ChatGPT

GLM-5に関する分析的・厳密報告書

Executive Summary

- GLM-5は、Zhipu AI [1]（Z.ai）が公開したフラッグシップ級の基盤モデルで、開発者向け公式ドキュメント上は**総計744B（=7,440億）／推論時アクティブ40B（=400億）**、**コンテキスト長200K**、**最大出力128K**、**事前学習トークン28.5T**が明示される。 [2]
- 公式の公開形態は「**商用API**（Z.ai API）＋**オープンウェイト配布**（Hugging Face等）」のハイブリッドである。モデルカードおよび公式リポジトリは「**DeepSeek Sparse Attention（DSA）**の統合」と、非同期RL基盤「**slime**」の利用を主要技術として掲げる。 [3]
- 指定記事（TECH NOISY）は「**総パラメータ745億／推論時44億**」「**Huawei Ascend＋MindSporeでトレーニング**」と記述するが、これは公式ドキュメントや公式リポジトリが示す“744B/40B”と**桁が1つ異なる**（745億≒74.5B）。本報告書では、一次ソース（公式ドキュメント／公式配布物）を優先しつつ矛盾点を明示する。 [4]
- 公開ベンチマークとして、Z.aiは**SWE-bench Verified 77.8**、**Terminal-Bench 2.0 56.2**等を含む推論・コーディング・エージェント系指標を提示している一方、ユーザ指定のMMLU/C-Eval/LAMBADA等の“古典的学術ベンチマーク”の全面的スコアは、少なくとも公開物上では体系的に提示されておらず、技術レポートも「coming soon」とされている。 [5]
- 環境負荷について、公式が明示する**アクティブ40B**と**28.5T tokens**を用い、学習計算量を近似式 **FLOPs ≈ 6 × N × D**（Nは“forward-activated”パラメータ、Dはトークン数）で推定すると、事前学習は概算で**6.84×10^24 FLOPs**規模となる（仮定と感度分析は後述）。 [6]

基本情報と公開形態

発表日・モデル名・公開チャネル

GLM-5は2026-02-11頃に公式GitHub組織・リポジトリが更新され、同日に「GLM-5」リポジトリが“Updated Feb 11, 2026”として確認できるため、少なくともこの時点で公式公開が進行したと判断できる。 [7] さらに、報道（Reuters [8]）はZhipuがGLM-5を発表し、オープンソース形態を採る旨を伝えている。 [9]

公開形態は、Z.aiの開発者ドキュメントが「GLM-5 API」の利用方法を示し、加えて公式がHugging Face上でモデル配布（オープンウェイト）を行っていることから、**商用API＋オープンウェイト配布**が一次情報で裏付けられる。 [10]

基本仕様表

項目	内容
発表日	**2026-02-11頃（公式GitHub更新時点で確認）**。報道でも発表が報じられる。
モデル名	GLM-5
パラメータ数	**公式：744B総計／40Bアクティブ**。一方、Hugging Face表示は**754B params**とも記載（整合性は後述）。指定記事は**745億／44億**と記述（公式と矛盾）。

項目	内容
コンテキスト長	**200K**（max_position_embeddings 202,752）
公開形態	**商用API（Z.ai）＋オープンウェイト（Hugging Face等）**
ライセンス	**ウェイト：MIT（Hugging Face）**。参考：公式GitHubリポジトリ（コード類）はApache-2.0。

上表の根拠：Z.ai公式ドキュメント（コンテキスト長、パラメータ規模、アクティブ規模、API利用）[2]、Hugging Faceモデルページ（ライセンスMIT、モデルサイズ表示、公開形態）[11]、公式GitHub（更新日・リポジトリ、Apache-2.0表記）[12]、指定記事（745億/44億等の主張）[13]。

