ChatGPT

Motif 3がAAII 47点を取った理由と、韓国LLMが日本LLMより国際評価で可視化される構造

エグゼクティブサマリー

2026年8月17日時点で、Motif Technologiesの **Motif 3** はArtificial Analysis Intelligence Index（AAII）v4.1.1で **47点**を記録している。これは韓国の「独自AIファウンデーションモデル」系モデルの中で、Upstage Solar Open2 250Bの37、SK Telecom A.X-K2の35、LG AI Research K-EXAONE 2.0の31を明確に上回り、Artificial Analysisの大型オープンウェイト群でもMiniMax-M3やDeepSeek V4 Proの45を上回る水準である。一方、Kimi K3の一部設定など、さらに高いモデルも存在するため、「世界最高」ではなく、**計算効率の高い大型MoEとして非常に強い frontier-adjacent model** と位置づけるのが正確である。Artificial Analysis自身もMotif 3を314B総パラメータ、13.2Bアクティブ、262kコンテキスト、AAII 47として掲載し、同クラス中央値27を大きく上回るとしている。 [1]

最も重要な結論は、**47点は単なる「韓国語能力の高さ」から生まれたものではない**、という点である。現行AAIIは英語・テキストのみの評価で、Agent系34%、Coding系24%、Scientific Reasoning系24%、General系18%という構成である。つまり58%がエージェント＋コーディングである。Motif 3は後学習で「agentic tool use」「professional work」「software engineering」「long-context reasoning」「math」「code/science」などを担当する専門教師を明示的に作り、Multi-teacher On-Policy Distillation（MOPD）で統合した。評価項目と開発目標の一致が非常に強い。これは不正なbenchmark targetingを意味しないが、**AAIIが測る能力空間とMotifのpost-training objectiveが高度に整合している**ことは47点の中心的説明である。Motif Technical Report（2026-08-10）およびArtificial Analysis Methodology（v4.1.1、2026-08時点）。 [2]

AAII 47はかなり正確に再現できる。Motifが公表する各AAIIサブテストを現行ウェイトで加重すると、

$$0.20(38.7) + 0.14(35.3) + 0.16(74.9) + 0.08(40.6) + 0.12(37.0) + 0.06(83.4) + 0.06(6.6) + 0.08(30.1) + 0.04(71.$$

となり、表示値 **47** に丸められる。最大の得点源はTerminal-Bench 2.1だけで約12.0点、GDPval-AA v2で7.7点、GPQA Diamondで5.0点である。したがってMotif 3の強さは、「百科事典的知識が突出している」というより、**端末操作・コード・ツール利用・長い推論・業務タスクを高確率で完遂する能力**にある。 [3]

技術面でもMotif 3は単純な既存モデルの微調整ではない。314B総パラメータのMoEでありながら推論時アクティブは13.2B、384 routed expertsから8 expertsを選択し、1 shared expertを併用する。53層、4096 hidden size、SuperBPE 220,160語彙、256K native context、GDLA、modified mHC、Expert-Specific PolyNorm、Multi-Token Prediction（MTP）を組み込む。約12.5兆トークンでpretrainされ、約70%は公開NVIDIA Nemotronデータ群を保持し、残りはMotif側の収集・処理データを中心とする。正確な言語別・ドメイン別比率は非開示である。 [4]

韓国と日本の差については、「韓国には良い言語データがあり、日本にはない」という説明は不十分である。日本のNII/LLM-jpは2026年4月に約12兆トークンのコーパスを整備し、政府・国会文書、公開Web、合成データ等を含むLLM-jp-4を公開している。日本にはGENIAC、NII、NTT、Preferred Networks、NEC、Fujitsu、SB Intuitionsなど相当な技術基盤がある。 [5]

違いはむしろ、**韓国が「世界共通のfrontier benchmarkで勝つ汎用・agenticモデル」を政策KPIとして明示し、大規模GPU・データ・競争的選抜・open-weight公開を一体化したのに対し、日本では、日本語性**

能、軽量化、オンプレミス、企業内データ、産業特化、Physical AIなどへ資源が分散していることにある。韓国政府自身がArtificial Analysisへの掲載を成果として政策発表で言及している一方、日本の主要モデルPLaMo 3.0 Prime、tsuzumi 2、LLM-jp-4などは2026年8月17日時点のArtificial Analysis検索で確認できない。これは日本に能力がないことではなく、**benchmark visibility、モデルアクセス形態、開発目標、リリース時期の差による強いselection effect**である。 [6]

結論を一文にすると、**Motif 3の47点は、MoEによる高いparameter efficiency、12.5T-token規模のpretraining、agentic/coding中心の専門教師RL、長コンテキスト訓練、open-weightによる評価可能性、そして韓国政府が世界ベンチマーク競争を政策的に強制したことが重なって生じたものであり、日本LLMの低い「見え方」は日本語能力の不足よりも、グローバル評価指標との目的関数・公開戦略の不一致による部分が大きい**、というのが本調査の判断である。 [7]

Motif 3の技術、学習、公開戦略

Motif 3の一次資料で最重要なのは、Motif Technologiesによる **“Motif 3: Technical Report”**（arXiv: 2608.09119、2026年8月10日）と公式Hugging Face model cardである。Technical Reportはモデル構造、pretraining、post-training、データ、評価、制約をかなり詳細に記載しているが、学習データの完全なmanifest、総学習計算量、GPU総数、総学習費用、安全性red-team結果などは公開していない。したがって「かなり透明性の高い技術報告」ではあるが、完全再現可能な研究公開ではない。 [8]

項目	Motif 3	分析上の意味
総パラメータ	314B	大規模MoE
推論時アクティブ	13.2B	総量の約4.2%のみアクティブ
層数	53	2 dense + 51 MoE
Routed Experts	384	tokenごとにTop-8
Shared Expert	1	routed expertsと併用
Hidden size	4,096	公式model card
Dense FFN	12,288	同上
Attention	80 Q / 16 KV heads	GDLAベース
Vocabulary	220,160	SuperBPE
Context	262,144 tokens ≒ 256K	native long context
Pretraining	約12.5T tokens	cutoffは2026年3月
Precision	MXFP8をpretrainingで活用	計算効率を重視
License	MIT	commercial self-host可能
Modality	Text → Text	現行モデルは非マルチモーダル
AAII	47	v4.1.1、英語text-only

出典：Motif Technical Report（2026-08-10）、Motif公式Hugging Face model card、Artificial Analysis（2026-08-17参照）。 [9]

アーキテクチャ。 Motif 3はdecoder-only autoregressive MoEで、384のrouted expertsから8を選択する。AttentionにはGated Decoupled Latent Attention（GDLA）を導入する。これはGated Delta Attention的要素、Multi-head Latent Attention、出力gateを組み合わせた設計として説明されている。約10B規模のcontrolled experimentでは、MLAがloss 3.2に到達するのと比べ、GDLAは約9.2%少ないtokenで同じlossに到達したと報告されている。ただしこれはMotif 3本体に対するend-to-end ablationではなく、小型モデルでの社内比較であるため、「Motif 3の47点の9.2%をGDLAが説明する」という読み方はできない。 10

modified mHC、Expert-Specific PolyNorm、Multi-Token Predictionも組み込まれている。MTPは次tokenだけでなく先のtokenも予測する補助目的で、pretrainingの学習効率や表現品質を改善する狙いがある。Attentionはfull-causalとsliding-windowを組み合わせ、256Kまでのcontextを計算可能にする構成である。 11

