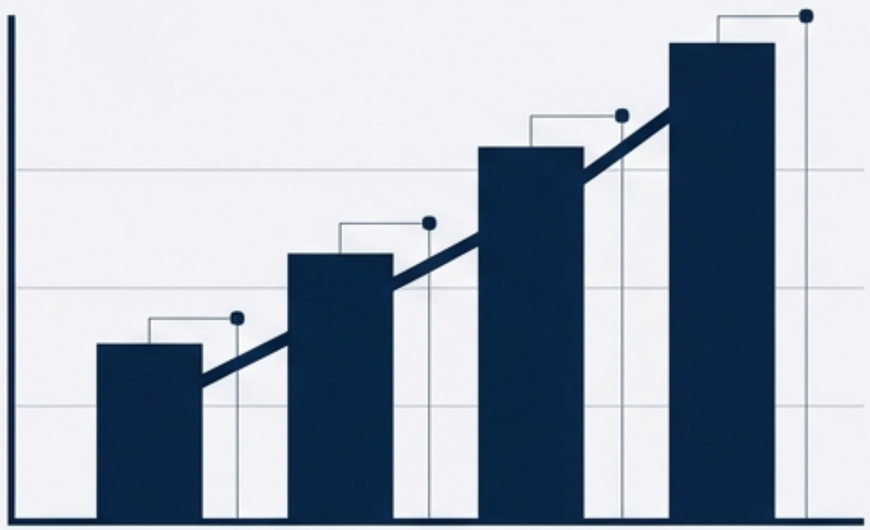
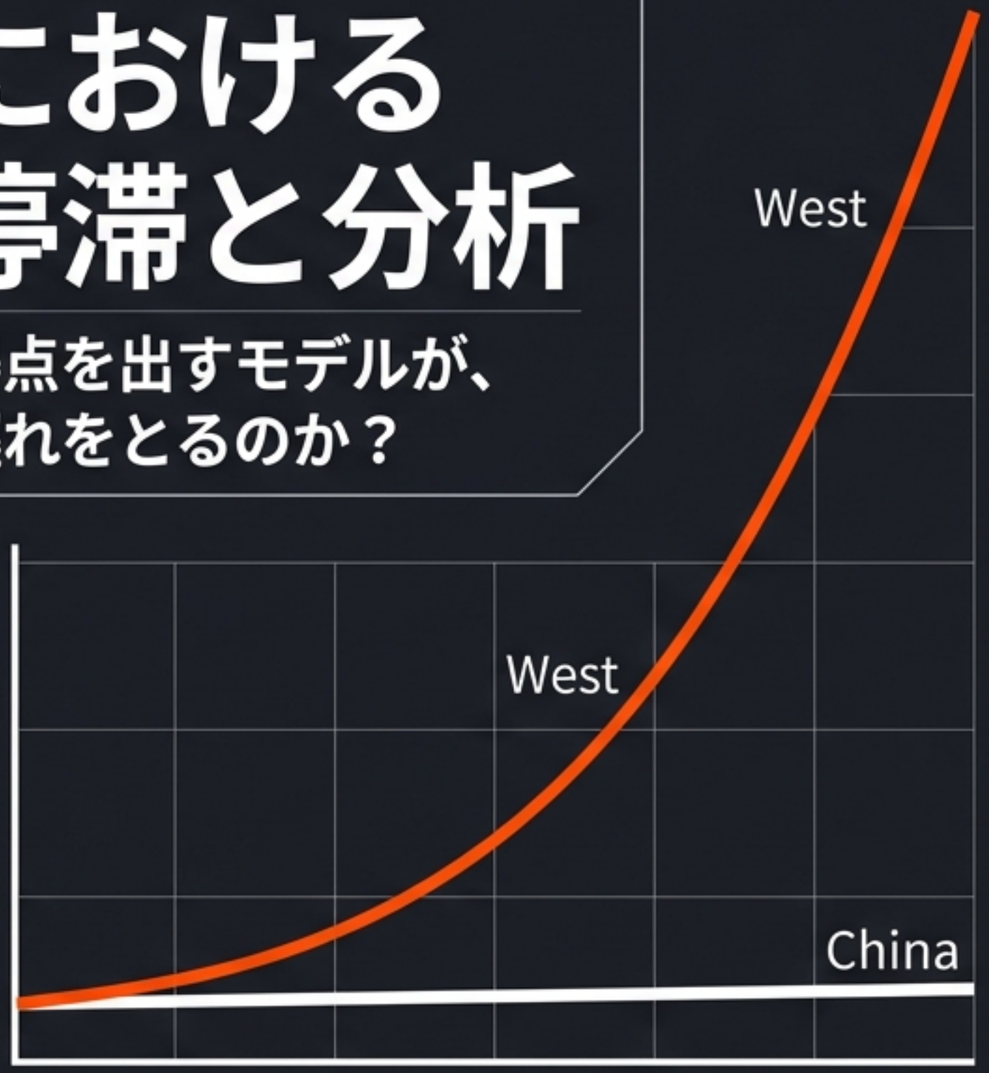
ARC-AGI-2における中国製LLMの停滞と分析

