

独Aleph Alphaの主権型言語モデル「Kolibri」の全貌と日本市場への波及分析

Gemini 3.8 Flash

ドイツのAIスタートアップ企業Aleph Alphaは、ドイツ統一の日にあたる2026年10月3日、新たな基盤言語モデル「Kolibri (Kolibri-1)」をApache 2.0ライセンスに基づくオープンウェイト形式で公開した¹。この発表は、欧州連合 (EU) が長年模索してきた「デジタルおよびAIの主権 (Sovereign AI)」を、実用的なオープンアーキテクチャとして具現化した試みとして国際的な注目を集めている⁶。米国および中国の巨大テック企業が市場を牽引するなか、自国・域内でのデータ統制や厳格な法規制遵守を最優先事項とする欧州のアプローチは、同様の地政学的・産業的課題を抱える日本にとっても示唆に富む⁷。

本レポートでは、公開された公式技術レポートおよびモデルカードの知見を基に、Kolibriの構造、性能特性、主権性設計の詳細を学術的・産業的視点から解剖する⁵。さらに、2026年9月に締結されたカナダCohere社との事業統合という地政学的再編を踏まえ、本モデルの登場が日本のAI政策、国内基盤モデル開発、およびエンタープライズ領域に及ぼす影響を多角的に検証する¹⁰。

開発の背景と「主権AI」の工学的再定義

Aleph Alphaは創業以来、行政、防衛、重工業、医療といったミッションクリティカルかつ規制の厳しい領域を標的としてモデル開発を進めてきた⁷。2026年10月3日というドイツ再統一を記念する象徴的な日に実施されたKolibriのリリースは、米中プラットフォームへの技術的従属から脱却し、欧州自前の計算資源と法規範の下で自立したAI基盤を保持できることを示す政治的・産業的マニフェストの意味合いを帯びている⁴。

本モデルにおいて「主権」という概念は、単なる政治的言説にとどまらず、厳密なエンジニアリング要件として定義されている⁷。第一に、モデルの全ライフサイクルが欧州域内で完結している点が挙げられる²。外部の先行モデルからの蒸留やファインチューニングに頼ることなく、データ収集から事前学習、事後学習に至る全工程がドイツおよびフィンランドに所在するインフラ上で、欧州およびドイツ法の下でスクラッチから実行された²。第二に、顧客組織への完全な運用管理権の移譲である⁹。利用規約上の制限や月間アクティブユーザー数 (MAU) による縛りを課さない完全なApache 2.0ライセンスを採用することで、機密データを外部に一切送信しない完全オフラインのエアギャップ環境での自社ホスティング運用を法的に保証している¹。

第三に、説明責任と規制適合性の担保が徹底されている²。Aleph Alphaは欧州AI規制法 (Regulation (EU) 2024/1689) の汎用AI (GPAI) 行動規範に署名しており、20兆トークンに及ぶ学習データ源の内訳、ベンチマーク汚染の排除手順、権利者対応窓口を網羅した詳細な技術文書を開示している¹。欧州や各国の公共調達において、単なるベンチマークスコアの多寡以上に「監査可能性」が選定の分水嶺となる現状を見据えた戦略的設計である²。

アーキテクチャと主要技術仕様

Kolibriは、総パラメータ数781億 (78.1B) を有しながら、推論時の計算負荷を劇的に削減する疎結合型Mixture-of-Experts (MoE) 構造を採用している¹。能力の絶対値のみを求めてパラメータを肥大化

