2026年 日本国産AIモデル（LLM）の到達点と未来展望：主権型AIの確立と産業実装の深化に関する包括的調査報告書

Gemini 3 pro

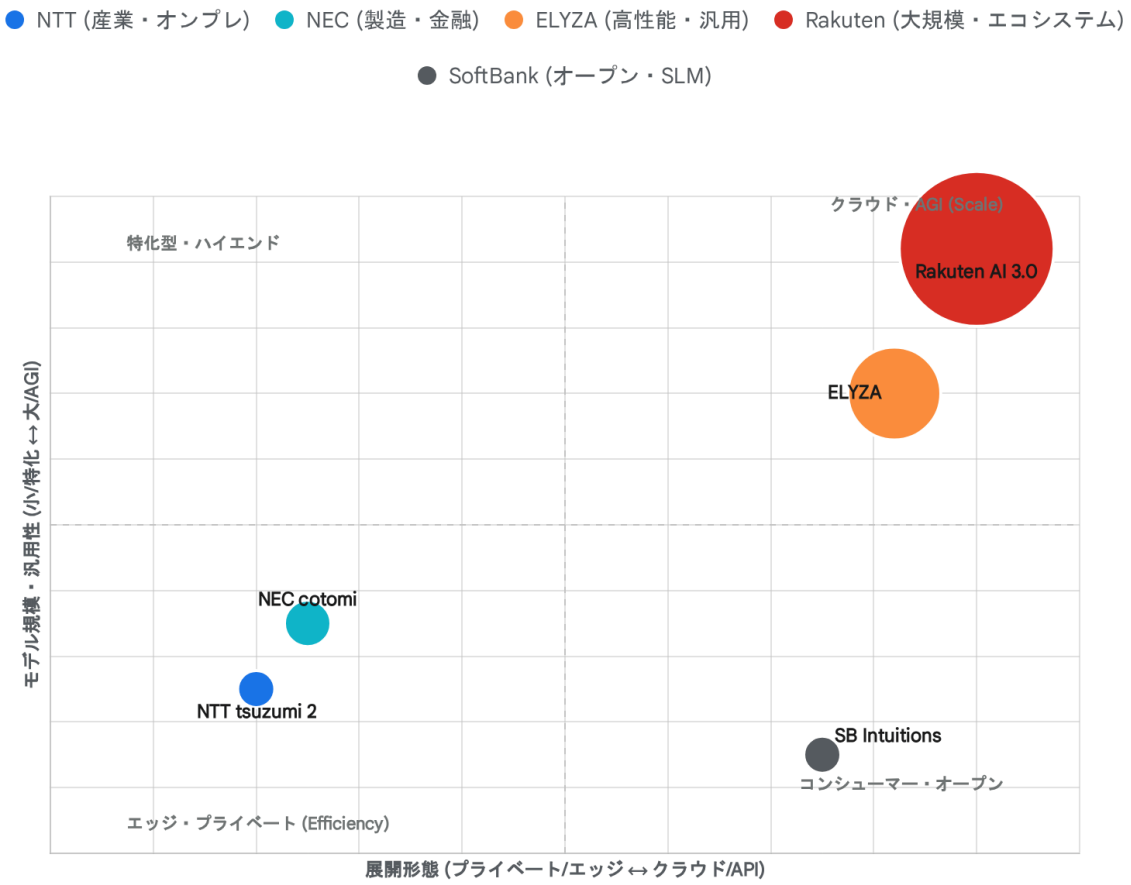
1. エグゼクティブサマリー

2026年1月現在、日本の人工知能（AI）産業、とりわけ大規模言語モデル（LLM）の開発競争は、かつてないほどの活況と構造的な転換期を迎えている。2023年から2024年にかけての「キャッチアップ期」を経て、2025年は各社が独自の強みを明確化した「差別化と実装の年」として位置づけられた。OpenAIやGoogleといった米国の巨大テック企業が汎用的な超巨大モデル（AGI）への道を突き進む一方で、日本の主要プレイヤーであるNTT、NEC、ソフトバンク、KDDI（ELYZA）、楽天などは、日本の商習慣、言語的ニュアンス、そしてデータ主権（Sovereign AI）という不可避なニーズに応える形で、独自の生態系を確立しつつある。

本報告書では、これら主要な国産モデルの現状（2025年後半のリリースを含む最新状況）を詳細に分析し、2026年以降の展望を予測する。特筆すべきは、各社が単なるパラメータ数の競争から脱却し、明確に異なる戦略軌道を描き始めた点である。NTTの「tsuzumi」は「軽量化・省電力化」を突き詰め、IOWN構想との融合を図る。NECの「cotomi」は「業種特化と高速推論」に活路を見出し、エージェント技術による業務代行へシフトしている。一方で、ソフトバンクの「Sarashina」や楽天の「Rakuten AI」は、圧倒的な計算資源を背景にグローバル水準の「規模」を追求し、超巨大MoEアーキテクチャへの挑戦を行っている。

また、経済産業省およびNEDOが主導する「GENIAC（Generative AI Accelerator Challenge）」プロジェクトが第3期を迎え、計算資源の提供が実用レベルの基盤モデル創出に寄与している事実も見逃せない。2026年は、これら「国産AI群」が、実証実験（PoC）のフェーズを完全に脱却し、金融、製造、自治体、医療といったミッションクリティカルな領域で本格稼働する「AI社会実装元年」となることが確実視される。

2026年 国産LLM 戦略ポジショニングマップ



各社の主力モデルの戦略的位置づけ（2026年1月時点）。Y軸はモデルのパラメータ規模と汎用性、X軸は展開形態（オンプレミス・エッジ志向か、クラウド・大規模API志向か）を示す。NTTやNECが産業特化・オンプレミス領域に注力する一方、ソフトバンクや楽天は大規模クラウド型モデルでグローバル水準を追随している。

Data sources: NTT R&D, TechPlay, DeBoNo, Ledge.ai, Note (ELYZA), Rakuten Group

本稿では、技術的仕様、ビジネスモデル、ベンチマーク性能、そして将来のロードマップを網羅的に紐解き、日本発のAIがグローバル市場の中でどのような立ち位置を築こうとしているのかを論じる。

