事等による誤報の可能性		config.json等の実コードに基づく数値	



アーキテクチャの全貌：巨大なMoEとスパース性



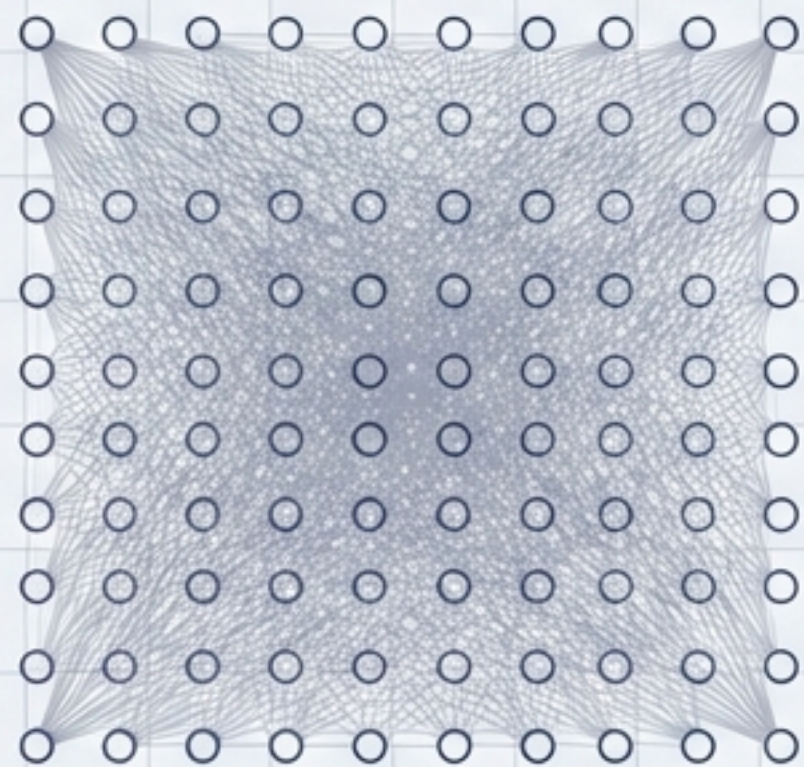
CONFIG.JSON SPECS	
Hidden Size	6144
Attention Heads	64
FFN Intermediate	12288
Total Params	744B
Active Params	~40B

総パラメータはGPT-4クラスだが、推論時の計算量はLlama-3-70Bクラス（40B Active）に抑え込まれている。「賢い間引き」による効率化。

DeepSeek Sparse Attention (DSA) の統合

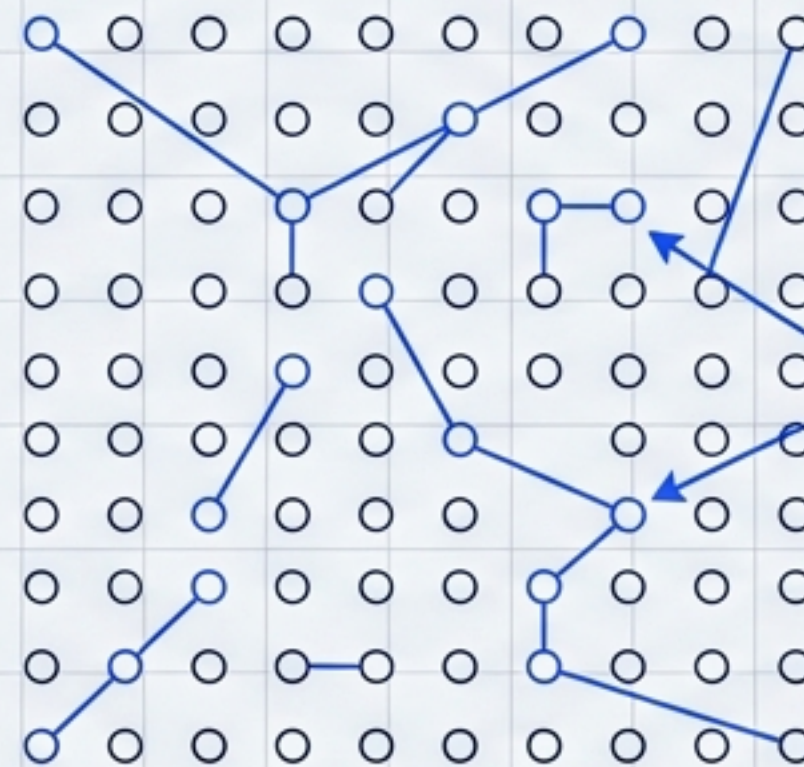
Attention Mechanisms

Standard Full Attention



Quadratic Complexity ($O(n^2)$)

