Motif 3 はなぜ 47 点なのか

Artificial Analysis の独立評価、韓国の開発環境、日本 LLM が「見えにくい」理由の検証

作成日：2026 年 8 月 17 日
作成者：Manus AI

要旨

ご提示の認識は概ね正しいですが、二点を正確化する必要があります。Motif 3 の Artificial Analysis Intelligence Index は **47 点**で、同社が定義する同等比較クラスでは **106 モデル中 7 位**、比較群の**中央値 27** を 20 点上回ります。したがって「平均を大きく上回る」より正確には、**中央値を約 74%上回る**水準です。[1] 一方、同サイトの Index ページで表示される現行最高値は 63 であり、Motif 3 は最上位陣に近い**第 1 グループの有力な open-weight モデル**と評価するのが妥当ですが、最先端の最高性能モデルと同格とまでは言えません。[1] [2]

Motif 3 が高評価を得た主因は、韓国語モデルであること自体ではなく、314B 総パラメータ・13.2B 活性化の大規模 MoE、約 12.5 兆トークンの事前学習、エージェント・コード・科学推論に特化した教師群からの蒸留、256K 文脈、そして MIT ライセンスの公開重みという組合せにあります。[3] [4] とりわけ Artificial Analysis の Index は英語テキストのエージェント実行とコーディングに 58%の重みを置くため、Motif 3 の訓練目標と評価設計がよく整合しています。[2]

日本の LLM がこの画面で目立たないことは、**日本の LLM に能力がないことの証拠ではありません**。Artificial Analysis の総合 Index は日本語性能を測るランキングではなく、英語・テキスト中心の実行型評価です。日本には、日本語の言語・文化・業務タスクを測る公開 leaderboard が別に存在し、PLaMo、llm-jp、ELYZA、ABEJA 等の日本発・日本語調整モデルが評価対象になっています。[8] [9] 観測される差は、能力の国籍比較というより、**評価ターゲット、公開・提供形態、英語でのエージェント／コード能力の実測、国際的な評価可能性**の差を強く含みます。

> ***結論：****Motif 3 の 47 点は十分に重要な成果です。しかしそれは「韓国語が強いから」でも「日本に LLM がないから」でもなく、韓国の一社が、国際的な英語・実行型ベンチマークに強く合う大規模モデルを、評価可能な形で公開した成果と読むべきです。*