なぜ従来のベンチマークで高得点を出すモデルが、「流動的知能」で欧米に遅れをとるのか？



Parity（拮抗）



Divergence（乖離）

2026年2月時点のデータに基づく分析

エグゼクティブサマリー：見かけのパリティと隠れた格差

The Status (2026年2月の現状)

DeepSeekやQwenなどの中国製モデルは、SWE-Bench、AIME、GPQA-Diamondなどの従来型ベンチマークにおいて、欧米のフロンティアモデルと完全に対等なスコアを記録している。

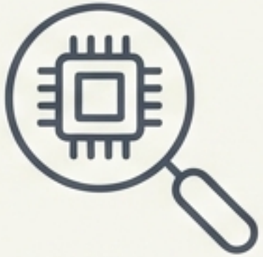
SWE-Bench / AIME = World Class

The Anomaly (ARC-AGI-2の壁)

しかし、未知のパターンへの適応を測定するARC-AGI-2において、Google（84.6%）や OpenAI (52.9%）が人間レベルに到達する一方、中国勢は ~1% 前後に留まっている。

Gap: >80 pts

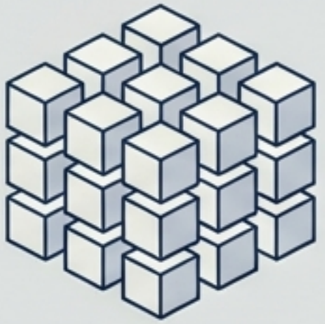
The Diagnosis (構造的要因)



この乖離は技術力不足ではなく、以下の要因に起因する：

1) 「テスト時計算（Test-Time Compute）」への投資不足
2) コスト効率（MoE）を最優先する市場戦略
3) GPU規制によるハードウェア制約

評価基準の転換：結晶性知能から流動的知能へ



結晶性知能 (Crystallized Intelligence)

- MMLU, GPQA, HumanEval

訓練データにある知識とパターンの再現。

China: High Performance ✅

流動的知能 (Fluid Intelligence)