させるフロンティアモデルとは一線を画し、推論コスト、実行速度、長文コンテキスト処理のトレードオフを実用的に最適化している点に技術的特異性がある²。

項目	仕様・構成値
モデル種別	英語・ドイツ語特化型 MoE トランスフォーマー推論モデル ⁵
総パラメータ数	78,103,074,560 (約78.1B) ⁵
トークンあたり有効パラメータ数	3,457,573,120 (約3.46B) ⁵
レイヤー構成	全50層 (50 MoEブロック、各層1共有エキスパート) ²
エキスパート構成	各層384エキスパート (1共有 + 6ルーティングエキスパート稼働) ⁵
アテンション構造	4:1 スライディングウィンドウ (SWA: 512) / グループドクエリアテンション (GQA) ⁵
コンテキスト長	ネイティブ学習長: 262,144トークン / 外挿最大長: 1,048,576トークン ⁵
語彙数およびトークナイザ	128,000語彙 (独自開発 UniBPE アルゴリズム) ²
量子化・配布精度	FP8 (float8_e4m3fn、128×128ブロック動的量子化) / FP8 KVキャッシュ評価 ¹
メモリ要件	FP8重みで約78GB (推奨: 2× H100 SXM、2× H200、1× B200等) ²
学習インフラ	768基のNVIDIA B200 (96 HGXノード)、事前学習に約21日間 (511時間) ¹
事前学習トークン量	20兆 (20T) トークン (英語約62.5%、ドイツ語約23.9%、コード約13.6%) ¹
ナレッジカットオフ	2026年6月18日 ²

ライセンス	Apache License 2.0 (商用利用・派生利用無制限) ¹
-------	--

疎結合MoEと周期的ハイブリッドアテンション

Kolibriのネットワーク設計は、各層に384個の微小エキスパート(隠れ層次元512)を配置し、トークンごとに常に作動する1つの共有エキスパート(Shared Expert)と、動的ルーティングにより選定される6つのエキスパートによって計算を分担する⁵。これにより、トークンあたりの有効パラメータ数を総量のわずか約4.4%にあたる3.46Bに抑え込み、小型モデル並みのデコードスループットを実現している¹。

長文文脈を支えるアテンション機構には、独自の周期的ハイブリッド設計が組み込まれた⁵。4層のスライディングウィンドウアテンション(窓幅512トークン、Rotary Position Embedding適用)に対して、5層ごとに全系列を俯瞰するフルアテンション層(位置埋め込み非適用)を1層挟む「4:1」の比率を全50層にわたり繰り返す⁵。この構造により、超長文処理時におけるKey-Value(KV)キャッシュの肥大化を抑えつつ、262,144トークンのネイティブ学習長と、位置スケールリングを介さない最大100万トークン(1,048,576トークン)への理論的コンテキスト拡張を達成している⁵。

形態素構造に適合する独自トークナイザ「UniBPE」

膠着語的性質や名詞の複合化が著しいドイツ語では、従来のByte-Pair Encoding(BPE)を用いると単語が無秩序に微細断片化され、トークン数の増大に伴うレイテンシの悪化とコンテキスト消費の加速が避けられなかった²。これに対しAleph Alphaは、BPEのボトムアップなマージ探索にUnigramの損失最適化を統合したハイブリッドトークナイザ「UniBPE」を設計した²。語彙数128,000の辞書上で言語的形態素を保ったトークン分割を行うことで、ドイツ語において4.90バイト/トークンという高いテキスト圧縮率を達成している²。これにより、数十ページに及ぶ行政通知や法的決定通知(Bescheid)の処理コストを抑制し、長大文書の実用的な推論スループットを底上げしている²。

ハルシネーションを理論的に抑止する「Merlin-Arthurプロトコル」

公共行政や産業分野でのAI導入を阻む最大の要因である虚偽生成(ハルシネーション)に対し、Kolibriは情報理論的な境界付けを行う「Merlin-Arthurプロトコル」を事後学習に採用した⁹。この仕組みでは、正解を導く文脈を提示する証明者役(Merlin)と、正解の根拠となる情報のみを周到に除去した文脈を提示する対抗者役(Morgana)を用意し、モデル本体(Arthur)を訓練する²³。モデルは、与えられた文脈内に入力質問を裏付ける客観的証拠が存在しない場合、無理に推測を行わず明示的に「回答を棄権(Abstain)」する振る舞いを獲得する⁹。この学習手法により、事実想起評価である「AA-Omniscience」において、旧世代モデルの15%から44%へと棄権成功率を向上させ、検索拡張生成(RAG)環境における接地性(Grounding)を高めている²³。

