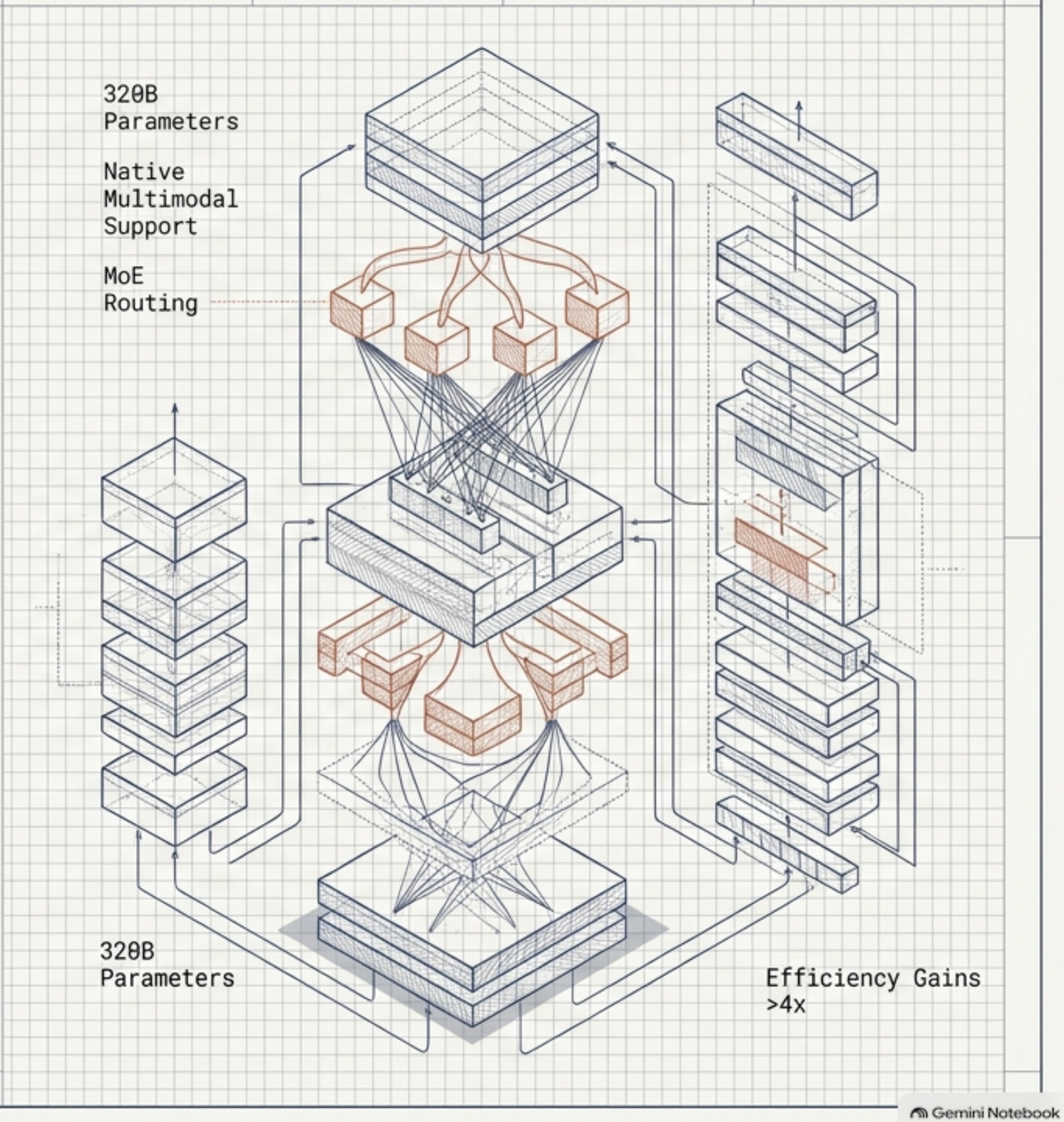
GLM-5.3-Flash: AIインフラの制約を破壊する320Bネイティブ・マルチモーダルMoE

匿名モデル「Ox Alpha」が証明した、ハードウェアの限界を凌駕するソフトウェア工学の極致



Gemini Notebook

The Engine Dashboard: 4つの破壊的ブレイクスルー

Intelligence

- 知能水準: **Claude Opus 4.8と同等**
- Artificial Analysis Index: **57**
- コーディング・エージェント実行に特化した極めて高い自律性。