2. 戦略的背景：なぜ今、「国産」なのか

2.1 データ主権と経済安全保障の要請

2023年の生成AIブーム初期、日本企業はOpenAI等の海外製モデルの利用に殺到した。しかし、2025年を経て2026年に至る現在、市場は明確な「使い分け」のフェーズに移行している。その最大の要因は、地政学的リスクとデータ主権（Sovereign AI）への意識の高まりである。金融機関の顧客

データ、医療カルテ、行政の住民情報、あるいは製造業の技術ノウハウといった機密性の高い「コアデータ」が、海外のサーバーに送信されることへの潜在的なリスク懸念は払拭されていない。特に、米国のAI規制やポリシー変更が日本企業のサービス継続性に直結する「カントリーリスク」は、経営層にとって無視できない課題となっている[1]。

内閣府や経済産業省が主導するAI戦略会議においても、重要インフラや機密情報を扱うシステムには、学習から推論までを国内で完結できる「純国産モデル」の採用が推奨されている。これが、NTTやNECが推進する「オンプレミス」や「クローズド環境」での提供モデルを強力に後押しするドライバとなっている。また、ソフトバンクが強調するように、通信インフラデータなどの国家的に重要なデータを国内で処理し、その知見を国内に還流させるサイクル（データエコシステム）の構築は、経済安全保障の観点からも不可欠とされている[4]。

2.2 日本語処理能力と文化的適合性の深化

海外製モデル（GPT-4oやClaude 3.5等）の日本語能力は飛躍的に向上しており、日常会話レベルでは遜色ない性能を発揮する。しかし、日本のビジネス現場における要求水準は、単なる翻訳や会話の成立を超えた領域にある。
例えば、稟議書の作成における特有の構成、取引先へのメールにおける微妙な敬語の使い分け（尊敬語・謙譲語・丁寧語の適切な混合）、さらには日本の法令やコンプライアンス基準（著作権法、景品表示法など）に準拠した出力においては、依然として国産モデルに一日の長がある。
2025年にリリースされた各社の最新モデル（tsuzumi 2, cotomi V2, Sarashina 2など）は、JGLUEやRakuda、MT-Bench (Japanese) といった日本語ベンチマークにおいて、パラメータ数で勝る海外モデルと同等以上のスコアを記録している[5]。これは、学習データの質とキュレーションにおける優位性を示している。日本のウェブデータ、書籍、公文書、判例などを網羅的に、かつ高品質に学習させることで、「日本人が違和感を持たない」だけでなく、「日本の業務にそのまま使える」レベルの出力を実現しているのである。

2.3 コスト構造とエネルギー効率の最適化

円安基調が続く経済環境下において、ドル建てのAPIコストは日本企業にとって無視できない負担となっている。また、学習・推論にかかる膨大な電力消費は、世界的な環境課題であると同時に、運用コストを押し上げる直接的な要因である。
これに対し、日本の開発勢は「パラメータ数を抑えつつ高性能を出す」という方向性で差別化を図っている。特にNTTの「tsuzumi」が提唱する「1GPUで動作する高性能モデル」は、高価なH100 GPUを大量に調達できない中小規模の企業や自治体にとって、現実的な導入の選択肢となっている8。計算資源の制約が厳しい日本だからこそ生まれた「軽量化技術」は、エッジAIの普及とともに、グローバルでも通用する独自の競争力となりつつある。

3. 主要プレイヤー別詳細分析

3.1 NTT「tsuzumi」：IOWN構想と連動する「軽量・高性能」の極致

3.1.1 技術的特徴と進化：tsuzumi 2の完成度

日本電信電話（NTT）は、長年の自然言語処理研究の蓄積を背景に、2023年11月に「tsuzumi」を発

表した。その後、2025年10月20日には、その性能を飛躍的に高めた最新版「tsuzumi 2」をリリースし、市場への浸透を加速させている[10]。

パラメータ戦略と軽量性

tsuzumiの設計思想の核心は「軽量性」にある。tsuzumi 2は、主に70億（7B）パラメータクラスと、さらに軽量な超軽量版（0.6B）で構成されており、依然として「1枚のGPU（コンシューマー向けハイエンドまたはエントリーサーバー向けGPU）で動作可能」という制約条件をクリアしている8。これは、数千億から1兆パラメータ規模を持つGPT-4等の巨大モデルとは対照的であり、計算リソースが限られたオンプレミス環境や、秘匿性の高い閉域網内、さらにはエッジデバイスへの搭載を前提とした戦略的なサイズ設定である。

2025年10月のリリースでは、約30億（30B）パラメータの中規模モデルの準備も進められており、推論性能とコストのバランスにおける選択肢を拡充している12。

性能評価と日本語能力

NTTのヒューマンインフォマティクス研究所による評価およびMT-Bench（Turn 1, Japanese）の結果において、tsuzumi 2はGPT-5（または同世代の競合モデル）に肉薄する高いスコアを記録している8。特に、日本企業の実務で多用されるRAG（検索拡張生成）を用いた文書要約や情報抽出といったタスクにおいて、独自の評価セットで顕著な性能向上を示している。

これは、パラメータ数という「量」ではなく、学習データの「質」を徹底的に磨き上げた成果である。

NTTは、日本語特有の文脈理解や、指示追従性（Instruction Following）の強化に注力しており、例えば「JSON形式で出力せよ」「特定のフォーマットで要約せよ」といったシステム連携に不可欠な指示に対しても、高い精度で応答することが可能となっている10。

マルチモーダル化とアダプタ技術

tsuzumi 2におけるもう一つの重要な進化は、マルチモーダル対応の強化である。テキスト情報だけでなく、図表やグラフを含むドキュメント全体の構造を理解する能力が向上した13。これにより、仕様書やマニュアル、請求書といった非定型ドキュメントの読み取りと理解が可能となり、BPO（ビジネス・プロセス・アウトソーシング）やバックオフィス業務の自動化に貢献している。