「745億」表記の矛盾整理（厳密性のため）

- 公式ドキュメント／公式GitHubは**“744B parameters (40B activated)”**という表記を一貫して用いる。[14]
- 一方、Hugging Faceのモデル表示には「Model size **754B params**」とある（同ページ内の説明文では744B表記も併存）。公式側の説明とUI表示が一致しないため、**(A) 744Bは概数／(B) 754Bは埋め込み等を含む別計数／(C) 更新差分**などが考えられるが、一次情報だけでは断定できない。[15]
- 指定記事は「総パラメータ745億、推論時44億」と記す。これは公式の“744B/40B”と**1桁（×10）**ずれるため、**“745億”が“7450億（=745B）”の誤記**、または別モデル／別設定の混同である可能性が高いが、記事自体は一次ソースではないため、本報告書では**“矛盾あり”として併記**する。[16]

アーキテクチャ

公式設定ファイルから読める構造（推定ではなく実データ）

Hugging Face上の`config.json`は、GLM-5が**“GlmMoeDsaForCausalLM”**という因果言語モデル（自己回帰）として定義され、`model_type`は`glm_moe_dsa`であることを示す。[17]
主要ハイパーパラメータ（“未指定”ではなく設定ファイル由来）は以下の通りである。

- 層数：`num_hidden_layers = 78`[17]
- 埋め込み次元：`hidden_size = 6144`[17]
- 注意ヘッド：`num_attention_heads = 64`、`head_dim = 64`[17]
- FFN：`intermediate_size = 12288`（dense部）[17]
- 位置エンコーディング：RoPE系（`rope_type="default"`、`rope_theta=1000000`、`max_position_embeddings=202752`、`rope_interleave=true`）[17]

この`max_position_embeddings=202,752`は、公式ドキュメントの「Context Length 200K」と整合する（“200K”は概数）。[18]

MoE（Mixture-of-Experts）構成

設定ファイルは、MoE関連として少なくとも以下を明示する。

- ルーティングエキスパート数：`n_routed_experts = 256`[17]
- 共有エキスパート数：`n_shared_experts = 1`[17]
- 1トークン当たり利用エキスパート：`num_experts_per_tok = 8`[17]
- MoE層頻度：`moe_layer_freq = 1`（毎層適用を示唆）[17]

- 先頭のdense置換：`first_k_dense_replace = 3`（冒頭3層をdense扱いにする設計を示唆） [17]

公式ドキュメントは、GLM-5が「**744B（40B　activated）**」へスケールし、長文性能とコスト低下の両立を狙うと説明している。 [2]
この“Activated　40B”は、MoEで**計算量がアクティブ部分に依存**するという一般的性質と整合的であり、後述の学習FLOPs推定でも重要な入力となる。 [19]

DSA（DeepSeek Sparse Attention）統合の位置づけ

Z.ai公式ドキュメントは、GLM-5が**DeepSeek Sparse Attention（DSA）を初めて統合**し、「長文性能を維持しつつデプロイコストを大幅に下げ、Token Efficiencyを改善する」と述べる。 [2]
DSAの一次ソースとしては、DeepSeek [20] の技術報告（arXiv）が、DSAを「長コンテキストで性能を維持しつつ計算複雑性を下げる注意機構」と位置づけている。 [21]

これにより、GLM-5のアーキテクチャ的独自性は「DSAそのものの新規提案」ではなく、**巨大MoE（744B/40B）にDSAを統合し、長文・エージェント用途へ最適化**した点にあると解釈できる（ただし統合の具体（どの層でどの疎性を採用等）は技術レポート未公開のため、詳細は未指定）。 [22]

概念図（Mermaid）

```
flowchart LR
  A[入力: 最大200K tokens] --> B[Tokenizer / Embedding]
  B --> C[Transformer Blocks ×78]
  C --> D{Attention}
  D -->|DSA: Sparse Attention| E[長文効率化]
  C --> F{FFN}
  F -->|MoE Router| G[256 routed experts<br/>top-k=8]
  G --> H[Activated params ≈ 40B]
  C --> I[出力: 最大128K tokens]
  J[Post-training] --> K[slime: async RL infra]
  K --> C
```