Architecture

- 320B-A18B MoE
- トランスフォーマー層を**半減**（**92層→45層**）。
- **1Mトークン**の長文脈を支える極細の計算パス。

Infrastructure

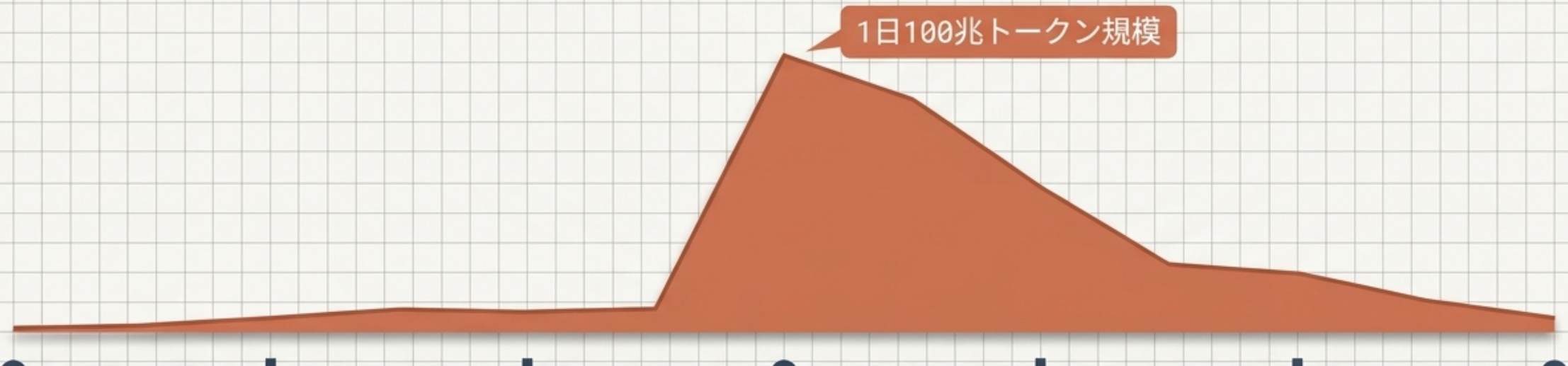
- **10万基**の**中国国産**チップで完全稼働。
- NVIDIA依存からの脱却。
- EPD分離とソフトウェア最適化による制約の突破。

Economics

- コスト: Opus 4.8の**1/40**（通常時）～**1/80**（プロモーション時）。
- オープンソース（MITライセンス）による重み公開。

Gemini Notebook

The 0x Alpha Phenomenon: 6日間の匿名プレビューが巻き起こした熱狂



2026.08.20

- OpenRouter / OpenCodeにて「stealth/ox-alpha」が突如無料公開。
- モデルカード・事前告知一切なし。

The Viral Metrics

- 6日間で23兆トークン以上を処理。
- DeepSWE検証: Ox Alpha 約80% vs Claude Fable 5 65%
- Stripe CEO Patrick Collisonが「very impressive」と言及。

2026.08.26

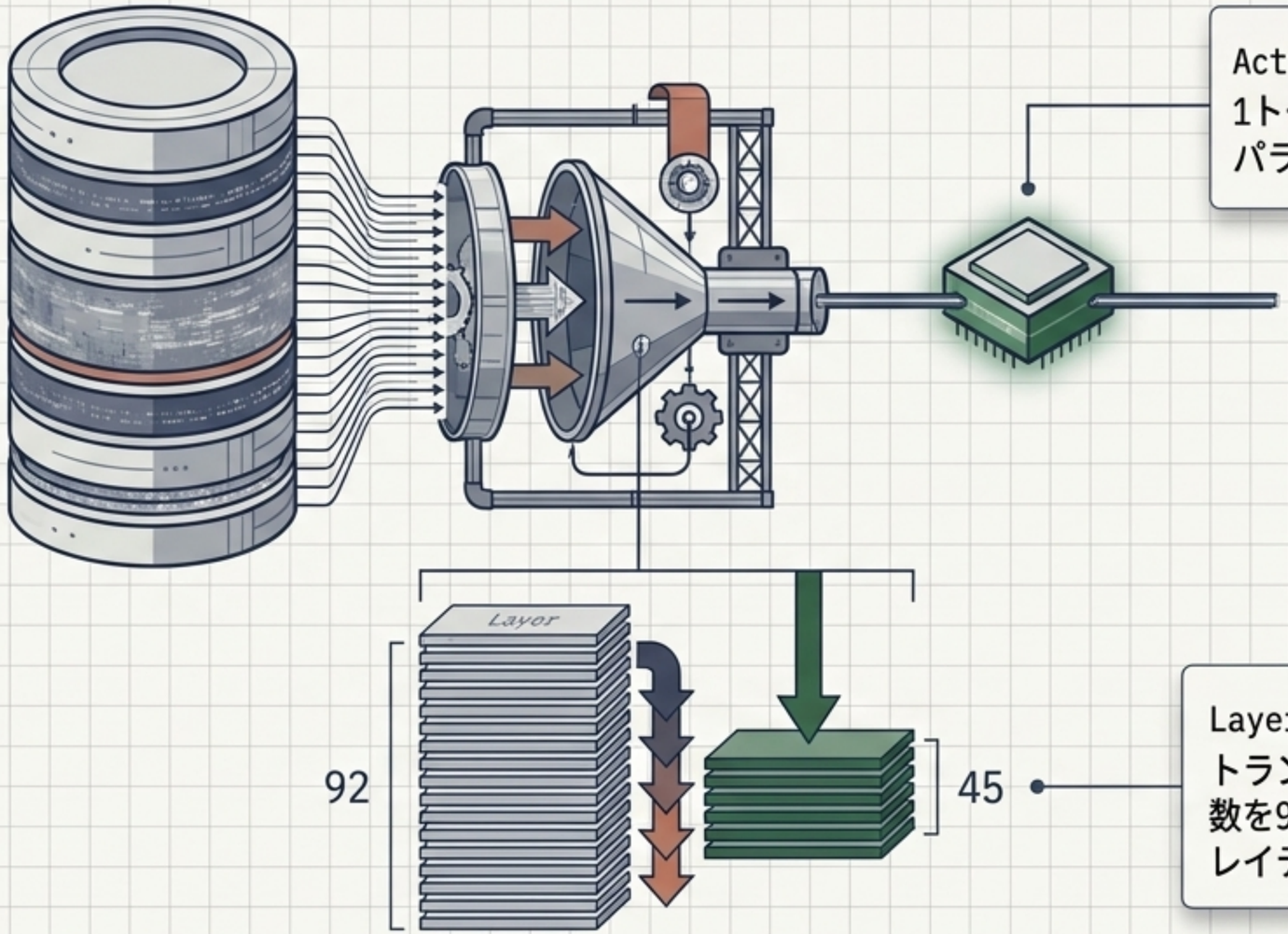
- トークナイザーの解析によりGLM-5.3系列と判明。
- 智譜AIが「GLM-5.3-Flash」として正式発表。

