SPECIAL REPORT

Motif 3 徹底調査

韓国 LLM が高評価を得た要因と
日本 LLM が Artificial Analysis で見えない理由

Artificial Analysis Intelligence Index 47 点の意味、技術構成、国家政策、公開・評価戦略の比較分析

要旨

Motif 3 は大型オープンウェイト LLM の先頭集団に入ったが、絶対的な世界首位ではない。高評価の核心は韓国語性能ではなく、英語のエージェント実行・コーディング・科学推論を重視する評価軸に、モデル設計と事後学習を正面から合わせた点にある。日本勢の不在は「モデルがない」ことを意味せず、評価軸・公開形態・政策設計の違いと、公開された汎用フロンティア能力の実質差が重なった結果である。

調査基準日

2026 年 8 月 17 日

GPT-5.6 Sol

公開情報に基づく調査・分析

エグゼクティブ・サマリー

Motif 3 は、Artificial Analysis Intelligence Index（以下「AAII」）で 47 点を記録し、同サイトの「大型・オープンウェイト」比較区分では 106 モデル中 7 位、区分中央値 27 を 20 ポイント上回った。全掲載モデルを横断した表示では 608 モデル中 28 位であり、公開モデルとして国際的な先頭集団に入ったと評価できる。ただし、同時点の主要オープンウェイトモデルには 53 点級も存在するため、「世界で最も高性能」ではなく、「フロンティア級に接近した上位オープンウェイトモデル」と表現するのが正確である。［1,6］

47 点は韓国語能力の評価ではない。AAII v4.1.1 は、テキストのみ・英語のみで実施され、エージェント 34%、コーディング 24%、科学的推論 24%、一般能力 18%で構成される。したがって、Motif 3 が高得点を得た直接の理由は、韓国語が優れていることよりも、ツール利用、端末操作、ソフトウェア修正、長期タスク遂行、科学推論といった国際共通能力が高いことである。［2,3］

Motif 3 の技術的な核は、総 314B・1 トークン当たり約 13.2B が動作する細粒度 MoE、384 のルーティング専門家から 8 専門家を選択する構造、262,144 トークンの長文対応、約 12.5 兆トークンの事前学習、そして 6 つの強化学習教師と 1 つのソフトウェア工学教師を統合する Multi-teacher On-Policy Distillation にある。特に、エージェント、専門業務、コード、長文、数学、科学、対話を別々の教師に専門化させた後、単一モデルへ統合する事後学習が、高い AAII スコアと整合する。［4,5］

韓国の浮上は Motif 3 だけの偶発的成功ではない。Artificial Analysis には、Solar Open2 250B が 37、A.X-K2 が 35、K-EXAONE 2.0 が 31 として掲載されている。比較区分やパラメータ規模が異なるため単純な横並びは禁物だが、複数の韓国モデルが国際評価に同時に現れた点は、モデル開発、計算基盤、公開、政策評価が一つの方向に揃ってきたことを示す。［7,8,9］

日本にも PLaMo 3.0 Prime、tsuzumi 2、LLM-jp など有力な国産 LLM・研究基盤が存在する。日本勢が Artificial Analysis で見えにくい主因は、英語・エージェント中心の評価軸との不一致、企業向け API・オンプレミス中心で独立評価しにくい提供形態、小型・省計算・日本語業務特化を優先する製品戦略、GENIAC の多様なテーマを広く育てる政策設計にある。一方、公表ベースでは、Motif 3 のような大規模 MoE、10 兆トークン超、専門教師型 RL、緩いライセンスのウェイト公開、国際独立評価 40 点台を同時に満たす日本モデルが確認できないという実質的な差も残る。［14-22］

> **本レポートの最終判断**
>
> 日本は国産 LLM を開発しているが、韓国は「国産 LLM を世界の公開フロンティア競争に載せる」制度設計を強め、その成果が Motif 3 として可視化された。日本の課題は、国内特化路線を捨てることではなく、国内実装型とは別に、世界標準の汎用エージェントモデルを狙う集中投資トラックを設けることである。

目次

1. 調査目的・方法と評価上の留意点

1.1 調査目的

本レポートは、Motif 3 の AAII 47 点が何を意味するのかを技術面・評価面・政策面から検証し、韓国の LLM が国際評価で急速に存在感を高めた理由と、日本の国産 LLM が同じ画面に現れにくい理由を明らかにすることを目的とする。単なるモデル性能比較ではなく、学習データ、事後学習、分散基盤、公開ライセンス、独立評価への接続性、国家プロジェクトの設計までを一体として分析した。

1.2 情報源と検証方針

調査基準日は 2026 年 8 月 17 日である。性能の中心指標には Artificial Analysis による独立評価を用い、技術仕様には Motif Technologies の技術報告および Hugging Face モデルカードを用いた。韓国政策については科学技術情報通信部（MSIT）、日本については Preferred Networks、NTT、国立情報学研究所の LLM-jp、経済産業省 GENIAC の一次資料を優先した。［1-5,10,11,14-21］

重要な区別として、AAII 47 点は Artificial Analysis が独立に測定した値である。一方、Terminal-Bench、SWE-bench、SciCode 等の詳細値は Motif 3 技術報告に掲載された開発元報告値であり、比較対象モデルの値も必ずしも同一条件で Artificial Analysis が再実行したものではない。本文では両者を混同せず、「独立評価」と「開発元報告値」を区別して記載する。［1,4］

