SYSTEM_READY // TELEMETRY: ONLINE

拡張するAIの知性領域：ローカルとクラウドの現在地

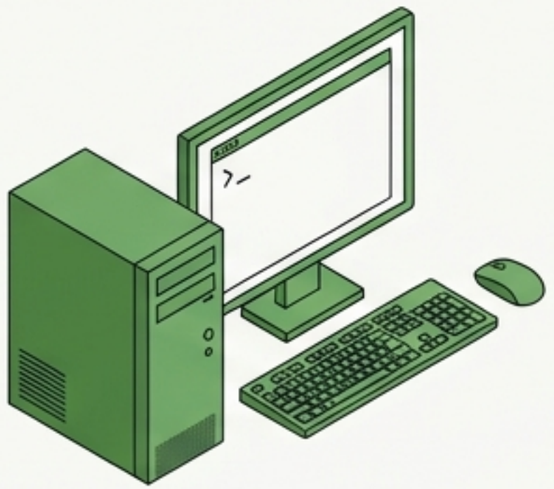
GPT-4登場から現在までの劇的なパラダイムシフトと、
次世代ハイブリッドAI戦略の全貌。



Gemini Notebook

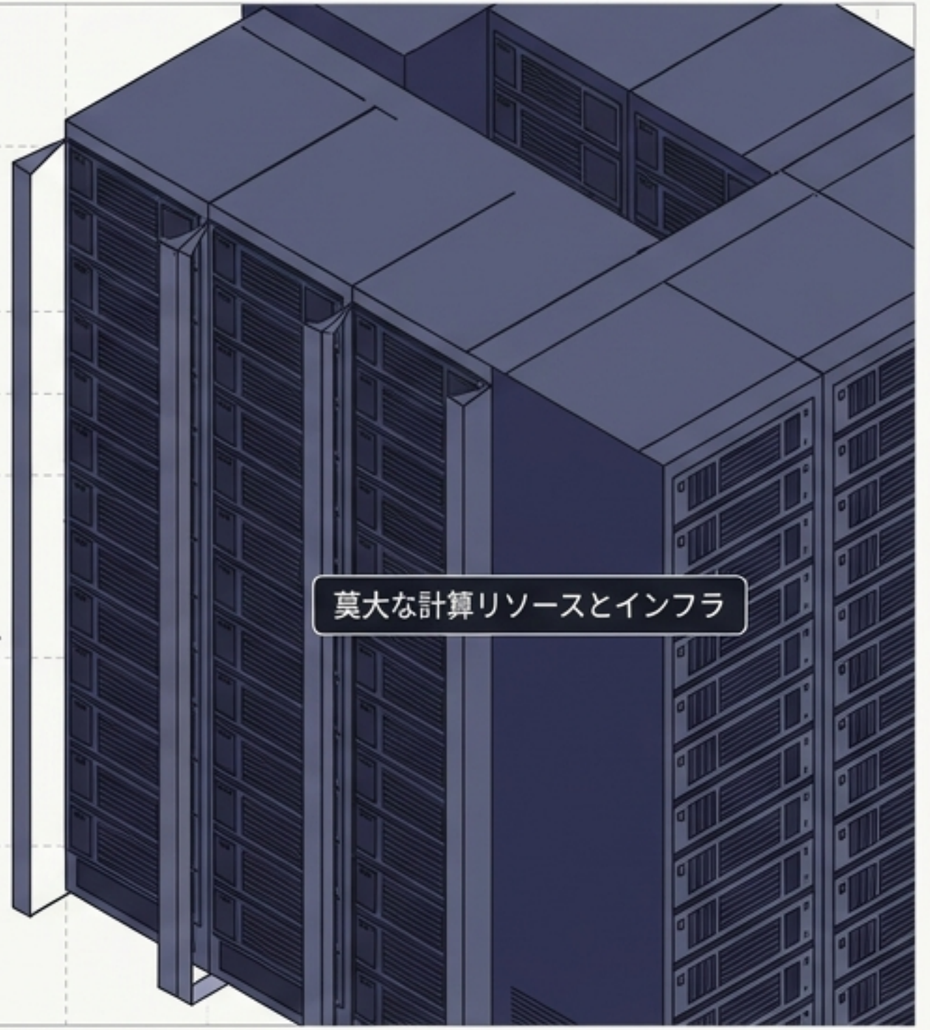
かつての常識：「最強」は常にクラウドの彼方に

2022年のGPT-3.5からGPT-4への進化。最強のフロンティアモデルを動かすには、巨大なデータセンターと莫大なコストが不可欠だと信じられてきた。



依存する旧パラダイム

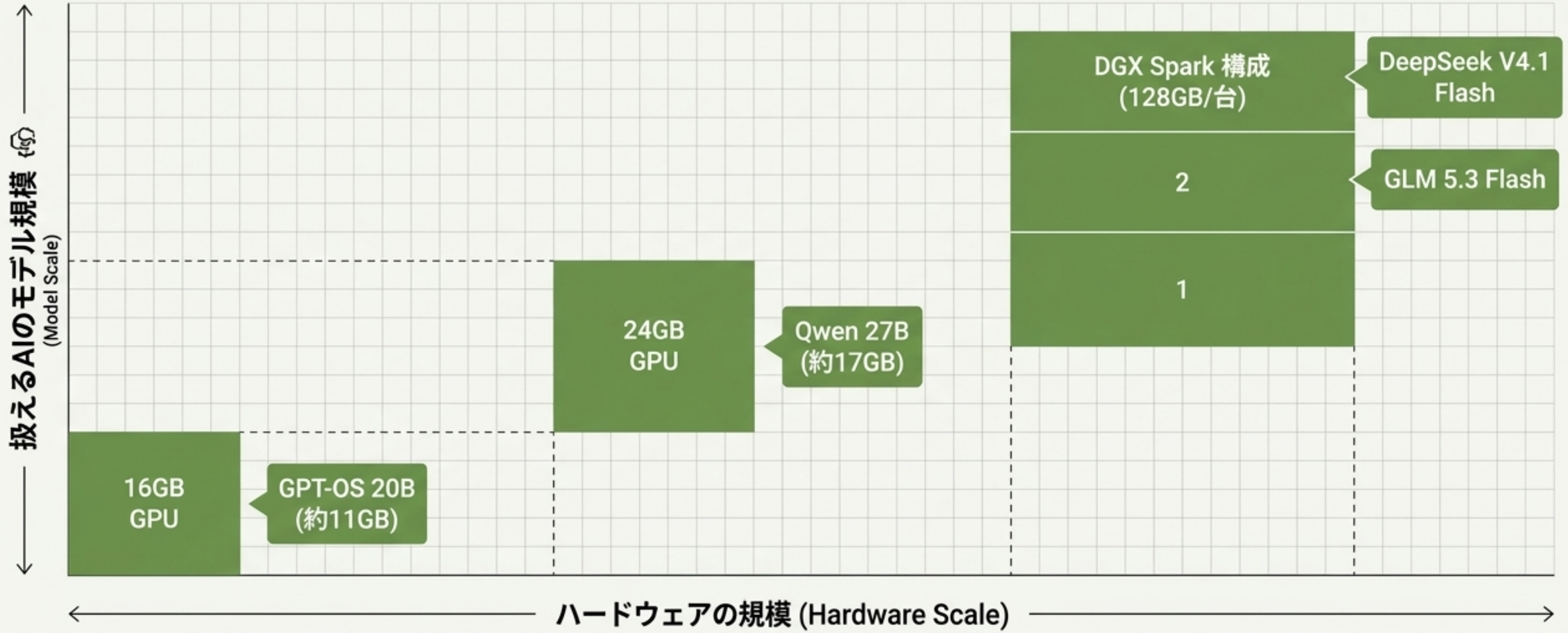
