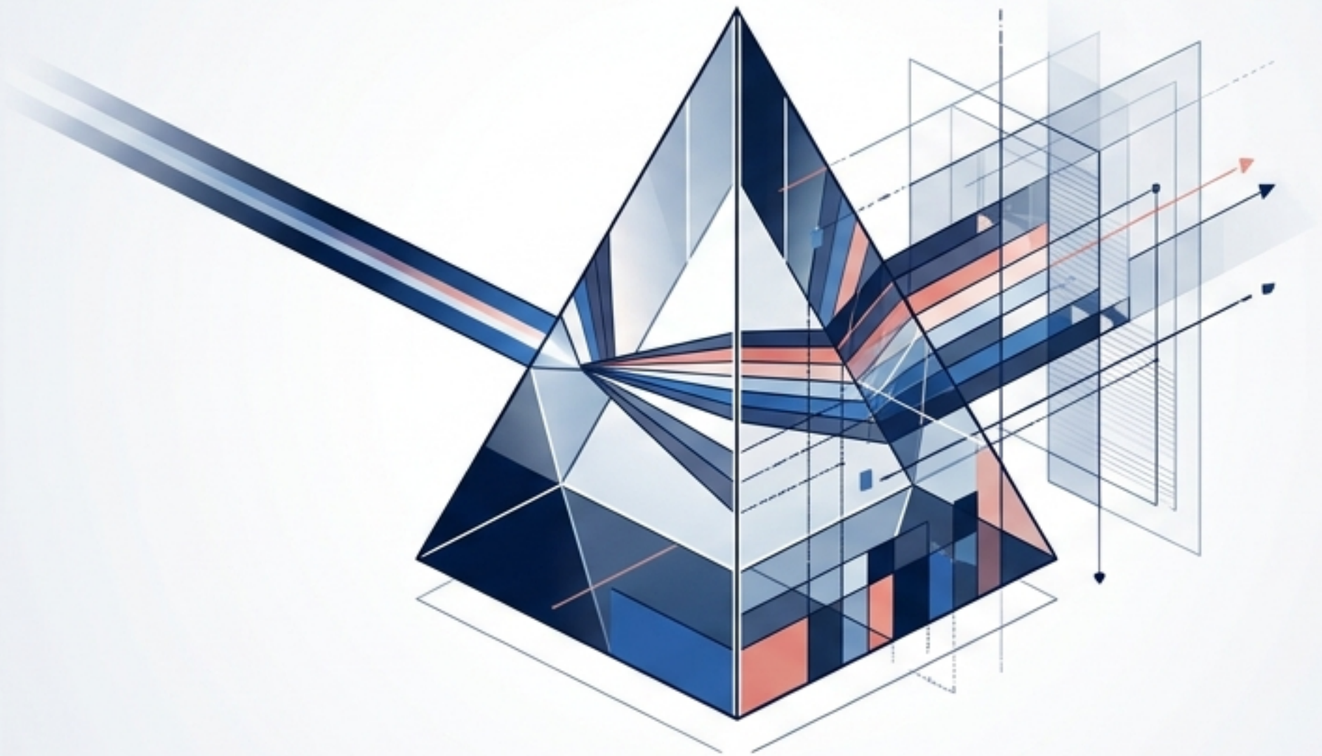
The DeepSeek V4 Paradox

解体される「8か月遅れ」のナラティブと非対称AIの導入戦略



2026年5月発行 | 対象読者：エグゼクティブ・エンジニアリング層 | CAISI 2026年5月評価に基づくインテリジェンス・ブリーフィング

The Executive Diagnostic

評価の歪み

CAISIの「8か月遅れ」は、5領域等重みと非公開タスクにより強調された数値。総合評価はGPT-5相当（Elo 800）。

非対称な実力

数学・公開コード領域では米国最先端に肉薄。一方、未知の抽象推論やサイバー領域では明確な断絶が存在する。

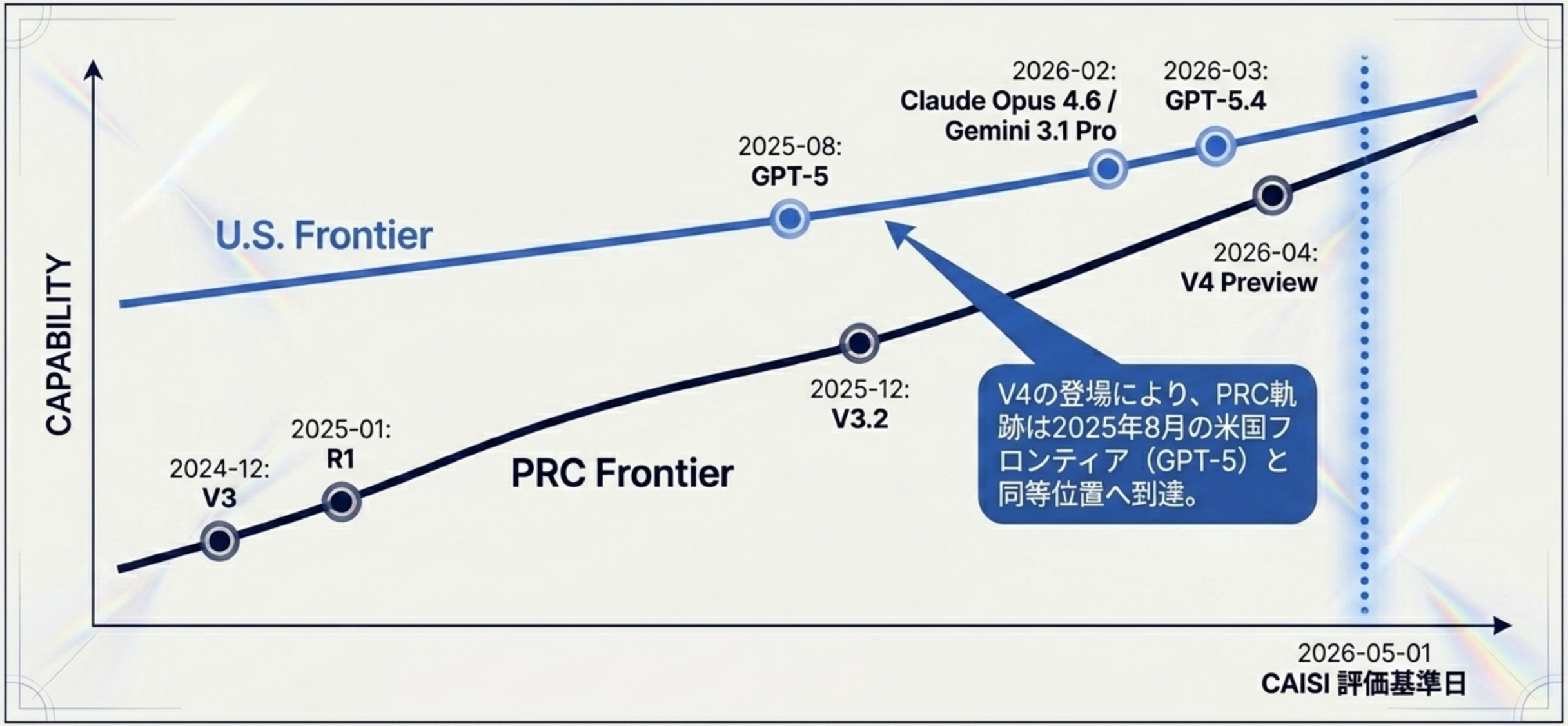
破壊的経済性

1Mトークン文脈、MITライセンス。GPT-5.4 miniを凌駕するコストパフォーマンスがエコシステムを席巻。

導入の前提条件

リーダーボードへの依存は危険。「ハルシネーション率94%」の特性を理解し、自社のHeld-out環境での検証が必須。

The Frontier Race: U.S. vs. PRC Trajectories



The Core Tension: Two Conflicting Narratives