Gemini Notebook

Architecture Shift: 極限まで研ぎ澄まされた計算パス

巨大な知識容量を維持しながら、推論時の計算負荷を劇的に削ぎ落とす「320B-A18B MoE設計」。

Total Capacity (320B):
前世代と同等水準の総パラメータを維持し、モデルの知識量を担保。



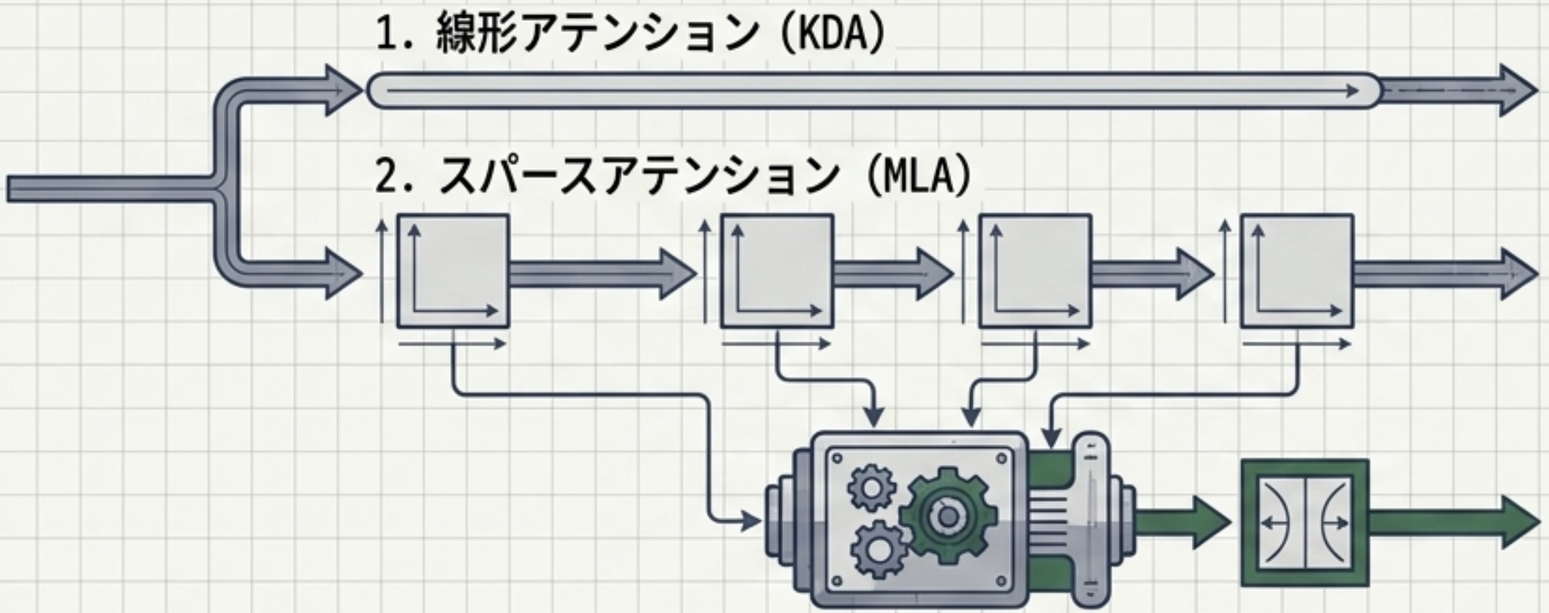
Active Parameters (18B):
1トークン処理時に稼働するパラメータを約半分に抑制。