DeepSeek Sparse Attention (DSA)



Focus on Relevant Tokens Only

Reduced Complexity (Sparse)

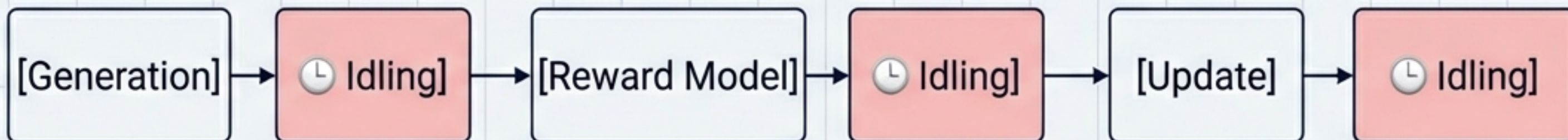
Context Length: **200K Tokens**

Max Position Embeddings: **202,752**

Strategic Note: Zhipu AIの独自発明ではなく、DeepSeekの技術 (DSA) を実用レベルで統合。長文・エージェント用途での「ロスレスな長文性能」と「デプロイコスト圧縮」を実現。

ポストトレーニングの革新：「slime」と非同期RL

**Traditional RLHF
(Synchronous)**



**Slime Framework
(Asynchronous)**

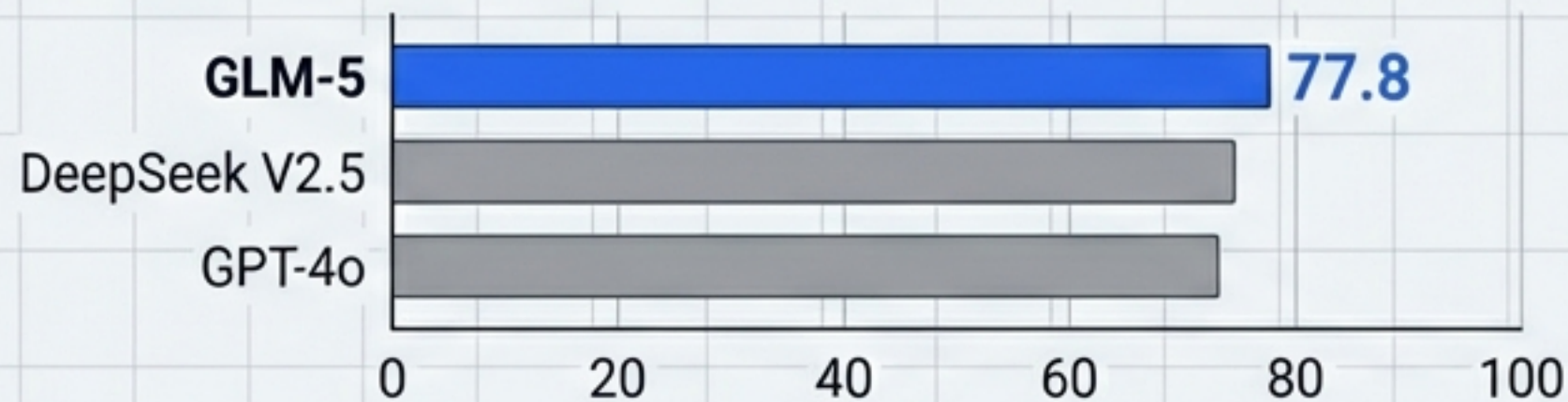


Key Takeaway: 単なる人間の好み（RLHF）の学習ではなく、長距離インタラクションや複雑なエージェント動作からの「継続学習」を可能にするインフラ。人間フィードバックのボトルネック解消。

ベンチマーク評価：知識から「行動」へのシフト

Action & Coding Capabilities

SWE-bench Verified



Terminal-Bench 2.0



GPQA-Diamond



The Missing Data (未公開データ)