ベンチマーク性能と実務運用上のトレードオフ

Kolibriは、特定の業務負荷に極めて明確な強みと弱みを持つプロファイルを示している²。同一の推論速度帯(アクティブ3B規模)のモデルと比較した場合、競技数学、ドイツ語および英語の多段論理推論、エージェント的コーディングにおいてクラストップの数値を記録する一方、世界知識を問うクローズドブック想起では密結合型の大規模モデルに劣後する²。

ベンチマーク指標	Kolibri-1 (78B-A3.46B)	Qwen3.6 35B-A3B	Nemotron 3 Super 120B-A12B
総合ベンチマーク平均(英語)	75.5 ²	71.4 ¹²	68.0 ²
総合ベンチマーク平均(ドイツ語)	70.8 ²	67.3 ¹²	67.9 ²
GPQA Diamond (高度科学推論・英語)	84.3 ¹²	83.4 ²¹	N/A
GPQA Diamond (高度科学推論・独語)	81.3 ¹²	84.2 ¹²	N/A
Humanity's Last Exam(英語難問)	21.5 ¹²	N/A	N/A
LiveCodeBench v6 (コード生成)	85.9 ¹⁴	N/A	N/A
SWE-Bench Verified(実バグ改修)	66.4 ¹⁴	N/A	N/A
MMLU / 知識想起 (ドイツ語)	77.5 ¹²	上位(高スコア傾向) ²	上位(高スコア傾向) ²

実務配備の観点において、Kolibriは4段階の推論深度(Reasoning Effort: none, low, medium, high)やHermes互換の外部ツール呼び出し(APIオーケストレーション)を同一モデル内で動的に制御できる柔軟性を備えている¹。しかし、運用計画上織り込むべき物理的制約も明確である²。アクティブパラメータが3.46Bであっても、モデル全体の重みとしてFP8精度で約78GBのVRAMを確保する必要があり、動作には最低でもH100 SXM(80GB)2基や単一のH200(141GB)といったデータセンター向けGPUクラスタが要求される²。また、標榜されている100万トークンはアテンションの理論的接続に基づく外挿であり、ネイティブ学習境界である262,144トークンを超えると長文情報抽出(RULER評価等)の精度が低下するため、Aleph Alpha自身も高精度が求められる業務処理では262k以下での運用を推奨している¹。

カナダCohere社との事業統合がもたらす地政学的変容

Kolibriの公開に先立つ2026年9月16日、Aleph AlphaはカナダのエンタープライズAI有力企業Cohere社との対等事業統合契約に最終合意した¹⁰。この動きは、一国のスタートアップ単独では耐えきれなくなった計算インフラ投資の負担を克服し、大西洋を跨ぐ新たな「主権AI連合」を形成する構造転換を意味している¹⁰。

統合組織はトロントとベルリンに本社を置く2本社体制をとり、グローバルブランドをCohereに統一する一方で、Aleph Alphaの本拠地であるハイデルベルクを中核研究拠点として維持する体制を敷いた¹⁰。Aleph Alphaの共同CEOを務めたIlhan Scheer氏が最高執行責任者(COO)に、共同創業者のSamuel Weinbach氏が最高研究責任者(CRO)に就任するなど、経営・研究の両輪で欧州側の主導権が組み込まれている¹⁰。