トレーニングデータ

データ量・モダリティ・言語分布

一次情報として、Z.ai公式ドキュメントおよび公式リポジトリは、事前学習データ（トークン数）が**23T→28.5T tokens**へ増加したと述べる。 [14]
モダリティは公式ドキュメント上で**Text入出力**とされ、少なくともGLM-5そのものはテキストモデルとして提供される。 [2]

一方で、データソース（ウェブ、書籍、コード等の内訳）、言語分布（中国語/英語比率など）、フィルタリング・クレンジング手法、データ透明性（再現可能なデータカード等）の詳細は、公開一次情報からは確認できず**未指定**である（技術レポートも“coming soon”とされる）。 [23]

透明性・再現性の評価（厳密な位置づけ）

- **トークン総量（28.5T）**は明示される一方、コーパスの出所・除外基準・重複排除・毒性/PII除去などのプロセスが公開されないため、外部研究者が学習データ面から性能・偏り・権利リスクを検証するのは困難である（現時点では再現性は低いと評価せざるを得ない）。 [10]
- ただし、**オープンウェイト**であり、推論挙動の追試・安全性評価・圧縮/量子化研究への利用は可能である。 [23]

トレーニング手法と最適化

事前学習目的関数・精度・分散学習

GLM-5は設定上`ForCausalLM`であり、モデル種別としては自己回帰型の因果言語モデル（次トークン予測）であると解釈できる。 [17]
混合精度については、公式配布の設定に`dtype = bfloat16`が記載されるため、少なくとも公開ウェイトはBF16を前提にしている。 [24]

学習率スケジュール、バッチサイズ、optimizer、分散学習方式（tensor/pipeline/data並列、expert並列など）は一次情報からは確認できず**未指定**である（今後の技術レポート待ち）。 [11]

ポストトレーニング（RLHF類似）の位置づけ

公式ドキュメントは、事前学習後の強化学習について「RLの非効率性」を課題とし、これに対して**“slime”という非同期RL基盤**を開発し、より細粒度のポストトレーニング反復を可能にしたと記述する。 [2]
slimeの公式リポジトリは、Megatron（学習）とSGLang（rollout/生成）を接続し、高スループット学習と柔軟なデータ生成を提供する“LLM post-training framework for RL scaling”であると説明している。 [25]

したがって、GLM-5の“RLHF等”は、一般的な同期型RLHF（PPO等）だけでなく、**非同期・分散ロールアウトを前提にしたRLスケーリング**志向である可能性が高い。ただし、採用アルゴリズム（PPO/DPO/RLAIF、報酬モデル構成、拒否訓練の有無等）は一次情報では**未指定**である。 [26]

使用ハードウェア

公式の一次情報（Z.aiドキュメント、Hugging Faceモデルカード、公式GitHub）からは、学習に使用したGPU/NPUの種類・台数は確認できず**未指定**である。 [27]

指定記事は「Huawei AscendチップとMindSporeでトレーニング」と断定的に述べるが、一次ソースの裏付けが本文中からは確認できないため、本報告書では**“二次情報（要検証）”**として扱う。 [13]
一方で、公式配布物は「Ascend NPU向けのデプロイ導線（xLLM等）」を明示しており、少なくとも推論・運用面でAscend系を視野に入れていることは一次情報で確認できる。 [28]

性能評価

公式が公開しているベンチマーク群（長期・エージェント寄り）

GLM-5の公式モデルカードは、従来型のMMLU等ではなく、**推論・コーディング・エージェント**寄りのベンチマークを中心に比較表を提示している。たとえば以下が明示される（抜粋）。

- SWE-bench Verified：**77.8**（GLM-5） [11]

- Terminal-Bench 2.0（Terminus 2）：**56.2 / 60.7** [11]
- GPQA-Diamond：**86.0** [11]
- Humanity's Last Exam（HLE）：**30.5**、HLE（w/ Tools）：**50.4** [11]

