ChatGPT

韓国政府「独自AI基盤モデル」第2次評価・全点数公開の検証報告

エグゼクティブサマリー

本調査は、KOREA WAVEが2026年8月31日に報じた「韓国政府、『独自AI基盤モデル』2次評価の全点数公開…総合1位はSKテレコム」を起点に、韓国科学技術情報通信部（과학기술정보통신부、MSIT）の公式発表、政府政策ブリーフィング、評価機関Artificial Analysis、各社のモデルカード・技術報告、韓国主要メディア、韓国の現行AI法制を突合したものである。調査対象期間は、制度設計・モデル開発の前史を必要に応じて2025年まで遡り、**2026年9月1日現在**までを対象とした。KOREA WAVEの記事は、MSITが8月27日に追加公開した点数をほぼ正確に伝えている。 [1]

結論を先に述べると、**「総合1位はSKテレコム」は事実だが、「モデルそのものの純粋な性能でSKTが1位」という意味ではない。** 第2次評価は100点中、ベンチマーク40点、外部専門家35点、ユーザー25点という政策選抜型の複合評価だった。SKTは70.6点、Upstageは69.9点、LG AI Researchは69.0点、Motif Technologiesは65.8点で、わずか0.7点差でSKTが首位になった。一方、グローバル能力指標Artificial Analysis Intelligence Index（AAII）ではMotifが11.9/25点で圧倒的1位、韓国NIA評価ではSKT、専門家評価と一般国民評価ではLG、AI専門ユーザー評価ではSKTがそれぞれ首位だった。 [2]

したがって今回の結果は、**「最高の基盤モデルはどれか」ではなく、「韓国政府が次段階で国家支援を継続するモデルとして、国際性能、韓国語・安全性、開発独自性、産業波及、実用性を合わせてどれを選ぶか」**という問いへの答えである。とりわけユーザー評価が順位を大きく動かした。MotifはベンチマークでSKTを2.4点上回ったが、専門家評価で2.2点、ユーザー評価で5.0点下回り、最終的に4.8点差で逆転された。これは公開点数から直接再計算できる。 [3]

一方で、評価の透明性には明確な限界が残る。MSITは8月27日、Motifの異議申立てと他チームの同意を受け、「詳細点数一切」を追加公開したが、公開情報から完全に再現できるのは主として**五つの大項目の点数**であり、NIAの個々の設問、専門家各委員の採点理由、ユーザーが入力した全プロンプト、ブランドを隠したブラインド評価の結果などは公開されていない。NIAは問題・正解を非公開にすることでデータ汚染を抑える性格を持つ反面、第三者による完全再現性を犠牲にしている。 [4]

技術的にも四者はかなり異なる。SKT A.X K2は688B総パラメータ／33B活性、約8.2兆トークンでゼロから事前学習した大規模MoE、Upstage Solar Open 2は250B／15B、約12兆トークン、1Mコンテキストを特徴とする効率重視MoE、LG K-EXAONE 2.0は750B／37Bで既存K-EXAONEをupcyclingし追加8兆トークンを継続事前学習した最大規模モデル、Motif 3は約314B／13.2Bで独自設計を前面に出すMoEである。 [5]

最も重要な政策的示唆は、韓国政府が「ソブリンAI」を単なる韓国語LLM競争から、**基盤性能＋国内データ＋独自技術＋安全性＋企業・国民が実際に使えること**を含む総合産業政策へ移している点である。ただし、この政策目的と「フロンティア基盤モデルの純粋性能競争」を一つの100点満点に混在させることが適切かは、今回の論争でなお未解決の論点となった。 [6]

公式原資料と評価制度の確認

担当省庁と原資料

担当省庁は韓国の**科学技術情報通信部（MSIT）**であり、担当部署は「人工知能技術基盤政策課（인공지능기술기반정책과）」である。MSITの公式電子政府サイトには、2026年8月27日付で、

> **原文:**「독자 AI 파운데이션 모델 프로젝트 2차 단계평가 결과를 상세히 공개합니다.」
> **訳:**「『独自AIファウンデーションモデル』プロジェクトの第2次段階評価結果を詳細に公開します」

という公式発表とHWPX/ODT原資料が掲載されている。 [7]

これはMotif Technologiesが同日、評価内容の詳細開示と再審議を求めたことを受け、残る参加チームの同意も得たうえで詳細点数を追加公開したものだった。KOREA WAVEの記事も、この8月27日の追加公開を直接のニュースソースとしている。 [8]

その前段としてMSITは8月18日に第2次段階評価の結果を発表し、8月20日にはメディアの「不透明な評価」「なぜAAII首位のMotifが脱落したのか」といった疑問を受け、

> **原文:**「2차 단계평가 기준과 결과에 대해 상세히 설명드립니다」
> **訳:**「第2次段階評価の基準と結果について詳細に説明します」

という公式説明資料を発表している。政府政策ブリーフィングには原資料のHWPX、ODT、HWPへのリンクも掲載されている。 [9]