Layer Compression:
トランスフォーマーレイヤー数を92層から45層へ半減し、レイテンシを極小化。

Gemini Notebook

1M Context Engine: ハイブリッドアテンションと「IndexPool」

オープンモデルとして初めてKDAとMLAを統合。メモリ爆発を防ぐ革新的なKVキャッシュ圧縮機構。



Metrics Output

- アテンション計算量: 3.01倍削減
- KVキャッシュサイズ: 4.44倍圧縮

1. 線形アテンション（KDA）: 極低コストで近傍の局所的依存性を追跡。

2. スパースアテンション（MLA）: 軽量なインデクサーを使用し、広範な文脈からの大域的検索を担う。

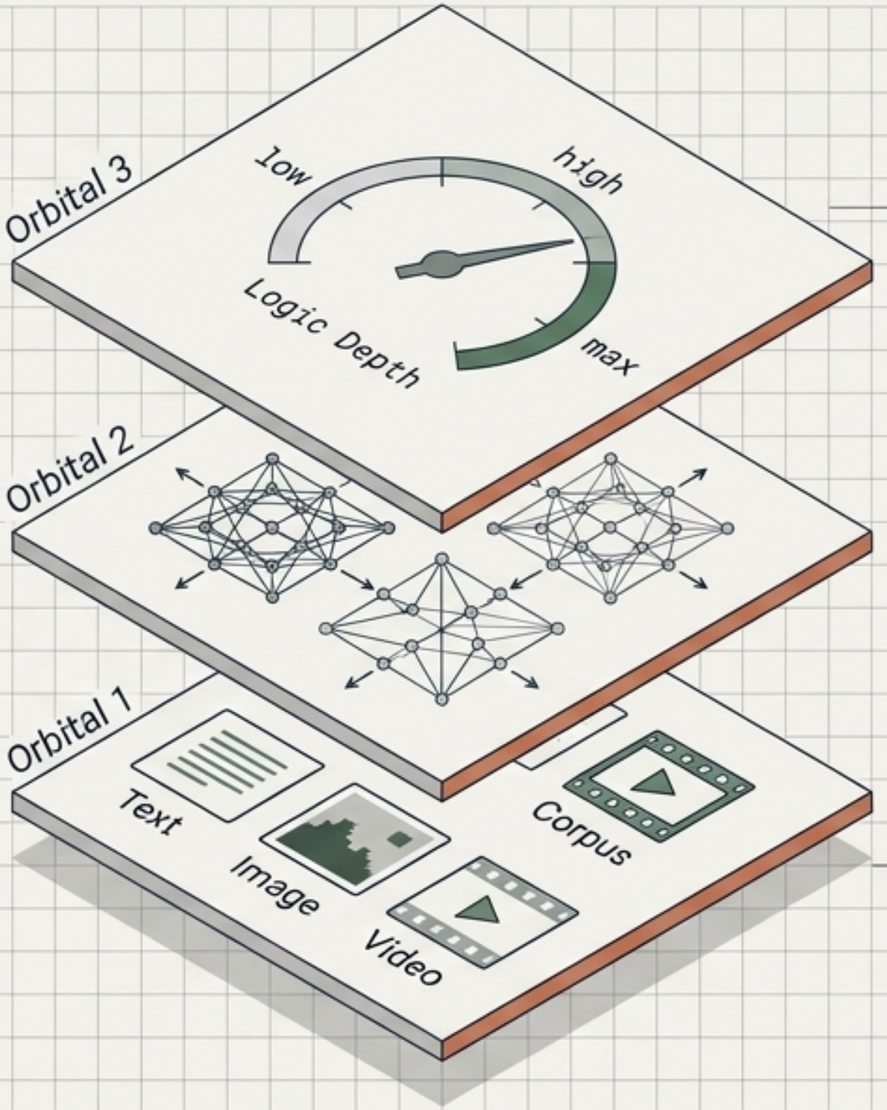
3. IndexPool（圧縮機構）: インデクサー内の4つのキャッシュキーベクトルを単一ベクトルに集約。

Gemini Notebook

The Neural Core: ネイティブ・マルチモーダルと常時稼働する思考

mHC (Manifold-Constrained Hyper-Connections)

