STRATEGIC INTELLIGENCE BRIEFING

TARGET: GLOBAL LLM EVALUATION

Motif 3の「47点」を解剖する

Artificial Analysis独立評価の深層と、日韓LLM戦略の分岐点

DATA SOURCE: Manus AI Report (2026.08.17)

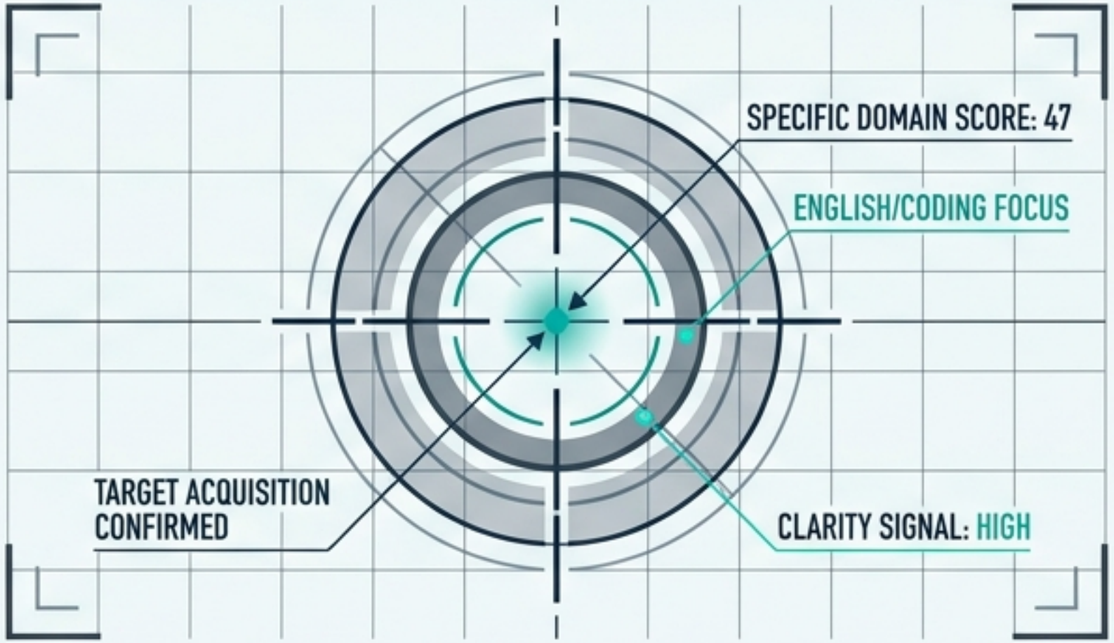
CLASSIFICATION: STRATEGIC INSIGHT

表面的なシナリオ

「韓国のLLMが世界トップクラスの47点を獲得し、日本のLLMはランキングに存在しない。日本は競争に敗北している。」

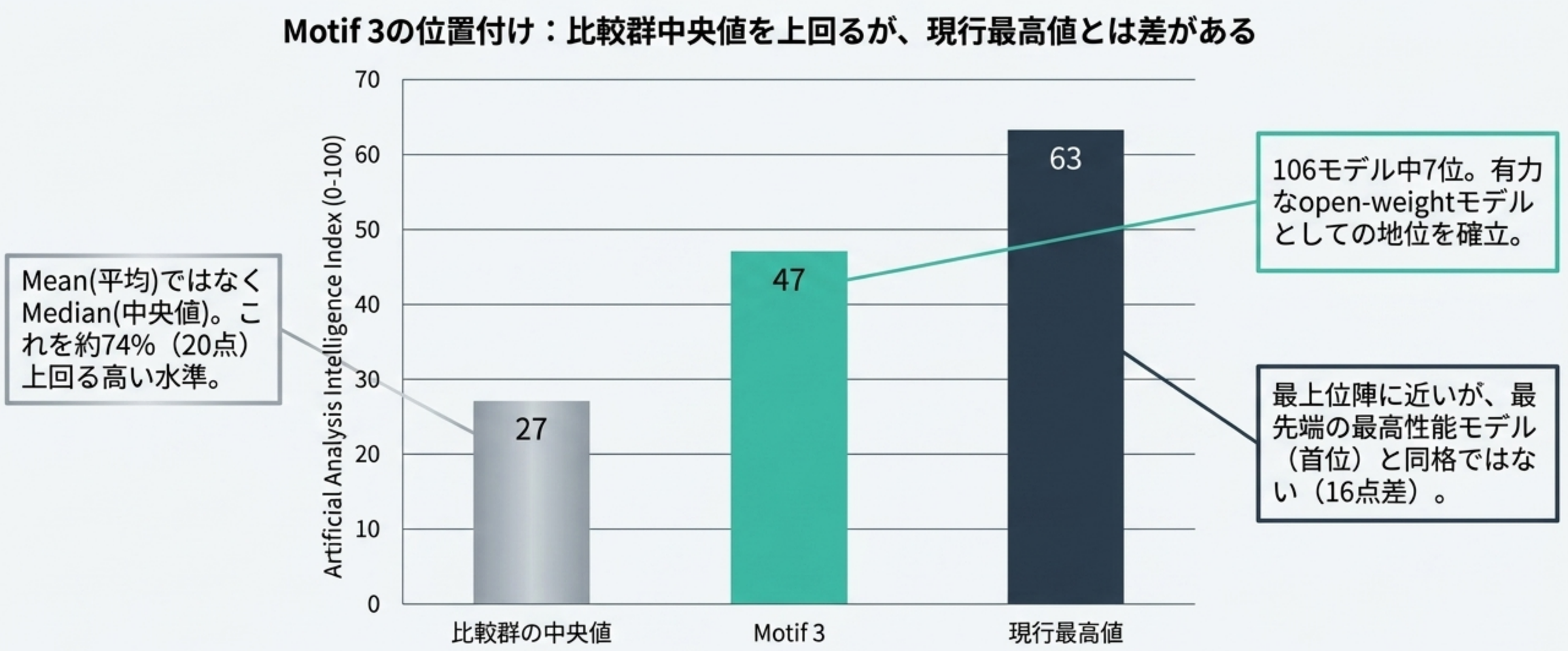