GPT-4　o1



Gemini Notebook

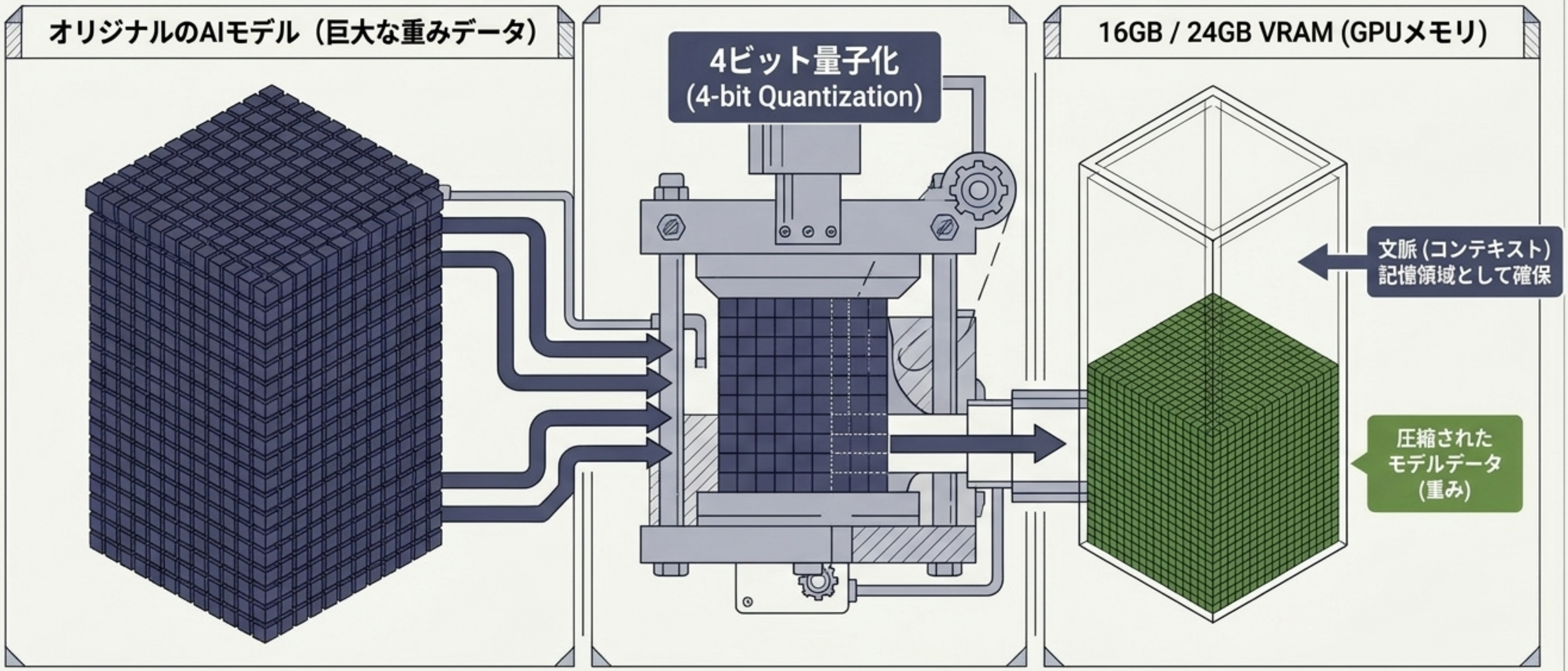
今、手元のデスクに収まる知性

ローカルハードウェアの進化と、それに伴うオープンモデルの巨大化。



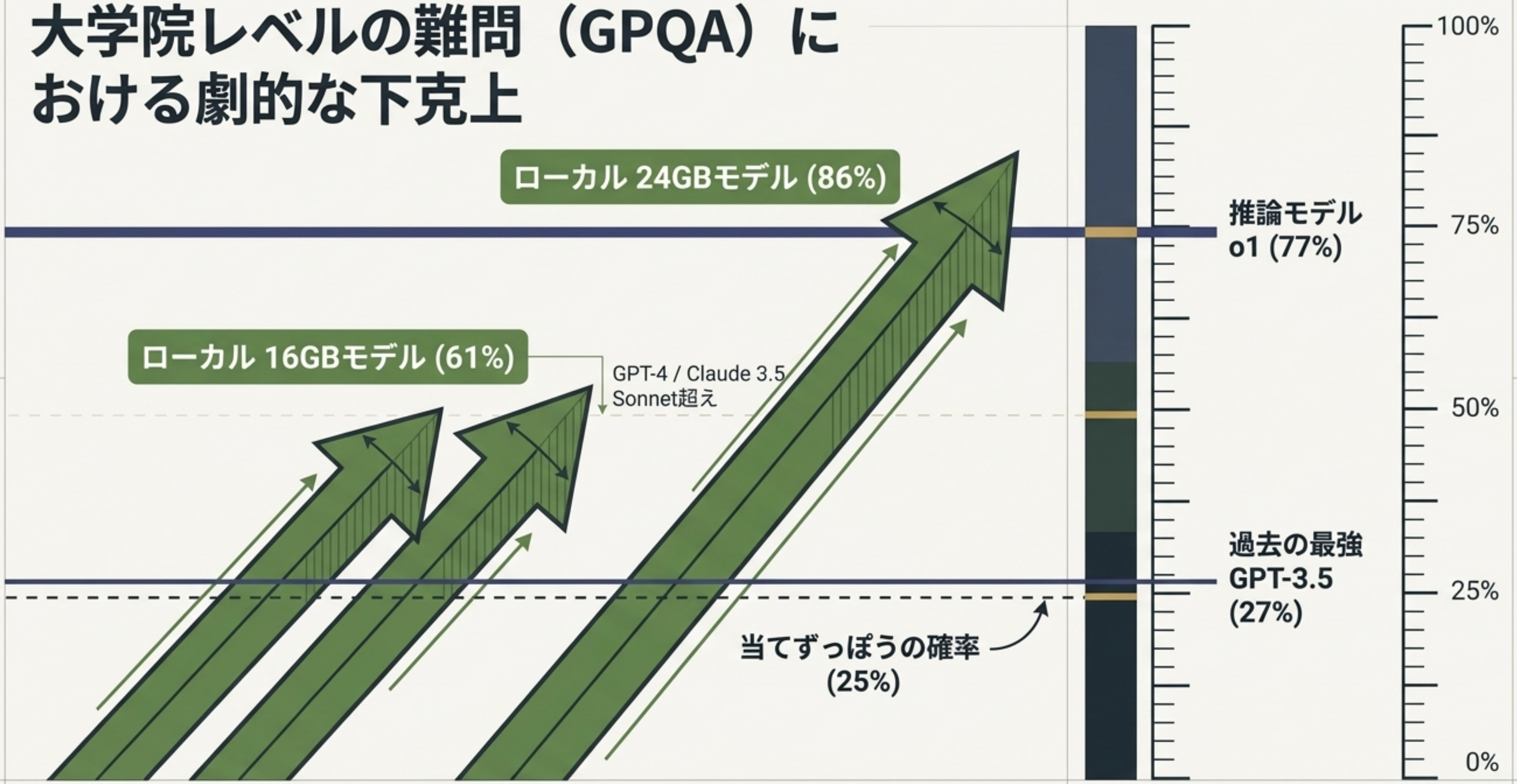
Gemini Notebook