- 残差接続の混合を特定の多様体に射影し、恒等写像を保持。
- 勾配爆発を抑え、スケーリングと学習の安定性を劇的に向上。



Always-On Thinking Mode

- タスクの複雑性に応じて思考深度 (low / high / max) を動的に切り替え。
- 推論の質を背後で自律的に高めるバックグラウンド論理展開。

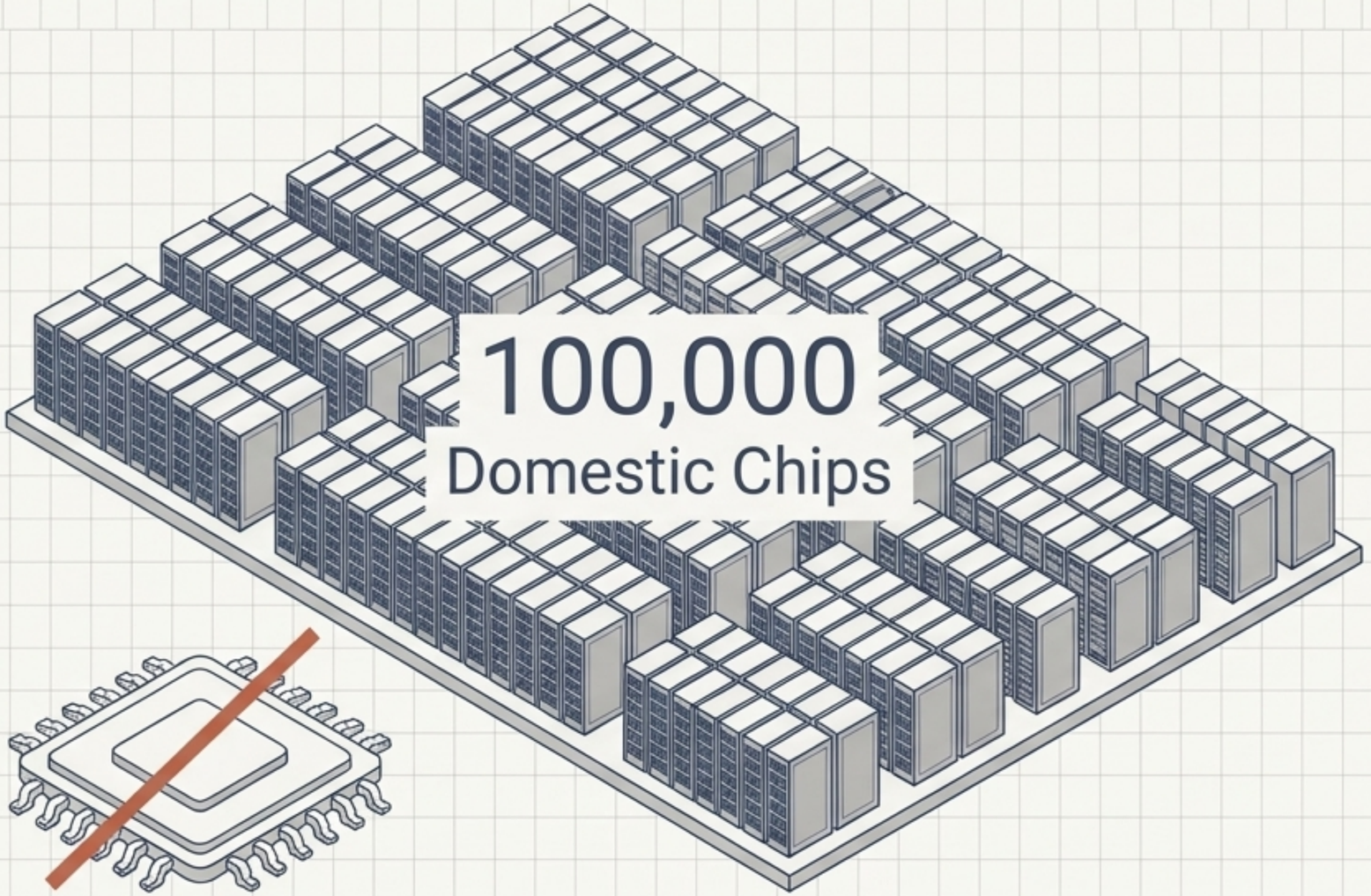
Training Corpus

- 30兆トークン規模のマルチモーダル事前学習。
- テキスト、静止画像、長尺動画をネイティブ入力として処理。最長128Kの長文出力。

Gemini Notebook

The Infrastructure Miracle: 10万基の「国産チップ」による完全稼働

GLM-5.3-Flashの真の破壊力。23兆トークンに及ぶ全トラフィックは、中国国産AIアクセラレータクラスタのみで処理された。



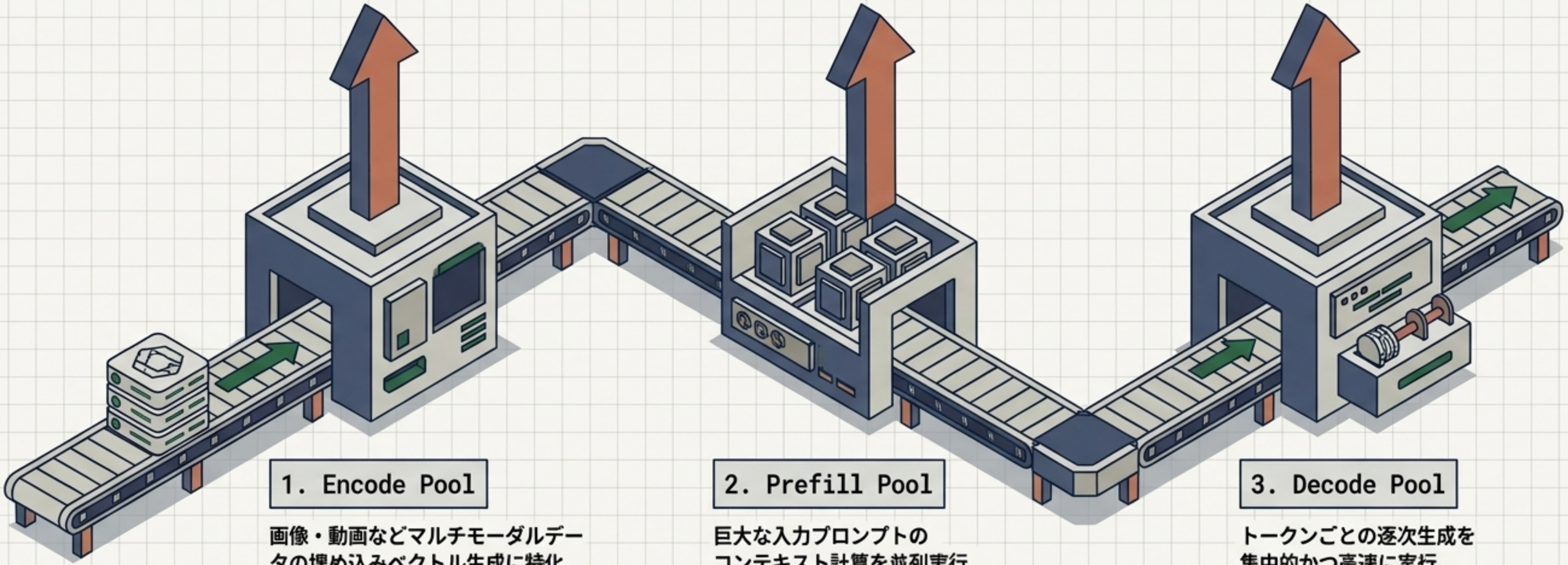
ソフトウェアによる制約突破の仕組み

- ハードウェアの制約を専用推論エンジン「SGLang」で突破：
  - **Infra Agentによる自動診断:** 先行モデルを用いたカスタムカーネル開発とオペレータの自動最適化。
  - **複合的な量子化・並列化:** ノード内テンソル並列、ReplaySSM、W8A8量子化、混合キャッシュ。
  - **Result:** エンドツーエンドのサービス性能を**3倍向上**。主流のGPUと**同等の処理コスト**を達成。

Gemini Notebook