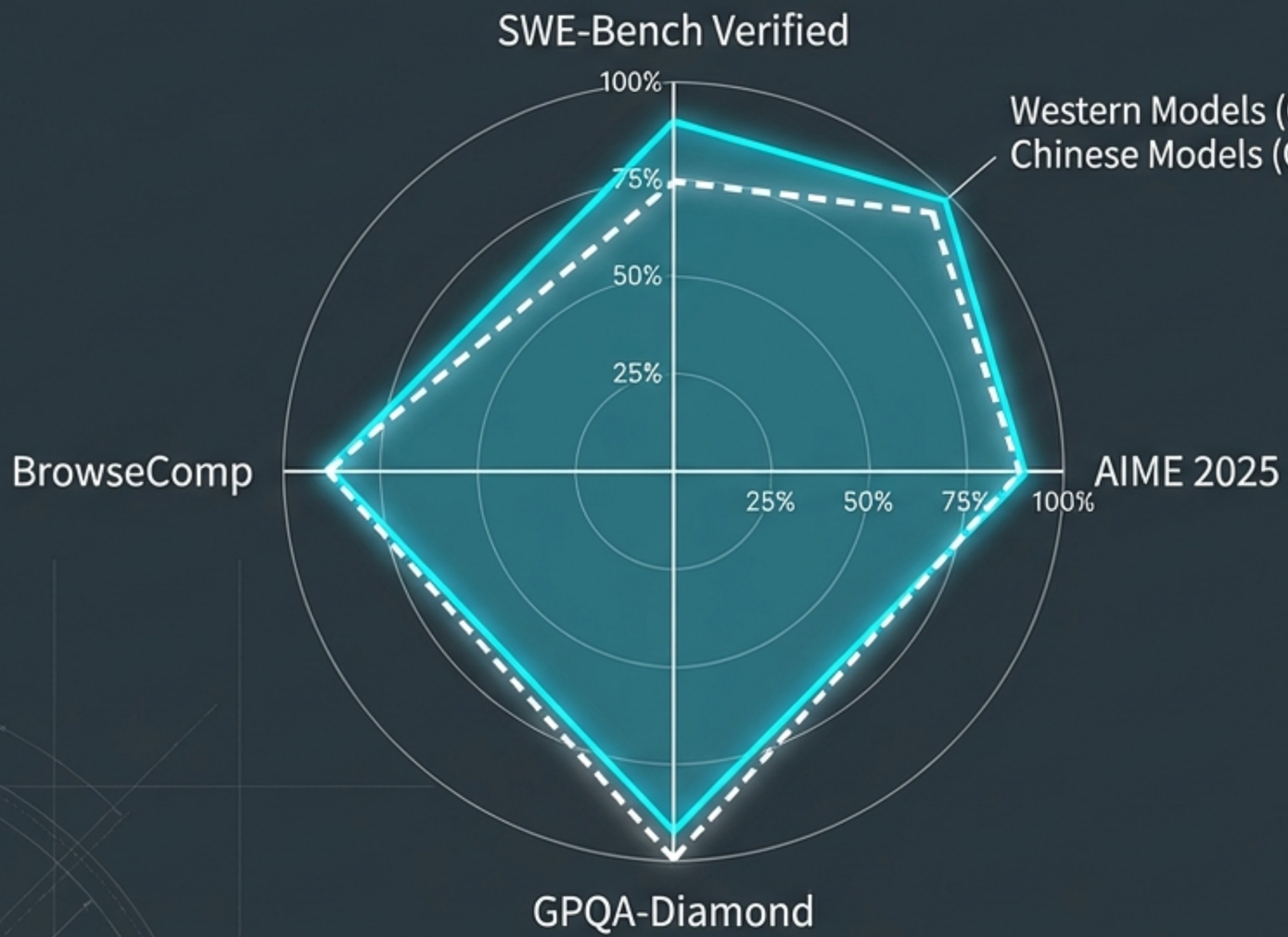
- ARC-AGI-2

未知の状況・パターンへの即時適応。記憶やブルートフォース（総当たり）が通用しない設計。

⚠️ **China: Low Performance**

ARC-AGI-2は、AIが『学習していないこと』にどう対処するかを測定する唯一の指標である。