このため、今回の情報公開は一回で完結したものではなく、**8月18日＝選抜結果発表 → 8月20日＝評価方法の追加説明 → 8月27日＝異議申立てを受け全チームの詳細点数追加公開 → 8月31日＝KOREA WAVE日本語報道**という流れで理解すべきである。 [10]

評価体系と配点

第2次評価は100点満点で、構造は次の通りであった。 [11]

評価群	評価項目	配点	主な意味
ベンチマーク	Artificial Analysis Intelligence Index（AAII）	25	国際的な一般能力、推論、科学、コード、エージェント等
ベンチマーク	NIA独自ベンチマーク	15	韓国語・国内文脈、知識、数学、長文、安全性・信頼性等
専門家	開発戦略・技術	10	技術戦略、独自性、設計能力等
専門家	開発成果・今後の計画	10	現時点の成果と次段階の開発計画
専門家	波及効果・貢献計画	15	国内AIエコシステム、産業AX、国家的波及
ユーザー	AI専門ユーザー	15	実際のモデル利用を通じた実用性・使用性
ユーザー	一般国民	10	一般利用者からみた使用性

評価群	評価項目	配点	主な意味
合計		**100**	

MSITはAIの進歩が速いため、約6カ月単位で開発目標を見直す「ムービング・ターゲット」方式を採用しており、第2次評価基準も参加チームと複数回協議して確定し、評価前に正式通知した時点では各チームから異議は出なかったと説明している。 [3]

AAIIのテスト内容

AAIIは民間の独立AI評価会社Artificial Analysisが運営する複合指数である。2026年6月時点のv4.1系は九つの評価を組み合わせ、GDPval-AA v2、τ^3-Banking、Terminal-Bench v2.1、SciCode、Humanity's Last Exam、GPQA Diamond、CritPt、AA-Omniscience、AA-LCRなどを使用する。Artificial Analysis自身は評価を独立実施し、指数全体について反復実験から95%信頼区間が概ね±1ポイント未満になると見積もっている。 [12]

ただしAAIIは固定不変の「IQテスト」ではない。指数の構成、評価タスク、採点器、正規化、重みはバージョン更新で変わり得るため、Artificial Analysis自身も異なる主要バージョン間の得点を単純比較すべきでないと説明している。 [13]

今回MSITが採用したAAII原点は、Motif 47.4、Upstage 37.4、SKT 35.0、LG 31.0で、25点満点に換算すると11.9、9.4、8.8、7.8となった。MotifはGDPval-AA v2、τ^3-Banking、Terminal-Bench v2.1、AA-LCR、AA-Omniscience、HLEなどで比較的強く、報道されたTerminal-Bench v2.1は74.9だった。 [3]

Artificial Analysisが現在掲載するMotif 3のAAIIは47で、同社は314B総パラメータ／13.2B活性と記載しているため、政府発表の47.4という評価時点スコアとも整合的である。 [14]

NIAベンチマーク

NIA（韓国知能情報社会振興院）側の評価は、報道説明上、**数学、知識、長文理解、指示履行、韓国語、信頼性、安全性、社会的安全性**の8領域を対象とした。Upstageは数学・指示履行・信頼性、SKTは知識・韓国語、LGは長文理解・安全性で相対的な強みを示したと説明されている。 [3]

ここには再現性上の重要な問題がある。NIA評価は問題・正解セットを一般公開しないことで事前学習への混入・ベンチマーク汚染を防止する考え方を採っており、その意味では「未知テスト」としてはAAIIより堅牢になり得る一方、**外部研究者が同一条件で完全再現することはできない**。 [15]

さらに公開された「原点合計」ではUpstage 732.4、SKT 728.7、LG 705.8、Motif 702.8なのに、15点換算後はSKT 13.4、Upstage 13.3、LG 12.8、Motif 12.7となっている。したがって単純に合計原点を15点満点へ比例縮小したわけではない。確認できた公開資料では、**この順位逆転を第三者が完全に再計算できるほど詳細な領域別正規化・重み付け式は示されていない**。これは本評価の再現可能性を検討するうえで重要な留保である。 [3]

専門家・ユーザー評価の手順

専門家評価は産業界3人、学界5人、研究界2人の外部専門家10人で行われ、提出資料について約5日間の書面審査と質疑応答を行った。各委員の最高点と最低点を除外し、残りを平均するtrimmed-mean方式が用いられた。また「独自性」の最低基準も確認された。 [3]

AI専門ユーザーは50人を選定し49人が実参加した。AIモデルを広く実利用した経験を持つスタートアップ代表らが、各チームの提供するWebインターフェースを使用し、1〜5点の絶対評価を行った。一般国民については住民登録人口の性別・年齢構成に沿ったクォータを設定して200人を無作為選定し、185人が実参加した。こちらも1〜5点評価を10点満点へ換算した。双方とも利益相反確認を行ったとMSITは説明している。[3]

しかし、**モデル名・企業ブランドを隠すブラインド試験ではなかった**ことが最大の争点となった。Motifは大企業と新興企業ではブランド認知に差があるため、一般利用者評価に先入観が入り得るとして、事前にブラインド化を求めていたと主張している。[16]