学習データ。 約12.5兆Motif-tokenizer tokensの最大カテゴリはgeneral webであり、そのほかSTEM、code、synthetic QA、math、reasoning、Korean/multilingual、legal/financial dataを含む。公開されたNVIDIA Nemotron pretraining componentをすべて取り込み、最終corpusでもNemotron系が約70%を占める。残りの大部分はMotifが自ら収集・処理したとされる。初期19 datasetから57 datasetへ拡張され、序盤はgeneral data比率を高く、後期にSTEM、math、code、synthetic QA、reasoningの比率を上げるdynamic mixtureを採用した。最新データは2026年3月までである。 12

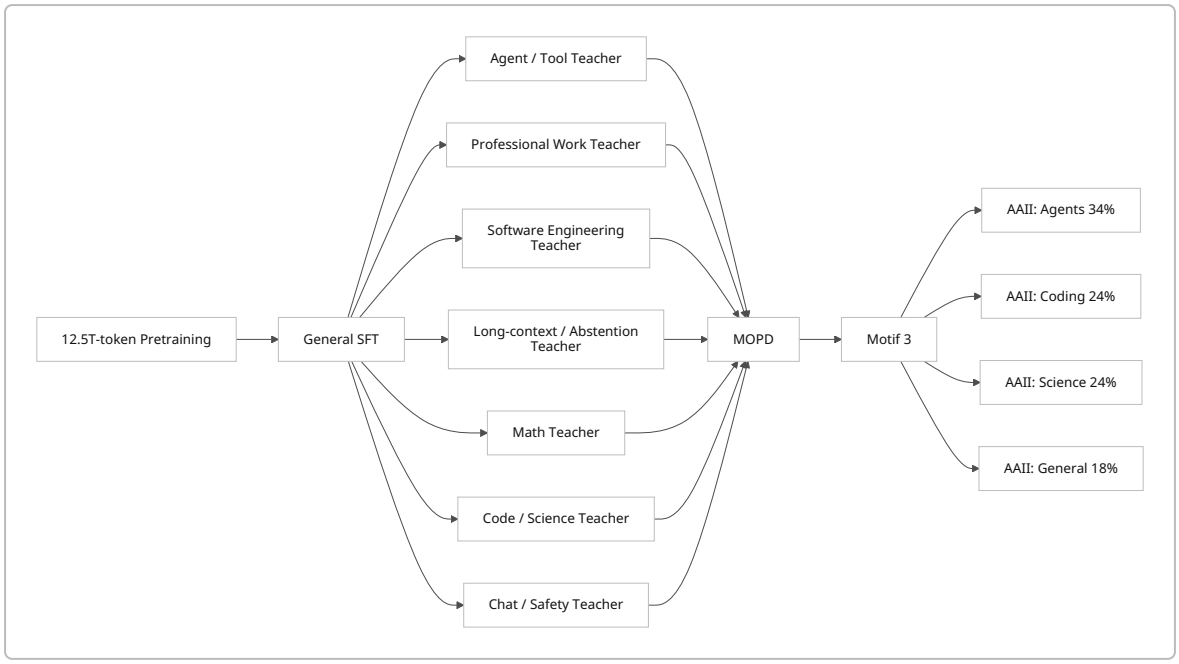
ここで重要な不明点がある。**英語、韓国語、日本語、中国語など言語別のtoken比率は公開されていない。各WebデータセットのURL/domain manifestも公開されていない。**「韓国語モデルだから韓国語中心に12.5T学習した」とみなす根拠はなく、むしろAAII性能を見る限り、広範な英語・STEM・codeデータが極めて重要だったと推測する方が自然である。この部分は公開情報からの推論である。 13

Pretrainingではcontext lengthを4Kから32K、最終的に256Kへ拡張した。専用long-context corpusは全体の約5%、reasoning-focused corpusは5%未満とされる。global batchは最大約75M tokens、learning rateは段階的に3×10^-4から7.5×10^-5、さらに3×10^-5へ調整された。 14

Post-trainingが47点の鍵である。 Motifは三段階を採用する。第一にgeneral SFT、第二にspecialist teacher training、第三にMulti-teacher On-Policy Distillation（MOPD）である。SFTにはinstruction、reasoning、coding、knowledge、agentic taskを含む広範なデータを使い、最長256Kのtrajectoryを扱う。 15

専門教師は6系統のGRPO RL teacherと1つのsoftware-engineering SFT teacherで構成され、agentic tool use、professional work、software engineering、long-context reasoning/abstention、math、code/science、chat等の能力を個別最適化する。rollout長はタスクにより8K〜192Kに達する。これをMOPDで単一モデルへ統合する。 16

この設計はAAIIとの整合が極めて高い。



したがってMotif 3は、「pretrainingで巨大な知識を詰め込み、そのままMMLUを解く」という2022〜2023年型のLLMより、**post-trainingでagentic trajectoryの成功確率を最大化する2025〜2026年型reasoning/agent model**に近い。AAII v4.1が従来のMMLU-Pro等を外し、GDPval、tau3、Terminal-Benchなど実行型タスクへ比重を移したことも、この設計を有利にしている。 17

Safety / guardrails。 Chat teacherはharmful、sensitive、unsafe requestに対するsafe responseを相当量含み、benign requestへのhelpfulnessを維持するよう学習したとされる。またlong-context teacherは「間違った回答」より「回答を控える」方に高いrewardを与え、hallucination回避を学習する。韓国語についてはcode-switching、script混在、言語不一致、反復、off-topic、極端に低品質な応答を排除するrubricも使っている。 18

しかし、**独立したsafety benchmark、jailbreak success rate、CBRN/cyber misuse評価、red-team methodology、deployment content filter、moderation APIなどはTechnical Report/model cardで十分には開示されていない**。したがって「安全なモデル」と一般化するには証拠が不足している。Motif自身も、評価範囲が全task/domain/languageを網羅せず、長期的state tracking、planning、recovery等には限界があると記載する。 19

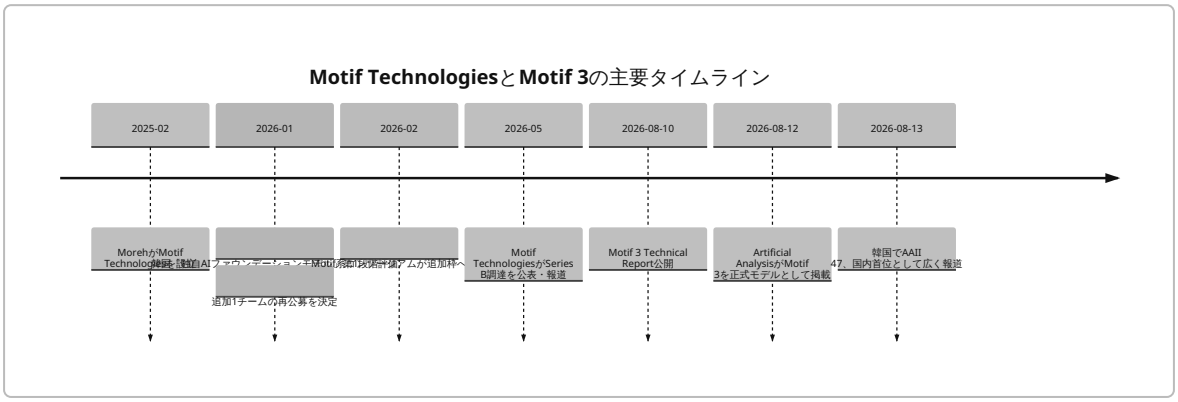
推論効率。 314B中13.2B activeなのでactive fractionは約4.2%である。これはdense 314BよりFLOPを大幅に抑え得るが、expert routing、memory bandwidth、all-to-all communicationがあるため、4.2%という数字をそのまま「dense 13Bと同じ速度」と解釈してはいけない。公式にはNVFP4 checkpointやvLLM-based deployment、8×H200/B200構成例が公開されているが、Artificial Analysisには2026年8月17日時点で標準化されたoutput tokens/secやlatencyが掲載されていない。 20