分析的な実態

「一社の韓国モデルが、特定領域(英語・コーディング・エージェント）に特化した指標で極めて高い適合性を示した成果である。
日本とは最適化の『目的関数』が異なる。」



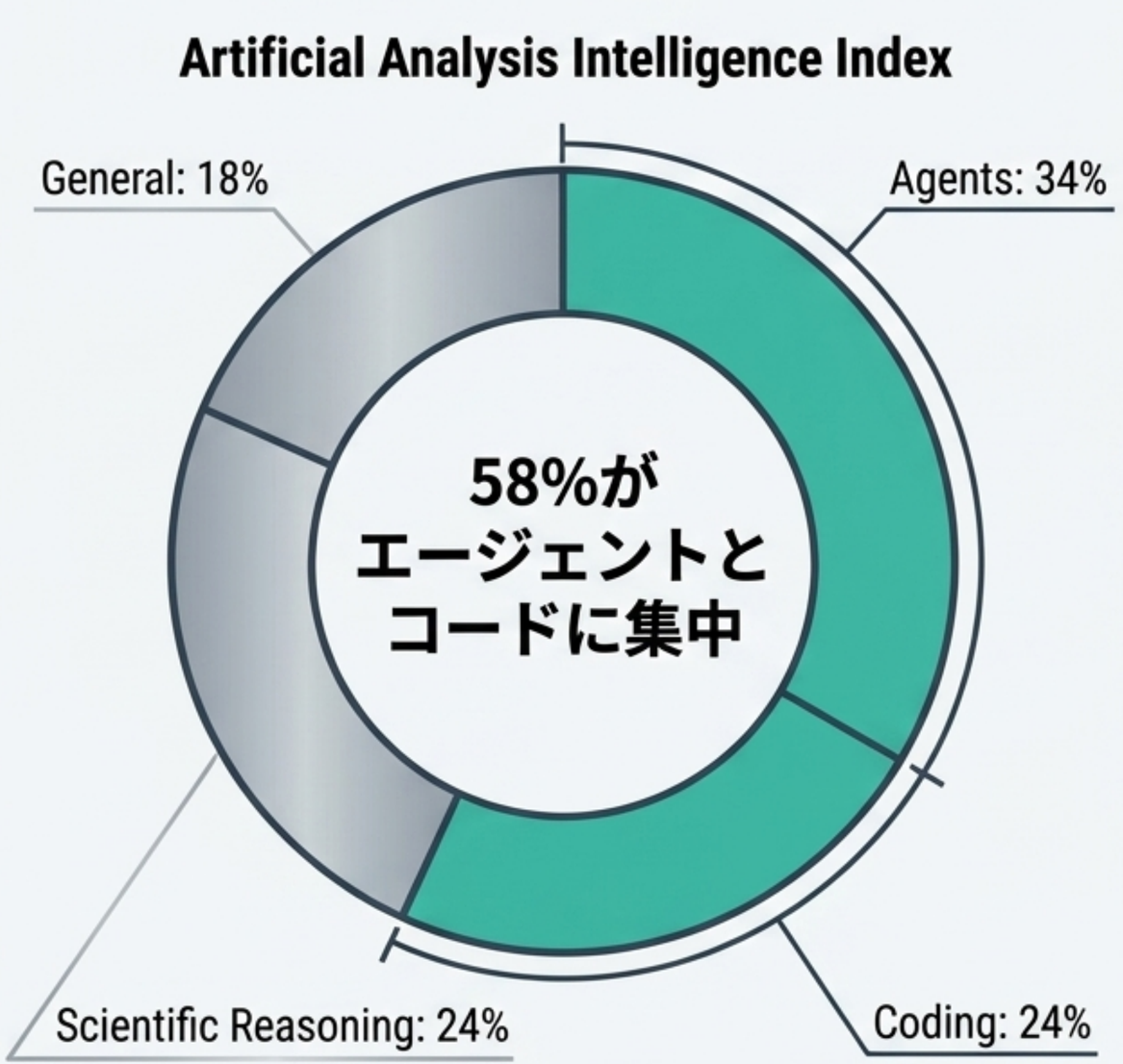
評価指標の構造と、日韓のエコシステムの真の差を検証する。

47点の相対的ポジション：中央値を大きく凌駕する第1グループ



結論：極めて重要な成果だが、万能の「知能指数」ではない。

ベンチマークの解剖：何が測定されているのか？



> Artificial Analysis Intelligence Index is a text-only, English language evaluation suite.