メモリの壁を突破する「4ビット量子化」



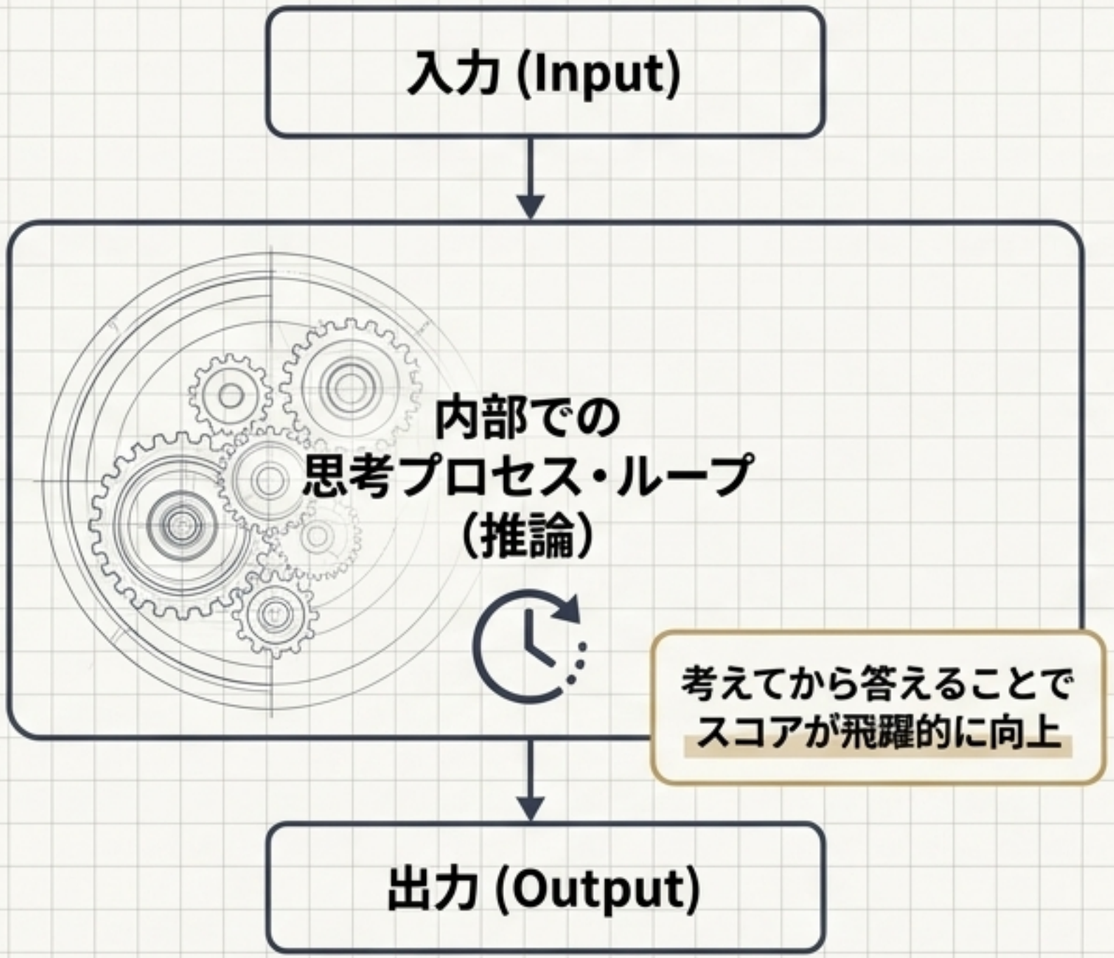
Gemini Notebook

大学院レベルの難問（GPQA）における劇的な下克上

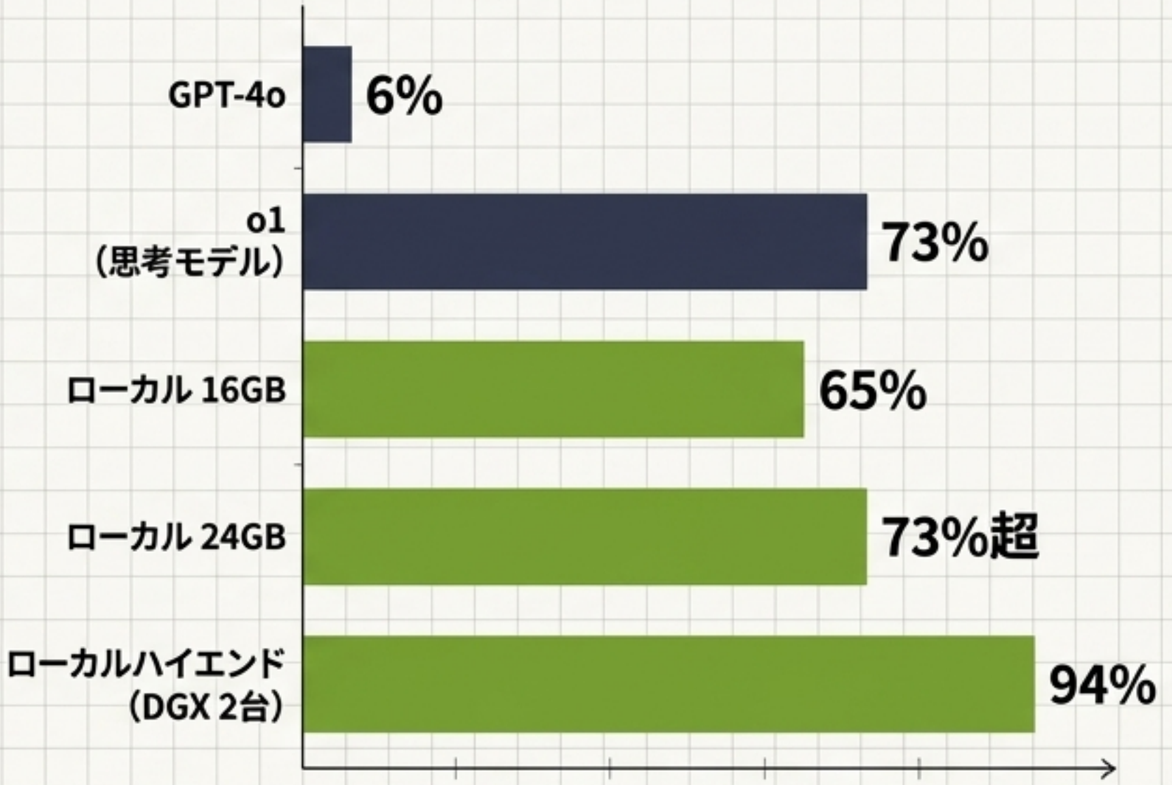


Gemini Notebook

「思考のタメ」が生み出す圧倒的な数学力