これらは、同表内でDeepSeek系・Kimi系・Claude系・Gemini系・GPT-5.2などと比較されているが、評価条件（最大生成長131,072 tokens、HLE-with-toolsでcontext 202,752等）が脚注で指定されており、**"長文・長生成"を前提にしたレジーム**である点に注意が必要である。 [29]

指定ベンチマーク（MMLU / C-Eval / HumanEval / LAMBADA等）に関する扱い

ユーザ指定の **MMLU / C-Eval / LAMBADA** 等について、GLM-5の一次情報は少なくとも公開ページ上で体系的スコアを提示しておらず、公式は技術レポートを「coming soon」としているため、現時点では**未指定**と扱うのが厳密である。 [23]

一方、HumanEvalについては、OpenAIのGPT-4技術報告に公式数値（HumanEval　67.0%）があり、比較の基準点としての利用は可能である。 [30]

ベンチマーク比較表（要求仕様：欠損は「未指定」）

（注）価格は「標準APIの**1M　tokensあたり**」を基本とし、各社が公開する価格表に基づく。GLM-5はZ.ai公式価格、OpenAIは公式Pricingページを採用。 [31]

モデル	MMLU	C-Eval	HumanEval	LAMBADA	コンテキスト長	推論コスト（入力/出力, 1M tokens）	レイテンシ/速度
GLM-5	未指定	未指定	未指定	未指定	200K	$1 / $3.2（cached input $0.2）	**未指定**（指定記事はTTFT<1s等を主張：非公式）
GPT-4	86.4%（5-shot）	未指定	67.0%（0-shot）	未指定	未指定	未指定（現行価格表に"gpt-4"が直接は見当たらない）	未指定
GPT-4o	未指定（※医療サブセット等は別掲）	未指定	未指定	未指定	未指定	$2.50 / $10.00	音声応答：最短232ms・平均320ms（音声入出力条件）

モデル	MMLU	C-Eval	HumanEval	LAMBADA	コンテキスト長	推論コスト（入力/出力, 1M tokens）	レイテンシ/速度
GPT-5	未指定（多言語MMLU 0-shotの表は公開）	未指定	未指定	未指定	400K（max output 128K）	$1.25 / $10.00	SWE-benchのブログ値74.9%言及あり（速度自体は未指定）

根拠：GLM-5のコンテキスト長・価格 [32]、指定記事の非公式レイテンシ主張 [13]、GPT-4のMMLU/HumanEval [30]、GPT-4oの音声レイテンシ [33]、GPT-4o価格およびGPT-5価格/コンテキスト [34]、GPT-5のMMLU多言語表およびSWE-benchブログ値言及 [35]。

価格・長文の観点での含意（コスト効率）

API価格だけを見ると、GLM-5（入力$1/出力$3.2）は、GPT-5（入力$1.25/出力$10）と比較して**出力単価が大幅に低い**（約3分の1）一方、GPT-5はコンテキスト長が400KでGLM-5の200Kを上回る。したがって、**長文入力の極限（>200K）**が必要な場合はGPT-5側に優位が出うるが、出力大量生成（コード差分大量、長期ログ解析など）ではGLM-5がコスト面で有利になりやすい。 [36]

安全性・倫理対策と商用化・エコシステム

安全性・倫理（有害出力抑制、RLHF、安全評価）

GLM-5について、OpenAIのSystem Cardに相当する**包括的安全性文書**（有害出力抑制、誤情報、バイアス評価、プライバシー、レッドチーミング結果など）は一次情報として確認できず、公式も技術レポートを「coming soon」としているため、詳細は**未指定**である。 [23]