むしろAAの実測で目立つのは**verbosity**である。AAII全体で約260M output tokensを生成しており、比較対象中央値約100Mを大幅に上回る。高い推論性能の一部を長いreasoningによって得ている可能性が高く、API提供時のcost/latencyは今後の重要な検証対象である。 21

ライセンスと商用化。 Hugging Face checkpointはMIT licenseで公開され、self-hostingが可能である。Artificial Analysisの「$0」やproviderなしという表記は、計算コストがゼロという意味ではなく、同社が測定できる商用APIが現時点で存在しないという意味に近い。自前GPU運用費は当然必要である。 20

Motif TechnologiesはMorehが2025年2月に設立したモデル開発子会社で、MorehのAI infrastructure/software能力と結びついている。MotifはPwC Samil、Microsoft Azure等との事業連携も示しており、研究モデル単独ではなく、企業導入を含むfull-stack commercializationを志向する。CEO Junghwan LimはMoreh AI責任者、Samsung Research等の経歴を持ち、Oxford PhD、KAIST学士と紹介されている。 22

Technical ReportではJoon Son Chung、Sungmin Lee、Junghwan Limがtechnical/management leadershipを担当し、Wai Ting Cheung、Gihun Cho、Minsu Ha、Sangho Kang、Beomgyu Kim、Dongseok Kim、Jangwoong Kim、Taehyun Kim、Taewhan Kim、Jeesoo Lee、Jeongdoo Lee、Junhyeok Lee、Dongpin Ohらがcore contributionとして挙げられている。研究は韓国科学技術情報通信部（MSIT）、NIPA、NIAの支援を受けた。 23



追加公募および国策プロジェクトの日付はNIPA・韓国政府資料、Technical Reportは2026年8月10日、Artificial Analysisのrelease dateは2026年8月12日である。資金調達額については一次資料で完全に検証できなかったため、ここではタイムラインに金額を入れていない。 24

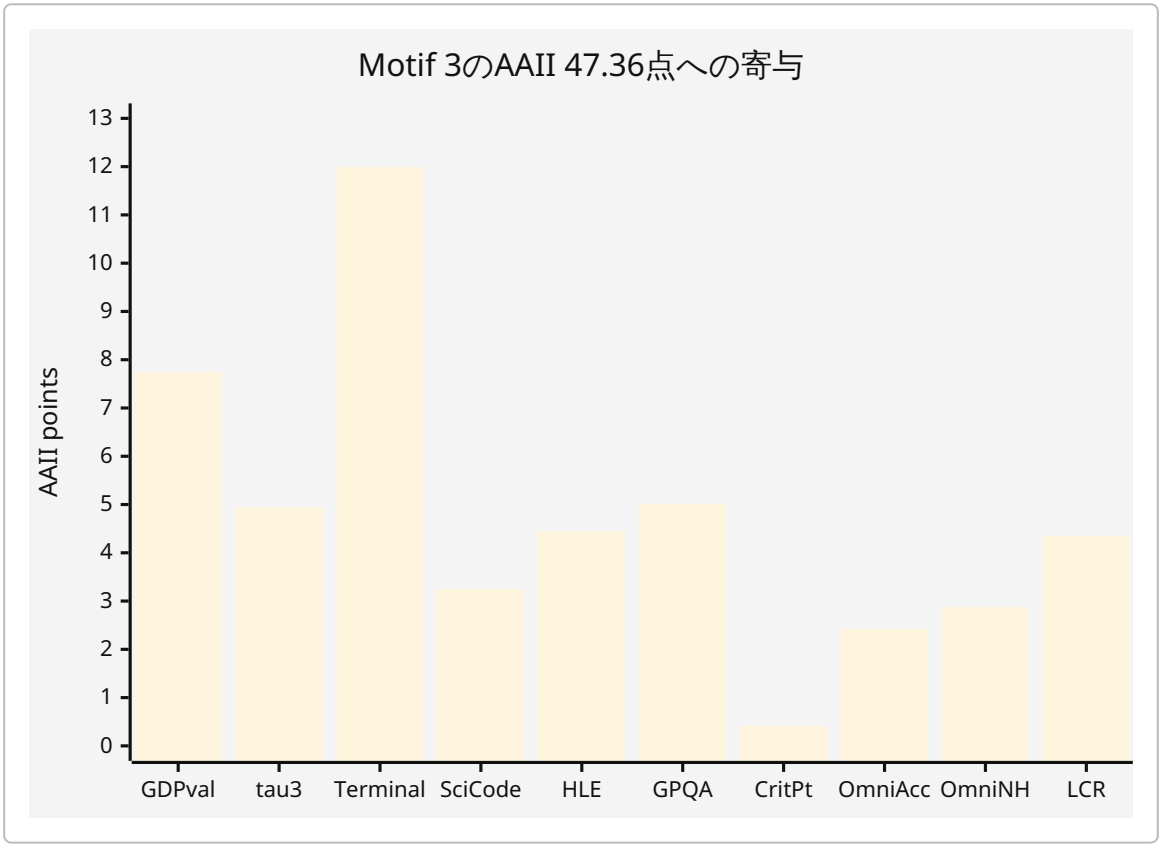
AAII 47点の再現と、評価指標そのものの偏り

現行Artificial Analysis Intelligence Indexは **v4.1.1** である。旧世代の「学術問題の正解率の平均」からかなり離れ、実際の業務、tool use、terminal操作、長文推論、scienceへ比重を移した指数になっている。Artificial AnalysisはAAIIを英語text-onlyで実施する一方、多言語能力はGlobal-MMLU-Lite等の別評価として扱う。したがって、**AAIIにおける「韓国モデルの高得点」は韓国語評価の高得点ではない**。 25

カテゴリ	Benchmark	AAII weight	Motif 3	Motif 3の得点寄与
Agents	GDPval-AA v2	20%	38.7	7.74
Agents	τ^3 Banking	14%	35.3	4.94
Coding	Terminal-Bench 2.1	16%	74.9	**11.98**
Coding	SciCode	8%	40.6	3.25
Scientific reasoning	Humanity's Last Exam	12%	37.0	4.44
Scientific reasoning	GPQA Diamond	6%	83.4	5.00

カテゴリ	Benchmark	AAII weight	Motif 3	Motif 3の得点寄与
Scientific reasoning	CritPt	6%	6.6	0.40
General	AA-Omniscience Accuracy	8%	30.1	2.41
General	AA-Omniscience non-hallucination	4%	71.6	2.86
General	AA-LCR	6%	72.3	4.34
合計		**100%**		**47.36 → 47**