また、これを支えるのが「アダプタ技術（Adapter Tuning）」である。巨大なベースモデル全体を再学習させるのではなく、特定のタスクや業界知識に対応した軽量な「アダプタ」モジュールを追加・交換するだけで、専門性の高いモデルへと変身させる技術だ。これにより、顧客ごとのカスタマイズコストを劇的に低減させている。

3.1.2 導入事例とビジネスモデル

2026年1月現在、tsuzumiの導入は「機密性」と「カスタマイズ性」を重視するセクターで加速している。

- 東京通信大学: 学生・教職員の学習データや個人情報を学内ネットワークに留めるという厳格な要件に対し、オンプレミス環境で動作するtsuzumi 2が採用された。クラウドへのデータ流出リスクを完全に排除したセキュアな教育AI基盤として稼働している[14]。
- 金融・自治体: NTTデータなどのグループ企業を通じて、金融機関や自治体への導入が進んでいる。例えば、ファイナンシャルプランニング技能試験2級レベルの専門知識を追加学習させる際、他社モデルと比較して圧倒的に少ないデータ量と計算リソースで合格基準に到達できることが実証されており、業界特化型モデルの構築プラットフォームとして選定されている[16]。

3.1.3 2026年以降のロードマップ：AIコンステレーションへ

NTTのAI戦略は、単体のLLMビジネスに留まらない。同社が全社を挙げて推進する次世代通信基盤「IOWN（Innovative Optical and Wireless Network）」構想と不可分に結合している。

- **AIコンステレーション（AI Constellation）**：今後、tsuzumiは単体で動作するだけでなく、IOWNの光ネットワークで結ばれた多数の「tsuzumi」が連携し、あたかも一つの巨大な知能のように振る舞う「AIコンステレーション」へと進化していく[17]。これは、個々のモデルは軽量でも、それらが高速な光通信で協調することで、巨大モデルに匹敵する、あるいはそれを凌駕する問題解決能力を持たせるという野心的なアーキテクチャである。
- 省電力化の追求：2026年から2027年にかけては、IOWNの中核技術である光電融合デバイスを活用した、さらに低消費電力な推論チップ上での動作検証が進むと見られる。データセンターの電力消費量が爆発的に増大する中、NTTは「光で動くAI」を武器に、グリーンAIの市場リーダーとしての地位を確立しようとしている。

3.2 NEC「cotomi」：SIerの強みを生かした「現場実装力」

3.2.1 技術的特徴：スピードスターとしての進化

NECの生成AI「cotomi」は、2023年の発表以来、実用性を最優先した進化を続けている。2025年後半には、機能強化版（cotomi V2相当）が登場し、特に「推論速度」において世界トップクラスの性能を打ち出した[5]。

圧倒的な速度とCotomi Fast
NECは「cotomi Fast v2」において、GPT-4oと比較して2.2倍以上の推論速度を実現したと発表している2。コンタクトセンターのオペレーター支援や、製造ラインでのリアルタイムトラブル対応など、人間の思考を妨げない即応性（低レイテンシ）が求められる現場において、この速度差は決定的な競争優位となる。単に速いだけでなく、標準的なGPUサーバー2台でGPT-4よりも87%～93%高速に動作するというインフラ効率の高さも特徴であり19、大量のリクエストをさばく大規模システムでの運用コスト削減に寄与する。
ロングコンテキストとMCP対応
実務対応力の強化として、入力トークン長が128k（約20万文字）まで拡張された2。これにより、数百ページに及ぶ社内マニュアル、契約書、仕様書などを分割することなく一度に読み込ませることが可能となり、文書全体を踏まえた矛盾のない回答生成や、複雑な条件抽出が可能となった。
さらに、2025年の重要な技術的マイルストーンとして、AIエージェントと外部システムを標準化された手順で接続する「MCP（Model Context Protocol）」への対応が挙げられる20。これにより、cotomiが社内データベース、SaaS、APIと連携し、情報の検索だけでなく、タスクの実行までを自律的に行う能力が飛躍的に向上した。

3.2.2 業種別特化戦略とcotomi Act

NECは、汎用的な基盤モデルを提供するだけでなく、同社が長年培ってきたSI（システムインテグレーション）の知見を活かし、特定業種向けにカスタマイズした「特化型モデル」の展開に注力している。

- 製造業モデル：設計図面や部品表（BOM）の構造を深く理解し、熟練技術者のノウハウ継承や、設計変更時の影響範囲分析、トラブルシューティングを支援するモデルを開発・提供してい

る[21]。

- 金融・自治体: 金融商品取引法などの法令知識や、自治体の行政手続きに関する知識を学習させたモデルを展開。複雑な金融商品の説明支援や、申請書類の自動作成支援において、高いシェア獲得を狙っている[5]。

さらに、2026年1月からは**「cotomi Act」**と呼ばれる新ソリューションの提供を開始する[23]。これは、従来の「人がAIに質問して答えを得る」スタイルを超え、AIが「業務ノウハウを自動的に抽出し、組織全体で共有・活用できる資産として蓄積する」ためのエージェント技術である。例えば、優秀な営業担当者のメール対応や提案書作成のプロセスをAIが学習し、それを他の社員が再利用できる形で提示するといった、組織知の形式知化を自動化する狙いがある。

3.2.3 今後の展望：500億円事業への道

NECは2025年度末までに生成AI関連事業で500億円の売上を達成するという野心的な目標を掲げている[24]。これを実現するために、NECは単なるLLMベンダーではなく、ハードウェア（Express5800シリーズ等）、ソフトウェア、コンサルティング、運用保守を一気通貫で提供する「AIトータルソリューションプロバイダー」としての立ち位置を強化している[25]。2026年は、cotomi Actによる自律型エージェントの実装が本格化し、国内エンタープライズ市場においてMicrosoft（Azure OpenAI）に対抗しうる強力な選択肢としての地位を固める年となるだろう。

3.3 SoftBank (SB Intuitions)「Sarashina」：圧倒的資本による「AGI」への挑戦

3.3.1 技術的特徴：真正面からのスケーリング

ソフトバンクグループのAI戦略の中核を担う子会社、SB Intuitionsは、国内他社とは一線を画す「規模の追求」を行っている。国内最大級の計算基盤（NVIDIA H100等を数千基規模で配備）を背景に、パラメータ数の大きい、真の意味での「基盤モデル」をゼロから開発している。