1. 47 点は何を意味し、何を意味しないか

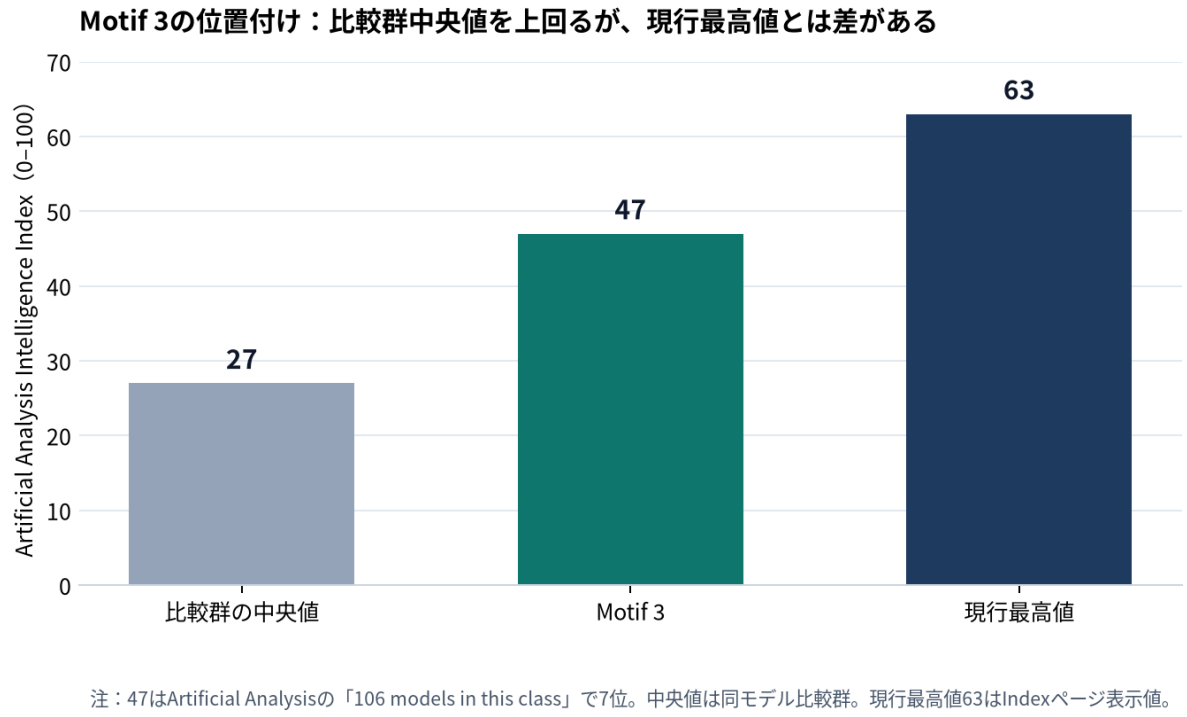

図 1：Artificial Analysis が表示する Motif 3 の 47 点、同等比較群中央値 27、Index ページ上の現行最高値 63。47 点は比較群で 7 位だが、最高値からは 16 点差がある。出典：1 2

Artificial Analysis の Motif 3 専用ページは、同モデルを **open weights** として掲載し、Intelligence Index を 47、比較クラス内の順位を 7/106、比較群中央値を 27 と表示しています。1 47−27 は 20 点であり、中央値比では約 1.74 倍です。ここで重要なのは、ページに書かれている基準が**mean（平均）ではなく median（中央値）**である点です。したがって、引用時には「中央値を 20 点上回る」とするのが最も忠実です。

ただし、この点数を一般的な「知能指数」や、日本語を含むあらゆる用途の総合順位と読むことはできません。現行の v4.1.1 は、GDPval-AA v2、τ³-Banking、Terminal-Bench v2.1、SciCode、AA-LCR、AA-Omniscience、HLE、GPQA Diamond、CritPt という 9 評価の加重平均です。[2] Artificial Analysis 自身が次のように明記しています。

> *“Artificial Analysis Intelligence Index is a text-only, English language evaluation suite.”* [2]

カテゴリ別の重みは、Agents 34%、Coding 24%、Scientific Reasoning 24%、General 18%です。[2] すなわち、エージェントとコーディングだけで 58%を占めます。日本語の敬語、国内固有の文書、商慣行、行政・法務・金融の制度知識、翻訳の精度といった、日本企業が重視しやすい能力は、この 47 点に直接は入っていません。

さらに、Motif 3 の同社ページは Index 評価時の出力トークンを 260M、比較群中央値を 100M と表示し、**very verbose** と注記しています。[1] これは不正を意味しません。むしろ高度な推論・エージェント評価では長い推論や試行が得点に寄与し得るという、利用コスト・速度とのトレードオフを示します。実務での採用判断は、47 点だけでなく、同じページのトークン量、推論速度、価格、実タスクの失敗モードを合わせて行うべきです。

観点	確認できる事実	解釈上の注意
独立スコア	47 点、106 モデル中 7 位、比較群中央値 27。[1]	「全世界の全モデルで 7 位」ではなく、Artificial Analysis の同一比較クラス内での順位である。
最高値との差	Index ページの現行最高値は 63、Motif 3 との差は 16 点。[2]	最上位モデル群に接近しているが、首位と同格ではない。
評価の中心	Agents 34%と Coding 24%で計 58%。[2]	日本語会話・日本国内業務の価値を完全には代表しない。
統計的な不確実性	Index 全体の 95%信頼区間は±1%未満と同社は推定するが、個別評価はより広い可能性がある。[2]	1 点程度の差の過大解釈は避けるべきである。

観点	確認できる事実	解釈上の注意
推論量	評価時の出力トークンは 260M、比較群中央値 100M。[1]	品質だけでなく、推論予算・レイテンシ・実行コストとの交換関係がある。

2. Motif 3 の実像：なぜこの評価軸に強いのか

Motif 3 は、Motif Technologies が公開した decoder-only MoE です。総パラメータ 314B に対し、各トークンで活性化するのは 13.2B で、各 sparse MoE 層では 384 experts のうち 8 を選択します。[3] 大きな専門家プールを持ちながら、トークン当たりの計算量を抑える設計です。重みは Hugging Face で MIT ライセンスとして公開され、英語・韓国語・多言語・コードを対象とするモデルとして掲載されています。[4]

技術報告によれば、事前学習は Web、STEM、コード、数学、合成 QA、韓国語・多言語、法務・金融を含む約 12.5 兆トークンです。このうち、NVIDIA Nemotron の公開事前学習データ群を加工した部分が約 70%を占め、残りの多くを社内で収集・処理したと報告しています。[3] 単に韓国語コーパスを増やしたのではなく、**英語の技術・コード・推論データを大きく取り込んだ多言語・専門領域型の訓練**です。