2026年春節の錯覚：従来指標における完全な拮抗



Data Callout

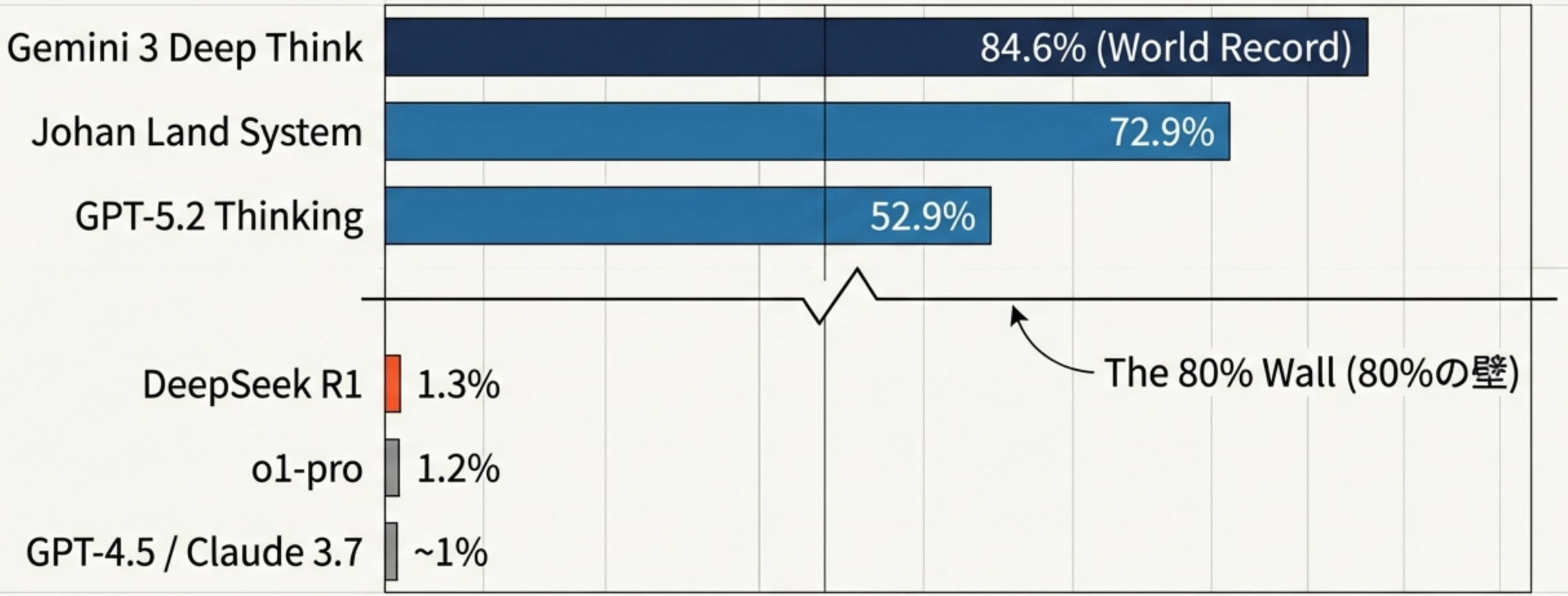
Score Comparison

SWE-Bench: Qwen 3.5 (76.4%) vs Claude Opus 4.6 (80.9%)

AIME 2025: Kimi K2.5 (96.1) vs GPT-5.2 (100)