- **Sarashina2**シリーズ: 2024年に公開された「Sarashina1」に続き、2025年には「Sarashina2」シリーズを展開している。特に注目されるのは、パラメータ数を巨大化させつつ推論効率を維持するためにMoE（Mixture of Experts）技術を採用した「Sarashina2-8x70B」などのモデル群である[26]。実質的な知識量は数千億パラメータ規模に達しており、汎用的な言語能力においてグローバルモデルを追随している。
- **Sarashina mini**: 一方で、商用利用のしやすさを考慮した700億（70B）パラメータクラスの「Sarashina mini」も2025年度中に商用提供が開始される予定であり[26]、高性能とコストのバランスを重視する法人顧客への浸透を図っている。
- マルチモーダル**VLM**: 2025年11月には、視覚情報を理解するVLM（Vision-Language Model）である「Sarashina2.2-Vision-3B」を公開した[28]。これは、画像認識と日本語処理を高度に統合したモデルであり、ロボティクスや監視システムへの応用が期待される。

3.3.2 孫正義氏の「ASI」ビジョンと2026年の位置づけ

ソフトバンクのAI戦略を理解する上で、孫正義CEOが掲げるビジョンは不可欠である。孫氏は「ASI（人工超知能）は2035年に実現し、人間の1万倍賢くなる」と予言しており[29]、Sarashinaはそのための

通過点に過ぎない。ソフトバンクにとっての国産LLMは、単なるチャットボット用エンジンではなく、グループ全体の未来を支える「共通知能基盤」である。

- **AI-RANとの統合**: ソフトバンクが世界に先駆けて推進する「AI-RAN（Radio Access Network）」構想において、Sarashinaは極めて重要な役割を果たす。AI-RANは、通信基地局に設置された高性能サーバーで、無線通信の信号処理とAI推論を同時に行う技術である。2025年9月の実証実験では、Sarashinaを組み込んだモデルが通信品質を90%以上の精度で予測することに成功している[4]。2026年以降、全国の基地局がそのまま「巨大な分散AIコンピュータ」となり、低遅延で高度なAIサービスを提供するインフラへと変貌する。
- **データエコシステムの構築**: LINEヤフー、PayPay、ソフトバンク通信事業など、グループが持つ膨大なユーザー接点から得られるデータを、国内のデータセンターで安全に学習し、その知見をSarashinaに還流させる。この「学習と還流」のサイクルを国内で完結させることは、経済安全保障上の要請に応えるとともに、AGI開発に必要な「データの量と質」を確保する戦略でもある[4]。

3.3.3 今後の展望：1兆パラメータへの道

2026年以降、ソフトバンクは計算資源のさらなる増強（北海道苫小牧などの巨大データセンター稼働）を進め、「1兆（1T）パラメータ級」のモデル開発を視野に入れるだろう。OpenAIへの巨額投資と連携を維持しつつも、自社でコントロール可能なコア技術としてのSarashinaを育成し、両者をハイブリッドで活用する戦略を採ると見られる。

3.4 ELYZA (KDDIグループ)：実用主義と「Smart Adapter」戦略

3.4.1 技術的特徴：最強の「日本語化」技術

東京大学松尾研発のスタートアップであり、現在はKDDIグループ傘下にあるELYZAは、ゼロからモデルを構築するのではなく、世界最高性能のオープンモデル（Meta社のLlamaシリーズ等）をベースに、独自の高度な事後学習（Post-training）を施すことで、効率的に高性能なモデルを開発する戦略を採っている。

- **Llama-3-ELYZA-JP**: Meta社のLlama 3をベースにした70Bクラスのモデルは、JGLUEやMT-Benchといった日本語ベンチマークにおいて、GPT-4等の商用クローズドモデルに匹敵、あるいは一部で上回るスコアを記録し、「日本語ならELYZA」というブランドを確立した[6]。英語圏の膨大な知識をベースモデルから継承しつつ、日本固有の言語的・文化的ニュアンスをファインチューニングで注入するこの手法は、開発スピードと性能のバランスにおいて極めて合理的である。
- **Smart Adapter戦略**: ELYZAもまた、汎用モデルに加え、特定タスクに特化した軽量モデルやアダプタの開発に注力している。これにより、計算コストを抑えながら、企業の個別ニーズに対応する柔軟性を持たせている。

3.4.2 ビジネス展開：WAKONXとSwing-by IPO

KDDIグループ入りしたことで、ELYZAの技術はKDDIの広範な顧客基盤と計算インフラを得て、社会実装のフェーズに入っている。

- **WAKONX**（ワコンクロス）: KDDIが2024年に始動したAIビジネスプラットフォーム「WAKONX」に

おいて、ELYZAのモデルは中核エンジンとして位置づけられている[30]。2026年は、このプラットフォーム上で、コンタクトセンター（アルティウスリンク等での活用）、法務、人事など、特定の業務ドメインに特化した「Micro-LLM」や「Agent」のラインナップが急速に拡充されるだろう。

- **Swing-by IPO**: ELYZAは、KDDIの大企業としてのリソースを活用して成長を加速させ（スイングバイ）、将来的には単独での上場（IPO）を目指すという独自の成長戦略「Swing-by IPO」を掲げている[30]。2026年はその実現に向けた事業基盤の確立期となる。