- **評価の焦点:** 英語テキスト中心の実行型評価（Terminal-Bench, SWE-bench等）。
- **測定外の領域:** 日本語の敬語、国内固有の文書、商慣行、行政・法務の制度知識。
- **戦略的意味:** この47点には、日本企業が重視する「国内業務への適合性」は直接反映されていない。

Motif 3のアーキテクチャ：計算効率と専門性の両立

THE DATA POOL

約12.5兆トークン
(Web, STEM, Code, Math, 専門領域)

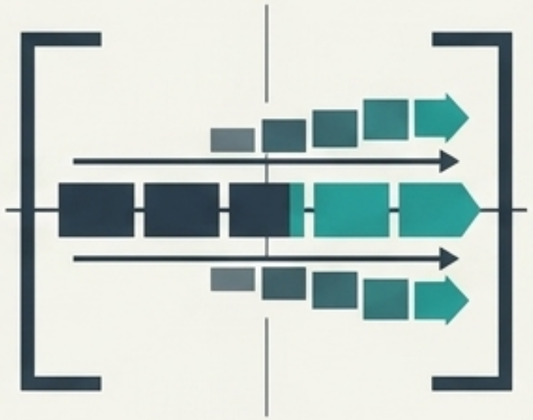
約70%がNVIDIA Nemotronの公開データ群を活用。英語の技術・推論データが中心。