公開された全点数と順位の再構成

「全点数公開」という見出しで注意すべきなのは、**企業別に公開されたすべての主要採点区分が明らかになったという意味**であって、NIA全設問の正誤、専門家10人×各小項目の個票、185人全員の採点票まで公開されたという意味ではないことである。MSIT公式原資料と複数の報道を突合すると、公開された企業別主要点数は以下の通りである。[17]

企業・モデル別の全公開得点

総合順位	主幹企業	第2次モデル	AAII /25	NIA /15	ベンチマーク / 40	専門家 / 35	AI専門ユーザー /15	一般国民 / 10	ユーザー計 / 25	**総合 / 100**
1	**SK Telecom**	**A.X K2**	8.8	**13.4**	22.2	29.3	**11.6**	7.5	**19.1**	**70.6**
2	**Upstage**	**Solar Open 2**	9.4	13.3	22.7	29.1	10.8	7.3	18.1	**69.9**
3	**LG AI Research**	**K-EXAONE 2.0**	7.8	12.8	20.6	**29.5**	11.3	**7.6**	18.9	**69.0**
4	**Motif Technologies**	**Motif 3**	**11.9**	12.7	**24.6**	27.1	8.4	5.7	14.1	**65.8**

出典：MSITの8月27日公開値を、政府説明および聯合ニュース、KOREA WAVE等と照合して再構成。[17]

評価項目ごとの首位は完全に分散している。

指標	1位	2位	解釈
AAII	Motif 11.9	Upstage 9.4	国際一般能力ではMotifが明確に優位
NIA	SKT 13.4	Upstage 13.3	韓国語・国内評価ではほぼ僅差
専門家	LG 29.5	SKT 29.3	戦略・成果・産業貢献ではLG
AI専門ユーザー	SKT 11.6	LG 11.3	実利用ではSKT
一般国民	LG 7.6	SKT 7.5	一般利用性はLGとSKTが僅差

指標	1位	2位	解釈
総合	SKT 70.6	Upstage 69.9	0.7点差

[18]

原点も含めた公開情報

公開された換算前データも重要である。

評価	SKT	Upstage	LG	Motif
AAII原点	35.0	37.4	31.0	**47.4**
AAII換算 /25	8.8	9.4	7.8	**11.9**
NIA原点合計	728.7	**732.4**	705.8	702.8
NIA換算 /15	**13.4**	13.3	12.8	12.7
AI専門ユーザー原点合計	**189**	176	184	137
AI専門ユーザー換算 /15	**11.6**	10.8	11.3	8.4
一般国民原点合計	698	675	**701**	528
一般国民換算 /10	7.5	7.3	**7.6**	5.7

[3]

この表から、KOREA WAVEの「総合1位はSKテレコム」という表現そのものは正確だが、「SKTが全性能で1位」という解釈は不正確であることが分かる。 [19]

なぜMotifが逆転されたのか

点数差をMotifとの差で分解すると非常に明確である。

SKT－Motif	得失点
ベンチマーク	**−2.4**
専門家	**＋2.2**
ユーザー	**＋5.0**
総合	**＋4.8**

つまりSKTは純ベンチマークではMotifに負けたが、**専門家評価でほぼ取り戻し、ユーザー評価で大きく逆転した**。UpstageもMotifに対して、ベンチマーク−1.9、専門家＋2.0、ユーザー＋4.0で最終＋4.1となる。

さらに首位争いのSKT対Upstageでは、

SKT − Upstage = ベンチマーク −0.5 ＋ 専門家 +0.2 ＋ ユーザー +1.0 = +0.7点

である。したがって最終首位を決めた最大の差分は、実質的にAI専門ユーザー・一般ユーザーを合わせた利用評価だった。公開得点から導かれるこの解釈は、韓国メディアが「使用性が勝負を分けた」と報じた内容とも一致する。 [20]

一方、**専門家35点の「開発戦略・技術10」「開発成果・計画10」「波及効果・貢献15」の企業別小項目点、NIA八領域の企業別数値、個々のユーザー評価票は、今回確認できた公開本文では数値表として完全公開されていない。** したがって「全点数」を「全採点セル」と解釈するのは過大である。

上位モデルの技術比較

第2次評価対象は4チームであり、「上位3〜5社」という条件に対しては**全4社を比較するのが最も完全**である。第2次に第5社は存在しないため、5社目は「該当なし」とする。

モデル仕様の比較

項目	SKT A.X K2	Upstage Solar Open 2	LG K-EXAONE 2.0	Motif 3
総パラメータ	**688B**	250.3B	**750B**	約314B
1 token当たり活性	33B	**15B**	37B	**約13.2B**
方式	Sparse MoE	Hybrid-Attention MoE	Fine-grained sparse MoE	Sparse MoE
層数	61	48	78	53
Expert	256 routed＋1 shared、top-8	320 routed＋1 shared、top-8	256 total、1 shared、top-8	384 routed＋1 shared、top-8
最大コンテキスト	256K	**1M**	256K	256K
主な言語	英・韓・中・日・西	英・韓・日	**10言語**	多言語（公開カードは英・韓タグ）
事前学習規模	約**8.2T tokens**	約**12T tokens**	K-EXAONEから継続＋**追加8T**	一次資料で総トークン数を確認できず
出自	ゼロから学習	Solar Open 1の一部重みを選択移植	K-EXAONEをupcycle	独自設計・ground-upを表明
公開ライセンス	Apache 2.0	Upstage Solar License	Apache 2.0	非商用中心、商用は許可必要
政府総合順位	**1**	**2**	**3**	4
AAII順位	3	2	4	**1**