3.5 楽天（Rakuten）：「Rakuten AI 3.0」とオープンウェイト戦略

3.5.1 Rakuten AI 3.0の衝撃：7000億パラメータ

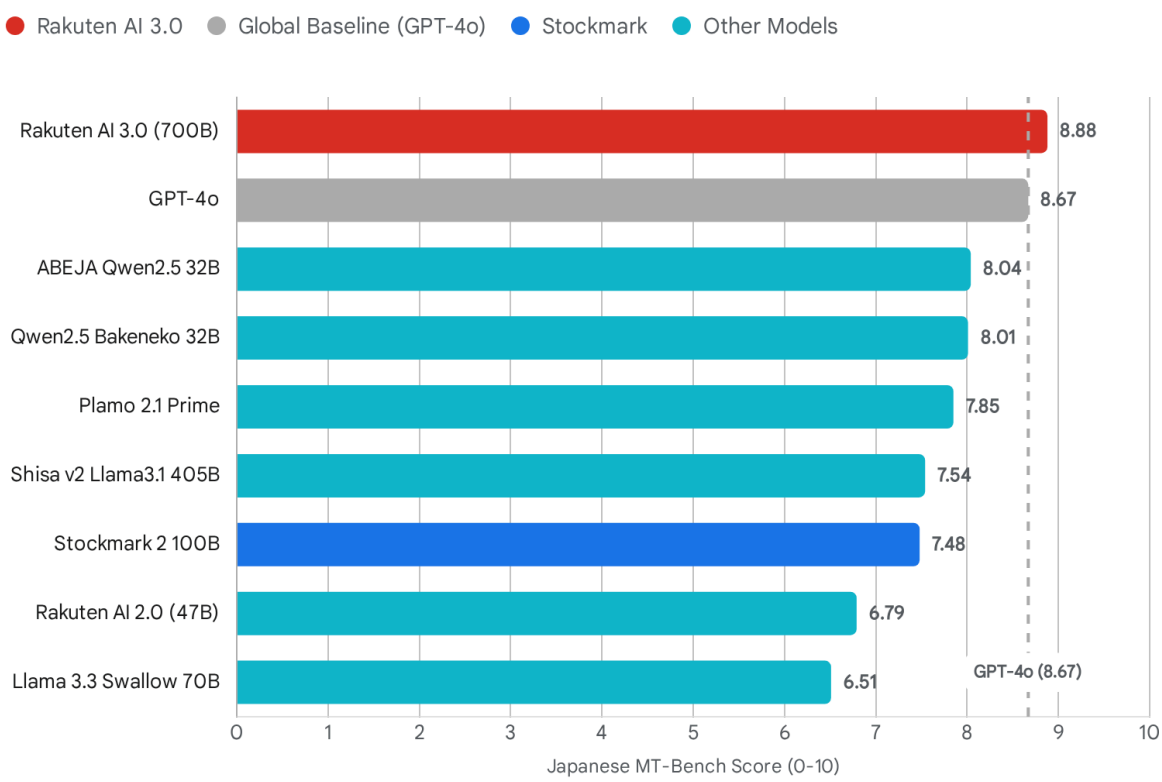
楽天は2025年12月18日、GENIACプロジェクトの成果として、国内最大規模となる約7000億（700B）パラメータを持つ「Rakuten AI 3.0」を発表した[7]。

- **MoE**アーキテクチャ: 700Bという数字は、従来の国産モデル（数百億～一千億クラス）とは桁が違う規模である。しかし、楽天はMixture of Experts (MoE) 技術を採用し、推論時にアクティブになるパラメータを約400億（40B）に抑えている。これにより、巨大な知識容量と推論効率（コストパフォーマンス）の両立を実現した[7]。
- 性能評価: 楽天の発表によれば、Rakuten AI 3.0は日本語MT-Benchにおいて「8.88」というスコアを記録し、GPT-4o（8.67）を上回る性能を示したとされる（ただし、ベンチマークは特定の条件下での評価であり、総合的な優劣を決定づけるものではない点に留意が必要）[7]。

3.5.2 オープンウェイト戦略とエコシステム

楽天の戦略における最大の特徴は、開発した巨大モデルを「オープンウェイト」として公開する点にある。

- **2026**年春の公開: Rakuten AI 3.0は、2026年春（3月～5月頃）にHugging Face等で公開される予定である[32]。これにより、日本の研究者やエンジニアは、世界トップレベルの日本語モデルを自由に使用、検証、チューニングできるようになる。これは日本のAI技術底上げに対する巨大な貢献となる。
- トリプルプレイへの適用: ビジネス面では、楽天モバイル、楽天市場、楽天カード・証券といった自社の「トリプルプレイ」エコシステム内でのAI活用を徹底する。マーケティングの自動化、通信ネットワークの運用効率化、顧客サポートの無人化などを自社モデルで行うことで、外部ベンダーへの依存を減らし、運用コストを90%削減することを目指している[34]。



2025年12月時点での主要モデルにおける日本語MT-Benchスコア比較。楽天やELYZAなどの国産（または国産チューニング）モデルが、グローバルトップモデルであるGPT-4oに肉薄、あるいは特定の評価基準で上回る成果を示している。
（注：スコアは各社の発表値および公開リーダーボードに基づく）