CAISI / NIST 評価

米国最先端から約8か月遅れ

アンカー基準: GPT-5相当

評価指標: 1PL-IRT推定能力尺度

総合スコア: IRT Elo 800 ± 28

> これまで評価した中国モデルの中で最高性能だが、GPT-5.4やOpus 4.6相当ではない。

VS

DeepSeek 公式発表

最先端から約3–6か月遅れ

アンカー基準: GPT-5.2 / Gemini 3.0 Proを凌駕

評価指標: 公開ベンチマーク

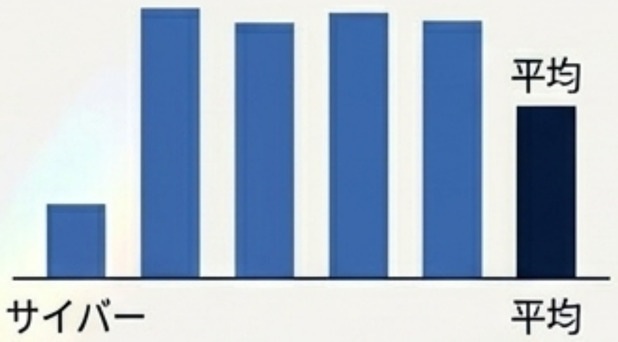
総合スコア: LiveCodeBench 93.5

> アーキテクチャ: 総1.6T/有効49Bパラメータ、32T超の学習トークンによる圧倒的推論力。

Anatomy of 8 Months Behind

1. 5領域の等重み集約

サイバー、ソフトウェア工学、
自然科学、抽象推論、数学



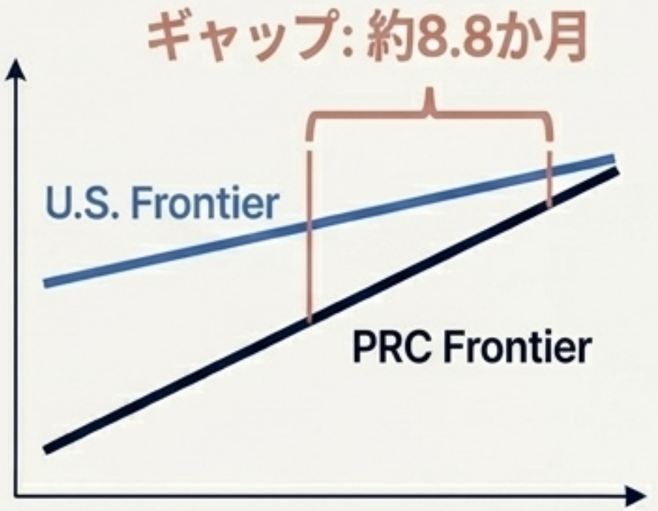
一部の極端な弱点（サイバー等）
が全体の総合スコアを強く押し
下げる構造

2. 1PL-IRT & Elo 推定

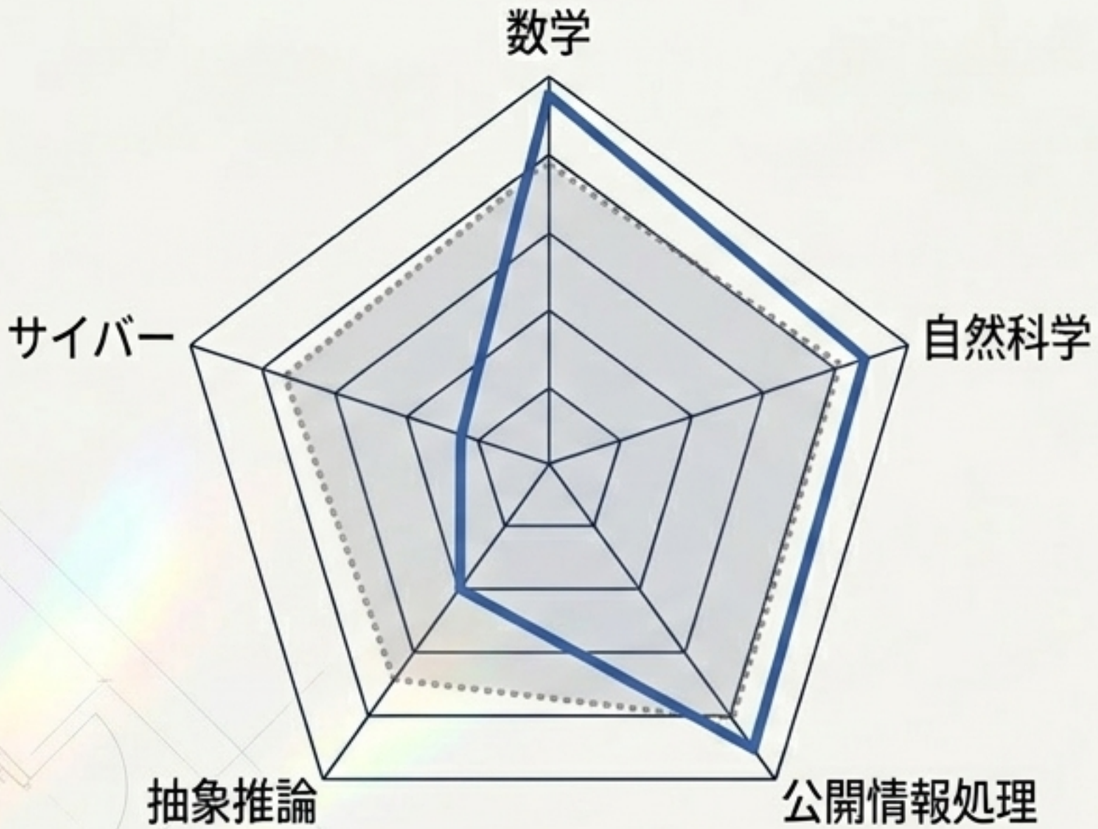
16ベンチマーク・35モデル
を統合し、能力をElo尺度に
マッピング

3. 時系列回帰ライン