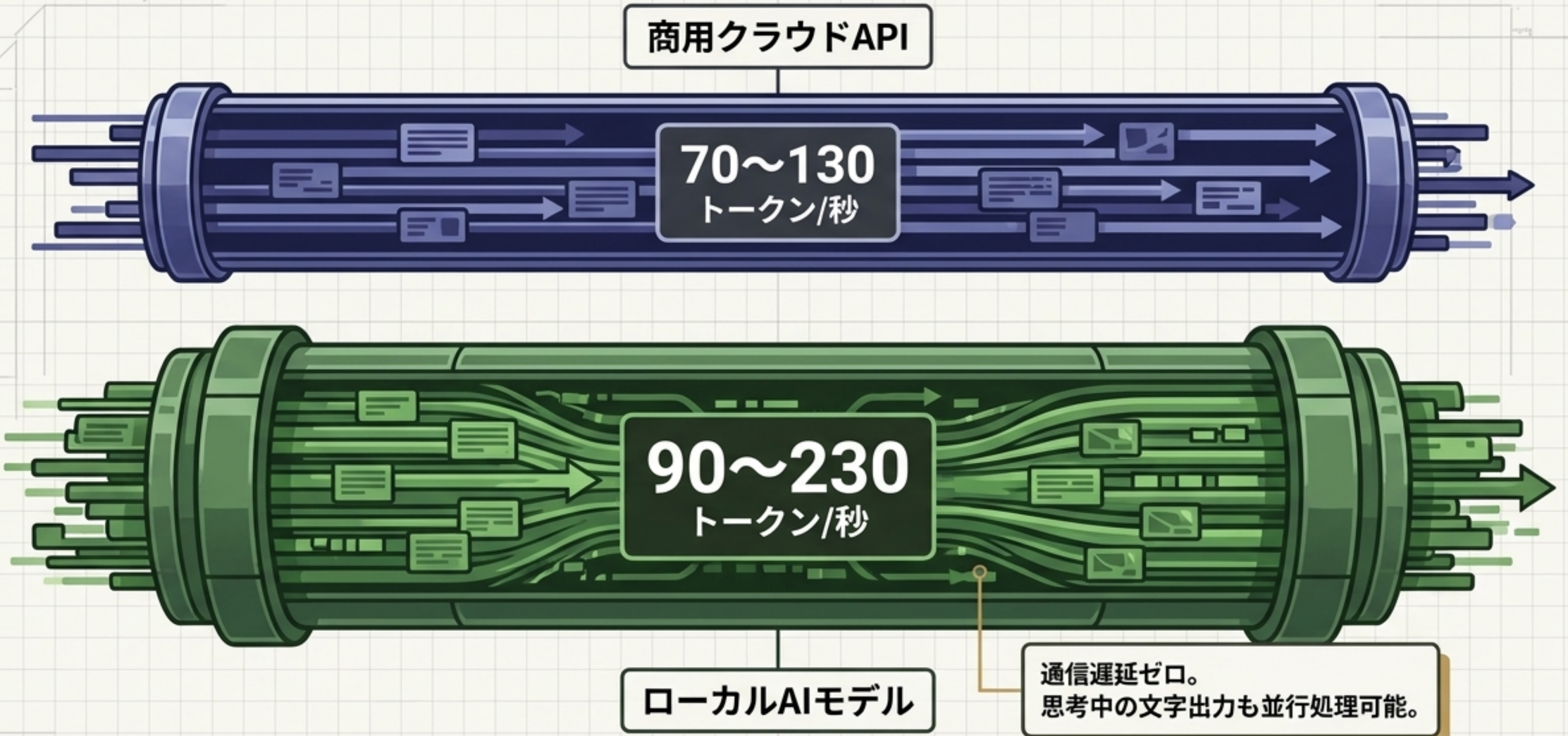


AIME（数学競技会）スコア比較



Gemini Notebook

クラウドAPIに匹敵、あるいは凌駕する生成速度



Gemini Notebook

体感評価でも旧フロンティアモデルと同等へ

LMSYS Arena（ブラインドテスト）における順位の割り込み