> **評価上の注意**
> ベンチマークはモデルの一側面を測るものであり、文章品質、創造性、韓国語・日本語能力、マルチモーダル能力、安全性、実運用コスト、再現性を一つの数値で表すものではない。特に AAII は英語・テキスト限定であるため、国語能力や国内業務適合性の総合ランキングではない。

2. Motif 3 の現在地：47 点をどう位置づけるか

2.1 主要仕様と独立評価

項目	内容
開発元	Motif Technologies
公開時期	2026 年 8 月 12 日
モデル種別	テキスト入出力の推論モデル／オープンウェイト
ライセンス	MIT
総パラメータ	約 314B
アクティブパラメータ	約 13.2B／トークン
コンテキスト長	262,144 トークン（256K）
AAII	47 点

項目	内容
同一比較区分	大型オープンウェイトモデル：7 位／106 モデル、中央値 27
全掲載モデル横断表示	28 位／608 モデル
AAII 評価時の出力トークン	260M、同区分中央値 100M

出所：Artificial Analysis および Motif Technologies モデルカードを基に作成。［1,5］

47 点は、大型オープンウェイトモデルの中で明確に上位であり、「韓国発として優秀」という限定付き評価ではない。公開された大型モデルの世界的な競争に参加できる水準である。ただし、Artificial Analysis のオープンウェイト比較では 53 点級のモデルも存在するため、絶対的な首位ではない。Motif 3 は「最先端モデルの一群に入った」という表現は妥当だが、「世界最高性能」と断定するのは過大評価となる。［1,6］

2.2 「同等のモデル」の正確な意味

Artificial Analysis の比較区分は、総パラメータ数により Tiny（4B 以下）、Small（4B 超 40B 以下）、Medium（40B 超 150B 以下）、Large（150B 超）に分類される。Motif 3 は総 314B のため Large に入り、同じ大型オープンウェイト区分の 106 モデルと比較されている。13.2B というアクティブパラメータ、推論コスト、学習費用、トークン消費量が同等という意味ではない。［1,6］

この点は重要である。Motif 3 は 314B の容量を持ちながら、各トークンでは 13.2B 程度だけを動かす MoE である。総パラメータ分類では大型だが、1 トークン当たり計算量は密な 314B モデルより大幅に小さい。他方、複数 GPU 上に全専門家を保持する必要があり、メモリ・通信・運用は「13.2B の小型モデル」と同じではない。［4,5］

2.3 高得点と推論効率は分けて見る

Motif 3 は AAII 評価全体で 260M 出力トークンを使用し、同区分中央値 100M の約 2.6 倍であった。これは、長く考えることで高得点を得る推論モデルであることを示す。性能の高さは事実だが、レイテンシ、電力、GPU 占有時間、エージェント実行コストの観点では、短い推論で同等性能を出すモデルより不利になる可能性がある。［1］

Artificial Analysis ページの入力・出力価格は 0 ドルと表示されるが、一般向け有料 API の価格が実測されたことを意味しない。モデルは自己ホスト可能であり、Hugging Face のデプロイ例には H200 または B200 を 8 基用いる構成が示される。したがって「無料モデル」ではなく、「ウェイトは公開されているが、運用計算資源は別途必要なモデル」と理解すべきである。［1,5］

3. Artificial Analysis Intelligence Index の構造

3.1 47 点は韓国語スコアではない

AAII v4.1.1 は、テキストのみ・英語のみの評価スイートである。画像入力、音声入力、多言語性能は別の指標で評価される。このため、Motif 3 の 47 点から「韓国語 LLM が英語圏モデルを凌駕した」と直ちに結論づけるのは不正確である。より正確には、「韓国の開発組織が、英語の国際共通タスクで上位に入る汎用モデルを構築した」という意味である。［2］

3.2 評価構成はエージェントとコードを重視

区分	重み	主な評価	測る能力
エージェント	34%	GDPval-AA v2、τ^3-Banking	実務成果物、ツール利用、銀行業務フロー
コーディング	24%	Terminal-Bench v2.1、SciCode	端末操作、ソフトウェア・科学コード
科学的推論	24%	HLE、GPQA Diamond、CritPt	広範な専門知識、大学院科学、物理推論

区分	重み	主な評価	測る能力
一般能力	18%	AA-LCR、AA-Omniscience	長文推論、知識正答・非幻覚

出所：Artificial Analysis Intelligence Benchmarking Methodology。［2］

エージェント 34%とコーディング 24%を合計すると 58%になる。従来型の知識問題や数学問題だけでなく、外部環境を操作し、複数段階のタスクを完了する能力が指数の過半を占める。2026 年 6 月の v4.1 改訂自体が、よりエージェント型の実務負荷へ重点を移すことを明示している。［2,3］

Motif 3 の事後学習は、エージェント型ツール利用、専門業務成果物、ソフトウェア工学、長文と棄権判断、数学、コード・科学、対話の七つの専門教師を統合する。この構成は、AAII の評価配分に偶然ではなく構造的に適合している。［4］