米中フロンティア系列の
相対的な水平距離を測定



The Capability Radar: Uneven Frontiers



接近・超越領域 (Strengths)		
数学 (OTIS-AIME)	**V4: 97%**	Opus 4.6: <97%
自然科学 (FrontierScience)	**V4: 74%**	GPT-5.4 mini: 74%
公開情報処理 (GPQA)	**V4: 90%**	Opus 4.6: 91%
断絶領域 (Weaknesses)		
抽象推論 (ARC-AGI-2 semi-private)	**V4: 46%**	GPT-5.5: 79%
ソフトウェア工学 (PortBench held-out)	**V4: 44%**	Opus 4.6: 60%
サイバー (CTF-Archive)	**V4: 32%**	Opus 4.6: 46%

インサイト: 全領域一様の遅れではなく、特定領域に鋭く特化した非対称な能力形状を持つ。

DeepSeek's Arsenal: Engineering for Extreme Context

DeepSeek V4 Pro Architecture

Attention Architecture

- 推論効率化の極致
- CSA / HCA ハイブリッド構造。
- V3.2比でsingle-token FLOPsを27%、KVキャッシュを10%へ劇的削減。

Context Horizon

- 長文脈とエージェント特化
- 1Mトークンの有効文脈。
- Mid-trainingでの Agentic dataの集中投下によるエージェント基盤の大幅強化。

Infrastructure Constraints

- 供給制約の突破
- Huawei Ascendへの最適化と制約下の学習。

資源制約が「汎用性」よりも「特定用途（長文脈・コード）への特化」を促進した可能性。

The Evaluation Prism: Diverging Methodologies

評価機関	評価手法・データ	主な結論	代表値
DeepSeek 公式	公開ベンチ / 自己報告	最先端から3-6か月遅れ	LiveCodeBench: 93.5
CAISI / NIST	非公開含む5領域等 重み・IRT集約	約8か月遅れ / GPT-5相当	IRT Elo: 800
Artificial Analysis	独立指標・実務評価	オープンウェイト最上位 級だが幻覚率高	幻覚率: 94%
LiveBench	公開・客観継続ベンチ	公開ベンチでは 最前線クラスに肉薄	Overall: 73.58
LM Arena	人間選好 (Blind Pairwise)	中上位止まり。 閉鎖型最前線には届かず	Text Rank: 25