Data sources: Rakuten Group, Tech in Asia, Hugging Face, Note.com

4. 技術アーキテクチャの分岐と進化

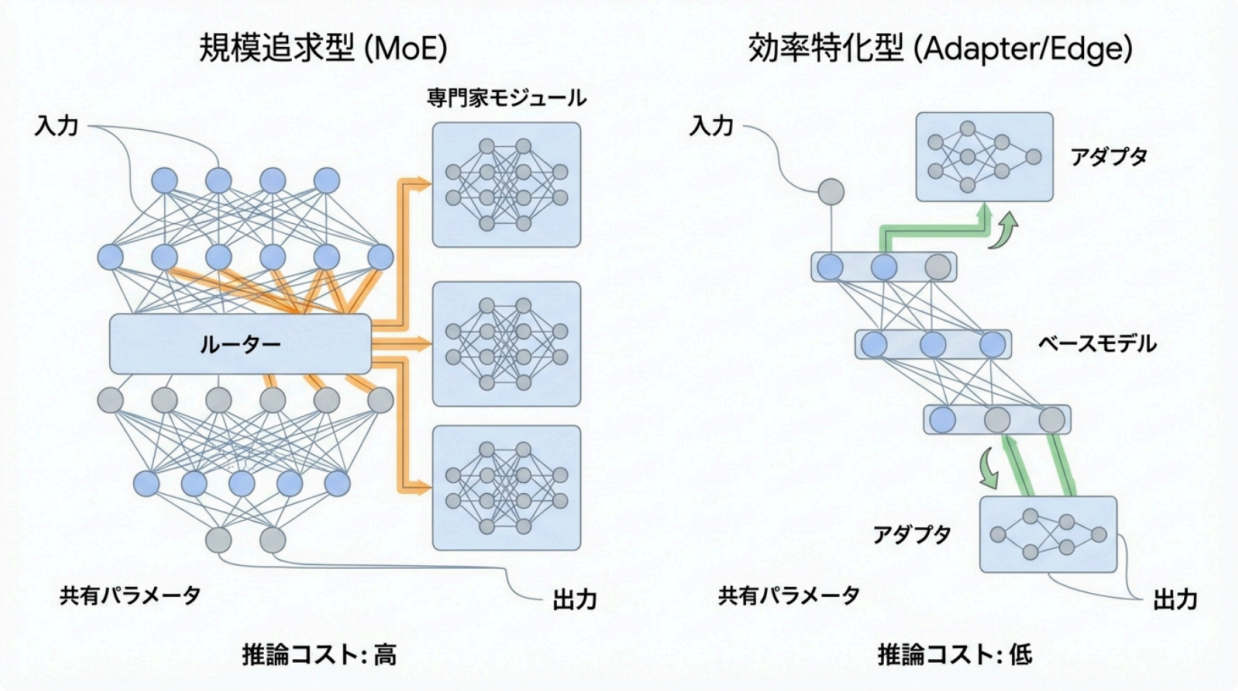
2026年の国産AI開発においては、大きく分けて「規模追求型（MoE）」と「効率特化型（Adapter/Edge）」という二つの技術的思想（アーキテクチャ）の分岐が見られる。

4.1 規模追求型（MoE）：SoftBank / Rakuten

ソフトバンクや楽天が採用するMixture of Experts (MoE) は、巨大なパラメータ数（知識の総量）を持ちながら、推論コストを現実的な範囲に抑えるための技術である。
ユーザーからの入力（プロンプト）に対し、「ルーター」と呼ばれるネットワークが、その処理に最適な「専門家（Expert）」ブロックだけを選択して起動する。例えば、「計算」のタスクなら数学担当のエキスパートだけが、「翻訳」なら言語担当のエキスパートだけが動く。これにより、全体で7000億パラメータを持っていても、一回の推論で計算するのは400億パラメータ分で済むといった効率化が可能になる。

このアーキテクチャは、汎用的な能力（AGI的な振る舞い）を高めるのに適しており、クラウド上での大規模サービス提供に向いている。

国産LLM 2つの設計思想：規模追求型 vs 効率特化型



左：ソフトバンクや楽天が採用する「規模追求型（MoE）」アーキテクチャ。巨大な知識ベースを持ち、ルーターが適切な専門家モジュールを選択する。右：NTTやNECが志向する「効率特化型（Adapter/Edge）」アーキテクチャ。軽量なベースモデルに、特定タスクごとの追加モジュール（アダプタ）を脱着し、省リソースで専門業務を遂行する。

4.2 効率特化型（Adapter/Edge）：NTT / NEC / ELYZA

一方、NTTやNECが採用するのは、比較的軽量なベースモデル（7B〜30B程度）に、追加学習用の軽量モジュール（アダプタ）を付加する方式である。
アダプタ技術（LoRAなど）を用いれば、ベースモデルの重みは固定したまま、タスク固有の知識だけを少量のパラメータで学習・保持できる。これにより、一つのベースモデルに対して「金融用アダプタ」「医療用アダプタ」などをカセットのように切り替えて使用することが可能になる。
この方式は、メモリ消費量が少なく、1枚のGPUで複数のタスクを切り替えながら処理できるため、オンプレミス環境やエッジデバイスへの実装に最適である。日本の製造業や自治体が求める「専用環境でのコスト効率の良いAI」に対する解は、こちら側にあると言える。

5. GENIACプロジェクトと政府の役割

5.1 計算資源の民主化とエコシステム形成

経済産業省とNEDOが推進する「GENIAC（Generative AI Accelerator Challenge）」は、日本のAI開発における最大の加速装置となっている。かつて、AI開発には数千億円規模のインフラ投資が必要とされ、それが参入障壁となっていたが、GENIACは国が確保したGPU計算資源（Google CloudやAWS、さくらインターネット等）を選抜された企業や研究機関に提供することで、この壁を取り払った。

2025年の「第3期」においては、単なるモデル開発だけでなく、「社会実装（活用実証）」が重視されたことが大きな特徴である[35]。楽天、Preferred Networks、Stockmark、Turing、Kotoba Technologies Japanなどの多様なプレイヤーが採択され、汎用モデルだけでなく、創薬、自動運転、建設、製造といった特定領域に特化した基盤モデルの開発が進められた。

5.2 評価指標の標準化：JGLUEとRakuda

GENIAC周辺のコミュニティ（LLM-jp等）は、モデルを作るだけでなく、それを正しく評価するための「物差し」作りも牽引している。「JGLUE」や「Rakuda」といった日本語独自のベンチマークの開発と普及が進んだことで、海外製のベンチマークでは測定できない「日本的な文脈理解」や「文化的妥当性」を定量化することが可能になった。これは、国産モデルが「GPT-4より賢いか？」という単純な問いに対し、「特定の日本的タスクにおいてはYesである」と客観的に主張するための基盤となっている[36]。

5.3 その他の注目プレイヤーと成果

GENIACの支援を受け、大手以外にもユニークな強みを持つ企業が台頭している。

- **Preferred Networks (PFN)**: マルチモーダル基盤モデル「PLaMo」を開発。2025年10月には、GENIACの成果として開発されたモデルの一部を公開し、独自のハードウェア（MN-Core）との垂直統合による高速化を目指している[39]。
- **Stockmark**: 製造業に特化した1000億パラメータ級の「Stockmark-LLM」を開発。社内文書や技術文書の検索・生成において高いハルシネーション（嘘）抑止性能を実現し、パナソニックスなどの大手製造業での採用実績を持つ[41]。
- **Turing**: 完全自動運転の実現を目指し、映像と言語を統合したマルチモーダルモデル「Heron」等の開発を進めている。