EPD Separation Architecture: クラスタ稼働率を最大化する分離パイプライン

100万トークンの負荷によるリソース競合を防ぐため、処理フェーズごとに独立したワーカープールを構築。



1. Encode Pool

画像・動画などマルチモーダルデータの埋め込みベクトル生成に特化。

2. Prefill Pool

巨大な入力プロンプトのコンテキスト計算を並列実行。

3. Decode Pool

トークンごとの逐次生成を集中的かつ高速に実行。

Key Benefit: 各ステージを個別にスケールアウトさせることで、国産チップクラスタ全体のハードウェア稼働率を限界まで引き上げる。

Gemini Notebook

Intelligence Benchmarks: トップティア・クローズドモデルとの対峙

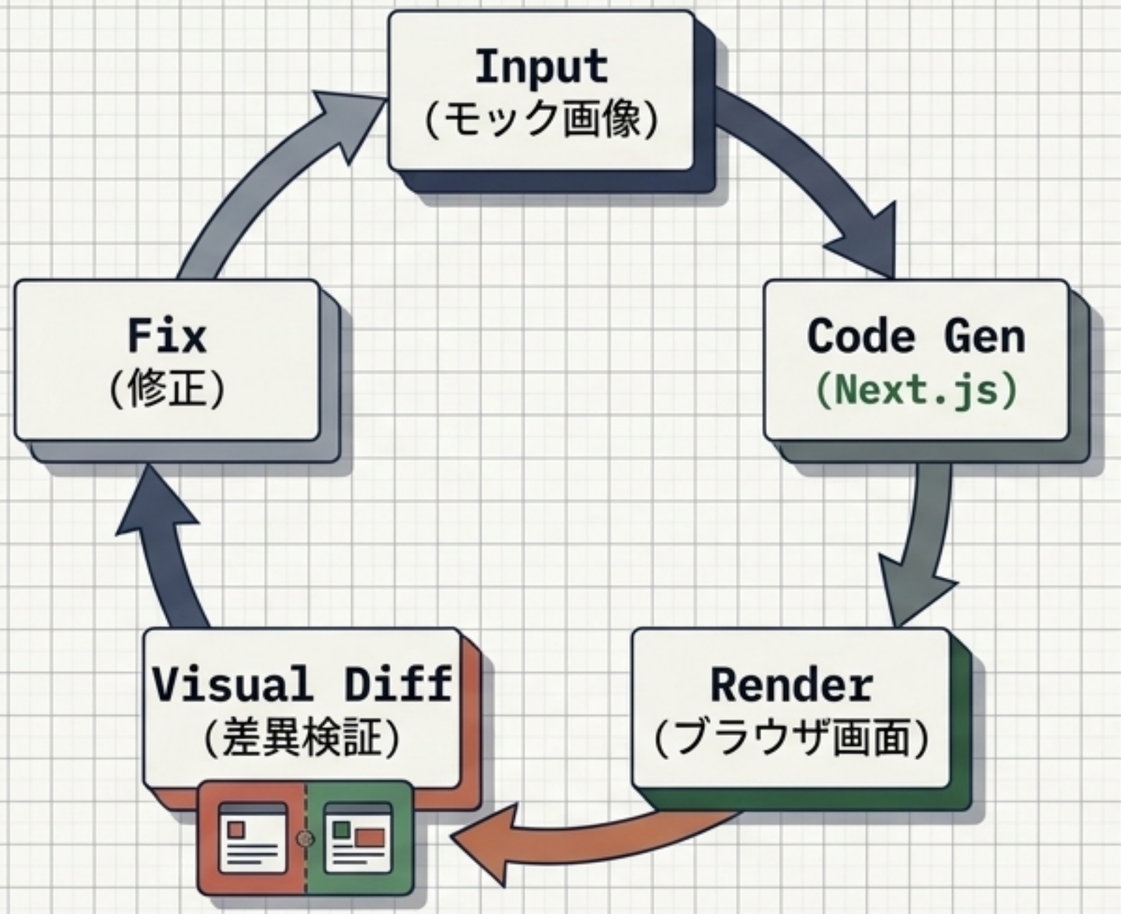
Benchmark Criterion	GLM-5.3-Flash	Claude Opus 4.8
総合知能 (Art. Analysis Index)	[GLM] 57	[Opus] 57 (Tie)
実務作業 Elo (GDPval-AA v2)	[GLM] 1773	[Opus] 1675 (GLM Win)
GitHub実課題 (DeepSWE v1.1)	[GLM] 63.4	[Opus] 46.2 (GLM Win)
超高難度推論 (Humanity's Last Exam)	[GLM] 55.3	[Opus] - (N/A)
空間認識・視覚 (BabyVision)	[GLM] 53.4	[Opus] 46.8 (GLM Win)