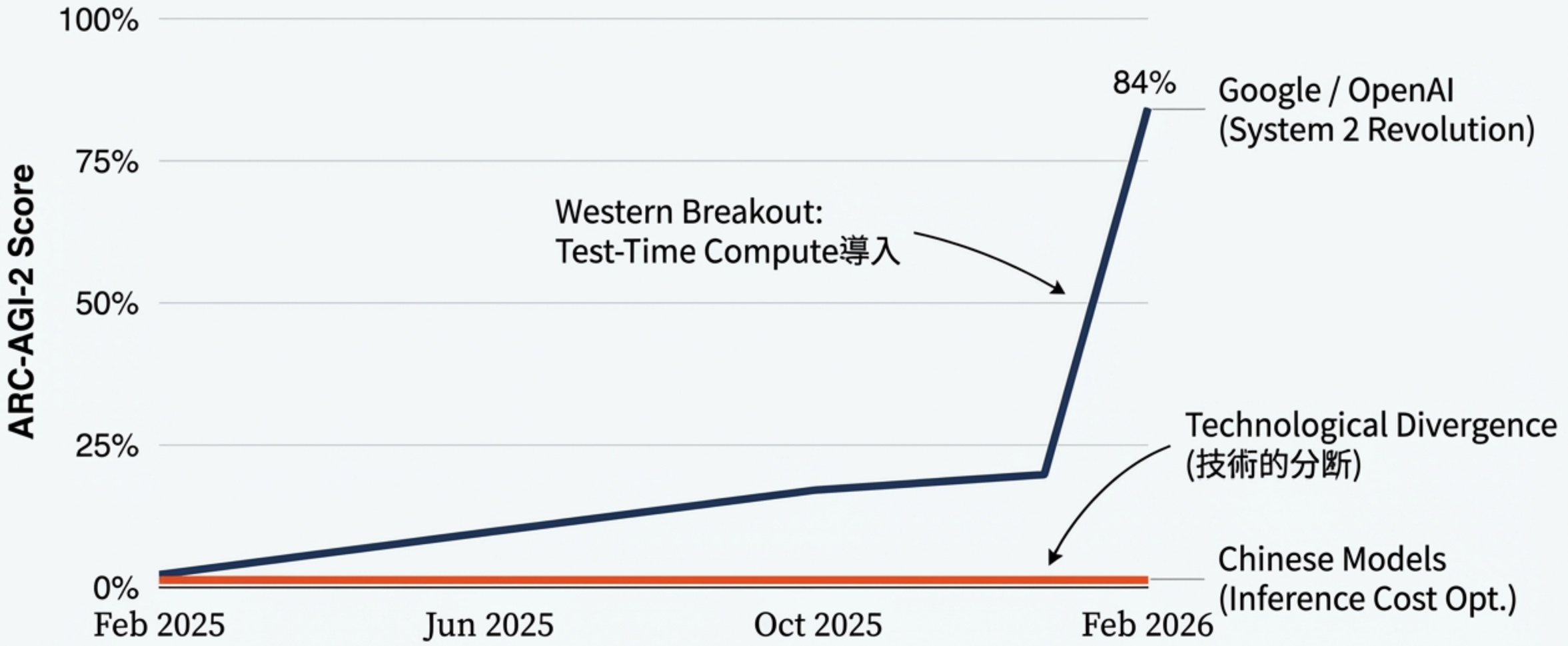
コード生成や数学において、中国製モデルはすでに世界最高水準にある。

現実の乖離：ARC-AGI-2 における「80%の壁」



春節にリリースされたQwen, GLM, Kimi などの最新モデルは、このリーダーボードにランクインすらしていない（または 1%圏内に留まっている）。

分岐のタイムライン：2025-2026年の技術的分断



4つの構造的障壁：なぜスコアが伸びないのか？

1. 技術戦略 (Technical Strategy)

Test-Time Compute: Deep Thinking vs. Fast MoE

2. 認知の性質 (Nature of Intelligence)

Memorization vs. Reasoning

3. ハードウェア (Hardware)

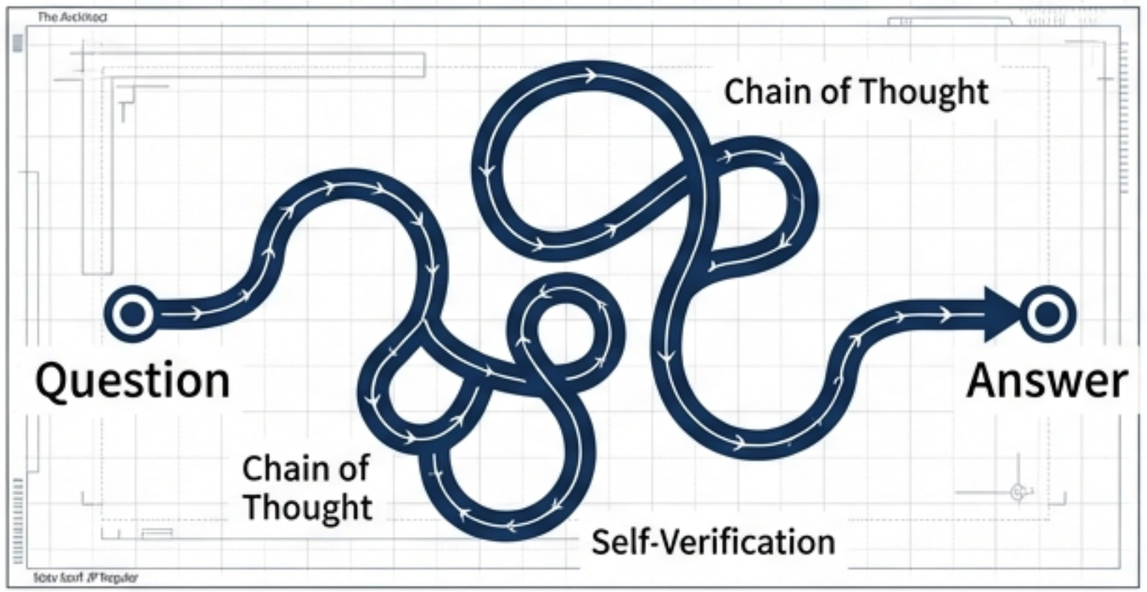
Chip Ban & Compute Constraints

4. 市場優先度 (Market Priority)