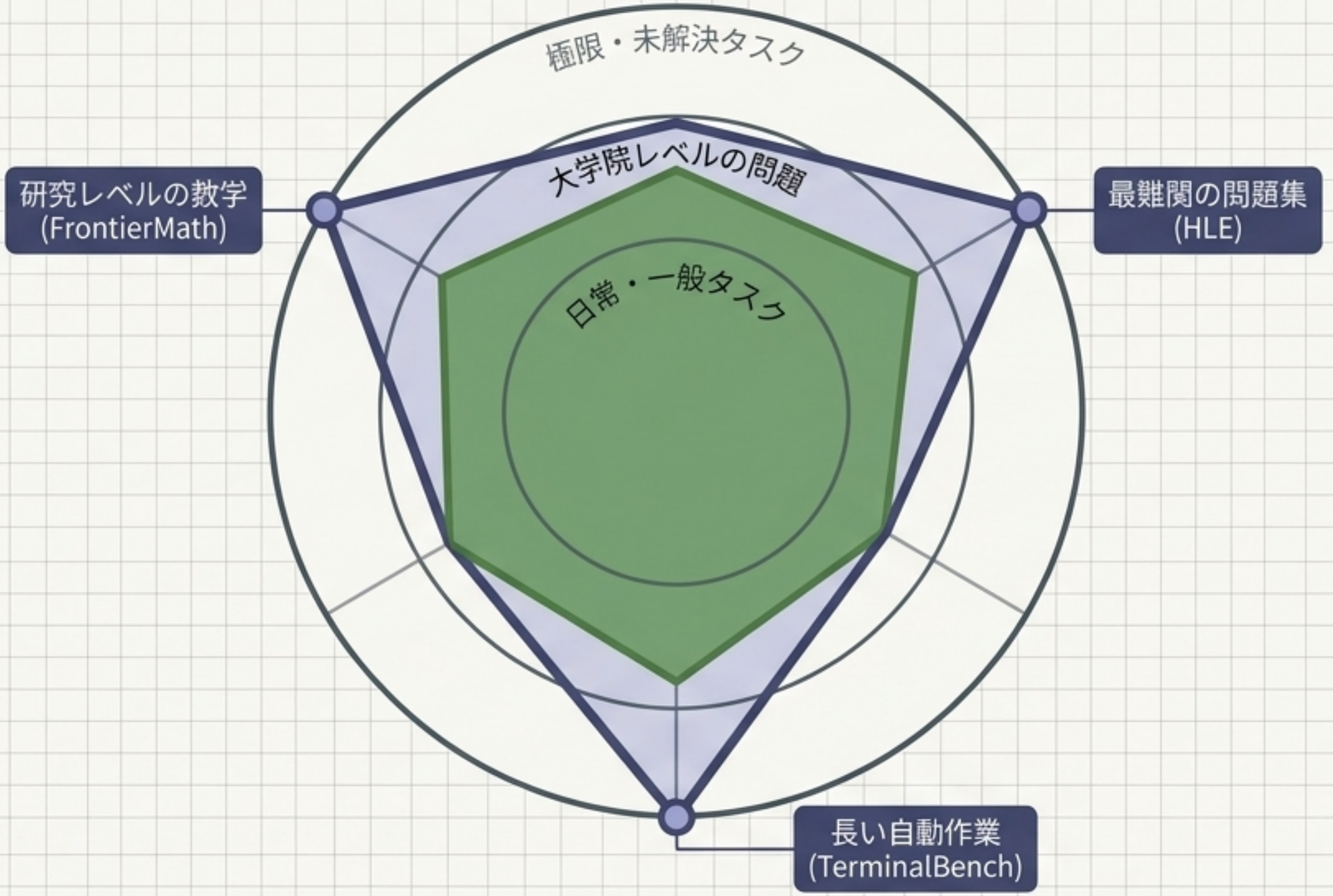
1	商用モデル：旧型最新GPT-4oクラス
2	ローカル：DGX構成（GLM 5.3 / DeepSeek V4.1）
3	商用モデル：GPT-4
4	ローカル：24GBモデル (Qwen 3.8 27B)
5	商用モデル：1年前の最新モデル群
6	ローカル：16GBモデル (GPT-OS 20B)