ただし、(i) ポストトレーニングで“slime”を用いること、(ii) モデルがエージェント用途を強く意識していること、は公式が明示しており、**ツール利用や長期タスクに伴う安全制御**（ツール権限、プロンプトインジェクション耐性等）が実運用上の論点になる。 [26]
（比較参考）OpenAIのGPT-5 System Cardは、MMLU多言語評価だけでなく、disallowed content評価、シコファンシー、プロンプトインジェクション、SWE-bench等を含む安全性・能力の体系評価を掲載しており、公開水準の差が際立つ。 [37]

商用API提供・価格・運用機能

Z.aiの公式価格表によれば、GLM-5は**入力$1 / 1M tokens、出力$3.2 / 1M tokens、cached input $0.2 / 1M tokens**で提供される（Cached Input Storageは期間限定Freeと表記）。 [38]
公式ドキュメントは、思考モード、ストリーミング、Function Calling、Context Caching、Structured Output等の機能を列挙し、OpenAI互換SDKで呼べる例も提示する（エンドポイント例：`/chat/completions`）。 [2]

また、Z.aiのCoding Plan（サブスクリプション）では、対応ツール（Claude Code等）と統合し、速度面で「55 tokens/sec超」を謳うが、これはプランの説明であり、**計測条件・対象モデル・ネットワーク条件が明確でないため性能指標としては限定的**に解釈すべきである。 [39]

エコシステム（ローカル推論・推論基盤・外部ベンチ）

公式モデルカードは、vLLM/SGLang/xLLMでのローカルデプロイ手順を示し、Hopper/Blackwell向けイメージ名まで記載する。さらに「Ascend　NPU向け」導線にも触れており、CUDA以外の推論基盤も視野に入れている。 28
また、Vending Bench 2については、実験がAndon Labs 40 により独立実施された旨が脚注で説明される。 11

（一次ソースURLは後掲）

研究的意義・市場影響と懸念点

研究的意義と競争力の分析

GLM-5の一次情報から読み取れる研究的ポイントは、概ね次の三点に集約できる。

第一に、**巨大MoE（744B総計／40Bアクティブ）**へのスケールと、**長文200K**の実用設計である。これは、単純な“密モデル大型化”よりも、推論コスト（アクティブ計算）を制御しながら能力を伸ばす路線であり、公開ベンチ（SWE-bench等）のスコア提示とも整合する。 10

第二に、**DSAの採用**である。DSAはDeepSeek側の一次報告で、長文での計算複雑性を下げつつ性能を維持する効率化として提示されている。GLM-5はこれを統合し“Token Efficiency”と“lossless long-text performance”を謳うため、長文エージェント用途での運用コスト低減が狙いだと解釈できる。 41

第三に、**RLスケーリング基盤（slime）**を前面に出し、「長距離インタラクションから継続学習する非同期エージェントRL」を提案している点である。これは“RLHFをやったか否か”より、**高スループットでポストトレーニングを回す工学**が中核にある。 26

市場面では、MITライセンスのオープンウェイト公開と、入力$1/出力$3.2という価格設定が、開発者にとって**「自前運用（オンプレ/クラウド）か、低価格APIか」**の選択肢を広げる。一方、OpenAI側はGPT-5を400Kコンテキストで提供し、価格体系も公開しているため、長文上限やエコシステム統合（コネクタ等）を含めた総合戦となる。 42

規制・地政学的影響（中国発モデルとして）

指定記事は「米国製GPUに依存しない国産スタック（Ascend/MindSpore）で学習」と述べ、これを半導体輸出規制文脈で強調するが、一次ソースで学習ハードウェアは確認できないため、ここは**“主張はあるが未検証”**として扱う。 13
ただし、報道はZhipuが中国製チップ（Ascend等）を推論に用いると伝えており、少なくとも“国産計算基盤での実運用”が地政学上の論点であること自体は裏付けられる。 9
加えて、公式配布がAscend向け推論導線を明示する点も、CUDA依存低減という戦略と整合する。 43