THE MOE ENGINE

Decoder-only fine-grained MoE

384 expertsのうち、各層で8つのみを選択し計算量を抑制。

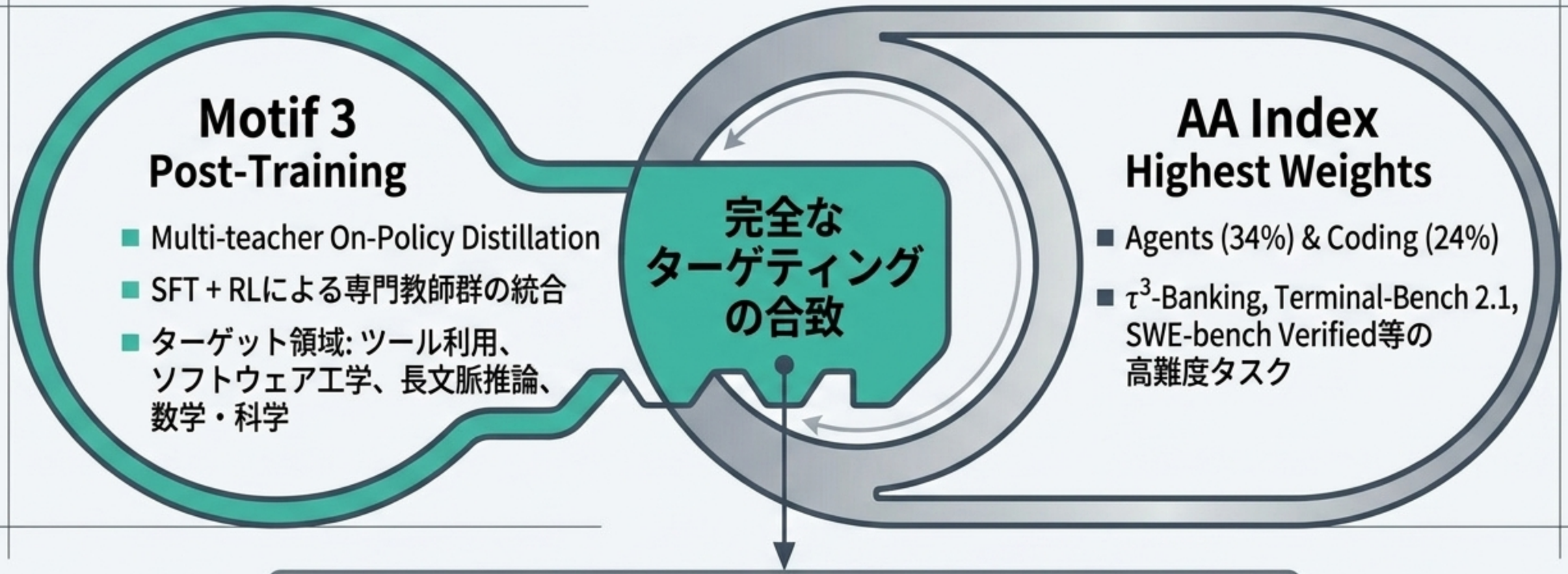
THE CONTEXT WINDOW



256K 長文脈
(GDLA, 文脈並列化)

