ChatGPT

2026年における日本の国産LLMの採用状況に関する調査レポート

エグゼクティブサマリー

本調査（更新日：2026-03-01　JST）は、2026年時点での「国産LLM（日本の企業・研究機関が主導して開発・提供し、日本市場における主権性／日本語最適化／セキュア運用を主要価値に据えるLLM群）」の**採用状況**を、公開一次情報（公式発表、政府資料、学術論文、信頼できる技術メディア等）を中心に整理した。全体像として、企業の生成AI利用は加速している一方で、**国産LLMの採用は「データ主権・機密性・ガバナンス要件が強い領域（公共、金融、医療、重要インフラ、製造の設計・品質等）」で特に伸びやすい構図**が見える。背景には、政府側の調達・利活用ガイドライン整備（行政向け）と、ベンダー側の「オンプレ／専有環境／国内運用」を前提にした提供形態の拡充がある。 [1]

企業利用の裾野を測る指標として、日本情報システム・ユーザー協会（JUAS） [2] の「企業IT動向調査2025」速報では**「言語系生成AI」導入（準備中含む）41.2%**とされ、導入済み企業の約7割が効果「あり」と回答した一方、**効果測定手法に課題**があることが示されている。 [3]　こうした「導入は進むが、本番化・定量効果の出し方が難しい」という状況は、国産LLM採用でも共通のボトルネックになりやすい。

主要プレイヤーの動向を俯瞰すると、**（1）比較的軽量〜中規模で推論コストを抑え、オンプレ／閉域で運用できる国産モデル（例：tsuzumi 7B、tsuzumi 2 30B、cotomi 130億級）**と、**（2）エンタープライズ向けに高精度・監査・統制を組み込み、プライベート環境提供を前提にするモデル群（例：Takane）**、**（3）オープンウェイトとして公開され、内製RAGや微調整の“素材”として使われる国産・日本語最適化モデル群（OpenCALM、Sarashina2.2、RakutenAI、tanuki等）**に大別できる。 [4]

技術面では、たとえばNTT [5] のtsuzumi　2は**30Bで「1GPUで推論可能」**を前提に、推論用GPUコストの試算として**30BはA100　40GB相当1基で約500万円**、比較対象として**400B級は約5千万円、700B級は約1億円**といった“桁感”を提示し、推論コストを**約10〜20分の1**に下げ得るという立て付けを示している。 [6]　同社資料では、tsuzumi発表（2023年11月）以降の相談・受注の累計として、FY25　1Q時点で**国内相談件数1,818、国内＋海外の受注件数1,827**などの数値が提示され、公共領域比率も示唆されている（ただし「生成AI関連」全体の集計であり、国産LLM単体の採用率ではない点に留意が必要）。 [7]

公共分野の具体例として、富士通 [8] は、**中央省庁（名称非公表）でのパブリックコメント業務PoC（2025年実施）**を発表し、約**12万字の実データ**を用いた分類・要約等を**約10分で完了**、さらに法案条文と意見の対応付けで**80%超**の意見について該当条文を正しく特定できたとしている。 [9]　このように、行政の「大量テキスト処理×説明責任×秘匿性」に適合しやすい業務から国産LLM採用が進む傾向が確認できる。

本レポートの結論として、2026年時点の国産LLM採用は、**①セキュアな専有環境（オンプレ／プライベートクラウド）**、**②日本語品質（敬語・公用文・専門語の安定性）**、**③推論コストと運用要件（GPU枚数・メモリ）**、**④ガバナンス（ログ、監査、脆弱性対策、リスク管理）**の4点で意思決定されやすい。行政向けにはデジタル庁 [10] が調達・利活用ガイドラインを整備しており、これが自治体・政府案件の標準的な評価軸になりやすい。 [11] 一方で、国産LLMの“シェア”や“採用割合”を横断的に示す公的統計は限定的であり、現時点ではベンダー公表の案件数・PoC事例、政府事業（GENIAC等）の成果公開、そして第三者ベンチマーク（Nejumi、JGLUEなど）を組み合わせた推定に留まる。 [12]

調査範囲と方法

本調査で扱う「国産LLM」は、厳密な法的定義ではなく、公開一次情報から以下の観点で実務的に分類した。

- **開発主導**：日本の企業・研究機関が基盤モデル開発、または基盤モデル選定後の大規模追加学習・最適化を主導していること（例：フルスクラッチ、または海外オープンモデルに日本語・業務特化を大規模に付与）。[13]
- **提供形態**：国内企業・行政の要件（閉域運用、専有環境、国内データセンター、監査）に適合する形で提供される、またはその選択肢が明示されていること。[14]
- **評価可能性**：JGLUE、Nejumi、Rakuda、Japanese MT-Bench等の日本語評価指標、またはそれに準ずる公開評価が確認できること。[15]