懸念点と未解決課題（透明性・悪用可能性・環境負荷）

透明性・再現性

- データ：トークン総量（28.5T）は明示されるが、データ内訳・権利処理・フィルタリングが未公開であり、再現性は限定的。 10
- 安全性：System Card級の体系評価が未公開で、どのような有害出力抑制・誤情報対策・バイアス評価が行われたかが不明（未指定）。 11

- ただし、オープンウェイトであるため、第三者が“実験としての安全性評価”を行う余地は大きい。[11]

推定トレーニング計算量・GPU日数・エネルギー・CO_2（計算と仮定を明示）

目的：公式が示す規模（40B activated, 28.5T tokens）から、事前学習の概算スケール感を推定する。[2]

計算式（一次ソースに基づく近似）
学習計算量は、しばしば **$C \approx 6 \times N \times D$**（FLOPs）で近似される。ここでNは“forward-activated”な非埋め込みパラメータ、Dは学習トークン数である、という定義が学術文献側で明示されている。[44]

GLM-5への代入（中心推定）
- N（activated params）：40B = 4.0×10^10[2]
- D（tokens）：28.5T = 2.85×10^13[2]
- よって **$C \approx 6 \times 4.0\times10^{10} \times 2.85\times10^{13} = 6.84\times10^{24}$ FLOPs**

GPU日数（例：NVIDIA H100想定の“尺度換算”）
ハードウェアは公式未指定のため、ここでは業界標準の参照としてNVIDIA[45] H100（SXM）の公称BF16 Tensor Core性能（最大1,979 TFLOPS）とTDP（最大700W）を用いる。[46]
また、実効利用率（MFU等）は環境に依存するため、**25%〜50%**の範囲で感度分析し、中心値として35%を採用する（これは仮定）。

- 公称性能：1.979×10^15 FLOPs/s[47]
- 実効性能（35%）：6.93×10^14 FLOPs/s（仮定）
- 総GPU日数（35%）：**約114,295 GPU-days**（= 114k H100日）
- 例：1,024GPUなら約112日、4,096GPUなら約28日規模（いずれも“理想化換算”）

エネルギー（中心ケース）
- GPU-hours：114,295×24 ≈ 2.74×10^6 GPU-hours
- GPU電力：700W = 0.7kW（上限TDPを採用、仮定として保守的）[48]
- IT電力量：≈ 1.92GWh
- PUE：1.2（仮定。参考としてGoogle平均1.09、Equinix 1.39等が公開される）[49]
- 施設電力量：≈ 2.30GWh

CO_2排出（中心ケース）
カーボン強度は地域依存だが、参考値として中国の電力排出原単位が**492 gCO_2/kWh**まで低下したとの報道があるため、0.492 kg/kWhを中心値として採用する（感度としてMEE公表値0.6205 $kgCO_2e/kWh$も併記）。[50]

- CO_2（0.492 kg/kWh, PUE1.2）：2.30GWh × 0.492 ≈ **約1.13千t-CO_2**
- CO_2（0.6205 kg/kWh, PUE1.2）：≈ **約1.43千t-CO_2**

推定トレーニングコスト（クラウド換算の一例）
AWSのH100価格は条件で変動するが、Capacity Blocks for MLの公表例では**約$3.933 / GPU-hour**が示される（地域表の“per accelerator”欄）。[51]
中心ケースのGPU-hours（約2.74M）を単純乗算すると、**約$10.8M**（GPU利用料のみの粗い下限推定）となる。
ただし、(i) 実際にはネットワーク/ストレージ/CPU等の費用、(ii) 学習は“6ND”より高い定数・追加計算（注意機構、MoEルータ、再学習、検証、RLポストトレ等）を伴う、(iii) 自社設備/国産NPUであればコスト構造が大きく異なる、ため、これは**“尺度感”**として提示するに留まる。[52]