一次資料：各モデルカード・技術報告。 [21]

SK Telecom — A.X K2

A.X K2は688B総パラメータ、1トークン当たり33B活性のMoEで、SKTは**「trained from scratch」**と明記している。256 routed expert＋1 shared expertのうち8 expert＋shared expertを活性化し、61層、256Kコンテキストを持つ。Sparse Gated Attention（SGA）によって長文時に関連トークンをtop-kで選択し、FP8ネイティブ学習とGated Normによって低精度推論を念頭に設計されている。 22

学習データは約8.2兆トークンで、6.4Tの一般知識、1.4Tの高品質推論、0.36Tの長文拡張という三段階カリキュラム。第1段階は英語72.7%、韓国語15.4%、コード8.3%、日本語・スペイン語・中国語が各約1%で、韓国語Webデータには約1.37兆トークンの自社クロールが含まれる。Nemotron-CC-v2.1、Nemotron Pretraining Code v2、FineWeb2なども利用する。 23

強みはNIA評価の知識・韓国語、AI専門ユーザー評価、韓国市場へのローカライゼーション、FP8＋SGAによるサービス展開志向である。総合1位はこの「性能と実運用のバランス」を反映している。 3

弱みはAAIIが4社中3位だったこと、およびSKT自身がモデルカードでTerminal BenchやBrowseCompなど長期エージェント系では最強モデルに後れを取るとしている点である。つまり「agentic foundation model」を標榜する一方で、最先端の自律エージェント能力はまだ改善余地がある。 23

Upstage — Solar Open 2

Solar Open 2は250.3B総パラメータながら活性パラメータは15Bに抑え、48層のHybrid-Attention MoEを採る。Softmax attentionを1層、linear attentionを3層というパターンで繰り返し、1Mトークンという四モデル最長のコンテキストを実現する。全48層中KV cacheを必要とするSoftmax層が12層だけであることも、長文推論時のメモリ効率を意識した設計である。 24

事前学習は約12兆トークン、英語・韓国語・日本語対応、NVIDIA B200で約200万GPU時間。Solar Open 1からアーキテクチャ変更後も残る約2.3%の重みだけを選択転送し、それ以外はランダム初期化したと説明する。 24

公開モデルカード上ではMMLU-Pro 86.2、GPQA-Diamond 86.3、HLE 28.8、LiveCodeBench v6 92.4、AIME2026 95.7、AA-LCR 62.3など高い自社評価値を提示している。これらは政府評価と条件が違うため直接加算できないが、「15B活性で高性能」というUpstageの技術的狙いを示す。 24

強みは四社中でも特に小さい活性パラメータ数、1Mコンテキスト、AAII2位、NIA原点合計1位であり、計算効率と基盤能力のバランスが良い。

弱みは専門ユーザー評価が10.8で3位だったことで、モデルスペック上の効率がそのまま評価時Webサービスの利用感で優位にならなかった。またApache 2.0ではなく独自のUpstage Solar Licenseであるため、二次利用者はApacheモデルよりライセンス条件を個別確認する必要がある。 24

LG AI Research — K-EXAONE 2.0

K-EXAONE 2.0は**750B総パラメータ／37B活性**と四モデルで最大であり、78層、256 expert級のMoEである。ゼロから再学習するのではなく、236B／23BだったK-EXAONEをdepthとexpert数の両方向に拡張する「upcycling」を行い、その後さらに8兆トークンのcontinual pre-trainingを行った。 25

さらに64Kと256Kへのmid-trainingをそれぞれ400Bトークンずつ行い、post-trainingのSFTでは推論、知識、指示追従、エージェント等を対象に350Bトークン規模のデータを集約・合成したと技術報告は説明す

る。韓国語についてはK-DATA、NIA、国立国語院、東北アジア歴史財団などの公的・専門データを組み込み、FineWeb2などを利用して10言語へ拡張した。 [26]

安全面では100以上の既存リスク領域を改訂し、新たに70領域を追加した「Korea-Augmented Universal Taxonomy」を構築し、レッドチーミングと専門家人手評価を実施したとする。 [27]

強みは専門家評価1位、一般国民評価1位、長文理解・安全性でのNIA上の強みである。モデル開発を単純なベンチマーク最適化ではなく、企業利用・安全性・多言語展開まで含めたシステムとして設計していることが政府評価と相性がよかったと考えられる。 [3]

弱みはAAIIが31.0、換算7.8で4社中最下位だったこと、および37Bという最大の活性パラメータ数から、同一ハードウェア条件ならUpstageやMotifより推論資源が重くなりやすいことである。ただしLGはMTPとDSparkの二つのspeculative-decoding経路を組み込み、生成速度上の不利を補う設計を採る。 [27]