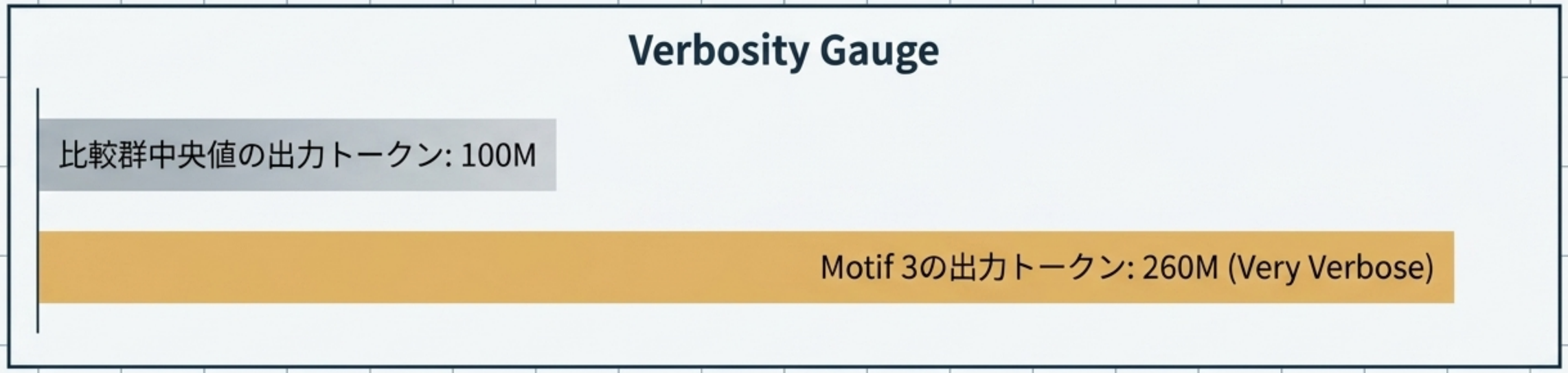
長い入力とエージェント軌跡を扱う基礎体力。

評価指標との「鍵と鍵穴」の適合



開発側の「ツール利用と端末型タスクに強い」という設計意図が、独立評価の配点構造と直接的に重なっている。

得点とコストのトレードオフ：Verbosity（多弁性）の代償



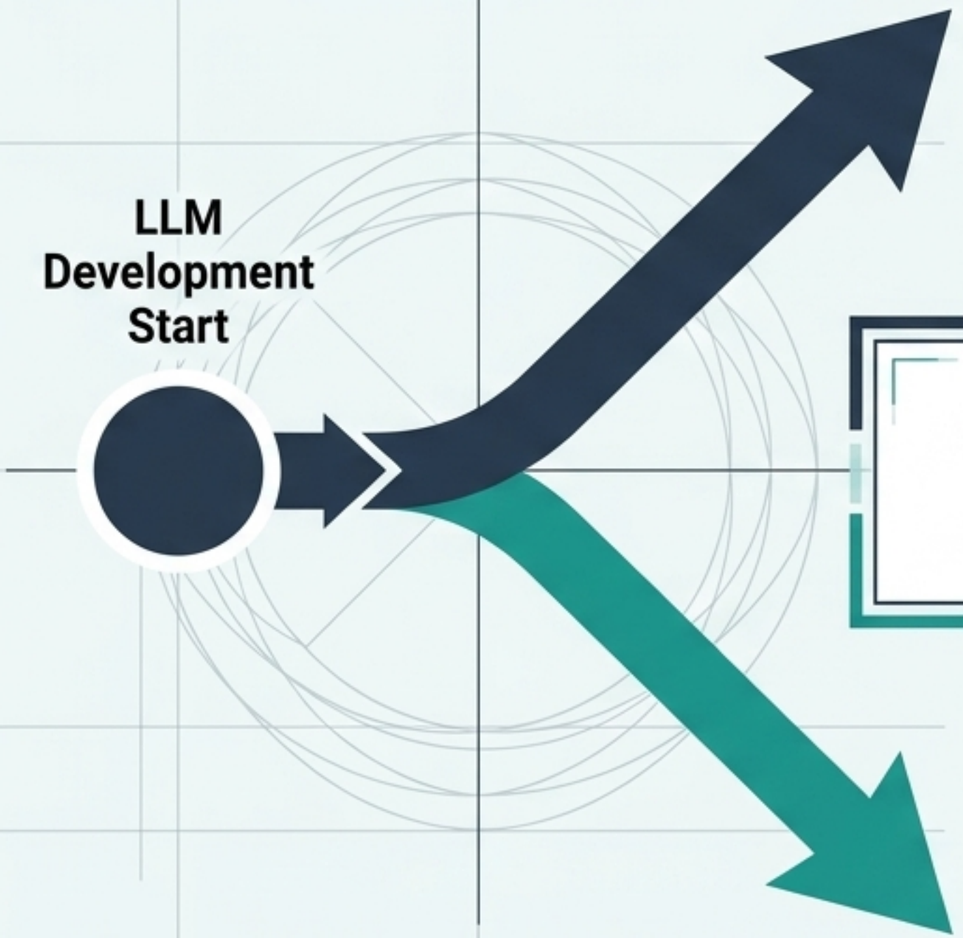
不正ではない戦略:
高度な推論・エージェント評価では、**長い試行と推論の連鎖**が得点に寄与する。

実務上の課題:
47点というスコアは、**推論予算（レイテンシ・実行コスト）**との**交換関係**の上に成り立っている。

**企業導入の判断基準は「スコア」だけでなく、
速度、価格、実タスクの失敗モードを含む総合的なROIに依存する。**

なぜ日本のLLMは「見えにくい」のか？

Diverging Objective Functions



The Global Evaluation Game (Motif 3 / Korea)

- **目的関数:** 国際ベンチマーク（英語・エージェント）での可視化。
- **形態:** 314B級の大規模モデル、Open-weight公開、英語偏重。
- **結果:** Artificial Analysis等の海外Leaderboardでの上位ランクイン。