Insight: CLI/エージェント環境（Terminalbench 2.1: 84.3）など、ソフトウェア開発と自律実行環境において突出した性能を発揮。一般的な視覚知覚よりもUI/UX制御に特化したチューニング。

Gemini Notebook

Autonomy in Action: 視覚フィードバックを伴う自己検証ループ

ソフトウェア開発とドキュメント解析における「視覚主導の自律性」



自律的UI/UX生成: デザイン画像から**Next.js**コードを生成。レンダリング結果を自身で撮影し、レイアウトの差異を反復修正する。

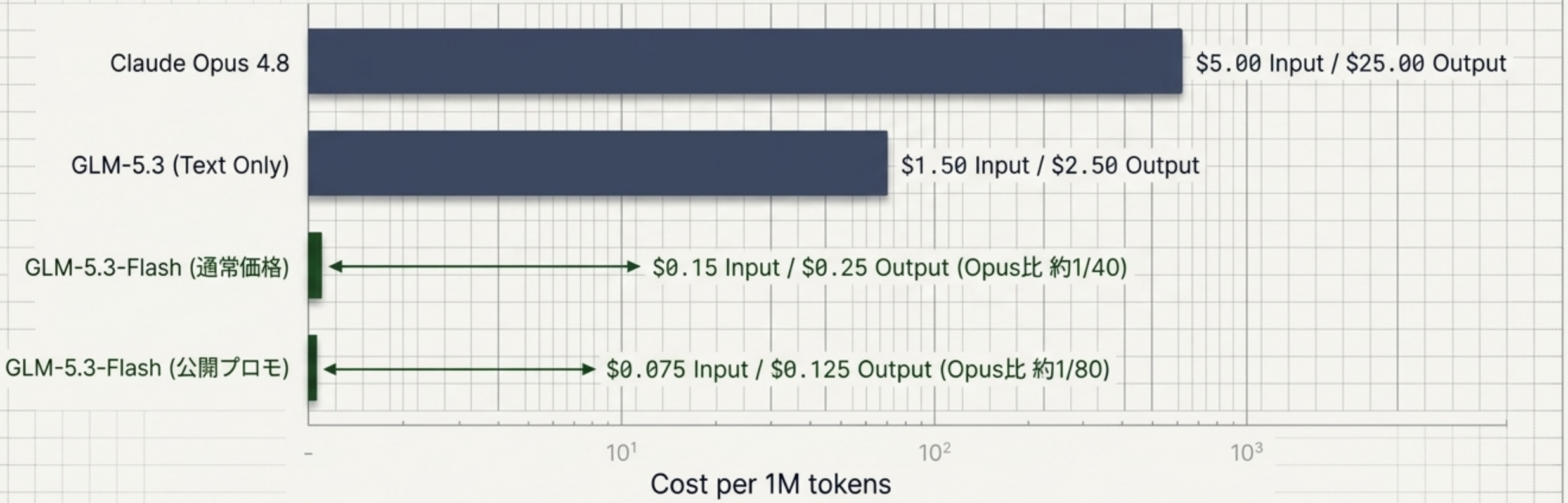
ダイレクト解析（No-OCR）: **PDF**、**XLSX**、**PPTX**を画像として直接解析し、財務分析や文書作成を完結。