3.3 AAII を過大解釈してはいけない理由

- 日本語・韓国語の自然さ、敬語、専門用語、文化的文脈は AAII 47 点に直接反映されない。［2］
- Motif 3 はテキスト専用であり、画像・音声・動画を扱うマルチモーダル能力は AAII から判断できない。［1］
- 出力トークン量が多いため、性能と費用対効果・応答時間は別軸で評価する必要がある。［1］
- ベンチマークへの適合が実務価値と重なる部分は大きいが、特定業界の正確性、セキュリティ、可用性、保守性を保証しない。

4. Motif 3 の技術的構成

4.1 細粒度 MoE：314B の容量と 13.2B の実行

Motif 3 はデコーダ専用の自己回帰型 MoE で、53 層（2 密層・51MoE 層）から成る。各疎な MoE 層には 384 のルーティング専門家と 1 つの共有専門家があり、各トークンについて 8 専門家が選択される。大量の専門家容量を持たせつつ、毎回すべてを計算しないことで、総容量とトークン当たり計算量を分離している。［4］

構成要素	Motif 3
総／アクティブ	約 314B／約 13.2B
Transformer 層	53 層（2 dense＋51 MoE）
専門家	384 routed、top-8、shared 1
注意機構	Grouped Differential Latent Attention（GDLA）
残差構造	修正版 manifold-constrained hyper-connections（mHC）
活性化	Expert-Specific PolyNorm
予測補助	1 層の Multi-Token Prediction head
最大文脈	262,144 トークン

出所：Motif 3 Technical Report, Table 1。［4］

4.2 GDLA：長文推論と KV キャッシュ効率

GDLA は、不要・冗長な注意を差し引く Differential Attention の考え方と、Key-Value 状態を低次元の潜在表現へ圧縮する Multi-head Latent Attention を組み合わせる。信号経路を多く、ノイズ推定経路を少なく割り当て、入力依存の係数でノイズ成分を抑制する設計である。長文推論時の KV キャッシュ負担を抑えながら、重要情報への選択性を高めることを狙う。［4］

4.3 mHC、Expert-Specific PolyNorm、MTP

修正版 mHC は、通常の単一残差ストリームを 4 本の並列ストリームへ拡張し、トークンごとに情報を選択・再配分する。深いモデルで表現力を高めつつ、学習後半に残差増幅が蓄積しないようスケールを調整する。［4］

Expert-Specific PolyNorm は、全専門家に同じ活性化関数を適用するのではなく、専門家ごとに異なる多項式係数を学習させる。異なる入力分布を受け持つ専門家の役割分化を促しつつ、正規化と係数制約で活性化の暴走を抑える。MTP は次の 1 トークンだけでなく複数トークン予測を補助目標として使い、自己投機的デコードにも利用できる。［4］

4.4 モデル研究とシステム工学の一体化

Motif 3 の報告は、モデル構造だけでなく、Expert Parallelism、低精度演算、通信最適化、融合カーネル、長文コンテキスト並列化、強化学習の非同期化を広く扱う。Moreh の MoAI 基盤も、数千 GPU クラスタの自動並列化、仮想化、動的割当を掲げてきた。Motif 3 の競争力は、アルゴリズム研究だけでなく、学習を止めずに回す分散システム能力を垂直統合した点にある。［4,13］

5. 事前学習・データ・事後学習

5.1 約 12.5 兆トークンと動的データ混合

Motif 3 は約 12.5 兆トークンで事前学習された。一般 Web 文書を最大成分とし、STEM、ソースコード、数学、合成 QA、多言語、推論重視データ、韓国語、法律・金融等を組み合わせる。公開された NVIDIA Nemotron 事前学習コレクションを追加フィルタしたデータが全コーパスの約 70%を占める。したがって、完全に韓国独自データだけで作られたモデルではないが、公開データを自ら選別し、独自トークナイザーと学習設計でゼロから重みを形成したモデルである。［4］

データ混合は固定ではなく、初期 19 データセットから 57 データセットへ拡張し、学習の進行に伴って一般データ中心から STEM、数学、コード、合成 QA、推論データの比率を高める。巨大なデータ量だけでなく、「どの段階で何を学習させるか」を制御するカリキュラム設計が重視されている。［4］

5.2 七つの専門教師を統合する事後学習

専門教師	担当能力
Agentic tool use	シェル・ツール環境、複数段階タスク
Professional work	表計算、スライド、報告書などの職業成果物
Software engineering	リポジトリ単位の修正とテスト実行
Long-context & abstention	超長文検索・統合、適切な棄権
Mathematics	検証可能な数学推論
Code & science	コード生成・科学的問題解決
Chat	対話品質、指示追従、安全性、韓国語品質

出所：Motif 3 Technical Report, Table 5 および関連記述。［4］

事後学習は、一般 SFT、6 つの GRPO 強化学習教師、SFT で作るソフトウェア工学教師、そして Multi-teacher On-Policy Distillation（MOPD）の順で構成される。一つのモデルを多数の報酬で同時に学習させるのではなく、能力ごとに専門教師を作り、最後に一つの学生モデルへ蒸留することで、報酬間の干渉を抑えながら能力を統合する。［4］