※点数が100点違うと、上位モデルが約64%の確率で人間に選ばれる

Gemini Notebook

しかし、極限領域（フロンティア）では商用モデルが圧倒的

日常タスクを完璧にこなすローカルに対し、未解決問題に特化し始めたクラウドAPI。



Gemini Notebook

フロンティア（最前線）における実力の壁

HLE
（専門家作成の最難関問題）



ローカル最大 40% /
商用トップ 61%
（約1.5倍の差）

TerminalBench
（長時間の自律作業）



ローカル最高 33% /
商用トップ 約60%
（約1.8倍の差）

※16/24GB機ではほぼ0〜6%で実行不可

FrontierMath
（研究レベルの未解決数学）



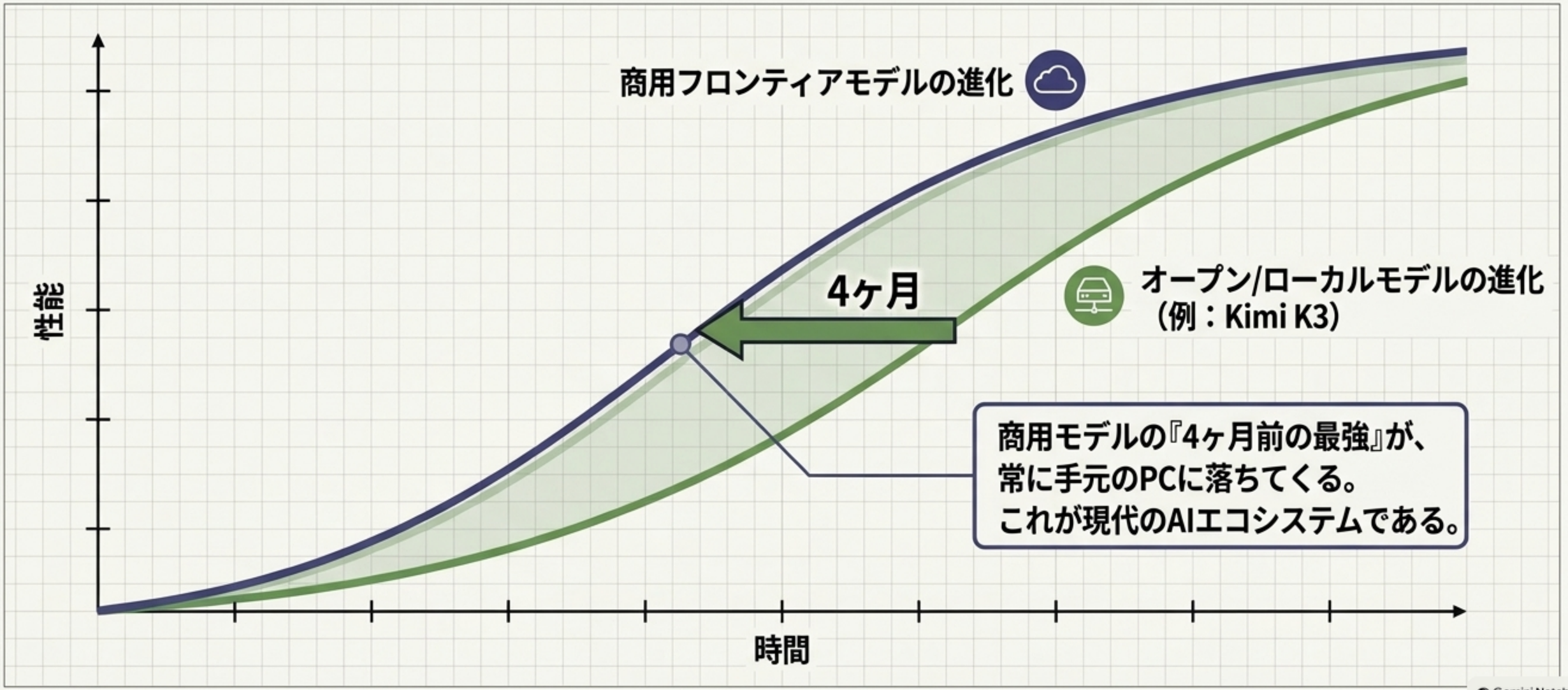
オープン最高 39% /
商用トップ 88〜98%
（圧倒的格差）

【APIコストの劇的な低下】かつて100万トークン60ドルだったGPT-4時代から、現在の最新モデルは10〜50ドルへ。同等の商用モデルは当時の1/3の価格へ下落。