ポストトレーニングも、Index と整合的です。一般 SFT の後に、RL で訓練した六つの専門教師と、SFT で訓練したソフトウェア工学教師を用意し、Multi-teacher On-Policy Distillation で一つの学生モデルに統合しています。[3] 開発者が掲げる専門領域は、エージェント型ツール利用、プロフェッショナル業務、ソフトウェア工学、長文脈推論、棄権の校正、数学、コード・科学、チャットです。これは AA Index の高配点領域とかなり直接的に重なります。

開発者の Table 6 では、Motif 3 は τ^3-Banking 35.3、Terminal-Bench 2.1 74.9、SWE-bench Verified 76.2、GPQA Diamond 83.4、AA-LCR 72.3 を報告しています。[3] 特にツール利用と端末型タスクに強いという開発側の説明は、独立 Index の構成とも方向が一致します。ただし、SciCode 40.6 と CritPt 6.6 は比較表の最強モデルを下回り、ITBench-AA 51.5 には public subset のみという注記があります。[3] これらは、Motif 3 が全領域で最強なのではなく、**エージェント実行・端末操作を明確な強みとする広能力モデル**であることを示します。

性能を支える要素	Motif 3 の実装・方針	47 点との関係
スケーリング	314B 総量、13.2B 活性化の fine-grained MoE。[3]	モデル容量を持ちつつ、推論計算を比較的抑えられる。
訓練データ	約 12.5 兆トークン。Web だけでなく STEM、コード、数学、推論、専門領域、多言語を混合。[3]	科学・コード・知識系の評価への基礎体力を形成する。
長文脈・システム	256K 文脈、GDLA、低精度計算・通信、文脈並列化。[3]	長い入力とエージェント軌跡を扱う能力・効率に寄与し得る。
ポストトレーニング	専門教師群と on-policy 蒸留で、ツール利用・SWE・推論等を統合。[3]	Agents/Coding を重く見る Index に特に整合する。
国際的な可視性	技術報告、モデルカード、公開重み、独立評価ページがそろう。[1][3][4]	第三者がモデルの存在・仕様・評価を発見し、再利用しやすい。

一方で、技術報告自体が限界も明示しています。実世界の全タスク・全言語・全相互作用を覆うわけではなく、モデルは主にテキストで、視覚入力の直接理解には向きません。また、より長いエージェント軌跡に必要な状態管理、計画、回復、環境相互作用も十分には解けていないとしています。[3] 47 点は強い証拠ですが、AGI 達成の証明ではありません。

3. 「韓国の LLM が強い」ことと、Motif 3 が強いことは同義ではない

韓国には、国際評価に出やすいモデルを生む環境が確かにあります。MSIT の Sovereign AI Foundation Model 事業は、海外モデルへの技術・文化・経済・安全保障上の依存を減らし、AI G3 を目指す施策として位置付けられています。[5] 2025 年に選定された当初の 5 チームは NAVER Cloud、Upstage、SK Telecom、NC AI、LG AI Research Institute でした。[5] 同事業は国際ベンチマーク、専門家評価、ユーザー評価を組み合わせ、モデルの規模、コスト効率、社会実装も見る制度設計です。[5]

この事実は、韓国で GPU・データ・人材・国際ベンチマークを結び付ける政策的な競争環境が強いことを裏付けます。しかし、**Motif Technologies はこの当初 5 チーム**

ではありません。MSIT 資料は Motif-2-12.7B を Artificial Analysis 掲載モデルとして別途挙げていますが、Sovereign AI 事業の採択成果としては扱っていません。[5] よって、Motif 3 の 47 点を「韓国政府プロジェクトが直接作った点数」と説明するのは不正確です。より慎重には、韓国に国際展開・大規模訓練・公開モデルを促す環境があり、その中で Motif のような独立ラボも可視化されやすい、と述べるべきです。

韓国の他社も、国際評価と公開戦略を意識しています。LG AI Research の K-EXAONE、Upstage の Solar 系列、SK Telecom の A.X 系列は、それぞれ open-weight または国際ベンチマークに接続する形で発信されています。実際、Artificial Analysis の日本語 Multilingual Index の表示候補にも K-EXAONE が含まれます。[6] これは「韓国語だけ」の競争ではなく、英語・コード・エージェント・多言語で国際比較できるモデルを出す競争です。

4. 日本 LLM が「いない」ように見える理由

日本にも国家支援とモデル開発基盤はあります。経済産業省・NEDO の GENIAC は、基盤モデルの開発に必要な計算資源、データセット、知識共有を支援し、アプリ開発者・ユーザー企業・VC/CVC との連携まで含めています。[6] 2026 年 8 月時点の参加主体には、ABEJA、ELYZA、Preferred Elements、楽天、Sakana AI、NII、松尾・岩澤研究室などが含まれています。[6] また METI は 2024 年、AI 開発に必要な計算資源を増強する 5 件の GPU クラウド事業に最大 725 億円の補助を決定しています。[7]