Professional work 教師は、隔離コンテナ内で表計算、スライド、報告書を作成・修正し、最大 100 回の対話ターンと 45 分の実行時間を使う長期課題で学習される。これは、単問正解ではなく、ファイル成果物を完成させる能力を直接訓練していることを示す。［4］

5.3 韓国語品質を鍛えながら英語指示追従も改善

韓国語対話教師では、不適切な英語混入、外国文字混在、指定言語との不一致、反復・退化、質問への未回答、脱線、低努力・過度に短い回答、非丁寧な韓国語という八つの反パターンを厳格に判定する。開発元は、この訓練が韓国語中心でありながら、韓国語と英語の双方で指示追従を改善したと報告する。韓国語品質と国際ベンチマーク能力を別々の製品に分けず、一つのモデルへ統合した点が特徴である。［4］

6. ベンチマークから見た強みと弱み

6.1 開発元報告の主要スコア

評価	Motif 3	測る能力	解釈
GDPval-AA v2	38.7	実務型の知識労働	競争力あり
τ^3-Banking	35.3	銀行ツール利用	比較表中の最高値
Terminal-Bench 2.1	74.9	端末・長期コーディング	最上位級
SWE-bench Verified	76.2	実リポジトリ修正	高水準
SciCode	40.6	科学コード	最上位モデルに劣る
GPQA Diamond	83.4	大学院科学推論	高水準
HLE	37.0	広範な難問	競争力あり
CritPt	6.6	研究レベル物理	明確な弱点
AA-Omniscience Accuracy	30.1	知識正答率	首位級ではない
Non-Hallucination	71.6	誤答回避・棄権	高水準
AA-LCR	72.3	長文推論	高水準
IFBench	78.2	指示追従	安定

注：数値は Motif 3 技術報告の開発元報告値。比較対象の測定条件が完全に統一されているとは限らない。［4］

6.2 最大の強み：環境を操作して仕事を完了する能力

Motif 3 は、τ^3-Banking、Terminal-Bench、SWE-bench で特に強い。これは、知識を文章で回答するだけでなく、外部ツールを選び、環境状態を確認し、失敗を修正しながら複数ステップを完了する能力を示す。AAII がエージェントとコードに 58%を割り当てるため、この強みが総合 47 点へ大きく寄与したと考えられる。［2,4］

したがって、Motif 3 の評価は単なるベンチマーク暗記だけでは説明しにくい。銀行業務、端末操作、ソフトウェア修正、ファイル成果物作成は、AI エージェントの経済価値と比較的近い。ただし、実企業環境では権限管理、監査ログ、データ機密性、再試行コスト、例外処理が加わるため、ベンチマーク性能をそのまま導入成果へ換算することはできない。

6.3 弱点：科学コード、専門物理、知識正答率

SciCode 40.6、CritPt 6.6 は比較表中の強豪より低く、科学コード生成と高度な物理推論には改善余地がある。AA-Omniscience Accuracy 30.1 も最高値ではない。一方、Non-Hallucination 71.6 は高く、確信がない場合に無理に答えない傾向がある。Motif 3 は「何でも最もよく知っているモデル」より、「間違いを抑えながらツールを使って作業を進めるモデル」に近い。［4］

7. 韓国 LLM が高評価を得た構造的要因

7.1 Motif 3 だけではない韓国モデル群

モデル	開発元	AAII	区分内表示	主な仕様
Motif 3	Motif Technologies	47	Large：7／106	314B／13.2B active
Solar Open2 250B	Upstage	37	Large：25／106	250B
A.X-K2	SK Telecom	35	比較区分：32／348	262K context
K-EXAONE 2.0	LG AI Research	31	Large：45／106	750B／37B active

注：比較区分・モデル規模が異なるため、国別の単純ランキングではない。2026 年 8 月 17 日時点。［1,7,8,9］

複数の韓国モデルが同時期に Artificial Analysis へ掲載されていることは、Motif 3 が単独の例外ではないことを示す。各社が異なるアーキテクチャと規模を採用しつつ、オープンウェイト、長文、推論、国際ベンチマークを共通の競争軸としている。［7,8,9］

7.2 要因 1：少数精鋭の勝ち抜き型国家プロジェクト

韓国の Sovereign AI Foundation Model Project は、15 チームから書類審査で 10、プレゼン評価で 5 チームへ絞り、Naver Cloud、Upstage、SK Telecom、NC AI、LG AI Research を選抜した。その後も段階評価でチームを絞り、追加枠として Motif Technologies 主導コンソーシアムを選定した。支援を受け続けるために、各チームが定期的に世界水準の成果を示す競争構造である。［10-12］

7.3 要因 2：評価指標そのものが世界市場を向く

第 1 段階評価は、ベンチマーク 40 点、外部専門家 35 点、ユーザー25 点で構成された。グローバル共通評価では、エージェント、数学、知識・推論、指示追従などを含む 13 の国際ベンチマークを使用し、個別評価では各チームのモデルを対応する世界 SOTA モデルと比較した。専門家は技術報告と学習ログを審査し、ユーザー評価では実利用性と推論費用対効果を検証した。［11］

この制度では、韓国語ベンチマークだけで高得点を取っても勝てない。国際リーダーボードに載る能力、学習プロセスの独自性、実利用性を同時に求めるため、開発チームの目標が自然に Artificial Analysis のような世界標準へ寄る。