この統合の背後には、ドイツの流通・ITコングロメリットであるシュワルツ・グループ(Schwarz Group)による大規模な資本・物理インフラの裏付けが存在する¹⁰。シュワルツ・グループは統合会社へ5億ユーロを出資するとともに、傘下のソプリクラウド「STACKIT」を通じて110億～130億ユーロを投じ、最大10万個のAIアクセラレータを収容する独国内データセンターを構築している¹⁰。Kolibriの公開形態においてパブリックAPIの従量課金メニューが設けられず、オープンウェイトの配布と自社運用環境の提供に特化している背景には、STACKIT基盤を軸とした規制産業向けプライベートホスティングへの誘導というビジネス戦略が連動している¹。

日本市場および国内AIエコシステムへの波及効果と戦略的示唆

Kolibri自体は英語とドイツ語に特化した二言語モデルであり、学習コーパスに日本語が含まれていないため、国内の言語処理タスクに本モデルがそのまま直接配備される可能性は低い²。しかし、その「アライアンスの波及」「アーキテクチャの示唆」「主権政策の実装モデル」という3つの側面において、日本のAIエコシステムに対して極めて重大な影響と示唆を与えている。

富士通を中心とする国内エンタープライズAI基盤への直接的な技術還流はその最たる例である¹⁰。富士通は2024年7月にCohereと戦略的提携を結んで出資を行い、Cohereの多言語モデルを基盤とした日本語特化型LLM「Takane(高嶺)」を共同開発してグローバル独占提供権を保持している¹⁰。今回のAleph AlphaとCohereの事業統合により、日本の重要インフラや大手企業に導入されているTakaneの背後にある開発体制が、欧州の主権AIの中核拠点と直接結びつく構図が成立した¹⁰。Aleph AlphaがKolibriで培ったUniBPEIによる形態素最適化、周期的アテンションによるKVキャッシュの圧縮、そしてMerlin-ArthurプロトコルによるRAGのハルシネーション抑制機構は、Cohereの基盤技術を通じてTakaneや富士通の企業向けソリューションへ迅速に還元される蓋然性が高い²。これは、国内ベンダーが米系ハイパースケーラーのクラウドAPIに対抗し、高度な説明責任と接地性を兼ね備えた企業向けAIを展開する上で強力な差別化要素となる¹⁰。

経済産業省とNEDOが進める「GENIAC(Generative AI Accelerator Challenge)」や、デジタル庁のガバメントAI「源内」など、日本独自の基盤モデル構築プロジェクトに対しても、Kolibriの構造は明快な設計指針を提示している¹¹。これまで国内で開発されてきた独自LLMの多くは数十億から千億パラメータ規模の密結合(Dense)モデルが主流であり、推論時のGPU計算資源と消費電力が運用上のボトルネックとなってきた¹¹。これに対しKolibriが実証した「総量78Bの表現力を確保しつつ、トークンあたり3.46Bのみを駆動させる」極小エキスパートMoEの設計は、限られた計算インフラで最大のコ

ストパフォーマンスを引き出す現実解を示している¹。さらに、膠着語であり複合名詞や助詞が連続する日本語のトークナイズ処理において、BPEの語彙爆発を防ぎつつ意味単位を保つUniBPEの知見は、日本語LLMの推論効率とコンテキスト圧縮率を改善するための有力な技術的参照点となる²。日本政府が人工知能基本計画において提唱する「開かれたAI主権」の制度設計においても、Kolibriの運用モデルは現実的な青写真を提供している¹¹。日本が目指す主権AIは、外国技術を排除する完全な自給自足ではなく、有事や供給停止時にも自律的代替を可能にする相互運用性の確保を前提としている¹¹。Kolibriが示した、完全なオープンウェイト形式によるエアギャップ(閉域網)運用、学習パイプラインの完全な自動化(Model Factory)、そしてEU AI法のGPAI規範に準拠した厳格なデータ追跡文書の開示は、日本の防衛、金融、重要インフラ領域における公共調達基準やAIガバナンスガイドラインを策定する上で、実質的な世界標準のモデルケースとして機能する¹。