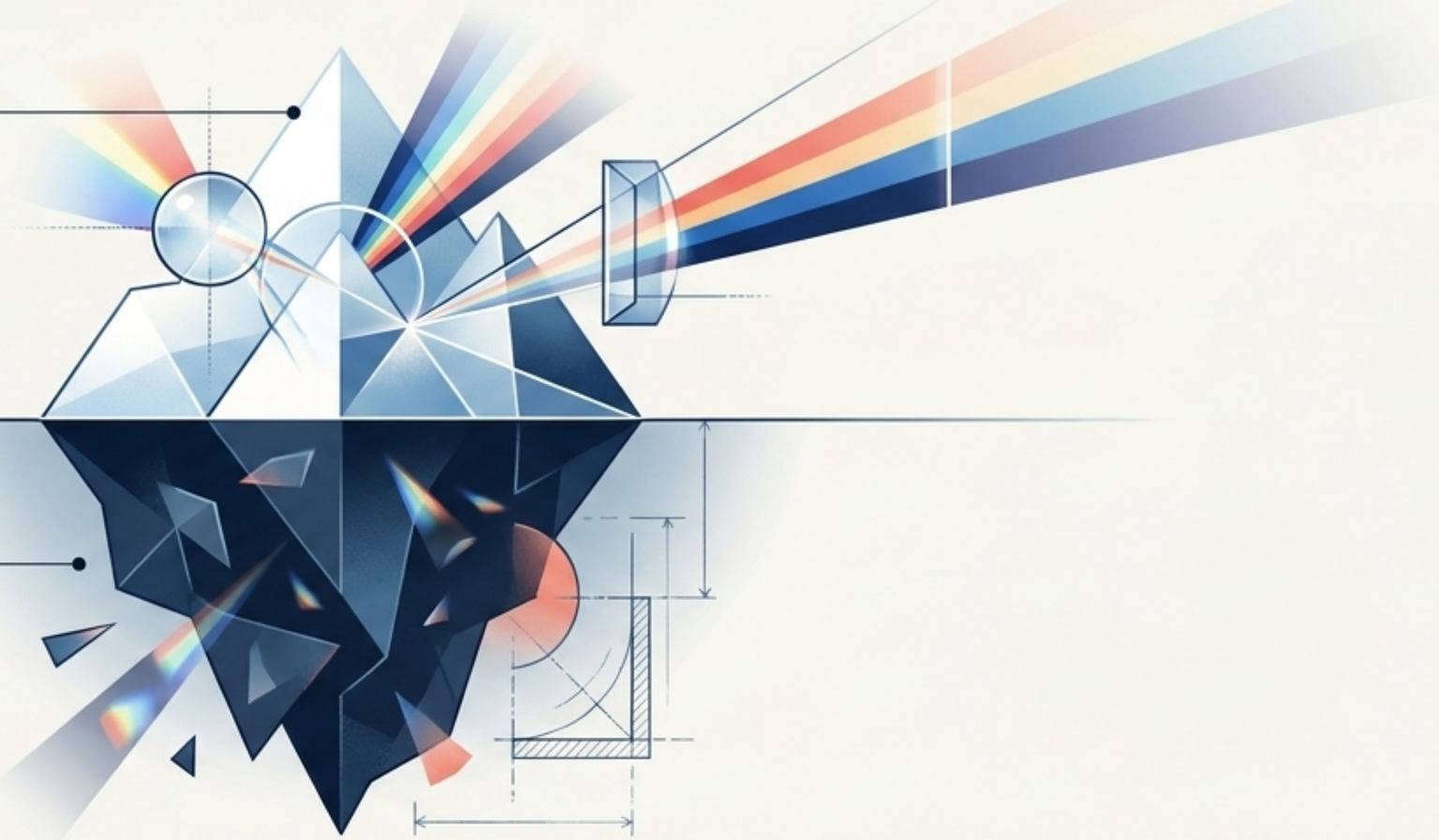
インサイト: 結論のブレはモデルの不安定さではなく、「公開ベンチ」「非公開タスク」「人間選好」という切り口（測定器）の違いによる必然的な結果である。

The Measurement Gap: Over-optimization vs. Generalization

The Surface: **公開ベンチマークの光**

LiveBench、自己報告ベンチ、
数学・コーディングタスク。

**V4はSOTA（最先端）として振る舞い、
「3-6か月遅れ」の姿を見せる。
公開データへの極めて高い最適化。**



The Subsurface:
非公開 Held-out **の深淵**

CAISIのPortBench、ARC-AGI-2
semi-private、サイバー演習。

**未知の抽象推論に対する一般化能力の欠如
が露呈し、「8か月遅れ」の姿が現れる。**

The Synthesis: V4は自己評価で嘘をついているわけでも、CAISIの評価が厳しすぎるわけでもない。
V4は公開データと特定タスクに過剰適合した『非対称なAI』であり、未知の運用環境において真の実力と脆さが露呈する。

Economic Disruption & Ecosystem Advantage

The Licensing Moat (ライセンスの堀)

- MITライセンスによる重みとコードの完全配布
- OpenAI / Anthropic 互換のAPI仕様
- 圧倒的な低コストと1M文脈の組み合わせによる、エコシステム拡張と移植性への極端な最適化。