7.4 要因 3：「ソブリン」をゼロからの設計・学習と定義

MSIT はソブリン基盤モデルを、海外モデルのファインチューニングではなく、国内で設計し、第三者モデルのライセンス制約から自由な形でゼロから事前学習したモデルと定義した。公開データやオープンソース技術を利用しても、初期重みを引き継がず、自らアーキテクチャ、データ処理、学習を管理することが核心となる。［11］

この定義は、国内企業が単なるアプリケーション層やファインチューニングにとどまらず、基盤モデルの設計、分散学習、事後学習を自前で持つ圧力として働く。Motif 3 の独自アーキテクチャと MOPD は、この政策要求と整合する。［4,11,12］

7.5 要因 4：計算資源をクラスタ単位で集約

韓国政府は、別枠の GPU 調達事業で 13,000 基の先端 GPU（B200 10,080 基、H200 3,056 基）を確保し、政府利用分を大規模クラスタとして配分する計画を示した。これらすべてが Motif 3 学習に使われたわけではないが、国家プロジェクトや産学研究へ数百〜数千 GPU 単位で資源を供給する政策環境は、巨大 MoE と長期 RL を実行する土台になる。［23］

重要なのは国全体の GPU 総数ではなく、一つの有力チームがまとまったクラスタを一定期間継続利用できることである。LLM 学習では、GPU を多数の小規模テーマへ細分化すると、通信最適化、障害復旧、長期実験の学習効果が失われやすい。

7.6 要因 5：公開・実行・評価までを製品として設計

Motif 3 は MIT ライセンスでウェイトを公開し、Hugging Face から申請なしで取得でき、vLLM と SGLang の実行方法、ツール呼出し用設定、H200／B200 での具体的なデプロイ例を提示する。英語の技術報告も同時公開され、海外の評価者が短期間で再現・検証しやすい。性能だけでなく、「世界に測ってもらうための流通設計」が整っている。［4,5］

7.7 要因 6：分散システム企業の垂直統合

Moreh は、MoAI を通じて数千 GPU クラスタの自動並列化、GPU 仮想化、動的割当を訴求してきた。Motif 3 の技術報告でも、MoE 通信、低精度、長文並列、強化学習基盤までが主要テーマとなる。韓国の成功を「優秀な研究者が大きなモデルを作った」とだけ捉えるのではなく、モデル、ランタイム、クラスタ運用、公開基盤を一体化した結果と見るべきである。［4,13］

8. 日本の LLM は本当に「全く顔を出さない」のか

日本の LLM が存在しない、あるいは研究能力がないという認識は誤りである。日本では、純国産の事前学習、企業向け運用、日本語業務、オンプレミス、研究基盤公開において複数の有力な取組が進んでいる。ただし、その評価目標と公開形態が、Artificial Analysis の主指標と一致していない。

8.1 PLaMo 3.0 Prime：日本語性能・コスト・エージェント対応

Preferred Networks は 2026 年 6 月 22 日、国内でフルスクラッチ開発する PLaMo 3.0 Prime を正式公開した。推論モデルと非推論モデルを用意し、コンテキストを 64K から 256K へ拡張し、強化学習を長期化し、構造化出力を追加した。提供は PLaMo Chat/API、Amazon Bedrock Marketplace、オンプレミス、Snowflake 等が中心である。［14,15］

PFN は、日本語の性能とコストの両立、指示追従、対話、ツール利用、医療、コード、安全性を強みとして報告する一方、Web 探索、長文、数学、STEM、日本法令等には改善余地があると開示している。これは国際汎用ランキングの絶対首位を狙うより、日本語の実用性能と利用費用の最適化を重視する戦略である。［15］

8.2 tsuzumi 2：単一 GPU、機密情報、業務文書

NTT の tsuzumi 2 は、2025 年 10 月に提供開始され、単一 GPU で推論できる軽量性、低コスト、オンプレミス・プライベートクラウドでの機密情報処理を主要価値とする。2026 年 5 月の更新では、日本語ビジネス文書に含まれる表、グラフ、チャートの理解を強化し、引き続き単一 GPU 環境を前面に出した。［16,17］

これは、巨大な国際ベンチマークで最高点を取ることよりも、企業が安全に導入し、既存システムへ組み込み、運用費を抑えることを優先した設計である。日本市場では合理的な選択だが、AAII の大型汎用モデル競争には映りにくい。

8.3 LLM-jp：モデルだけでなく共同研究基盤を公開

国立情報学研究所の大規模言語モデル研究開発センターが主宰する LLM-jp は、日本発 LLM 開発プロジェクトであり、モデル、事前学習コーパス、データ構築スクリプト、評価・学習ツールを公開している。目的には、オープンかつ日本語に強い大規模モデルの構築、組織横断の研究連携、成果物公開が明記される。［18］

LLM-jp の価値は、一つの商用モデルの点数だけではなく、日本国内で再利用可能なデータ・評価・学習基盤を作ることにある。2026 年 6 月には LLM-jp-4 Thinking の量子化版も公開されている。［18］