結論

Aleph Alphaが2026年10月3日に発表した「Kolibri」は、単なる特定言語モデルの性能更新にとどまらず、巨大クラウドベンダーの閉鎖的APIモデルに対峙する「欧州発・主権型オープンウェイトモデルの成熟形」を示した²。3.46Bアクティブパラメータによる低推論コスト、UniBPEによる高い形態素圧縮、周期的アテンションによる長文耐性、そしてMerlin-Arthurプロトコルによる接地性の担保は、実務特化型AIの設計思想として極めて洗練されている¹。

Cohereとの統合を通じて形成された大西洋主権AI連合は、富士通の「Takane」を介してすでに日本のエンタープライズ市場と構造的に接続されている¹⁰。日本が自国のデータと計算基盤の主権を確保しつつ実用的なAI実装を加速させる上で、Kolibriが打ち立てた高効率MoEアーキテクチャ、厳格な監査適合性、そして閉域自社運用の枠組みは、技術と政策の双方において追従すべき不可欠な道標となる²。

引用文献

1. Kolibri: Aleph Alphaの78Bオープンウェイトモデルを解説 - OrcaRouter, <https://www.orcarouter.ai/ja/blog/kolibri-release-explained>
2. Kolibri 1: Aleph AlphaのソブリンなオープンウェイトLLM - DataCamp, <https://www.datacamp.com/ja/blog/aleph-alpha-kolibri-1>
3. Aleph Alpha's Kolibri: what the Apache 2.0 license guarantees - ActuaIA, <https://www.actuia.com/en/news/aleph-alphas-kolibri-what-the-apache-20-license-guarantees/>
4. Aleph Alpha Kolibri: 78B Open-Weight German MoE (Apache 2.0), <https://www.explainx.ai/blog/aleph-alpha-kolibri-sovereign-open-weight-moe-german-english-2026>
5. <https://aleph-alpha.com/en/blog/kolibri-has-landed-a-sovereign-open-weight-model/>
6. Aleph Alpha Kolibri、最先端規模よりもソブリンAIの制御を優先, <https://www.remio.ai/ja/post/aleph-alpha-kolibri-puts-sovereign-ai-control-ahead-of-frontier-scale-ja>
7. Germany's Aleph Alpha releases Kolibri, a sovereign AI model for, <https://www.eu-startups.com/2026/10/germanys-aleph-alpha-releases-kolibri-a-sovereign-ai-model-for-mission-critical-government-and-industry-applications/>