Motif Technologies — Motif 3

Motif 3は約314B総パラメータ／13〜13.2B活性、384 routed expertから8個を選ぶMoEで、53層、256Kコンテキスト。Motifは「既存オープンソースアーキテクチャのre-parameterizationではなく、完全な自社設計」と明記し、Grouped Differential Latent Attention、Grouped PolyNorm、modified mHC、MTPなど独自構成を採用する。 [28]

技術評価上の最大の強みはAAIIである。政府評価時47.4で他三社の37.4、35.0、31.0を大幅に上回り、Artificial Analysisの現在の独立ページでも47と評価されている。 [29]

一方、NIAは12.7で最下位、専門ユーザー8.4、一般国民5.7とユーザー評価で大差を付けられた。したがって政府選抜での弱点は**基盤モデル能力そのものというより、韓国語・国内評価と評価時サービスの利用体験**だった可能性が高い。 [3]

また公開Hugging Faceカードは調査時点で「Motif-3-Beta」「intermediate checkpoint」と記載されており、商用利用にはMotifの書面許可を要求する非商用中心のライセンスである。政府が評価したチェックポイントと現在公開されているBetaチェックポイントがビット単位で同一かは公開資料だけでは確認できない。 [28]

推論効率について分かることと分からないこと

四社ともMoEで、総パラメータより**活性パラメータ**を抑えて推論費を下げる設計を採っている。ただし政府の第2次評価表には、同一GPU・同一batch・同一contextで測定したtokens/sec、time-to-first-token、joule/token、1M token当たりコストといった標準化された推論性能指標がない。したがって、

「13B活性のMotifが33B活性のSKTより必ず2.5倍速い」

といった結論は導けない。Attention方式、Expert Parallelism、量子化、KV-cache、通信帯域、speculative decoding、生成長により実効速度は大きく変わる。各社の一次資料から確認できるのは、SKTのSGA＋FP8、Upstageのlinear attention＋少数KV-cache層、LGのMTP/DSpark、MotifのMTP self-speculative decodingという**推論効率化の設計方針**までである。 [30]

評価の妥当性、再現性、専門家・メディアの反応

「何を選ぶ評価か」が論争の中心

今回の論争には二つの異なる評価哲学がある。

一方はMotif側の考え方で、国家的な基盤モデル事業ならまず**foundation modelそのものの能力差を強く反映すべき**だという立場である。Motifは、AAIIの47と31という大差が25点換算では11.9対7.8、すなわち約4点差に圧縮されるため、技術的な性能差が最終順位に十分反映されないと批判した。 16

> **Motif側原文:**「실제 성능 차이를 … 보다 적절하게 반영돼야 한다」
> **日本語訳:**「実際の性能差が、より適切に反映されるべきだ」

というのが核心である。 16

他方、MSITは総合75点に相当するAAII以外の評価で、韓国語・信頼性、独自技術、開発計画、産業貢献、現実の使用性を確認することを重視した。つまり「フロンティアLLM競争」だけでなく**国家が資金・GPUを追加投入する産業政策案件**として選考したわけである。MSITは評価項目を事前に各社と協議し、最終案通知時点では異議がなかったと説明している。 3

この意味で、政府評価が「間違っている」と単純には言えない。**評価関数そのものがAAIIとは違う**のである。

AAIIの強みと限界

AAIIの強みは、外部独立機関が九つの難しい評価を共通条件で実施し、世界の多数モデルと横比較できることにある。Motifの47という値は政府内部だけで算出された数字ではなく、Artificial Analysisの現在の公開ページでも確認できる。 31

さらにMSITのブリーフィングによれば、ベンチマーク特化学習、いわゆる「benchmaxxing」があったのではないかという疑いについてArtificial Analysis側に確認し、**組織的な暗記・評価過適合を示す証拠は確認されなかった**との説明がなされた。 32

一方でAAIIには韓国固有の言語・歴史・法制度・文化、安全上の文脈を測る目的はない。また指数自体がバージョンアップされるため、47という数字は絶対的能力値ではない。Artificial Analysis自身が指数の構成変更・正規化・重み変更を明記している。 13

ユーザー評価のブランドバイアス

Motifの最も妥当性がある批判の一つはブラインド化である。同社は、

> **原文:**「대기업 브랜드가 영향을 미치지 않았나」
> **日本語訳:**「大企業ブランドが影響したのではないか」

という問題を提起し、評価前から匿名化を要請していたと報じられている。 16

この批判は統計的には検証可能である。最良の設計は、同じ回答を「企業名表示」と「企業名非表示」でランダム化して比較することであり、それを行えばブランド効果を推定できる。しかし今回、そうしたA/B試験の結果は公開されていない。

それでもユーザー評価を無意味とみなすのも適切ではない。Webインターフェース上では単一QAベンチマークでは拾えない応答安定性、指示遵守、対話継続性、待ち時間、拒否挙動、ツール機能などが露出するためである。ただし、**どの要素を何点として採点したかの細粒度ルーブリックが十分公開されなければ、「モデル品質」と「製品UI品質」が混ざる**という方法論上の問題が残る。評価が基盤モデル開発支援であるほど、この区別は重要である。 [33]