8.4 日本勢の位置づけ

取組	主な強み	Artificial Analysis 上の見えにくさ
PLaMo 3.0 Prime	日本語性能、低コスト、推論／非推論、256K、構造化出力	API・商用提供中心。AAII の公開評価に接続しにくい
tsuzumi 2	単一 GPU、オンプレミス、機密情報、日本語業務文書	小型・業務特化で大型汎用ランキングと目的が異なる
LLM-jp	モデル・コーパス・ツールの公開、研究共同基盤	研究エコシステムの価値が単一指数に表れにくい

出所：各開発元・研究プロジェクトの公表資料を基に作成。［14-18］

9. なぜ Artificial Analysis で日本勢が見えないのか

9.1 理由 1：評価軸と設計目標の不一致

AAII は英語のみで、エージェントとコーディングが 58%を占める。日本語の敬語、行政文書、法令、金融・医療用語、製造現場の文脈、オンプレミス安全性を改善しても、AAII には直接加点されにくい。日本モデルが国内業務で価値を持っていても、英語の端末・コード・科学課題を優先しなければ、同じランキングでは上位に出にくい。［2］

9.2 理由 2：未掲載は低性能の証明ではない

Artificial Analysis の日本語 Multilingual Index ページで評価済みと表示されるのは 110 モデル中 13 モデルに限られ、上位表示も海外モデルである。同指標自体も Global-MMLU-Lite による単一言語の一般推論を中心とする。日本の主要国産モデルが載っていないことは、少なくとも現在の公開ページでは評価対象が網羅的でないことを示す。［22］

独立評価機関がモデルを実行するには、安定した API、ウェイト、推論ランタイム、価格情報、モデル固有のプロンプト設定が必要である。企業向け契約やオンプレミス提供を中心とするモデルは、性能が低くなくても評価パイプラインへ入りにくい。したがって「未掲載」と「低スコア」は明確に区別すべきである。

9.3 理由 3：日本は小型・業務特化・省計算を重視

tsuzumi 2 は単一 GPU を、PLaMo 3.0 Prime は日本語性能とコストの両立を強調する。これらは、企業の機密性、導入費、応答速度、国内業務適合性を考えれば合理的である。しかし、大型 MoE、長期エージェント、科学推論を重視する AAII では、314B・12.5 兆トークン・専門教師型 RL を投入した Motif 3 の方が有利になる。［4,14-17］

9.4 理由 4：GENIAC は多様な社会実装を広く育てる

GENIAC は、国内の基盤モデル開発力を底上げし、企業の創意工夫を促し、社会実装と海外展開を支援するプロジェクトである。第 3 期には 24 社、第 4 期には AI 基盤モデル開発テーマ 16 件を採択している。領域特化、ロボット、製造、医療、業務エージェント等の多様なテーマを育てるポートフォリオ型である。［19-21］

韓国のように少数チームを世界 SOTA と比較し、段階的に脱落させる方式と比べ、日本は資源と成果の分散が起こりやすい。その代わり、産業別の多様な技術・企業を育て、失敗リスクを分散できる。両制度は目的が異なるが、国際リーダーボード上のフラッグシップを生む確率は、集中・勝ち抜き型の方が高い。［10,11,19-21］

9.5 理由 5：可視性だけでは説明できない実質差

日本勢の不在を「評価対象になっていないだけ」と説明し切ることもできない。2026 年 8 月 17 日時点の公表情報では、日本から、300B 超級の疎な MoE、10 兆トークン超の事前学習、複数の専門教師による大規模 RL、強い端末・ツール・ソフトウェア評価、MIT 等の緩いライセンスによるウェイト公開、独立 AAII 40 点台を同時に満たすモデルは確認できない。

したがって、日本勢が見えない理由は三層に分ける必要がある。第一に評価軸の不一致、第二に公開・接続性の不足、第三に公開された汎用フロンティア能力の実質的な差である。どれか一つだけを原因とする説明は不十分である。

> **整理**
>
> 未掲載 ≠ 低性能。しかし、未掲載であることだけを理由に能力差を否定することもできない。日本には国内業務で競争力のあるモデルが存在する一方、世界の公開エージェントモデル競争では、韓国勢が先に可視化された。

10. 韓国と日本の戦略比較

比較項目	韓国	日本
主要目標	世界水準のソブリン汎用モデル、国際可視性	国内基盤力、社会実装、業務特化、産業裾野
支援方式	少数精鋭、段階評価、脱落を伴う勝ち抜き	多数テーマを採択するポートフォリオ型
評価	国際 13 ベンチ、世界 SOTA 比較、学習ログ、ユーザー評価	テーマ別成果、社会実装、企業・研究者の多様性
計算資源	大規模クラスタを国家プロジェクトへ集中	複数事業者・研究テーマへ幅広く提供
モデル方向	大型 MoE、長文、推論、エージェント	小型・省計算、日本語・業界特化も重視
事後学習	専門教師型 RL、ツール・端末・成果物タスク	各社個別。日本語・安全・業務調整に強み
公開	オープンウェイト、英語報告、vLLM／SGLang	API、オンプレミス、企業契約、研究公開が混在
強み	世界リーダーボードで早く可視化、開発能力を集中	国内顧客適合、機密性、コスト、産業応用の多様性
主なリスク	集中投資の失敗、巨大計算費、ベンチマーク偏重	資源分散、フラッグシップ不在、国際認知不足