観測される「不在」は、能力の欠如ではなく、製品設計と露出経路の意図的な相違である。

The Local Enterprise Utility (Japan Ecosystem)

- **目的関数:** 国内産業への実装と、日本語文化・商慣行の高精度な処理。
- **形態:** 軽量・省電力モデル、特定業界向け、API・オンプレミス中心。
- **結果:** 独自の国内Leaderboardでの評価、エンタープライズの堅牢な導入。

日本のLLM開発エコシステムの実態

Ecosystem Node Map

[GOVERNMENT & INFRASTRUCTURE]

GENIAC (経産省・NEDOによる計算資源・データ共有支援)

METI GPUクラウド補助 (最大725億円、5件の計算資源増強)

日本の基盤モデル開発網

[DOMESTIC EVALUATION BOARDS]

Open Japanese LLM Leaderboard (LLM-jp / 20超のデータセット)

Swallow LLM Leaderboard (日本語・英語の分離測定)

[KEY DEVELOPERS]

PLaMo, llm-jp, ELYZA, ABEJA, Sakana AI,
Preferred Elements など多数

「日本は基盤モデルに投資していない」は明白な誤り。
集中型ではなく、社会実装を目指した分散型の開発が進んでいる。

Strategic Intelligence Dossier

韓国の競争環境：「主権AI」と国際評価へのアラインメント

国家主導の枠組み（MSIT Sovereign AI Foundation Model）

- 目的: AI G3を目指し、海外モデルへの依存を低減。
- 初期5チーム: NAVER Cloud, Upstage, SK Telecom, NC AI, LG AI Research。
- 評価基準: 国際ベンチマークと専門家評価を制度設計に組み込む。

独立系ラボの台頭 (Motif Technologies)

- 注目点: Motif 3は上記「当初5チーム」の直接的成果ではない。
- 意味: 政府プロジェクトだけでなく、独立ラボが国際評価に合わせた大規模モデルを開発・公開できる厚い競争環境が存在する。

グローバル展開への共通意識

- K-EXAONE, Solar系列, A.X系列なども、国際ベンチマークやOpen-weightでの発信を強く意識。

戦略的ダイバージェンス：観測される差異の構造

評価次元 (Dimensions)	Motif 3 / 韓国側のアプローチ	日本LLMエコシステムのアプローチ
総合評価の言語基準	英語テキスト・コード中心 (AA Index適合)	日本語固有の評価を別立て (LLM-jp, Swallow等)
最高配点の能力	エージェント実行・端末操作 (58%配点)	日本語業務、軽量性、特定産業への適応
提供・公開形態	技術報告、モデルカード完備のOpen-weight公開	商用API、企業向けオンプレミス、研究限定など多様
国家戦略の重心	国際指標を意識した集中競争とGPU・データの結集	GENIACを通じた社会実装重視・分散型のエコシステム

「モデルの国籍」と「能力」の混同を避ける

[STEP 1: THE FACT]

Motif 3は特定指標で世界7位

Artificial Analysis v4.1.1（英語/エージェント特化）において上位の成績を収めた公開重みモデルである。

[STEP 2: THE MISINTERPRETATION]