Gemini Notebook

真実：「4ヶ月のタイムラグ」というエコシステム

Epoch AI 総合指数の推移：オープンモデルは劣っているわけではない



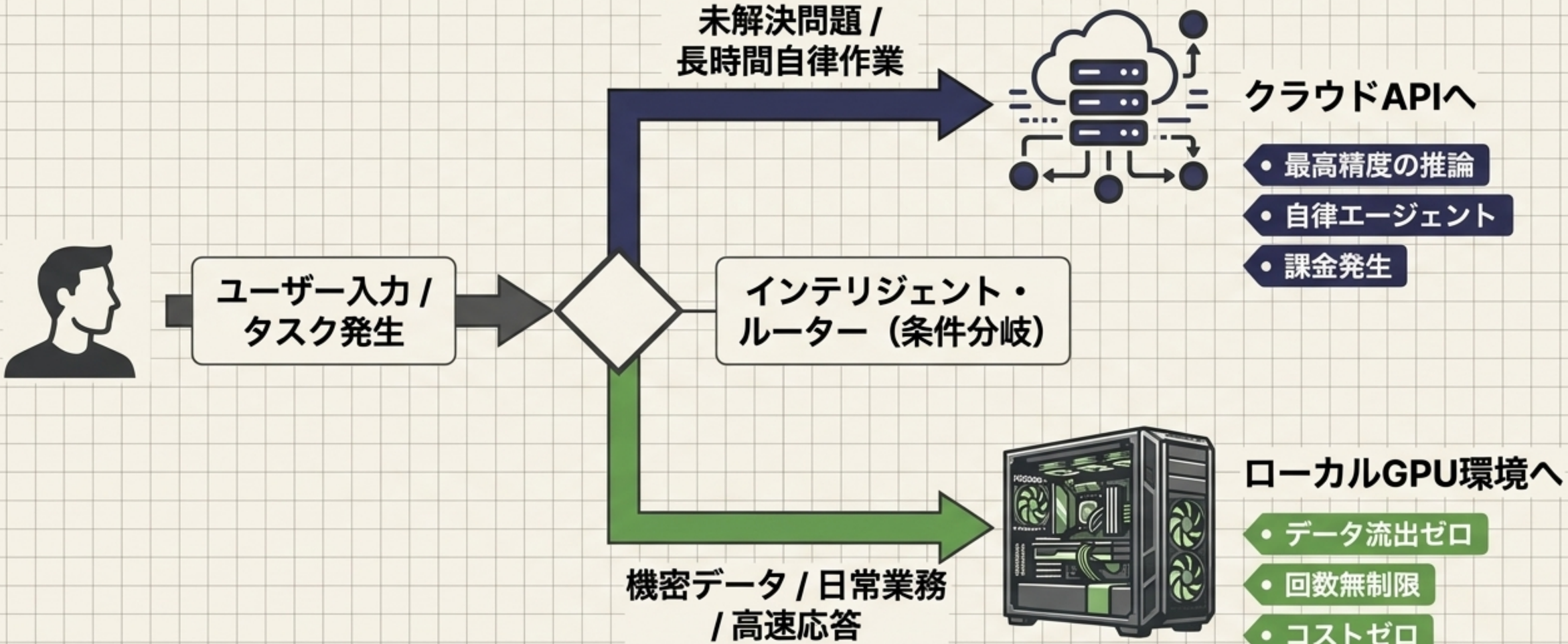
Gemini Notebook

AI実力診断マトリクス：どこで切り替えるべきか？

	ローカル （16GB〜24GB）	ローカルハイエンド （DGX等）	商用API （フロンティアモデル）
日常タスク （文章・コード作成）	完璧にカバー	完璧にカバー	完璧にカバー
大学院レベル （GPQA）	十分	余裕	完璧
長時間自律作業 （TerminalBench）	実行不可	限定的 33%	必須 60%
研究レベル数学 （FrontierMath）	実用外	限定的	必須 98%

Gemini Notebook

究極の解答：適材適所のハイブリッド・ワークフロー



Gemini Notebook

ローカル vs 商用API：運用モデル決定の4条件

Decision Axis

1. プライバシー (Privacy)

機密データを外部に
出せないか？

YES → ローカル一択

Decision Axis

2. コスト (Cost)

トークン単価を気にせず
大量試行したいか？

YES → ローカル

Decision Axis

3. 自律性 (Autonomy)

長時間の自動作業を
丸投げしたいか？

YES → 商用API

Decision Axis

4. 極限推論 (Frontier Logic)

研究レベルの数学や
最難関の精度が必須か？

YES → 商用API

**常に4ヶ月遅れの最強AIが手元にある事実を活かし、
アジャイルに使い分けることが次世代の標準となる。**

Gemini Notebook