Motif各スコアは公式model card、weightはArtificial Analysis v4.1.1 Methodologyに基づく。[3]



この分解から三つの事実が見える。

第一に、**Terminal-Benchだけで全AAIIの約25%に相当する12ポイント近くを稼いでいる**。Motif 3は74.9であり、software engineering teacher、code/science teacher、agentic trainingがそのまま効く領域である。[26]

第二に、GDPvalとτ³を合わせると約12.7ポイントである。つまり「terminal coding」だけでなく、professional workとagentic workflowの実行能力が重要である。Motifが「professional work teacher」「agentic tool-use teacher」を独立して訓練したのはAAII v4.1の設計と非常に相性がよい。これは**評価への不正なtrain leakageを示す証拠ではなく、task-distribution alignment**を示す証拠である。[27]

第三にCritPtは6.6と弱く、SciCodeも40.6でfrontier最高水準ではない。つまり47点は全能力が均等に強いからではなく、**高weight領域を強く、低い領域の弱さを他で相殺した結果**である。 [28]

GDPval-AA v2は少し特殊である。44職種のprofessional taskについてモデル出力を比較し、複数のfrontier LLM judgeによるpairwise preferenceをBradley–Terry系Eloへ変換する。AAIIに入れる際はおおむね

$$\mathrm{GDPvalNormalized} = \mathrm{clamp}\left(\frac{\mathrm{Elo} - 500}{2000}\right)$$

としている。したがってMotifの38.7を0.387と解釈すると、対応するEloは概算

$$500 + 2000(0.387) \approx 1274$$

となる。これはAAIIの変換式からの**逆算値であり、Artificial AnalysisがMotifのElo=1274と直接公表した数字ではない**。 [29]

Artificial Analysisは各testについて複数runを行う。例えばτ^3 Bankingは97 tasksを5回、Terminal-Bench 2.1は89 tasksを3回、GPQA Diamondは198問を5回、CritPtは70問を5回実施する。AAは代表モデルを10回以上測った結果などから指数の95% confidence intervalを概ね±1未満と推定しているが、完全なvariance matrixはまだ公開されていない。 [30]

AAIIには少なくとも五つの構造的バイアス／限界がある。

英語バイアス。 AAIIは英語text-onlyである。日本語特化モデルの最大の比較優位をほぼ測らず、逆に英語・code・math・agent taskへのpost-trainingが厚いモデルを評価する。Korean modelが有利なのではなく、**韓国企業が英語global frontierを狙うモデル設計をしていることが有利**なのである。 [30]

Agentic-taskバイアス。 Agents 34%＋Coding 24%で58%。2026年のLLMの実用性を測る合理的な選択ではあるが、翻訳、日本語敬語、文化知識、要約、RAG、低latencyなどを重視するモデルの価値を過小評価し得る。 [31]

LLM-as-a-judgeバイアス。 GDPvalはfrontier model judgesを使い、HLE、AA-LCR、Omniscience等にもLLM graderが入る。v4.1.1ではHLE/LCR/Omniscience graderも更新された。judge model固有の文体嗜好や推論スタイルがゼロとは言えない。 [32]

benchmark curationバイアス。 Humanity's Last Examは複数の既存frontier modelsが解けない問題を選ぶ過程を持つため、問題選定時に使用されたモデルと後発モデルの直接比較には注意が必要だとArtificial Analysis自身が記載している。 [33]

verbosity / compute-budgetバイアス。 Motif 3はAAIIで260M output tokensを生成した。reasoning budgetを大量に使うモデルと簡潔なモデルを単純に「知能」だけで比較すると、実運用コストが隠れる。現行MotifにはAA上の標準速度・API価格もないため、47点は**intelligence scoreであってcost-adjusted intelligenceではない**。 [21]

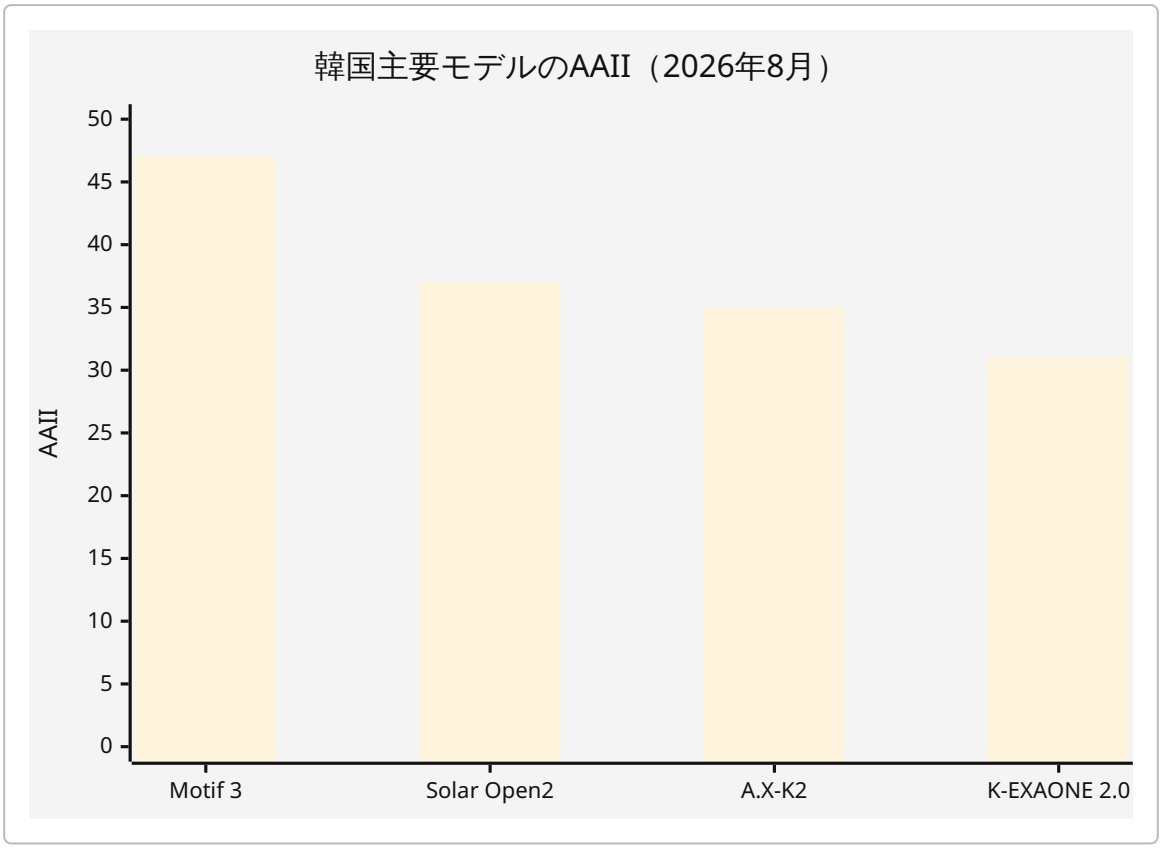
また、MMLUは現在のAAIIには入っていない。Motifのpretrained base modelではMMLU 86.20（5-shot）、MMLU-Pro 68.56、ARC-C 94.71、GSM8K 93.93、MATH 70.58、HumanEval 73.70、MBPP 84.60が報告されているが、これは**最終post-trained Motif 3のAAII評価と条件が違う**。よってMMLU 86.2を他社chat/reasoning modelのMMLUと無条件に横並びにすべきではない。 [14]