6. 2026年以降の展望と課題

6.1 開発ロードマップとマイルストーン (2024-2027)

各社の発表やこれまでの動向を総合すると、2026年から2027年にかけての国産AI開発は、以下のようなロードマップで進行すると予測される。

時期	フェーズ	主要イベント・マイルストーン (予測含む)

2024年	基盤構築期	・NTT tsuzumi 1.0 リリース ・NEC cotomi 発表 ・Sarashina1 公開 ・GENIAC 第1期・第2期 実施
2025年	性能強化・特化期	・**NTT tsuzumi 2** リリース (10月) [10] ・**NEC cotomi V2 / Fast** 展開 ・**Sarashina2 / Vision** モデル公開 [28] ・**Rakuten AI 3.0** (700B) 発表 (12月) [7] ・GENIAC 第3期 成果報告
2026年	社会実装・エージェント期	・**NEC cotomi Act** (エージェント) 提供開始 (1月) [23] ・**Rakuten AI 3.0** オープンウェイト公開 (春) [32] ・**Sarashina mini** 商用提供開始 [26] ・各社による自律型AIエージェントの実装本格化 ・OpenAI GPT-5 (仮) への対応と棲み分け明確化
2027年以降	インフラ融合・**AGI**期	・**IOWN (NTT)** 上でのAIコンステレーション実用化 ・**AI-RAN (SoftBank)** の全

		国展開と分散推論 ・1兆パラメータ級モデルの登場 (SoftBank等) ・専用ハードウェア (PFN MN-Core等) の普及

6.2 「チャット」から「エージェント」へ

2025年までのLLMは「質問に答える（チャットボット）」が主用途であったが、2026年は「タスクを実行する（エージェント）」への進化が決定的となる。NECの「cotomi Act」やソフトバンクの構想に見られるように、AIが自ら計画を立て、社内システムを操作し、メールを送り、決済を行う世界観である。ここでは、推論の正確性だけでなく、外部ツールとの接続性（Connectivity）や、誤作動を防ぐガードレール機能が競争軸となる。

6.3 ハードウェアの制約と主権

最大の課題は「GPUの確保」である。NVIDIA製GPUの争奪戦は続いており、円安も相まって調達コストは上昇している。これに対し、Google (TPU) やAmazon (Trainium) のような独自チップを持たない日本企業は不利な立場にある。
しかし、NTTの光プロセッサや、PFNの独自チップ（MN-Coreシリーズ）など、ハードウェアレベルでの差別化を図る動きも出てきている。2027年以降、これらの国産ハードウェアが実用化されれば、ソフトウェア（モデル）とハードウェアの垂直統合によるブレイクスルーが期待できる。

7. 結論

2026年、日本の国産AIモデル群は、もはや「海外の劣化コピー」ではない。NTTの省電力技術、NECのシステム統合力、ソフトバンクのインフラ投資、楽天のエコシステム実装といった、各社のDNAを色濃く反映した「ソリューション」へと進化している。

ユーザー企業にとっての最適解は、単一のモデルに依存することではなく、これらを適材適所で組み合わせる「マルチモデル戦略」にある。クリエイティブなタスクや英語での情報収集にはGPT-5を、社外秘情報の要約や国内業務の自動化にはオンプレミスのtsuzumiやcotomiを、工場内の制御には特化型モデルを使う——そのような使い分けが当たり前になる中、国産モデルは日本のDX（デジタルトランスフォーメーション）を支える不可欠な「社会インフラ」として定着していくだろう。
日本は今、GENIAC等の政策的支援と民間企業の技術革新が噛み合い、真の意味での「AI主権」を確立する瀬戸際にある。2026年の各社の動向は、日本の産業競争力の未来を占う試金石となる。

引用文献

1. NTT版LLM「tsuzumi」 図表読解、GPU不要の超軽量版も, 1月 11, 2026にアクセス、https://www.watch.impress.co.jp/docs/news/1543861.html
2. NEC開発の生成AI「cotomi」, 1月 11, 2026にアクセス、https://jpn.nec.com/LLM/cotomi.html