再現性の評価

本調査の判断をまとめると次の通りである。

評価	第三者再現性	理由
AAII	**中〜高**	方法・構成指標・公開スコアあり。ただし一部評価環境は外部サービス依存、バージョン変動あり
NIA	**低**	問題・正解が非公開。汚染防止には有利だが完全再現不能
専門家評価	**低〜中**	配点、委員構成、trimmed meanは公開。個々の判断理由・採点票は不十分
AI専門ユーザー	**低〜中**	人数・1〜5点方式・合計点は公開。全プロンプト、全個票、ブランド統制なし
一般国民	**低〜中**	標本設計は説明されているが、全試行ログ・個票・匿名化対照実験なし
総合点	**高**	公開された五大項目から70.6等を再計算可能

[34]

したがって**総合得点そのものの信頼性は高いが、「その点数になった原因」を独立監査できる程度の透明性はまだ低い**、というのが最も妥当な評価である。

韓国国内メディアの反応

聯合ニュースは、AAII＝Motif、NIA＝SKT、専門家・国民＝LG、専門ユーザー＝SKTという「評価軸ごとに首位が異なる」構図を強調した。 [35]

韓国のテック・経済メディアでは特に、AAII25点に対し47→11.9という単純換算をした結果、グローバル性能差が相対的に縮小した点、専門家35点の企業間差が小さかった一方でユーザー25点が順位を大きく動かした点が論争になった。 [36]

Motif自身は8月27日に正式な異議を申し立てたものの、

> **原文:**「탈락 번복 요구 아냐」
> **日本語訳:**「脱落結果の撤回を求めるものではない」

とし、評価詳細の開示と今後の制度改善を主目的とした。また異議申立ての結果にかかわらず第3次段階には参加しないと表明した。 [16]

その結果、政府は同日に全チームの同意を得て詳細点数を開示した。透明性を求める異議申立てが実際に追加情報公開を引き出したという点では、制度的には一定の成果があった。 [37]

国外での受け止め

2026年9月1日までに確認できた範囲では、**選抜手続きそのものを詳細に検証する英語圏主要メディア報道は韓国国内報道ほど多くない**。日本語圏ではKOREA WAVEの記事がAFPBB News等を通じて「総合1位SKT」「AAII首位Motif」という構造を紹介した。[19]

一方、技術的な国際評価ではArtificial AnalysisがMotif 3をAAII 47、A.X K2を35、K-EXAONE 2.0を31などとして独立評価しており、韓国勢が国際ベンチマークの対象として普通に比較される段階には入っている。[38]

これは重要である。韓国内の政府ランキングと国外の技術ランキングは**競合するランキングではなく、目的関数の異なる二種類のランキング**とみるべきである。

技術差が商用化・規制・安全性に与える影響

商用化の優位性は総パラメータ数だけでは決まらない

商用化では、総パラメータ数より、活性パラメータ、GPUメモリ、KV-cache、context length、量子化、ライセンス、韓国語品質、サービス運用の成熟度が重要になる。

この観点ではUpstageは15B活性＋1M contextで効率面、Motifは13.2B活性で小さい計算量、SKTはFP8ネイティブ＋SGAと大規模通信サービス運用、LGはApache 2.0＋企業向け安全性・長文処理で、それぞれ異なる商用上の優位点を持つ。[30]

ライセンスも差が大きい。SKT A.X K2とLG K-EXAONE 2.0はApache 2.0で公開されており、企業が改変・再配布を行いやすい。一方Motif 3の公開カードは個人・教育・非商用研究を許可し、商用利用には事前の書面許可を要求する。Upstageは独自Solar Licenseである。したがって「open weights」という言葉だけでは商用自由度を比較できない。[39]

韓国AI基本法との関係

韓国ではAI基本法「人工知能の発展と信頼基盤造成等に関する基本法」が2026年から施行され、2026年9月1日現在、生成AIや高影響AIに関する透明性・安全性義務が実際の事業環境となっている。国家法令情報センターには現行法および施行令が掲載されている。[40]

現行法31条は、生成AIまたは高影響AIを使った製品・サービスを提供するAI事業者に、そのAIに基づいて運用される事実を事前告知することを求め、生成AIの結果物についてもAI生成である旨の表示義務を設けている。[41]

高影響AIについては、リスク管理、主要な判断基準や学習データ概要の説明、人による管理・監督、利用者保護、関連文書の保存等が要求される。[42]

したがって今回の政府評価に「信頼性」「安全性」「リスク管理」「実用性」が含まれることには、単なる政策的好み以上の意味がある。**次世代K-AIモデルを医療、金融、雇用、公共サービス等へ展開する場合、基盤モデルのベンチマーク性能だけでは法的・社会的要件を満たさない**からである。[43]

ただし今回の第2次評価を「AI基本法適合認証」と理解してはならない。AI基本法30条は安全性・信頼性の検証・認証を別制度として位置づけており、第2次評価で高得点だったことは特定の高影響AIサービスに対する法令適合性を保証しない。[44]