同時代の韓国・グローバルモデルとの比較

2026年8月のAA leaderboardを見ると、韓国勢の位置づけはかなり明瞭である。Motif 3は韓国モデルの中で突出しているが、「韓国全体が一様に47点級」というわけではない。

モデル	組織 / 国	総 / Active params	Context	AAII v4.1.1	特徴
Motif 3	Motif Technologies / 韓国	**314B / 13.2B**	262k	**47**	MoE、agentic、MIT open weights
Solar Open2 250B	Upstage / 韓国	250B / 非開示	1M	37	大型open-weight
A.X-K2	SK Telecom / 韓国	688B / 33B	262k	35	MoE、agentic、Apache-2.0
K-EXAONE 2.0 0803	LG AI Research / 韓国	750B / 37B	262k	31	reasoning MoE、open weights
Kimi K3 (low setting)	Moonshot / 中国	2.8T / 104B	1M	48	超大型MoE
MiniMax-M3	MiniMax / 中国	428B / 23B	1M	45	reasoning / multimodal
DeepSeek V4 Pro, max	DeepSeek / 中国	1.6T / 49B	1M	45	大型MoE reasoning

Artificial Analysis、2026-08-17参照。Solar Open2のactive parameterは参照したAAページでは明記されないため「非開示」とした。 34



特筆すべきは**active parameter efficiency**である。Motifは13.2B activeで47、A.X-K2は33B activeで35、K-EXAONEは37B activeで31である。もちろんAAII/active-Bを単純な効率指標にするのは不適切で、architecture、context、reasoning token量、hardware utilizationが異なる。それでもMotifがかなり小さいactive footprintで高scoreに達している点は、設計上の実質的な強みである。[35]

A.X-K2も弱いモデルではない。SK Telecomによれば688B-total / 33B-active MoEで、Sparse Gated Attentionを導入し、long contextとagent taskを強化している。14 benchmark平均でA.X K1より32.2 percentage points改善したと同社は報告し、Apache-2.0 open weightsとして公開している。ただしこの「32.2ポイント」はSKT独自のbenchmark aggregationであり、AAIIとは別物である。[36]

Motif公式model cardが同時代の主要モデルと同じプロトコルで掲載するサブベンチマークを比較すると、Motifの得意・不得意がさらに分かる。

Benchmark	Motif 3	MiniMax-M3	GLM-5.1	Kimi-K2.6	Qwen-3.7 Max	DeepSeek-v4-Pro
GDPval-AA v2	38.7	**44.4**	37.8	34.4	39.0	40.2
τ^3 Banking	**35.3**	15.3	13.6	23.3	12.0	30.1
ITBench public	**51.5**	—	40.3	31.2	42.5	38.3
SWE-Bench Verified	76.2	75.0	76.4	76.2	**80.4**	77.4
Terminal-Bench 2.1	74.9	65.2	61.8	65.9	**75.0**	64.0
SciCode	40.6	45.4	43.8	**53.5**	**53.5**	50.0
GPQA Diamond	83.4	**92.9**	86.8	91.1	92.4	88.8

Benchmark	Motif 3	MiniMax-M3	GLM-5.1	Kimi-K2.6	Qwen-3.7 Max	DeepSeek-v4-Pro
HLE	37.0	39.0	30.1	37.5	**41.4**	37.5
AA-LCR	72.3	**80.3**	68.0	76.7	75.0	70.0
OmniScience accuracy	30.1	16.7	23.7	32.6	31.0	**42.9**
OmniScience non-hallucination	71.6	**81.6**	70.1	59.5	74.0	5.9

出典：Motif公式model card。なおモデル名・設定は同カードの比較設定に従い、現在のArtificial Analysis上の最新モデル名と完全には一致しない。 [28]

ここから分かるのは、MotifはGPQAやSciCodeで絶対首位ではないということだ。**τ^3 Banking、ITBench、Terminal-Benchで非常に強い**。つまり「純粋なscience IQ」より、「複数stepを踏んで環境を操作し、仕事を完遂する」側に最適化されている。

この傾向はAAIIのweightingと一致するため、Motifは個々のacademic benchmarkでQwen/Kimiに負けても、composite AAIIでは競争力を保てる。

MMLUについては、Motifのbase model 86.2という一次値はあるが、この表のfinal reasoning modelsについて同一prompt、same-shot、same-thinking-budgetで揃ったMMLUはMotif資料にはない。Artificial AnalysisもAAII v4系ではMMLU/MMLU-Pro中心の古い設計から移行しているため、**現代frontier model比較ではMMLUを主指標にしない方が分析上適切**である。 [37]

韓国と日本のLLMエコシステムの構造差

韓国の最大の特徴は、政府が単なる研究助成ではなく、**「frontier modelを作り、世界leaderboardで競争させる」こと自体を産業政策化したこと**にある。

2025年の韓国政府AI追加予算は約1.8兆ウォンで、そのうち1.46兆ウォンを用いてadvanced GPU 10,000枚を確保、1723億ウォンで民間GPU約2,600枚を追加レンタルし、約1936億ウォン規模の「World Best LLM」系プロジェクトで最大5チームを3年間競争的に支援する方針を打ち出した。これらの項目は一部同じ政策パッケージ内にあるため単純加算すべきではない。 [38]

その後の「独自AIファウンデーションモデル」プロジェクトでは、GPU、data、talentを与えたうえで、benchmark、expert evaluation、user evaluationを組み合わせてstage-gate方式でチームを絞った。2026年1月の第1段階ではLG、SKT、Upstageが次段階へ進み、空いた枠について追加公募が行われ、Motif系が加わった。政府はこの場で、K-EXAONE、HyperCLOVA X、Solar、Motif、KT系モデルなどがArtificial Analysisに掲載されたことを明示的に「成果」として挙げている。 [39]

この一点は日本との差を理解するうえで極めて重要である。韓国では**Artificial Analysisへの可視化それ自体が国家プロジェクトの成功指標の一部になっている**。

韓国政府はさらに2025年中に10,000 GPUを確保する政策を進め、2026年7月には50,000 GPU目標のうち35,000を確保したと報告している。2026年以降も追加GPU、supercomputer、national AI computing centerへの投資を継続している。 [40]

領域	韓国	日本
政策の中心	Sovereign / frontier general-purpose foundation model競争	GENIAC＋産業特化＋Physical AI＋国内実装
開発者選抜	少数チームをstage-gateで競争	多数案件へ計算資源・R&D支援
Global benchmark	政府発表でAA掲載を成果として言及	国内/日本語/用途別評価が相対的に多い
GPU政策	国家による大量GPU直接確保・配分	GENIAC、ABCI等で計算資源供給
代表企業	Motif/Moreh、SKT、LG、Upstage、NAVER	NTT、PFN、NII/LLM-jp、NEC、Fujitsu、SB Intuitions等
Open weights	国家プロジェクトで強く促進	LLM-jp等は極めてopen、一方企業モデルはclosed/on-premも多い
商用戦略	API、open weights、cloud、telco、global leaderboard	国内企業、on-prem、security、業種特化が大きい
2026政策方向	General frontier + sovereign AI	General AIも支援しつつPhysical AI/industrial dataへ拡張