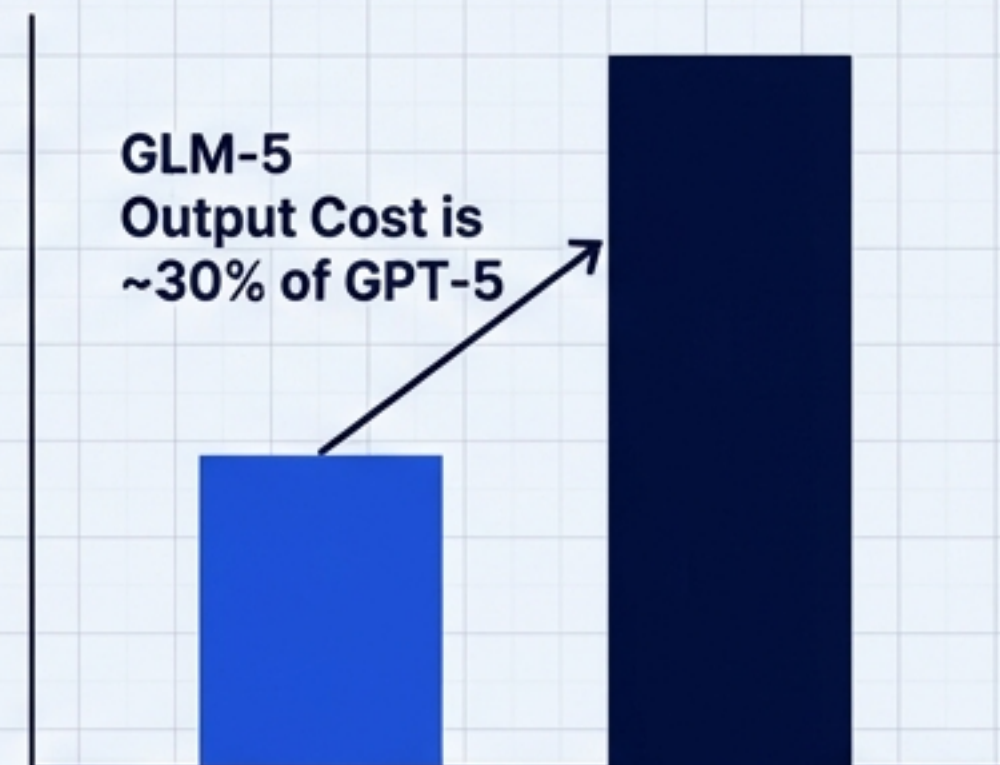
MMLU / C-Eval / LAMBADA

📄 Systematic scores not presented in public documents.

「百科事典的知識」のスコアよりも、ツール利用やコーディングといった「問題解決能力」の数値を優先して公開。実用特化の姿勢が鮮明。

圧倒的な価格競争力とコスト構造

Cost per 1M Tokens (USD)			
	Model Name	Input Cost	Output Cost
1	GLM-5	\$1.00	\$3.20
2	GPT-5 (OpenAI)	\$1.25	\$10.00
3	GPT-4o (OpenAI)	\$2.50	\$10.00



- **Cached Input (\$0.20)** も提供されており、定型プロンプト運用も安価。
- **大量のコード生成、ログ解析など「出力重視」**のタスクでコストメリットが最大化する。

推定環境負荷と学習計算量 (Estimated Environmental Impact)

The Formula

Estimated FLOPs $\approx 6 \times N$ (Active Params) $\times D$ (Tokens)

$N = 40B, D = 28.5T$

$\approx 6.84 \times 10^{24}$ FLOPs

Resource Equivalency



Hardware: 1,024 NVIDIA H100s

Time: ~112 Days (at 35% utilization)

Environmental Impact



Energy: ~2.30 GWh



Carbon: ~1,130 t-CO2

Takeaway: オープンウェイトモデルとして公開されているが、再学習 (Pre-training from scratch) には巨大な資本と計算リソースが必要。

透明性と安全性の課題

RISK ASSESSMENT (Inter Tight Bold)

UNKNOWNNS (Black Box Risks)