したがって「日本は基盤モデルに投資していない」「日本には担い手がいない」は事実ではありません。より本質的な違いは、国際総合ベンチマークから見た場合の**製品設計と露出経路**です。日本の代表的な取り組みには、軽量・省電力・企業導入、日本語や国内業務の高精度、特定業界向け、API・オンプレミス中心、研究用の小中規模 open-weight など、異なる最適化目標があります。こうした選択は合理的ですが、314B 級の英語中心・エージェント／コード型の実行評価に、そのまま最大化されるとは限りません。

日本語性能には、別の厳密な評価経路があります。Hugging Face と LLM-jp による Open Japanese LLM Leaderboard は 20 超のデータセットを基に、NLI、多段 QA、要

約、数学、コード生成、JMMLU、金融文書の感情分析などを評価します。[8] Swallow LLM Leaderboard は、日本語の JamC-QA、M-IFEval-Ja、JHumanEval に加え、英語の MMLU-Pro、GPQA、AIME、LiveCodeBench も併記し、日本語・英語を分けて測定しています。[9] これらの board には、PLaMo、llm-jp、ELYZA、ABEJA、Sarashina 等の日本関連モデルが掲載されています。[9]

見え方を左右する要因	Motif 3／韓国側で観察できること	日本側で観察できること	何を意味するか
総合 Index の言語	AA Index は英語テキスト中心。[2]	日本語固有の評価は LLM-jp・Swallow 等で別立て。[8][9]	AA Index 非掲載・低順位だけで日本語実用性を判断できない。
高配点能力	Agents+Coding が 58%。[2] Motif 3 はツール利用・端末型課題を主要な訓練・評価対象にする。[3]	多くの日本モデルは日本語業務、軽量性、特定産業、導入形態も重要視する。GENIAC も社会実装を重視する。[6]	同じ「LLM」でも最適化している目的関数が異なる。
評価可能性	公開重み、技術報告、明確なモデルカード、AA 個別ページがある。[1][3][4]	商用 API、企業向け導入、オンプレミス、研究限定など提供形態が多様である。	評価サイトが同一条件で実行・比較できるモデルほど国際表で見えやすい。
資源・政策	韓国は主権 AI を掲げ、国際ベンチマークと GPU・データ・人材を結ぶ競争枠を運用する。[5]	日本も GENIAC、GPU クラウド支援、研究・産業の広い参加者を持つ。[6][7]	「韓国だけ支援がある」という説明は誤り。集中競争と分散的社会実装の発信差はあり得る。
国籍と能力の区別	Motif 3 は Motif Technologies の一モデルである。[3][4]	日本は複数企業・大学・コンソーシアムが別のモデル目的を持つ。[6][8][9]	一社・一モデルのベンチマークを国家全体の能力に一般化してはならない。

ここで、個別の日本モデルが Artificial Analysis に載らない理由を「API がないから」「申請していないから」「性能が低いから」と断定することも避けるべきです。今回確認した公開資料は、AA Index が英語中心で、モデルの評価可能性と実行

型能力が重要であることを示しますが、各日本モデルの非掲載理由を AA が一覧で公表しているわけではありません。非掲載の理由は、モデルの提供可否、レート制限、評価実行コスト、リリース時期、比較対象の選択、提供者との連携など、モデルごとに異なり得ます。

5. 総合判断

Motif 3 は、少なくとも Artificial Analysis v4.1.1 が測る英語テキストのエージェント、端末タスク、コード、科学推論、長文脈で、世界上位の open-weight モデルです。47 点・7/106・中央値 27 という独立評価と、技術報告にあるモデル規模・訓練・ポストトレーニングの方向性は整合します。[1] [2] [3] その意味で、「最先端モデル群に入る」という評価は**条件付きで妥当**です。

ただし、より正確な表現は次の通りです。

> ***Motif 3 は、英語中心でエージェント／コーディングの比重が高い Artificial Analysis Intelligence Index において、比較群中央値を大幅に上回り、同一クラスで 7 位に入る、世界上位の公開重みモデルである。これは韓国の LLM 開発力の重要な成果だが、韓国全体の優越や日本 LLM 全体の劣位を証明するものではない。***

日本の課題を挙げるなら、研究・企業用途として優れた日本語モデルを作ることと並行して、国際的に比較可能な**英語・コード・エージェント性能、再現可能な評価、公開モデルカード、継続的なグローバル leaderboard 参加**をどう作るかです。他方、韓国側にとっても、英語 Index の高得点を日本語・韓国語の文化的適合性、低コスト運用、企業の安全な導入、長期エージェントの信頼性へ変換できるかが次の競争になります。

参考文献