8. Aleph AlphaのKolibriモデル、外国AIへの依存に対抗するドイツの、
<https://www.remio.ai/ja/post/aleph-alpha-kolibri-model-puts-german-sovereignty-against-foreign-ai-dependen-ja>
9. <https://gigazine.net/news/20261006-kolibri/>
10. 富士通も出資のCohereが独Aleph Alphaと統合 - ビジネス+IT,
<https://www.sbbt.jp/article/cont1/187003>
11. ソブリンAIとは？日本の現在地を公的資料で確認【2026】 - Mihata,
<https://mihata.jp/column/sovereign-ai-japan>
12. Aleph-Alpha/Kolibri-1 - Hugging Face,
<https://huggingface.co/Aleph-Alpha/Kolibri-1>
13. Kolibri – Tech Report [pdf] - Hacker News,
<https://news.ycombinator.com/item?id=49946069>
14. Aleph Alpha open-sources Kolibri-1: 78B MoE (3.46B ... - AICoder,
<https://aicoder.com/news/news-20261004-aleph-alpha-kolibri-1-open-weight>
15. Aleph Alpha Becomes Cohere: What Is Left of Sovereign AI,
<https://everbright-it.de/en/blog/aleph-alpha-cohere-deal-sovereign-ai/>
16. Cohere, Aleph Alpha combine to target enterprise AI market | live,
<https://live.euronext.com/en/financial-news/cohere-aleph-alpha-combine-target-enterprise-ai-market>
17. Cohere Signs Definitive Combination Agreement with Aleph Alpha,
<https://www.unite.ai/cohere-signs-definitive-combination-agreement-with-aleph-alpha/>
18. Aleph Alphaとは？意味をわかりやすく解説 - AI経営手帖,
<https://keieiax.jp/glossary/aleph-alpha/>
19. (PDF) Building Europe's Frontier AI Kolibri-1 -the Full 78B Mixture-of-
https://www.researchgate.net/publication/415363162_Building_Europe's_Frontier_AI_Kolibri-1_-the_Full_78B_Mixture-of-Experts_Language_Model
20. Aleph Alpha Releases Kolibri: A 78.1B Open-Weight English,
<https://www.marktechpost.com/2026/10/04/aleph-alpha-releases-kolibri-a-78-1b-open-weight-english-german-moe-model-with-only-3-46b-active-parameters/>
21. Kolibri 1: Aleph Alpha's open-weight AI model | Synthex Prompts AI,
<https://www.synthexprompts.ai/blog/kolibri-1-aleph-alpha>
22. Kolibri-1 必要スペック早見表 - 株式会社オブライト,
<https://www.oflight.co.jp/ja/columns/kolibri-1-aleph-alpha-requirements-vram-local-2026>
23. First German LLM : Kolibri-1 - Medium,
<https://medium.com/data-science-in-your-pocket/first-german-llm-kolibri-1-1d4368a1a649>
24. Merlin-Arthur Protocols for Mutual-Information Bounds in Language,
https://www.reddit.com/r/TecnologiaArtifziala/comments/1wwga6m/bounding_hallucinations_merlinarthur_protocols/
25. We trained Kolibri with abstention data and with our Merlin-Arthur,
<https://news.ycombinator.com/item?id=49946138>
26. Merlin-Arthur Training and Evaluation EXPLAINED - YouTube,
<https://www.youtube.com/watch?v=796GVFTFiB0>

27. New MoE Model - Aleph-Alpha Kolibri-1 78B-A3.5B : r/LocalLLM,
https://www.reddit.com/r/LocalLLM/comments/1wxjuim/new_moe_model_aleph_alpha_kolibri1_78ba35b/
28. Cohere Acquires Aleph Alpha: Deal Born of Sovereignty, Necessity,
<https://futurumgroup.com/insights/cohere-acquires-aleph-alpha-a-deal-born-of-sovereignty-necessity/>
29. Cohere and Aleph Alpha Sign Agreement to Become the First,
<https://aleph-alpha.com/en/news/cohere-agreement-transatlantic-sovereign-ai/>
30. Cohere、エージェントAI「Command A+」をオープンソース化,
<https://innovatopia.jp/ai/ai-news/104815/>
31. 富士通とCohere 日本語強化版LLM「Takane(高嶺)」を共同開発,
https://ledge.ai/articles/fujitsu_cohere_partnership
32. Takane,
<https://docs.fujitsu/documents/003331/fujitsu-generative-ai-for-cohere-catalog-jp.pdf>
33. How Fujitsu uses Cohere | UsedBy.ai,
<https://www.usedby.ai/case-studies/how-fujitsu-uses-cohere-ba861d8c>
34. ソブリンAI・国産LLMとは | データ主権とAI選び | 株式会社Sei San Sei,
<https://www.sei-san-sei.com/blog/blog-0423.html>
35. 国産AIはどのレベルにあるのか？政府調達データで読む国際AI事情,
<https://www.seo-research.co.jp/japan-domestic-ai-level-2026/>
36. 国産LLMにはどんなものがある？主要モデルの比較と企業の選び方,
<https://exawizards.com/column/article/ai-article/japanese-domestic-llm/>
37. 週刊 | 話題のAI新製品・新機能ニュース(2026/10/4～10/10号) - note,
<https://note.com/yasuhitoo/n/n0e6f5d55c206>
38. 日本発の「ソブリンAI」として独自開発の日本語LLM「Stockmark-2」,
<https://stockmark.co.jp/r-and-d/topics/20250925/>