Deployment Vulnerabilities: The Risk Surface

High Hallucination Rate

知らないことでも沈黙せずに回答を生成する傾向。Artificial Analysisにより幻覚率94%が報告されており、自律的な情報提供における最大のリスク。

The Text-Only Constraint

マルチモーダル（画像・動画）非対応。視覚情報を前提とした業務フローや次世代の複合インターフェース構築から排除される懸念。

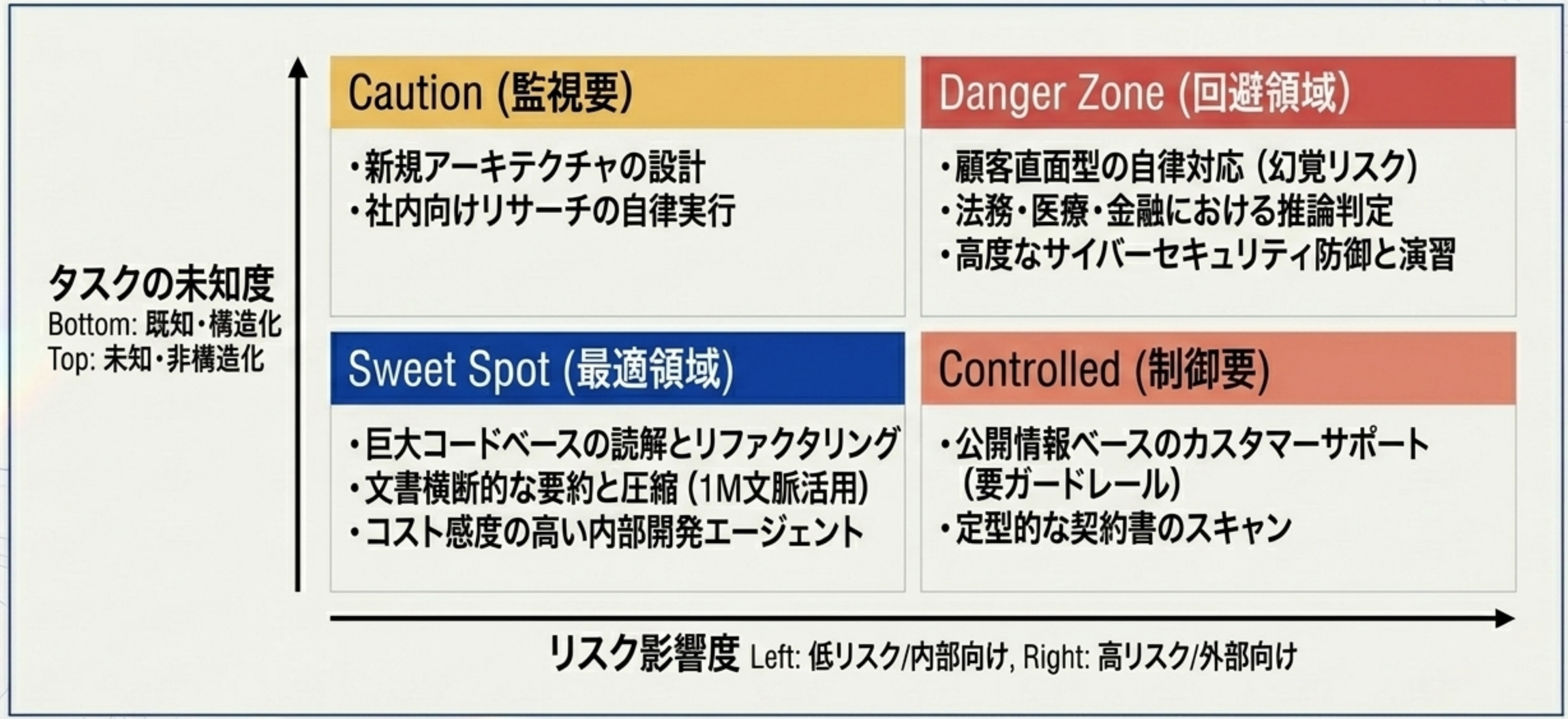
Governance & Data Privacy

プライバシーポリシー上、中国国内でのデータ処理・保管の可能性が明記。詳細なSystem Cardやリスクレポートが欠如しており、規制産業での利用障壁となる。

Fragility in Transfer（移植の脆さ）

公開コードの読解には強いが、未知のコード移植や高度なサイバーセキュリティ演習（CTF等）における適応力と一般化能力が著しく低い。

Risk vs. Reward: Actionable Capability Matrix



Strategic Recommendations

For Developers (開発者)

- リーダーボードへの盲信を脱却する。
- 自社独自の Held-out課題（未知推論・コード移植など未公開タスク）での事前検証を、パイプラインの必須要件として組み込む。

For Enterprise Deployers (導入企業)

- ユースケースに応じ、MIT重みの自前環境運用（セルフホスト）とAPI利用を切り分ける。
- API利用時のデータ所在リスクと、セルフホスト時の安全フィルタ実装責任を明確にコントロールする。

For Regulators (規制当局)

- 能力を「○○か月遅れ」という単一指標で語る危険性を認識する。
- 事前固定された非公開スイートを用いた評価と、Elo推定の前提条件の併記を標準化し、多角的な評価を要請する。

測定器を自ら設計する者だけが、真の能力を捉えることができる。