主要一次ソースURL（ユーザ要件により明記）

```
【Zhipu / Z.ai 公式】
https://docs.z.ai/guides/llm/glm-5
https://docs.z.ai/guides/overview/pricing
https://github.com/zai-org/GLM-5
https://huggingface.co/zai-org/GLM-5
https://github.com/THUDM/slime

【指定記事】
https://tech-noisy.com/2026/02/12/zhipu-glm-5/

【関連一次（DSA）】
https://arxiv.org/abs/2512.02556

【比較（OpenAI公式）】
https://cdn.openai.com/papers/gpt-4.pdf
https://openai.com/gpt-5/
https://developers.openai.com/api/docs/pricing/
https://cdn.openai.com/gpt-4o-system-card.pdf
https://cdn.openai.com/gpt-5-system-card.pdf
```

（注）本報告書は一次情報（公式ドキュメント／公式配布物／公式価格表／公式論文）を優先し、未公開・不明点は“未指定”と明示した。指定記事由来の主張は、一次情報と矛盾する場合は矛盾として扱い、検証不能な断定は回避した。 53

---

1 2 3 4 10 14 18 19 20 22 26 27 32 41 53 https://docs.z.ai/guides/llm/glm-5
https://docs.z.ai/guides/llm/glm-5

5 11 15 23 28 29 42 https://huggingface.co/zai-org/GLM-5
https://huggingface.co/zai-org/GLM-5

6 44 https://proceedings.neurips.cc/paper_files/paper/2024/file/8066ae1446b2bbccb5159587cc3b3bcc-Paper-Conference.pdf
https://proceedings.neurips.cc/paper_files/paper/2024/file/8066ae1446b2bbccb5159587cc3b3bcc-Paper-Conference.pdf

7 12 https://github.com/zai-org
https://github.com/zai-org

8 33 https://cdn.openai.com/gpt-4o-system-card.pdf
https://cdn.openai.com/gpt-4o-system-card.pdf

9 40 https://venturebeat.com/technology/z-ais-open-source-glm-5-achieves-record-low-hallucination-rate-and-leverages
https://venturebeat.com/technology/z-ais-open-source-glm-5-achieves-record-low-hallucination-rate-and-leverages

13 16 https://tech-noisy.com/2026/02/12/zhipu-glm-5/
https://tech-noisy.com/2026/02/12/zhipu-glm-5/

17 24 https://huggingface.co/zai-org/GLM-5/blob/main/config.json
https://huggingface.co/zai-org/GLM-5/blob/main/config.json

21 https://arxiv.org/abs/2512.02556
https://arxiv.org/abs/2512.02556

25 https://github.com/THUDM/slime
https://github.com/THUDM/slime

30 https://cdn.openai.com/papers/gpt-4.pdf
https://cdn.openai.com/papers/gpt-4.pdf

31 36 38 https://docs.z.ai/guides/overview/pricing
https://docs.z.ai/guides/overview/pricing

34 https://developers.openai.com/api/docs/pricing/
https://developers.openai.com/api/docs/pricing/

35 37 https://cdn.openai.com/gpt-5-system-card.pdf
https://cdn.openai.com/gpt-5-system-card.pdf

39 https://docs.z.ai/devpack/overview
https://docs.z.ai/devpack/overview

43 https://github.com/zai-org/GLM-5
https://github.com/zai-org/GLM-5

45 51 52 https://aws.amazon.com/jp/ec2/capacityblocks/pricing/
https://aws.amazon.com/jp/ec2/capacityblocks/pricing/

46 47 https://www.megware.com/fileadmin/user_upload/LandingPage%20NVIDIA/nvidia-h100-datasheet.pdf
https://www.megware.com/fileadmin/user_upload/LandingPage%20NVIDIA/nvidia-h100-datasheet.pdf

48 https://www.nvidia.com/en-us/data-center/h100/
https://www.nvidia.com/en-us/data-center/h100/

49 https://datacenters.google/efficiency
https://datacenters.google/efficiency

50 https://climatecooperation.cn/climate/china-releases-2025-report-on-product-carbon-footprint-management/
https://climatecooperation.cn/climate/china-releases-2025-report-on-product-carbon-footprint-management/