本表は公表資料に基づく分析整理。国・企業内のすべての取組を一括して評価するものではない。［4,10-21］

韓国方式が全面的に優れ、日本方式が誤っているわけではない。日本の企業導入では、小型、オンプレミス、業界知識、日本語文書、法令・商慣行への適合が重要であり、国内特化路線には明確な経済価値がある。一方、世界の基盤モデル技術へ影響を与え、研究者・開発者・投資・データを引き寄せるには、国際評価に耐える公開フラッグシップも必要である。

最適解は二者択一ではなく、「世界フロンティアを狙う少数集中トラック」と「日本の産業を支える多数の特化・省計算トラック」の二層構造である。

11. 日本への提言

1 **二層型の国家戦略へ移行** GENIAC の多様な社会実装支援を維持しつつ、1〜2 チームを選び、世界の公開汎用モデルを狙う別枠を設ける。

2 **GPU を数百〜千基単位で複数年配分** 単年度・小口の利用枠ではなく、長期事前学習と RL を反復できるクラスタ利用権を成果連動で付与する。［23］

3 **世界標準ベンチマークを継続条件に組み込む** AAII 相当、Terminal-Bench、SWE-bench、GDPval、τ^3、SciCode、HLE、GPQA、長期エージェント評価を公開マイルストーンにする。［2,3,11］

4 **事後学習基盤を共同資産化** 検証可能な RL 環境、端末コンテナ、業務成果物タスク、専門教師、報酬モデル、蒸留基盤を各社が重複開発せず共有する。 [4]

5 **国際的に実行できる公開形態を標準化** 英語技術報告、Hugging Face ウェイト、商用利用しやすいライセンス、vLLM／SGLang 対応、API、再現手順を支援条件とする。 [5]

6 **独立評価と費用・速度を同時公開** 開発元ベンチだけでなく、第三者評価、タスク当たり時間、出力トークン、GPU 要件、電力・費用を公表し、過度な長考依存を抑える。 [1-3]

7 **日本語・産業特化を世界的な差別化へ変換** 日本語だけで閉じず、特許、製造、材料、ロボット、医療、行政、金融の高品質環境を多言語エージェント評価として公開する。

8 **公開フラッグシップと国内安全運用を分離しない** 高性能モデルから小型・オンプレミス版へ蒸留し、世界評価と国内導入を同一技術系列で循環させる。

11.1 実行優先順位

期間	優先施策
0〜6 か月	独立評価の実施、主要モデルの API／ウェイト接続、共通エージェント評価環境の設計
6〜18 か月	少数集中チームの選抜、GPU 長期枠、専門教師・RL 基盤の共同整備
18〜36 か月	300B 級 MoE または同等能力の公開モデル、国際リーダーボード継続参加、産業版への蒸留

12. 結論

Motif 3 の AAII 47 点は、韓国語能力の高さを直接示す値ではなく、英語のエージェント、コーディング、科学推論、一般能力を組み合わせた独立評価で、世界の大型オープンウェイトモデル上位に入ったことを示す。47 点は区分中央値 27 を大きく上回り、先頭集団入りを裏付けるが、53 点級の上位モデルが存在するため絶対的な世界首位ではない。 [1,2,6]

高評価の主要因は、314B／13.2B active の細粒度 MoE、12.5 兆トークンの動的事前学習、七つの専門教師を統合する MOPD、端末・ツール・業務成果物を直接扱う長期エージェント学習、分散システム基盤、MIT ライセンスと国際標準ランタイムによる公開を一体化したことにある。 [4,5,13]

韓国は、少数精鋭の勝ち抜き、世界 SOTA との比較、学習ログ審査、ユーザー評価、クラスタ型計算資源、ゼロからのソブリン開発、オープンウェイト公開を同じ方向へ揃えた。Motif 3 はこの制度と技術蓄積が結びついた成果である。 [10-12,23]

日本にも PLaMo 3.0 Prime、tsuzumi 2、LLM-jp 等があり、国内業務、安全運用、省計算、研究基盤で重要な強みを持つ。しかし、英語・エージェント中心の評価との不一致、独立評価へ接続しにくい提供形態、資源分散型の政策、公開された大型汎用エージェント能力の差により、Artificial Analysis では存在感が見えにくい。 [14-22]

> **一文でまとめると**
>
> 日本は「国産 LLM を作る」段階には達している。韓国はさらに、「国産 LLM を世界の公開フロンティア競争で測らせ、勝たせる」段階へ進み始めている。日本は国内特化の強みを維持しながら、国際公開フラッグシップを狙う集中トラックを追加すべきである。

参考文献

本文中の［ ］内番号は、以下の参考文献番号に対応する。ウェブ資料の最終参照日は、特記のない限り 2026 年 8 月 17 日である。

［1］Artificial Analysis, “Motif 3 – Intelligence, Performance & Price Analysis,” 2026 年 8 月公開（2026 年 8 月 17 日参照）。
URL: https://artificialanalysis.ai/models/motif-3

［2］Artificial Analysis, “Artificial Analysis Intelligence Benchmarking Methodology: Artificial Analysis Intelligence Index v4.1.1,” 2026 年（2026 年 8 月 17 日参照）。
URL: https://artificialanalysis.ai/methodology/intelligence-benchmarking