安全性評価でLGが示す方向性

技術資料上、四社のなかで最も体系的に安全データを公開しているのはLGである。K-EXAONE 2.0は既存100以上のリスク領域を改訂し70領域を追加、韓国社会文化に合わせたtaxonomyを構築し、安全性を後付けフィルタではなく学習・評価全体へ組み込んだとしている。 [27]

SKTも、A.X K2は高リスク分野で自律判断に用いるための認証を受けておらず、医療、金融資格、雇用、法務、法執行、安全クリティカル判断の唯一・主要根拠にすべきでないと明示する。さらに事実誤り、バイアス、adversarial promptingによる安全性逸脱が残ることをモデルカード上で認めている。 [23]

この種の**limitationsを明示するモデルカード文化**は、韓国AI基本法下でのリスク説明・企業導入において、単純なベンチマーク順位以上に重要になっていくと考えられる。これは法制度と各社技術資料から導く推論である。 [45]

政策的意義、国際競争力、今後の展望

今回の評価から読み取れる第一の意味は、韓国の「独自AI」がもはや単に「韓国企業が外国モデルに依存しない」ことだけを指していないことである。A.X K2の自前事前学習、Motifの独自アーキテクチャ、Upstageの効率的ハイブリッドAttention、LGのupcycling＋安全性体系のように、**モデル設計、訓練データ、GPU運用、推論技術、韓国語・文化データ、サービス化まで国内に技術蓄積すること**が政策目標となっている。 [46]

第二に、世界トップモデルとの距離はなお大きい。Artificial AnalysisでMotif 3が47という点は世界的に競争力のあるopen-weightモデル水準ではあるが、同時期の最上位モデルは60点台に達している。Artificial Analysisの最新ページでは指数首位層が63前後であり、Motif自身も「47でもフロンティアには届いていない」という問題意識を示した。 [47]

したがって「韓国が米中と並ぶAI超大国になった」という評価は時期尚早である。他方、四つの国内チームすべてが数百B級MoEを構築し、国際ベンチマークに投入し、モデルウェイトや技術報告を公開する段階に来たことは、**単なるLLM API利用国から、frontier-scaleモデルを設計・訓練・評価できる国へ技術層が厚くなっている証拠**と評価できる。 [48]

第三に、今回の最大の制度的教訓は「性能」と「国家代表選抜」を一つのスカラー得点に押し込める難しさである。AAII首位、NIA首位、専門家首位、利用者首位が全部別企業だったという事実は、ある意味では評価制度が複数の価値を実際に測っていた証拠である。しかし同時に、「世界最高性能を追求するR&D支援」と「国内に広く展開できる国家AI基盤を選ぶ産業政策」が同じ順位表で決められたことが論争を生んだ。 [49]

政策設計上は、今後、

フロンティア性能トラックと**実用・エコシステムトラック**を分ける、あるいは総合順位だけでなくPareto frontier形式で複数のチームを支援することが合理的である。これは今回の公開点数から導かれる分析上の提案であり、現行の公式方針そのものではない。

さらに次回評価の信頼性を高めるには、少なくとも次の情報が重要になる。NIAについては設問そのものを秘密に保ったままでも、領域別重み・正規化式・信頼区間を公開できる。専門家評価は匿名化した委員別小項目点と評価理由を公開できる。ユーザー評価は企業名・UIを可能な範囲で統一した**二重ブラインドまたはブランド表示／非表示A/B試験**を追加し、「基盤モデル品質」と「製品UI」を分離すべきである。AAIIについては採用した正確なバージョンと評価日時を固定し、将来の指数更新による順位変化と区別する必要がある。これらは再現可能性を高めつつ、秘密テストの汚染耐性も維持できる。

本調査における事実確度

論点	確度	判断
SKT 70.6で総合1位	**非常に高い**	MSIT公式＋複数メディア一致
公開五大項目の全点数	**非常に高い**	公式追加公開値と聯合/KOREA WAVE等が一致
MotifがAAII首位	**非常に高い**	MSIT＋Artificial Analysis双方で確認
SKTが「純性能」1位	**否定**	AAIIではMotif、NIAでも複数指標が混在
ユーザー評価が順位を大きく動かした	**高い**	得点分解で直接確認可能
ブランド認知がMotifの低得点を生んだ	**未証明**	Motifの妥当な仮説だが対照実験なし
NIA評価を第三者が完全再現可能	**否定**	問題・正答非公開
専門家評価を完全監査可能	**否定**	個票・詳細理由が不十分
四モデルの同一HW推論速度順位	**不明**	政府は統一tokens/secを公開していない

主要一次資料

最重要原資料は、MSITの2026年8月27日「第2次段階評価結果の詳細公開」で、公式ページからHWPX/ODT原文を取得できる。 [7] 8月20日の「第2次段階評価基準と結果の詳細説明」は政府政策ブリーフィングに原ファイル付きで掲載されている。 [50] 8月18日の結果発表・記者質疑も政府政策ブリーフィングに記録されている。 [32]