「韓国全体の優越」ではない

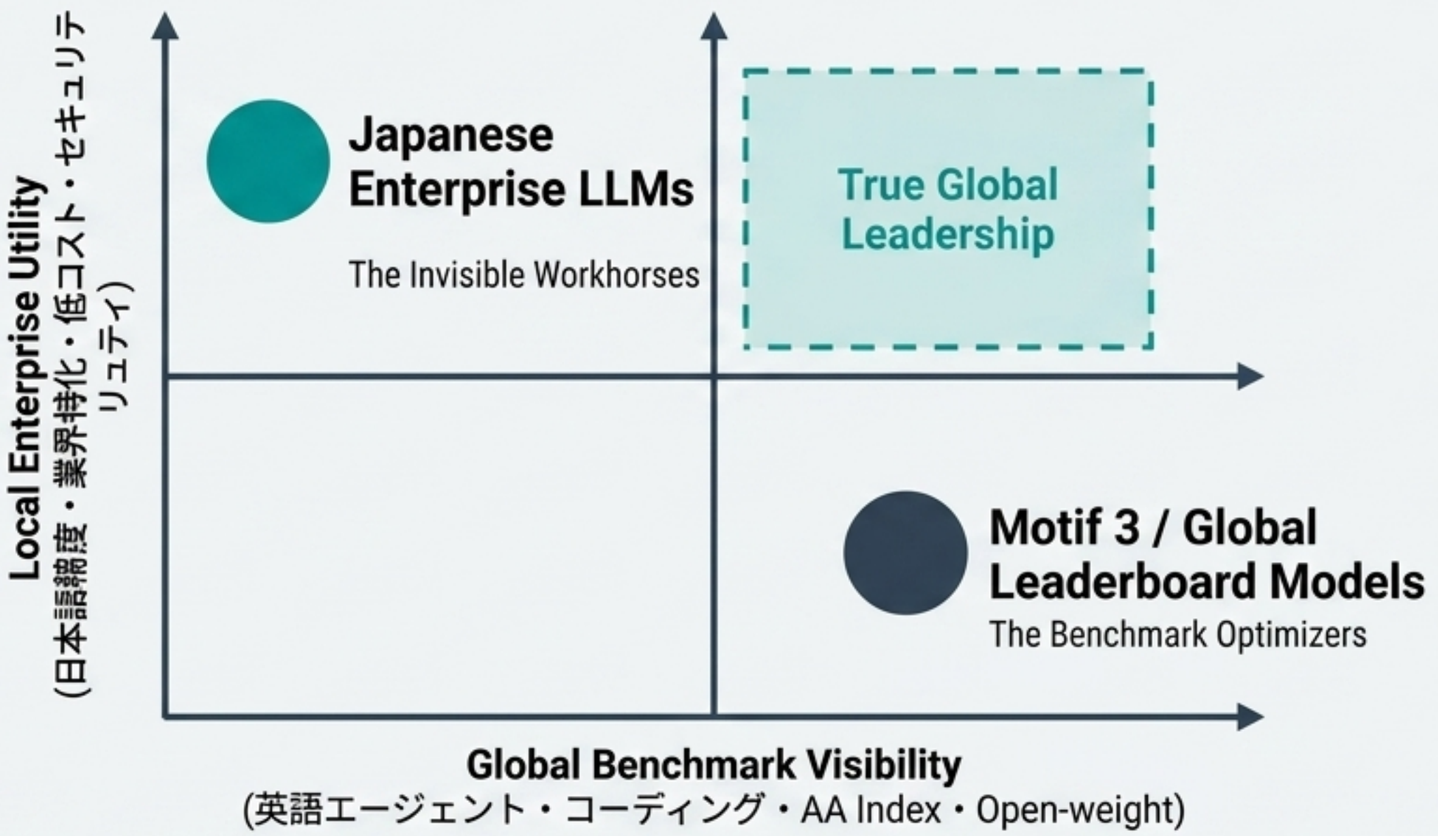
これは韓国の一企業（独立ラボ）の戦略的成功であり、韓国の全モデルが日本を凌駕しているという証明にはならない。

[STEP 3: THE STRATEGIC REALITY]

最適化の違い

評価サイトが同一条件で比較できるモデル（グローバル可視性）と、国内の商用環境で最適化されたモデル（ローカル実用性）を単一の物差しで測ることは無意味である。

戦略的バランシング：次世代AI競争の座標軸



韓国の次の課題は「高得点の実務的・文化的適合」への変換。
日本の課題は「実用モデルの国際的ベンチマークへの接続」である。

日本のAI戦略へのインペラティブ（必須課題）

01

[グローバル・リーダーボードへの意図的参加]

国内用途（API・軽量化）の開発と並行し、**国際比較可能な英語・コード・エージェント性能**を持つモデルを意図的にパブリッシュする。透明性のある公開モデルカードの整備が急務。

02

[評価可能性（Evaluability）の確保]

AIモデルは「存在する」だけでは**世界から認知されない。第三者の評価サイト**（AA等）が同一条件で実行・比較できるインターフェースとアーキテクチャを戦略的に提供する。

03

[「見えない価値」の指標化]

日本モデルの強みである「**低コスト運用**」「**特定業務の堅牢性**」「**長期エージェントの信頼性**」を、独自のベンチマークから国際的なスタンダード指標へと昇華させる発信力を持つ。

私たちは能力で負けているのではない。
グローバルな『評価のゲーム』の戦い方をアップデートする時が来ている。