韓国政府、日本経産省・NEDO・NII・NTT・PFN資料に基づく整理。 41

韓国は言語データ基盤も長期的に構築している。MSIT/NIA運営のAI HubはAI学習用data platformを提供し、国立国語院の「모두의 말뭉치（みんなのコーパス）」は新聞、日常会話、online text、構文解析、意味役割、parallel corpus等を体系的に提供している。ある公開時点で134種の韓国語corporaが整備されたと国立国語院は報告している。2025年分には書籍、新聞、online posts、daily dialogue、syntax、semantics、image/table descriptions等も含まれる。 42

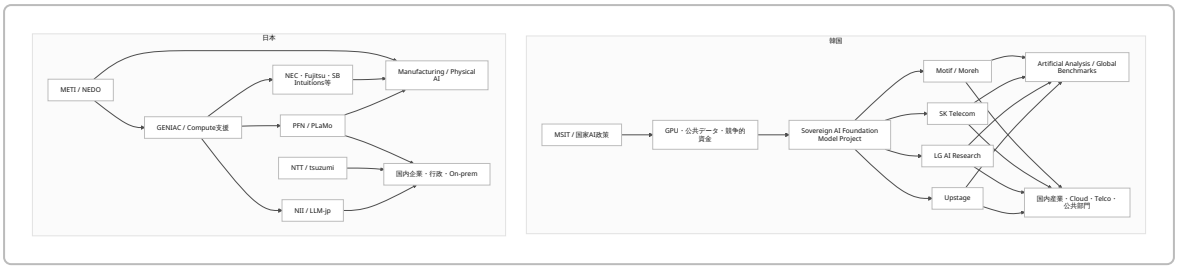
ただし、これもMotif 3のAAII 47を直接説明する主因ではない。AAIIは英語だからである。韓国語data infrastructureの価値は、国内サービス、post-training、culture alignment、韓国語品質、および企業利用を支える側にある。

日本にもデータ不足を単純に当てはめることはできない。NIIのLLM-jp-4は2026年4月、公開Web、政府・国会文書、synthetic data等を含む**約12兆tokens**のcorpusでfull-scratch pretrainingした8Bと32B-A3B MoEを公開した。第三者が入手可能なデータを重視し、透明性を強く意識した設計である。 43

その前のLLM-jp-3では約172B modelをfull-scratchで開発し、モデルだけでなくtraining dataまで公開する方向を採った。つまり、**日本にはfull-scratch LLM開発能力も、巨大日本語corpus構築能力も存在する**。 44

日本政府のGENIACは2024年2月から、foundation model開発に必要なcompute、dataset accumulation、knowledge sharing、community運営を支援してきた。2026年6月の第4期ではAI foundation model関連16テーマを採択している。さらに2026〜2030年はmanufacturing、multimodal foundation models、robotics/Physical AI、AI-ready industrial dataへ支援領域を広げている。 45

したがって、政策資金の有無そのものではなく、**集中度と目的関数が違う**と見るべきである。



産業プレイヤーの構造も異なる。韓国ではSK Telecomというtelecom、LGというindustrial conglomerate、NAVERというportal/cloud、UpstageというAI startup、Moreh/MotifというAI infrastructure/model startupが同じ国策競争に直接参加する。SKTはmodel、GPUaaS、AI data center、consumer AIまでをfull-stackとして位置づける。 46

日本ではNTTのtsuzumi 2が典型的で、NTT自身が「軽量で高い日本語能力」「機密情報を扱うon-prem/private cloud」を主要用途として説明する。これは日本企業にとって合理的な製品戦略だが、AAIIのような「巨大英語agent modelの能力競争」とは違う方向である。 47

PFNのPLaMo 3.0 Primeは2026年6月にreasoning/non-reasoning両モデル、256K context、tool calling、multi-step processing、coding、agent用途まで拡張しており、日本勢の中ではAA型評価とかなり近づいている。ただし2026年8月17日時点で今回のArtificial Analysis検索から同モデルのAAIIページは確認できなかった。 48

規制環境についても「韓国は規制がなく、日本は規制が厳しい」という単純図式ではない。韓国ではAI Basic Actが2026年1月22日に施行され、AI R&D、training data infrastructure、adoption supportを法制化する一方、生成AIの透明性、高影響AI、安全性等の義務も導入した。ただし政府は少なくとも1年間のregulatory grace periodとsupport deskを設けて企業移行を支援している。 49

日本でも「人工知能関連技術の研究開発及び活用の推進に関する法律」が2025年9月に全面施行され、研究開発・利用促進とrisk responseの両立を掲げている。したがって現状では、規制差だけで韓国・日本のleaderboard差を説明する証拠は弱い。 50

人材フローについては、日韓のfoundation-model researcherの純流入・純流出を直接比較できる最新の一次統計を今回確認できなかった。この点は不明とするのが妥当である。ただし韓国政府は海外のtop AI researchersを誘致する場合、最大3年間、年20億ウォンまで支援するAI Pathfinder構想を2025年追加予算で打ち出しており、「国外人材を直接買いに行く」政策手段を取っている。 38

なぜ日本LLMはArtificial Analysisで見えにくいのか

ここは最も誤解しやすい部分である。

Artificial Analysisの2026年8月のcountry-level trend画面には、主要国として米国、中国、韓国、フランス、カナダが表示され、日本はその可視化に現れていない。今回、Artificial Analysis上でPLaMo、tsuzumi、cotomi等を検索した範囲でもモデルページを確認できなかった。これは**「日本にはLLMが存在しない」という証拠ではなく、「AAが比較可能な形で評価済みの日本モデルが少ない」という証拠**である。 51

その原因は、証拠の強さが高い順に以下のように整理できる。

最重要要因は評価目的のミスマッチである。

AAIIは英語text-onlyで、agent34%、coding24%。一方、日本企業の公式発表は日本語品質、on-prem/security、enterprise workflow、translation、domain adaptation、manufacturingといった価値を強調する傾向が強い。NTT tsuzumi 2は軽量・日本語・private deployment、PLaMoは日本語cost-performanceと企業利用、LLM-jpは研究透明性を強く打ち出す。これらの強みの相当部分はAAIIで得点にならない。 52

対照的にMotif 3はagentic tool use、professional work、SWE、math、science、long-contextをpost-trainingの中心にしている。**同じ「LLMを開発する」でもobjective functionが違う。** 16

第二はスケールとリリース時期である。

2026年4月時点のLLM-jp-4公開モデルは8Bと32B-A3Bで、より大きいモデルは2026年度中に開発・公開予定だった。対して同時期の韓国国策勢はMotif 314B、A.X-K2 688B、K-EXAONE 750B、Solar Open2 250Bと、最初から数百B級frontier competitionへ投入している。 53

パラメータ数が能力を決定するわけではないが、large-scale MoEと大量post-training rolloutsを行えることはAAII上のagentic reasoningでは大きな余地を与える。

第三はアクセス可能性である。

Artificial Analysisが独立評価するには、安定したAPIか、再現可能にself-hostできるweightsが必要になる。Motif 3はMIT open weights、A.X-K2はApache-2.0 open weights、K-EXAONEもopen weightsであり、標準的なHugging Face/vLLM ecosystemへ入りやすい。 54

一方、NTT tsuzumi 2は企業・行政のon-prem/private cloud利用が重要な販売形態である。これはsecurity上は長所だが、第三者benchmark事業者から見れば評価導入のfrictionが大きい。 55