Cost Efficiency vs. Research Capability

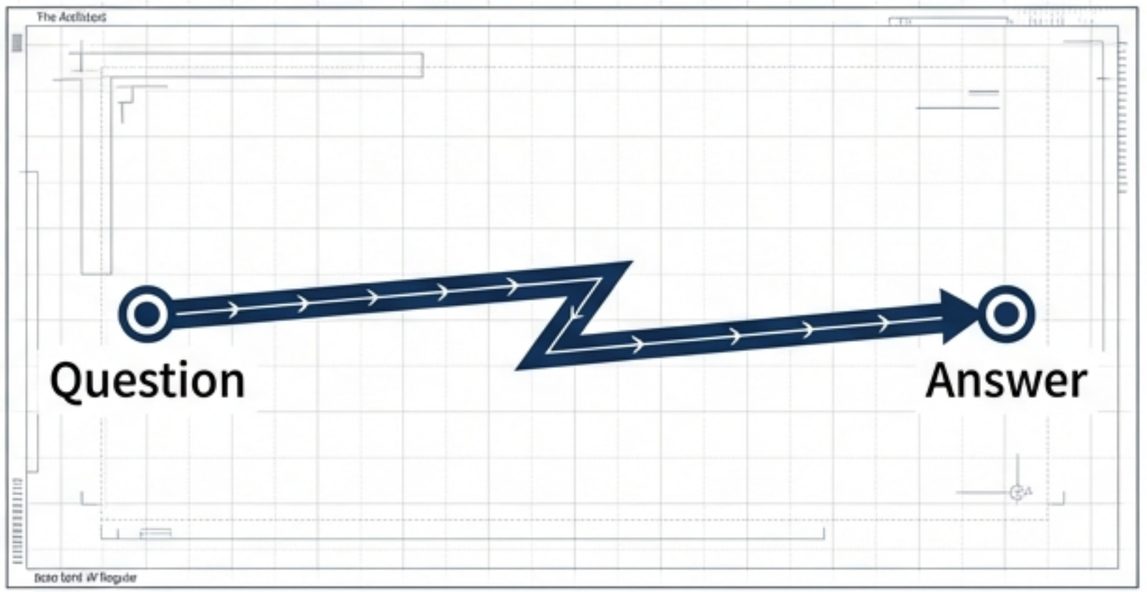
要因１：テスト時計算（Test-Time Compute）の欠如

Western Approach - Gemini 3 Deep Think



System 2: Thinking Longer (高コスト・高精度)