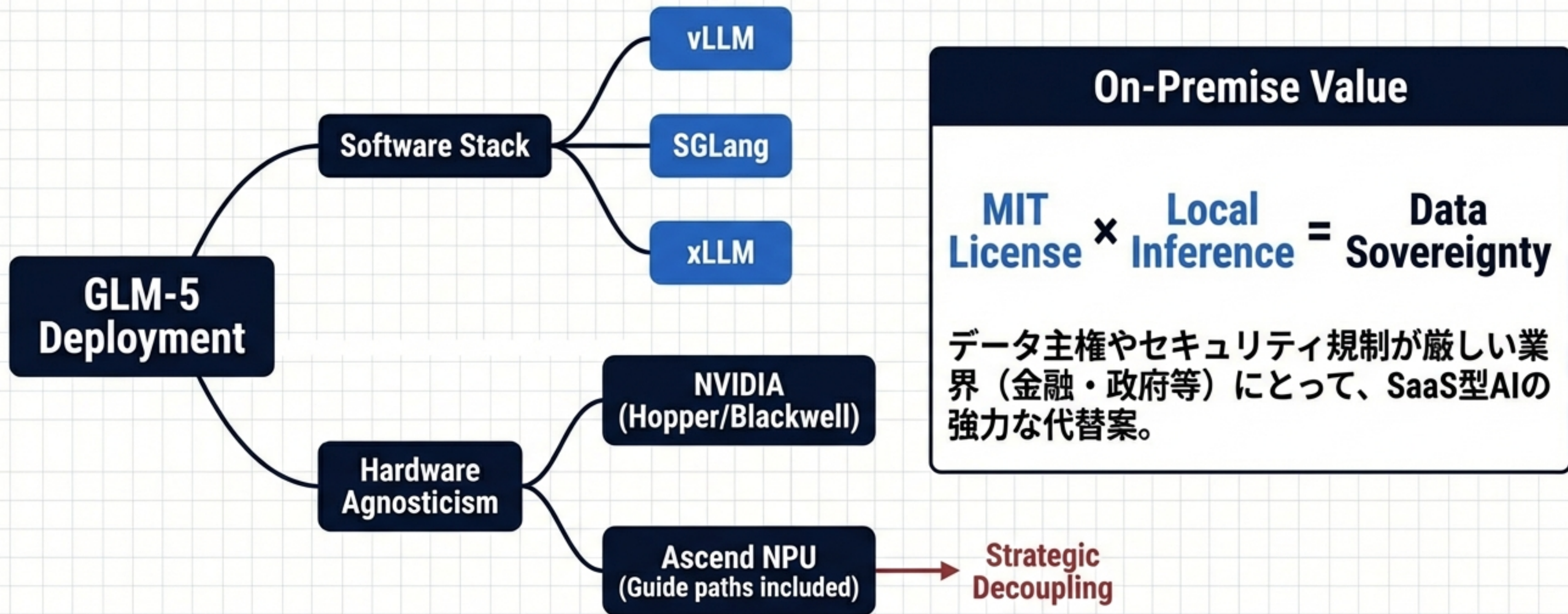
- ❗ **Training Data Composition:** 28.5T tokens confirmed, but rights/PII status undisclosed.
- ❗ **Safety & Red Teaming:** No comprehensive System Card available (unlike OpenAI).
- ❗ **Bias & Toxicity:** Official benchmarks absent.

KNOWNNS (Mitigations)

- ✓ **Weight Access:** Open weights allow independent safety testing.
- ✓ **Local Control:** Possibility for private fine-tuning.

再現性は低い。企業利用の際は、自社基準での安全性評価（Red Teaming）が必須となる。

エコシステムと地政学的含意



再現性は低い。企業利用の際は、自社基準での安全性評価 齟齬する時全堀が必要となる。

結論：誰がGLM-5を選ぶべきか？



The Smart Saver (賢い節約家)

- **744B** Performance at **40B** Cost.
- Output cost is **1/3** of competitors.



The Pragmatic Pro (実務のプロ)

Specialized in Coding & Agent Actions.
Not for encyclopedia trivia,
but for getting work done.



The Sovereign (自律の追求者)

Open Weights / MIT License.
Hardware independence & On-premise capable.

GPT-4/5の完全な代替（知識量）ではないが、コスト感度が高い「大量生成タスク」や、機密性を要する「社内エージェント構築」において、現時点で最も合理的な選択肢の一つ。

参考文献・データソース (References & Data Sources)

Primary Sources (Zhipu AI)	
Z.ai Docs (Pricing/Guides)	
GitHub (GLM-5, slime)	
Hugging Face (Model Card)	
Technical Context	
DeepSeek Sparse Attention (arXiv Paper)	
OpenAI GPT-4 System Card	
Media Check	
Tech Noisy (Contrast Source)	

本分析は2026年2月時点の公開情報を基に作成されています。将来のアップデートにより数値や仕様は変更される可能性があります。