3. NECが開発・提供する生成AI「cotomi」とは？具体的な特徴や導入 ..., 1月 11, 2026にアクセス、https://craftai.jp/nec-cotomi/
4. 通信業界向け生成AI基盤モデル「Large Telecom Model」が国産AI ..., 1月 11, 2026にアクセス、https://www.softbank.jp/corp/news/press/sbkk/2025/20251029_01/
5. NECのLLMを技術的優位性から導入まで徹底解説！ - debono, 1月 11, 2026にアクセス、https://debono.jp/7148
6. 【ELYZA】日本語LLMで未来と市場を切り拓く！ - note, 1月 11, 2026にアクセス、https://note.com/unicorn_startup/n/n7321c8e13878
7. Rakuten Unveils Japan's Largest High-Performance AI Model ..., 1月 11, 2026にアクセス、https://global.rakuten.com/corp/news/press/2025/1218_01.html
8. NTT's Large Language Models 'tsuzumi 2', 1月 11, 2026にアクセス、https://www.rd.ntt/e/research/LLM_tsuzumi.html
9. NTT's Next-Generation LLM "tsuzumi 2" Now Available, 1月 11, 2026にアクセス、https://group.ntt/en/newsrelease/2025/10/20/251020a.html
10. NTT版大規模言語モデル「tsuzumi 2」| NTT R&D Website, 1月 11, 2026にアクセス、https://www.rd.ntt/research/LLM_tsuzumi.html
11. NTT「tsuzumi 2」の真価：国産・軽量・高性能AIが実現するコスト ..., 1月 11, 2026にアクセス、https://note.com/eiji71/n/n3ec64647af89
12. 【NTT tsuzumi Open Days〜キャリア編〜】LLM独自モデル開発 ..., 1月 11, 2026にアクセス、https://techplay.jp/column/2040
13. NTT's LLM “tsuzumi” - NTT Technical Review, 1月 11, 2026にアクセス、https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202408fa1.html
14. NTT「tsuzumi 2」発表：30Bパラメータで1GPU動作、RAG・業界 ..., 1月 11, 2026にアクセス、https://gai.workstyle-evolution.co.jp/2025/10/28/ntt-tsuzumi-2-japanese-llm-30b-parameters-one-gpu-rag-industry-specific-tokyo-university-adoption/
15. NTTが大規模言語モデル「tsuzumi 2」を提供開始 世界トップクラス ..., 1月 11, 2026にアクセス、https://k-tai.watch.impress.co.jp/docs/news/2056478.html
16. NTT「tsuzumi 2」発表 RAG強化でビジネス利用が加速する理由とは, 1月 11, 2026にアクセス、https://www.pc-webzine.com/article/3344
17. NTT AI最前線！次世代技術「tsuzumi 2」と「AIコンステレーション ..., 1月 11, 2026にアクセス、https://corp.omake.co.jp/ntt-ai%E6%9C%80%E5%89%8D%E7%B7%9A%EF%BC%81%E6%AC%A1%E4%B8%96%E4%BB%A3%E6%8A%80%E8%A1%93%E3%80%8Ctsuzumi-2%E3%80%8D%E3%81%A8%E3%80%8Cai%E3%82%B3%E3%83%B3%E3%82%B9%E3%83%86%E3%83%AC%E3%83%BC%E3%82%B7/
18. NECが生成AI事業をアピール、「業界特化型の国産LLMが競争力の ..., 1月 11, 2026にアクセス、https://cloud.watch.impress.co.jp/docs/news/1664788.html
19. NEC Develops High-speed Generative AI Large Language Models ..., 1月 11, 2026にアクセス、http://en.iccsz.com/News.Asp?id=68
20. NEC、生成AI「cotomi」の性能強化でAIエージェントの活用を加速 ..., 1月 11, 2026にアクセス、https://prtimes.jp/main/html/rd/p/000000989.000078149.html
21. NEC、独自開発 LLM「cotomi」のエージェント性能を強化 128K ..., 1月 11, 2026にアクセス、https://ledge.ai/articles/nec_cotomi_agent_upgrade_128k_mcp
22. 【発表！】2025年度上期「ものづくりの未来」人気記事ランキング, 1月 11, 2026にアクセ

ス、https://jpn.nec.com/manufacture/monozukuri/iot_mono/2025-10/03.html
23. NEC、企業ノウハウをAIエージェントで自動抽出・組織資産化し ..., 1月 11, 2026にアクセス、https://jpn.nec.com/press/202512/20251203_01.html
24. 2024年の生成AI関連株投資戦略：市場動向と注目企業分析, 1月 11, 2026にアクセス、https://media.buzzconne.jp/generative-ai-stocks-investment-opportunities-2024/
25. 大塚商会、オンプレミス環境に閉じて運用できる生成AIサーバー ..., 1月 11, 2026にアクセス、https://it.impress.co.jp/articles/-/27395
26. うさぎでもわかる 10億AIエージェントが創る未来！SoftBank World ..., 1月 11, 2026にアクセス、https://note.com/taku_sid/n/nbc50b0ce9dba
27. 本のヘルスケア分野におけるLLMの現状と課題 - IPA, 1月 11, 2026にアクセス、https://www.ipa.go.jp/digital/chousa/j5u9nn0000009rgh-att/aiws3_20250911_keynote2_Kakizaki.pdf
28. “目”を持つAIの新モデル「Sarashina2.2-Vision-3B」、SB Intuitions ..., 1月 11, 2026にアクセス、https://www.itmedia.co.jp/aiplus/articles/2511/25/news116.html
29. SoftBank's Masayoshi Son Says Artificial Superintelligence to Exist ..., 1月 11, 2026にアクセス、https://japan-forward.com/softbank-masayoshi-son-says-artificial-superintelligence-to-exist-by-2035/
30. KDDIとELYZAの提携で進む―日本語特化LLM開発と生成AIの社会 ..., 1月 11, 2026にアクセス、https://biz.kddi.com/beconnected/feature/2024/241018/
31. 楽天、7000億パラメータの日本語LLM「Rakuten AI 3.0」, 1月 11, 2026にアクセス、https://www.watch.impress.co.jp/docs/news/2072366.html
32. Japan's Rakuten is going to release a 700B open weight model in ..., 1月 11, 2026にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1pr20el/japans_rakuten_is_going_to_release_a_700b_open/
33. Rakuten launches open-weight LLM, claims beat GPT-4o, 1月 11, 2026にアクセス、https://www.techinasia.com/news/rakuten-launches-open-weight-llm-claims-beat-gpt-4o
34. Rakuten develops lower-cost AI models for ecommerce - Tech in Asia, 1月 11, 2026にアクセス、https://www.techinasia.com/news/rakuten-develops-lowercost-ai-models-ecommerce
35. GENIAC（METI/経済産業省）, 1月 11, 2026にアクセス、https://www.meti.go.jp/policy/mono_info_service/geniac/index.html
36. NTT's Large Language Model “tsuzumi”: A High-performance and ..., 1月 11, 2026にアクセス、https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202408fr1.html
37. LLM Leaderboard 2025 - Vellum AI, 1月 11, 2026にアクセス、https://www.vellum.ai/llm-leaderboard
38. JGLUE: Japanese General Language Understanding Evaluation, 1月 11, 2026にアクセス、https://github.com/yahoojapan/JGLUE
39. PFN、CEATEC2025に出展・登壇 - PR TIMES, 1月 11, 2026にアクセス、https://prtimes.jp/main/html/rd/p/000000018.000156310.html

40. GENIAC第2サイクルに継続採択 - 株式会社Preferred Networks, 1月 11, 2026にアクセス、https://www.preferred.jp/ja/news/pr20241010
41. Rakuten has developed a new high-performance AI model ..., 1月 11, 2026にアクセス、https://www.moomoo.com/news/post/63005554/rakuten-has-developed-a-new-high-performance-ai-model-rakuten
42. パナソニックHDとストックマーク、国内最大規模（1000億 ..., 1月 11, 2026にアクセス、https://news.panasonic.com/jp/press/jn240702-3