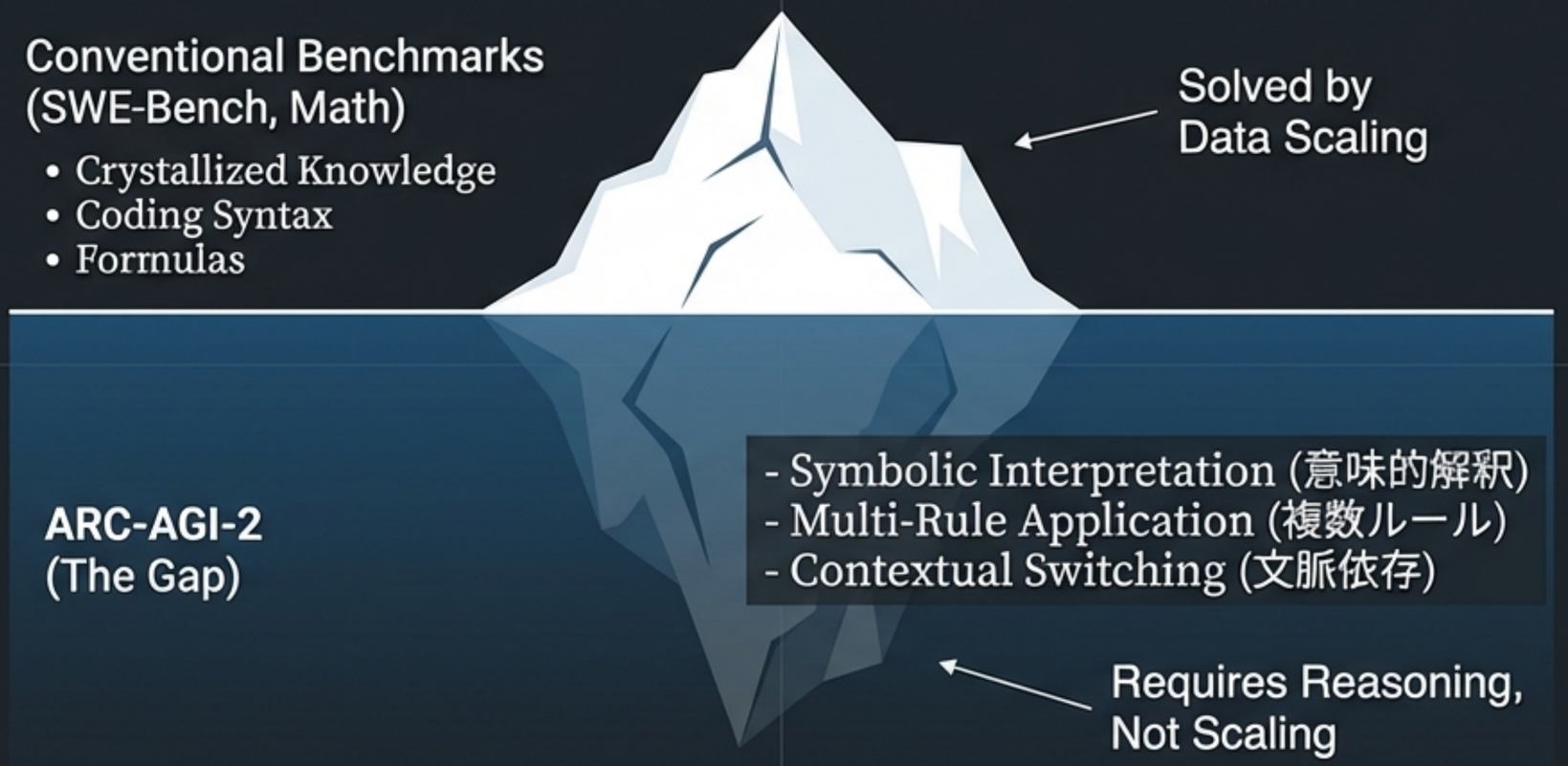
Chinese Approach - MoE Architecture



System 1: Thinking Faster (低遅延・低コスト)

“Gemini 3 Deep Thinkの84.6%は、応答生成前により長く『考える』能力による成果である。”

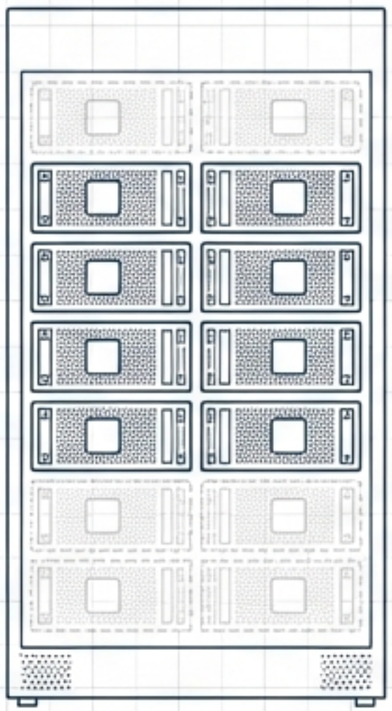
要因 2：スケーリング則が通用しない「質的な壁」



これらはデータ量の拡大（スケーリング）では解決できない。
表面的なショートカットではなく、根底にある選択原理の理解が必要。

要因３＆４：ハードウェア制約と「安さ」への最適化

Hardware Constraints



米国の輸出規制により、推論時に大量の計算資源（H100等）を「燃やす」アプローチが困難。

Example: GLM-5 trained on Huawei Ascend.

The Price War Strategy

Western API Cost: $$$

Chinese API Cost: ¢

中国市場の競争軸は「実用性」と「価格」。
API価格は欧米の1/5〜1/40。

“ARC-AGI-2 is research. SWE-Bench is revenue.”