PLaMo 3.0 PrimeはAPI提供があるため潜在的には評価可能だが、今回確認した範囲ではAA掲載を確認できない。したがって「日本モデルはclosedだから評価されない」と一括りにすることも誤りで、**モデルごとに事情が異なる。** 48

第四は韓国政府がglobal benchmark visibilityを意図的に作っていること。

韓国のMSITは2026年1月の公式briefingで、国内モデルがArtificial Analysis leaderboardへ掲載されたことを明示的に評価している。さらに国策チーム同士をbenchmark、expert、user evaluationで競わせる。これにより企業には「国内で使えるモデル」だけでなく「国際leaderboardに投入できるモデル」を作る強いincentiveが生じる。 56

日本のGENIACは非常に重要な政策だが、2026年の公式資料を見ると、foundation model開発だけでなくdomain-specific AI、robotics、physical AI、manufacturing AI-ready dataへ対象が広がっている。日本の産業構造上は合理的だが、**AAIIのような単一general-intelligence scoreboardへの集中度を弱める。** 57

第五はopen-weight release cadenceとbenchmark engineeringである。

韓国の国策主要モデルが2026年夏にほぼ同時期にArtificial Analysisへ投入されたことは偶然とは考えにくい。Motif、Upstage、SKT、LGのAAIIが短期間に可視化され、韓国国内でもその差が報道された。これは政府競争プログラムが「release → independent benchmark → comparison」のfeedback loopを作ったことを示す。 58

第六に、韓国語と日本語の言語資源差を主因とする説は弱い。

韓国にはAI Hub、国立国語院corpusがあるが、日本にも12T-token LLM-jp corpusがある。したがって「日本語データが少ないから日本モデルが弱い」という単純説明は、少なくとも2026年には成立しない。 59

むしろ問題は、**そのデータを何B parameterのモデルに、何GPUで、何tokenまでpretrainし、その後何百万trajectory規模のagentic RLを行い、どのbenchmarkへ公開するか**である。

第七に、市場文化・commercial strategyの違いがある可能性が高い。

これは一次資料を組み合わせた推論である。日本企業は機密データ、オンプレ、社内業務、行政、翻訳、製造など「国内顧客が買う理由」を強く重視している。韓国勢もB2Bを狙うが、それと同時に「Global Top 3」「Sovereign AI」「global benchmark」という外向きのbrandingを政府・企業双方が繰り返している。 60

これを「日本企業が内向き、韓国企業が外向き」という文化論だけで説明するのは粗雑だが、**製品KPIが違えば研究KPIも変わる**という経営上の効果は十分に考えられる。

第八に、資金不足という説明も単純すぎる。

日本政府はGENIACでcompute supportを継続し、2026年第4期だけでも16のAI foundation-modelテーマを採択している。NIIは172B full-scratchも既に経験している。したがって「日本には政府支援がない」は事実ではない。 61

ただし韓国は1.46兆ウォンのGPU確保などを含む極めて明確な国家compute acquisitionを行い、少数frontier teamsへ競争的に資源を集中させている。日本は支援対象がより分散的である。**総額よりも、frontier general-purpose model一社当たりに集中できるcomputeとpost-training budgetの差**を追う必要がある。日本GENIAC各社と韓国各チームの実GPU-hours比較は公開情報からは算出できず、重要なopen questionである。 62

リスク、未解決論点、日本への実行提言

まずMotif 3そのものについて、評価を過大解釈してはいけない。

AAII 47は強いが、**AAIIはAGIの絶対尺度ではない**。英語、text、特定のagent/coding/science benchmarkを一定weightで合成したindexである。weightを変えれば順位も変わる。例えば日本語customer support、translation、低latency、on-prem inference、privacy、vision/audioなどを高く評価すれば、Motif 3の相対順位は大きく変わり得る。 30

Motif 3はtext-onlyであり、現在のKimi/MiniMax等にあるimage/video能力を持たない。Technical Report自身も今後image/videoと1M超contextを方向性として挙げる。 63

さらに、全AAII benchmarkに対する完全なtraining-data decontamination reportは公開されていない。Motifはsoftware-engineering trainingでSWE-bench baseとの重複やgold patchを除外したと説明しているが、全AAII datasetについて同程度のauditが開示されているわけではない。**benchmark contaminationがあったとする証拠はないが、ないと独立に証明できる情報も不足している。** 18

Safetyについても同様で、safety-oriented chat trainingは存在するが、独立red-team結果が不足する。MIT open weightsで自由度が高いことはcommercialization上の長所である一方、deployment safetyは利用者側にも大きく委ねられる。 64

日本側について最大のリスクは、「AAIIに載っていない＝技術力がない」と誤診し、単純にbenchmark chasingへ走ることでも、逆に「日本語さえ強ければglobal benchmarkは不要」と無視することでもある。

必要なのは**二層戦略**である。

ステークホルダー	推奨アクション	理由
日本のLLM研究者	日本語benchmarkだけでなくAAII、SWE-Bench、Terminal-Bench、tau系をrelease時に標準公開	global comparabilityを確保
Model developers	Agent/tool/SWE/long-context専用teacherとon-policy RLを大幅強化	現行frontier差はpost-trainingで大きく開く
企業	APIまたはopen-weight checkpointを国際評価機関が利用可能にする	「評価される可用性」を作る
GENIAC/NEDO	少数のgeneral frontier trackを別枠化し、GPUを長期集中	domain特化だけではglobal leaderboardが空白になる
政府	独立海外benchmarkの定期受験を政策KPI化	韓国のvisibility loopを再現
NII/大学	LLM-jpの透明性を維持しつつ100B〜数百B MoE、256K+、agentic post-trainingへ拡張	transparencyとfrontier capabilityの両立
産業界	日本語・企業特化の優位を維持しながら英語agent capabilityを別軸で持つ	国内価値と海外比較を両立
政策担当	GPU枚数ではなく「usable training FLOPs / team」とpost-training rollout budgetを追跡	nominal GPU countだけでは競争力を測れない

研究者への第一提言は、benchmark portfolioを変えることである。 Japanese MT-Bench、JGLUE、JMMLU等だけでは、海外から見たfrontier capabilityを説明できない。一方でAAIIだけでも日本語能力を説明できない。したがってrelease cardには、①日本語、②英語general reasoning、③coding/SWE、④agent/tool use、⑤long context、⑥safety、⑦tokens/sec/costという七つの評価系をセットで載せるべきである。AAIIが英語text-onlyであることを逆手に取り、「global能力」と「Japanese sovereign capability」を分離表示すべきだ。AAのmultilingual evaluationがGlobal-MMLU-Liteを別建てにしているのも、この考え方と整合する。 30

企業への第二提言は、post-trainingにpretraining並みの戦略的重要性を置くことである。 Motif 3の技術報告から得られる最大のlessonは、新しいattention architectureそのものより、specialist teachers＋on-policy trajectories＋distillationである。2026年の高性能LLMは、pretrained knowledge modelではなく**policy model**として競争している。日本の基盤モデル開発でも、SWE、terminal、browser、spreadsheet、enterprise workflow、long-horizon recoveryを独立teacherとして鍛え、その後統合する設計が有力である。 65