［3］Artificial Analysis, “Announcing Artificial Analysis Intelligence Index v4.1: a shift toward agentic workloads,” 2026 年 6 月 15 日。
URL: https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-1

［4］Lim, Junghwan, et al., “Motif 3: Technical Report,” arXiv:2608.09119v1, 2026 年 8 月 10 日, DOI: 10.48550/arXiv.2608.09119。
URL: https://arxiv.org/abs/2608.09119

［5］Motif Technologies, “Motif-Technologies/Motif-3,” Hugging Face Model Card, MIT License（2026 年 8 月 17 日参照）。
URL: https://huggingface.co/Motif-Technologies/Motif-3

［6］Artificial Analysis, “Comparison of Open Source AI Models across Intelligence, Performance, Price, Context Window, and more,” 2026 年（2026 年 8 月 17 日参照）。
URL: https://artificialanalysis.ai/models/open-source

［7］Artificial Analysis, “Solar Open2 250B – Intelligence, Performance & Price Analysis,” 2026 年 8 月公開（2026 年 8 月 17 日参照）。
URL: https://artificialanalysis.ai/models/solar-open2-250b

［8］Artificial Analysis, “A.X-K2 – Intelligence, Performance & Price Analysis,” 2026 年 8 月公開（2026 年 8 月 17 日参照）。
URL: https://artificialanalysis.ai/models/a-x-k2

［9］Artificial Analysis, “K-EXAONE 2.0 – Intelligence, Performance & Price Analysis,” 2026 年 8 月公開（2026 年 8 月 17 日参照）。
URL: https://artificialanalysis.ai/models/k-exaone-2-0-0803

［10］Ministry of Science and ICT, Republic of Korea, “MSIT Selects Five Elite Teams for the ‘Sovereign AI Foundation Model’ Project,” 2025 年 8 月。
URL: https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&nttSeqNo=1152&sCode=eng

［11］Ministry of Science and ICT, Republic of Korea, “Phase 1 Evaluation Results Released for Sovereign AI Foundation Model Project,” 2026 年 2 月 10 日。
URL: https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1212&sCode=eng

［12］Kang, Yoon-seung, “Gov’t picks Motif team as new contender for homegrown AI model development,” Yonhap News Agency, 2026 年 2 月 20 日。
URL: https://en.yna.co.kr/view/AEN20260220009300320

［13］Moreh, “Introducing Motif: A High-Performance Open-Source Korean LLM by Moreh,” 2024 年 12 月 2 日。
URL: https://moreh.io/blog/introducing-motif-a-high-performance-open-source-korean-llm-by-moreh-241202/

［14］Preferred Networks, 「国産生成 AI 基盤モデル PLaMo 3.0 Prime を正式リリース」2026 年 6 月 22 日。
URL: https://www.preferred.jp/ja/news/pr20260622

［15］今村英明, 「PLaMo 3.0 Prime をリリースしました」Preferred Networks Tech Blog, 2026 年 6 月 22 日。
URL: https://tech.preferred.jp/ja/blog/plamo-3-0-prime-release/

［16］NTT, Inc., “NTT’s Next-Generation LLM ‘tsuzumi 2’ Now Available,” 2025 年 10 月 20 日。
URL: https://group.ntt/en/newsrelease/2025/10/20/251020a.html

［17］NTT, Inc., “NTT’s LLM tsuzumi 2 Updated,” 2026 年 5 月 19 日。
URL: https://group.ntt/en/newsrelease/2026/05/19/260519a.html

［18］LLM-jp／国立情報学研究所 大規模言語モデル研究開発センター, 公式ウェブサイト（2026 年 8 月 17 日参照）。
URL: https://llm-jp.nii.ac.jp/

［19］経済産業省,「GENIAC」公式ウェブサイト（2026 年 8 月 17 日参照）。

URL: https://www.meti.go.jp/policy/mono_info_service/geniac/index.html

［20］経済産業省,「『GENIAC』における計算資源の提供支援（第 4 期）において、AI 基盤モデル開発テーマ計 16 件を採択しました」2026 年 6 月 4 日。

URL: https://www.meti.go.jp/press/2026/06/20260604003/20260604003.html

［21］経済産業省,「生成 AI 基盤モデル開発 第 3 期採択事業者 中間報告会を開催」2026 年 4 月 16 日。

URL: https://www.meti.go.jp/policy/mono_info_service/geniac/geniac_magazine/interimreport_3.html

［22］Artificial Analysis, “Japanese Language – AI Models Benchmark | Multilingual LLM Performance,” 2026 年（2026 年 8 月 17 日参照）。

URL: https://artificialanalysis.ai/models/multilingual/japanese

［23］Ministry of Science and ICT, Republic of Korea, “GPU Procurement Project (First Supplementary Budget, KRW 1.46 Trillion) Selects Participants,” 2025 年。

URL: https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1155&pageIndex=&sCode=eng&searchOpt=ALL&searchTxt=

注記：本レポートは公開情報に基づく分析であり、各社の非公開学習条件、実際の計算資源、製品ロードマップを完全に反映するものではない。ベンチマーク値、モデル順位、提供条件は更新される可能性がある。