次なる技術的転換点：ギャップを埋めるための鍵

1

Deep Test-Time Reasoning

深層テスト時推論。自己検証ループと、より長い思考時間。

2

Visual Native Processing

視覚ネイティブ処理。テキストを介さず、視覚パターンのまま推論する。

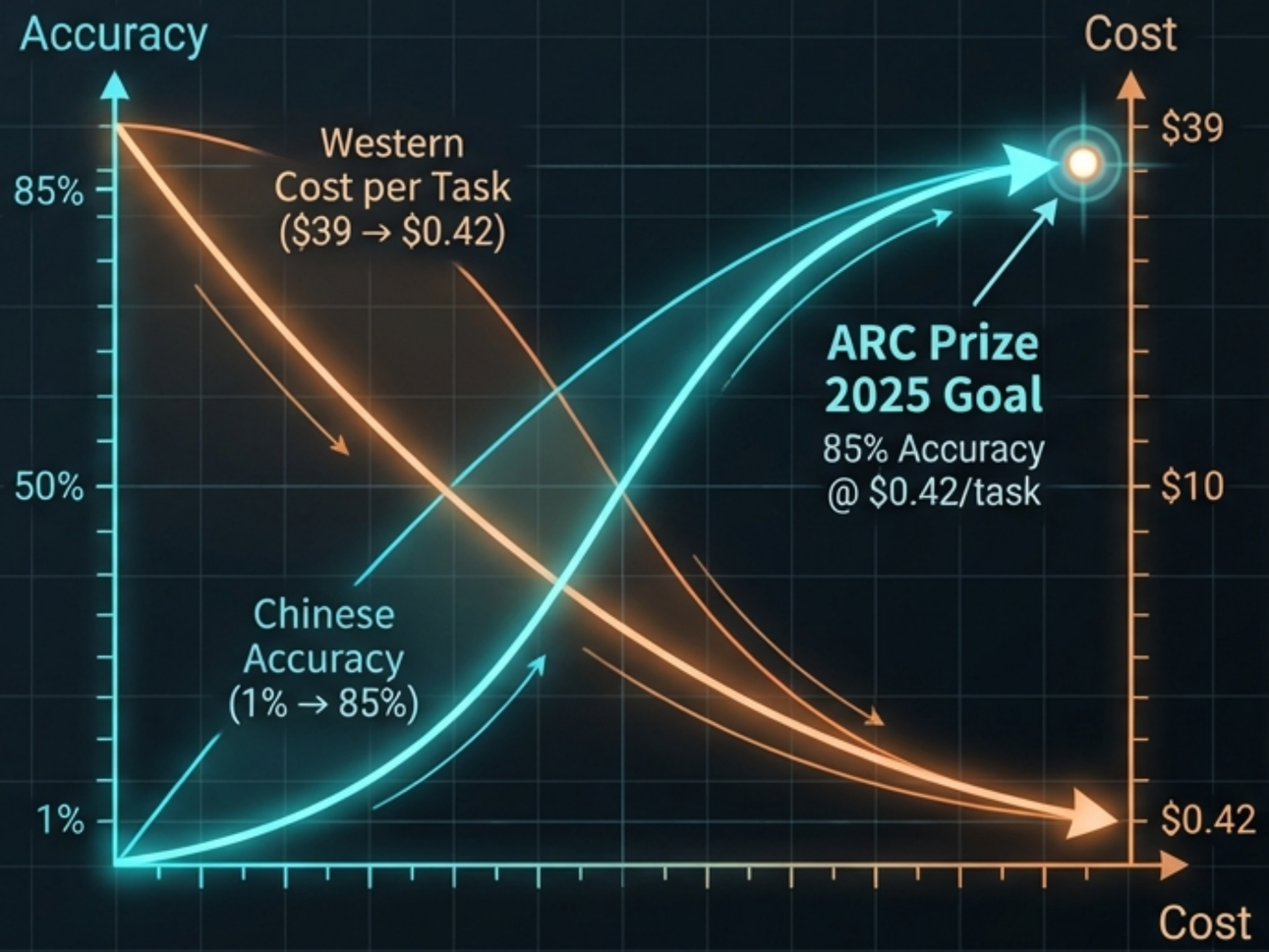
3

Reflective Reasoning

マルチモデル反射的推論。複数のモデルが回答を批評し合うシステム。

今後の展望：DeepSeek V4と目標の収斂

DeepSeek V4 Rumors: Engram Technology & Deep Reasoning.



Convergence: If China solves reasoning, reasoning, their efficiency expertise will dominate.

結論：評価と推奨事項

01. Commercial Readiness（商用利用）

ARC-AGI-2のスコアだけで中国製モデルの全体能力を判断してはならない。ビジネス用途（コーディング、数学）において、これらは十分に強力かつ安価である。

02. The Watchlist（監視対象）

2026年後半に登場する中国の「推論特化型」モデル（DeepSeek V4等）に注目せよ。ギャップが急速に縮まる可能性がある。

03. The Era of Inference（推論の時代）

AI開発の競争軸は、「学習（Training）」から「推論（Inference）」へと完全に移行した。テスト時計算能力が次の覇権を握る。