技術一次資料としては、SKT A.X K2公式モデルカード・技術報告、Upstage Solar Open 2公式モデルカード、LG K-EXAONE 2.0 Technical Report、Motif 3公式モデルカードが最も重要である。 [30] 国際ベンチマークについてはArtificial Analysis公式の指数ページ・評価方法を参照した。 [12] 現行規制については韓国国家法令情報センターのAI基本法・施行令を参照した。 [40]

総合すると、KOREA WAVEの「**総合1位はSKテレコム**」という報道は数値として正しい。しかし今回の資料群から得られる、より厳密な結論は、**「国際ベンチマーク最強はMotif、韓国国内ベンチマークと専門利用者ではSKT、専門家・一般利用者ではLGが首位となり、その複合政策評価を0.7点差でSKTが制した」**である。したがって、このニュースを韓国AIの技術力比較として読む際には、「総合順位」と「基盤モデルそのものの能力順位」を明確に分離する必要がある。 [51]

---

[1] [8] [19] 韓国政府、「独自AI基盤モデル」2次評価の全点数公開…総合1位はSKテレコム　写真枚　国際ニュース：AFPBB News
https://www.afpbb.com/articles/-/3650569

[2] 과기정통부, 독파모 2차 평가 점수 공개…SKT 1위
https://stock.mk.co.kr/news/view/1149472?utm_source=chatgpt.com

[3] [11] [20] [29] [33] [51] 과기부, 독파모 2차 평가 세부점수 공개…SKT 1위, 업스테이지 2위, LG AI연구원 3위
https://v.daum.net/v/5kg0nbDgsN?f=p

[4] [17] [37] (보도참고)「독자 AI 파운데이션 모델」프로젝트 2차 단계평가 ...
https://www.msit.go.kr/bbs/view.do?bbsSeqNo=94&mId=307&mPid=208&nttSeqNo=3187709&sCode=user&utm_source=chatgpt.com

5 21 22 23 30 39 45 46 48 skt/A.X-K2 · Hugging Face
https://huggingface.co/skt/A.X-K2

6 32 독자 AI 파운데이션모델 프로젝트 2차 단계평가 결과 - 정책뉴스 | 뉴스 | 대한민국 정책브리핑
https://www.korea.kr/news/policyNewsView.do?newsId=156774781

7 보도자료 - 과학기술정보통신부
https://www.msit.go.kr/bbs/view.do?bbsSeqNo=94&mId=307&mPid=208&nttSeqNo=3187709&sCode=user

9 10 50 [보도설명] 「독자 AI 파운데이션 모델」 프로젝트 2차 단계평가 기준과 결과에 대해 상세히 설명드립니다(다수매체) - 사실은 이렇습니다 | 브리핑룸 | 대한민국 정책브리핑
https://www.korea.kr/briefing/actuallyView.do?newsId=148970354

12 47 Artificial Analysis Intelligence Index v4.1.1
https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index?utm_source=chatgpt.com

13 Data API docs
https://artificialanalysis.ai/data-api/docs?utm_source=chatgpt.com

14 31 38 Motif 3 - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/motif-3?utm_source=chatgpt.com

15 독파모 2차전 '모티프' 탈락…업스테이지·SKT·LG 진출
https://zdnet.co.kr/view/?no=20260818110107&utm_source=chatgpt.com

16 모티프 '독파모' 탈락 공식 이의 제기…왜?
https://tv.edaily.co.kr/News/NewsRead?Kind=257&NewsId=02663366645551584

18 35 49 [AI프리즘] 독파모 2차 성적표 공개…항목별 1위는 달랐다
https://www.yna.co.kr/view/AKR20260820140800017?utm_source=chatgpt.com

24 upstage/Solar-Open2-250B · Hugging Face
https://huggingface.co/upstage/Solar-Open2-250B

25 26 27 K-EXAONE 2.0 Technical Report Journey to Global Frontier-Scale Foundation Models
https://arxiv.org/html/2608.04505v1

28 Motif-Technologies/Motif-3-Beta · Hugging Face
https://huggingface.co/Motif-Technologies/Motif-3-Beta

34 Artificial Analysis Intelligence Benchmarking Methodology
https://artificialanalysis.ai/methodology/intelligence-benchmarking?utm_source=chatgpt.com

36 흔들리는 '국가대표 AI' 선발… 평가 방식·배점 놓고 잡음 이어져
https://v.daum.net/v/0AnVtA0HA8?f=p&utm_source=chatgpt.com

40 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법
https://www.law.go.kr/LSW//lsInfoP.do?lsId=014820&utm_source=chatgpt.com

41 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 (약칭: ...
https://www.law.go.kr/LSW//lsSideInfoP.do?
docCls=jo&joBrNo=00&joNo=0031&lsiSeq=282791&urlMode=lsScJoRltInfoR&utm_source=chatgpt.com

42 43 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 (약칭: ...
https://law.go.kr/lsLinkCommonInfo.do?chrClsCd=010202&lsJoLnkSeq=1031810839&utm_source=chatgpt.com

44 인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 (약칭 ...
https://law.go.kr/lsLinkCommonInfo.do?chrClsCd=010202&lsJoLnkSeq=1031810687&utm_source=chatgpt.com