政策への第三提言は、GENIACに「Frontier General Model Track」を明確に設けることである。 日本の製造業・Physical AI重視は合理的であり、捨てるべきではない。しかし16社・多数テーマへ分散する支援と、数千〜数万GPUを長期間一つの汎用モデルへ投入する支援は別の政策である。韓国は後者を競争形式で実施した。日本も「日本語特化の小型モデル支援」「産業特化」「frontier general model」を同一KPIで管理せず、別トラック化する方がよい。GENIACがPhysical AIとAI-ready manufacturing dataへ拡大している現状では、特にその区別が重要である。 66

第四提言は、海外独立評価を補助金の成果条件にすることである。 Artificial Analysisだけを唯一のjudgeにする必要はない。しかし「releaseから30日以内に少なくとも二つの海外独立leaderboardで評価可能にする」「APIまたはweightsとserving instructionsを提供する」「evaluation configを公開する」といった条件は、日本モデルのvisibilityを大きく改善する可能性がある。韓国政府がArtificial Analysisへの掲載を政策成果として明示した事例は、その効果を示している。 56

第五提言は、NII/LLM-jpのopen-scienceモデルを戦略資産として扱うことである。 日本が韓国より明確に強い可能性のある一面は、training corpusを含む高い透明性である。LLM-jp-3 172BやLLM-jp-4の公開姿勢は研究再現性という意味で国際的価値が高い。ここに256K以上のcontext、MoE scaling、agentic RL、inference-efficient quantizationを足せれば、「透明性が高いだけの研究モデル」ではなく、AAでも競争できるreference modelになり得る。 67

第六提言は、日本企業の既存強みを捨てないことである。 NTTのlightweight/on-prem戦略やPFNのJapanese cost-performanceは、日本の機密性の高い企業市場では実際に重要である。必要なのはMotifをコピーして314B modelを全社が作ることではない。国として少数frontier modelsを世界比較可能にしつつ、各企業は小型・on-prem・domain modelsでcommercial advantageを取るportfolio strategyが合理的である。 68

最後に、現時点で決着していない重要な問いを明確にしておく。

Motif 3の総training FLOPs、GPU種類・GPU-hours、総開発費、言語別token構成、完全なdecontamination結果、安全性red-team、real-world inference tokens/secは非開示または十分に公開されていない。日本と韓国についても、各frontier teamが実際に使った**effective GPU-hours、post-training rollout token数、研究者一人当たりcompute、海外からのnet talent flow**を比較できる統一統計は確認できない。これらなしに「韓国人研究者の方が優秀」「日本はデータ不足」「韓国政府は日本より多額を使った」などと断定するのは証拠不足である。 69

現時点で最も証拠に整合する説明は、もっと制度的である。

韓国は、compute・data・企業・大学・open weights・global benchmarkを一つの競争システムに接続した。日本は、かなりの技術資産と公的支援を持ちながら、それを日本語性能、企業導入、軽量化、研究透明性、製造業、Physical AIへ分散している。AAIIという物差しで見ると、その違いが「韓国47、日本不在」という極端な形で可視化される。

したがって日本に必要なのは「韓国語LLMの作り方を模倣すること」ではない。**世界共通benchmarkで測れるfrontier-general能力と、日本独自の言語・産業・安全性の強みを、同じモデル開発エコシステムの中で両立させること**である。Motif 3の47点は、その技術的lesson以上に、**研究目標、公開方法、政策インセンティブ、第三者評価を一つのfeedback loopにした韓国AI政策の成果**として読むのが最も有益である。 70

---

1 21 34 35 Motif 3 - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/motif-3?utm_source=chatgpt.com

2 4 8 9 10 11 12 13 14 15 16 18 19 23 27 37 63 64 65 69 Motif 3: Technical Report
https://arxiv.org/html/2608.09119v1

3 20 26 28 54 Motif-Technologies/Motif-3 · Hugging Face
https://huggingface.co/Motif-Technologies/Motif-3

5 43 約12兆トークンの良質なコーパスで学習した新たな国産LLM ...
https://www.nii.ac.jp/news/release/2026/0403.html?utm_source=chatgpt.com

6 39 56 70 독자 AI 파운데이션 모델 프로젝트 1차 단계 평가 결과
https://www.korea.kr/briefing/policyBriefingView.do?newsId=156740468&utm_source=chatgpt.com

7 38 41 62 정부, AI 추경 1조 8000억 원 편성…'국가 AI역량 강화' 본격 ...
https://www.korea.kr/news/policyNewsView.do?newsId=148942035&utm_source=chatgpt.com

17 31 Artificial Analysis Intelligence Index v4.1: a shift toward agentic workloads
https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v4-1

22 About - Moreh
https://moreh.io/about/?utm_source=chatgpt.com

24 독자 AI 파운데이션 모델 프로젝트 추가 공고
https://www.nipa.kr/home/bsnsAll/0/nttDetail?bbsNo=4&bsnsDtlsIemNo=&nttNo=16449&tab=2&utm_source=chatgpt.com

25 29 30 33 52 Intelligence Benchmarking | Artificial Analysis
https://artificialanalysis.ai/methodology/intelligence-benchmarking

32 Changelog | Artificial Analysis
https://artificialanalysis.ai/changelog

36 SK Telecom Unveils A.X K2,Driving AI Adoption in Industry ...
https://news.sktelecom.com/en/3204?utm_source=chatgpt.com

40 정부 확보 GPU 1만 장, 산·학·연에 푼다…"AI혁신 본격 지원"
https://www.korea.kr/news/policyNewsView.do?newsId=148956711&utm_source=chatgpt.com

42 59 AI허브
https://aihub.or.kr/?utm_source=chatgpt.com

44 67 "llm-jp-3-172b-instruct3" Now Publicly Available- Achieving ...
https://www.nii.ac.jp/en/news/release/2024/1224.html?utm_source=chatgpt.com

45 61 66 「GENIAC」における計算資源の提供支援（第4期）において
https://www.meti.go.jp/press/2026/06/20260604003/20260604003.html?utm_source=chatgpt.com

46 SK Telecom to Showcase Full-Stack AI at World IT Show ...
https://news.sktelecom.com/en/3006?utm_source=chatgpt.com

47 55 60 68 NTT's LLM tsuzumi 2 Updated ~Achieving World-Class ...
https://group.ntt/en/newsrelease/2026/05/19/260519a.html?utm_source=chatgpt.com

48 国産生成AI基盤モデルPLaMo 3.0 Primeを正式リリース
https://www.preferred.jp/news/pr20260622?utm_source=chatgpt.com

49 '인공지능기본법' 22일 시행…생성형 AI 결과물 '워터마크 ...
https://www.korea.kr/news/policyNewsView.do?newsId=148958380&utm_source=chatgpt.com

50 ［政策お知らせ］人工知能関連技術の研究開発及び活用の推進 ...
https://www.gov-online.go.jp/hlj/ja/november_2025/november_2025-08.html?utm_source=chatgpt.com

51 AI 洞察与趋势
https://artificialanalysis.ai/zh/trends?utm_source=chatgpt.com

53 Release of New Japanese LLMs, "LLM-jp-4 8B" and "LLM ... - NII
https://www.nii.ac.jp/en/news/release/2026/0403.html?utm_source=chatgpt.com

57 本公募「AIロボット・フィジカルAIを見据えたマルチモーダル基盤 ...
https://www.nedo.go.jp/koubo/CD2_100431.html?utm_source=chatgpt.com

58 독자AI 모티프 47점…업스테이지 37·SKT 35·LG 31 순
https://ichannela.com/news/detail/000000544922.do?utm_source=chatgpt.com