長尺動画の自律編集: **48**分間の講演動画から、**57分**間で**17箇所**の重要発言を抽出し、字幕付きクリップとして自律出力。

Gemini Notebook

The Cost Disruption: 破壊的な経済インプリケーション

Opus 4.8クラスの最高峰の知能を、極小コストで提供。



Insight: DeepSeek等が牽引してきたオープンソース価格競争を、ネイティブ・マルチモーダルの領域でさらに一段階押し下げる結果に。

Gemini Notebook

Open Source Ecosystem: ローカル推論への道

MITライセンスによる完全な重み公開。商用利用・改変・再配布が自由に。

Native Release

- 約328GB（FP8）
- フル精度運用にはNVIDIA Sparkなどの大規模クラスタが必要。

Community Quantization

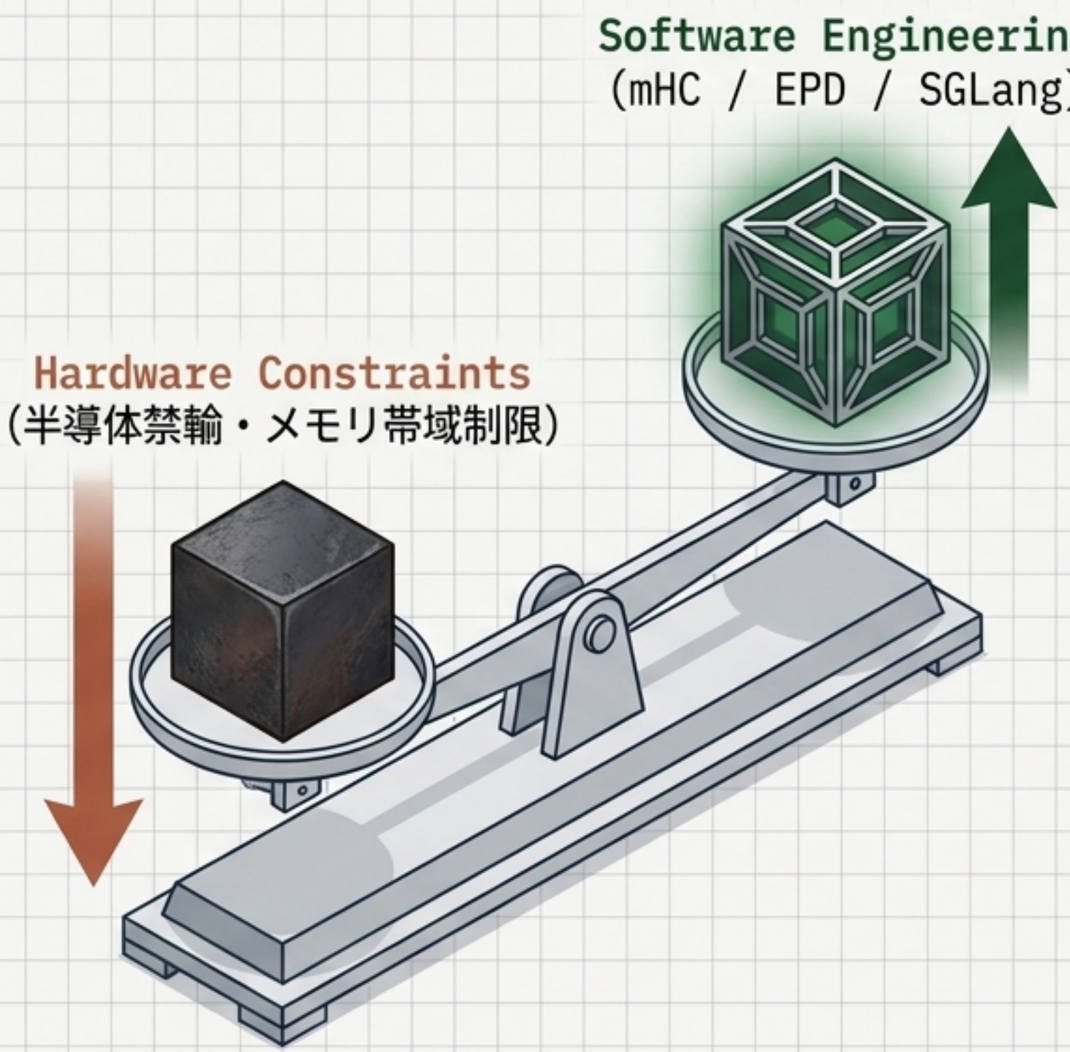
- **Unsloth**によるDynamic GGUF
- **4-bit / 2-bit**量子化への即時対応。

Local Agent Enablement

- **UD-Q2_K_XL**
- **Mac Studio**等の統合メモリ環境を用い、ワークステーション規模での**高度なローカル推論**が可能に。

Gemini Notebook

Strategic Impact: 「ソフトウェア」が「ハードウェアの制約」を喰らう



Core Insight

GLM-5.3-Flashの真の功績は、半導体供給制約という『ハードウェアの致命的ビハインド』を、アーキテクチャの革新という『ソフトウェア工学』によって完全に相殺できることを実証した点にある。

Strategic Takeaways

- **The End of the Hardware Monopoly:** 巨大なAIサービスの基盤が、特定ベンダー（NVIDIA）の最先端GPUのみに依存する時代からのパラダイムシフト。
- **Algorithmic Resilience:** ハードウェアの堀（Moat）を、アルゴリズムの力で越える新たな地政学的アプローチの確立。

Gemini Notebook

Conclusion: The Blueprint for the Next Era

Intelligence vs Cost

フラッグシップ級（Claude Opus 4.8）の知能と自律性を、Flashクラスを下回る圧倒的な低コスト(1/40)で提供する新基準の確立。

Architectural Elegance

320Bの知識容量と18Bのアクティブ計算を両両立するMoE、そしてハイブリッドアテンションによる「極限の推論効率化」の実現。

Infrastructure Independence

最先端ハードウェアの不在をソフトウェア・アルゴリズム工学で突破。オープンソースAIと地政学的インフラ競争における歴史的転換点。

Gemini Notebook