情報源は「一次情報（開発元公式、導入組織の公式発表、政府資料、学術論文）」を優先した。一次情報で不足する箇所のみ、信頼できる技術メディア等で補った。数値・採用状況は、**公表値のみ**を原則とし、未公表は「未指定」として扱った。

重要な制約として、**国産LLMの“採用割合（市場シェア）”を直接示す統計は、2026-03-01時点で横断的に整備されていない**。そのため本レポートでは、（a）企業一般の生成AI導入率、（b）主要国産LLMの案件数・PoC公表、（c）政府事業の成果公開、（d）第三者ベンチマークと運用要件、を統合して「採用状況」を分析した。[16]

主要な国産LLMの整理

国産LLMの主要候補一覧（公開情報に基づく）

下表は、2026年時点で参照頻度・導入実績公表・政府事業への登場等から「主要」と判断した国産LLMを、要求項目に沿って整理したものである（未公表は未指定）。

系列/モデル（代表）	開発主体	初回公開/提供開始	モデルサイズ（公表）	ライセンス形態	商用可否	API/オンプレ提供	日本語性能評価指標（公開範囲）
tsuzumi（MaaS版）	NTTデータ[17]	2024年版ガイド公開（MaaS提供）	7B	未指定（MaaS提供、モデル配布ではない）	未指定	Azure MaaS上のAPI提供が前提	RakudaでGPT-3.5を上回る旨、同規模国産を上回る旨（定量値は当該抜粋範囲では未指定）[18]
tsuzumi 2	NTT	2025-10-20 提供開始	30B	プロプライエタリ（商用提供）	商用提供（顧客導入）	オンプレ/プライベートクラウド運用を想定、1GPU推論	llm-jp-eval、M-IFEval_Ja、AnswerCarefully等を用いた比較評価の枠組み提示（詳細スコアは図表中心）[19]

系列/モデル（代表）	開発主体	初回公開/提供開始	モデルサイズ（公表）	ライセンス形態	商用可否	API/オンプレ提供	日本語性能評価指標（公開範囲）
cotomi（v1系）	NEC [20]	2023-12-15（戦略発表）	130億級（「従来の…130億クラス」）	プロプライエタリ	商用提供	マネージドAPI、セキュア環境、業種特化モデル方針	Rakudaで国内外トップクラス群を上回る旨、長文処理30万字（値は機能として公表） [21]
cotomi Pro / Light	NEC	2024-04-24	未指定	プロプライエタリ	商用提供	GPU1〜2枚/2枚で高速推論、用途別モデル	ELYZA Tasks 100：3.87（Pro）、3.53（Light）／Japanese MT-Bench：7.71（Pro）、6.61（Light） [22]
cotomi v3（推論API基盤提供）	NEC（モデル）＋さくらインターネット [23]（提供基盤）	2025-12-02	未指定	未指定（価格・ライセンスは承認ユーザーに表示）	未指定	推論API（さくらのAI Engine）	「自治体・金融・医療などの分野で得た実証知見」を根拠に実務最適化を謳う（指標値は未指定） [24]
Takane（基盤）	富士通＋Cohere [25]	2024-09-30グローバル提供	「中規模」明記、ベースはCommand R+（104Bオープンウェイトの研究公開例あり）	プロプライエタリ（富士通が独占提供）	商用提供	DI PaaS/Kozuchi統合、セキュアなプライベート環境	JGLUE平均0.92（各タスクも提示）、Nejumi Leaderboard3のカテゴリスコア（0.862/0.773）等 [26]
Takane（オンプレ/展開用 小・中規模）	富士通（展開）	2025-10-10（PRIMERGY基盤側の記載）	Takane 7B / Takane 32B	未指定	未指定	オンプレAI基盤での対応モデルとして提示	未指定 [27]
Sarashina（系列）/ Sarashina API（mini）	ソフトバンク [28]＋SB Intuitions [29]	2025-11-28（API提供開始）	Sarashina：4,600億パラメータ（miniは未指定）	API提供（商用サービス）	商用提供	Chat Completion / Embeddings API	法人向け提供開始・社内約2万人トライアル等（指標値は未指定） [30]

系列/モデル（代表）	開発主体	初回公開/提供開始	モデルサイズ（公表）	ライセンス形態	商用可否	API/オンプレ提供	日本語性能評価指標（公開範囲）
Sarashina2-70B（系列）	SB Intuitions（言及）	2026-02-18時点で実在言及	70B	未指定	未指定	未指定	「日本語性能No.1を目指す」等、性能値は未指定 [31]
Sarashina2.2（0.5B/1B/3B）	SB Intuitions	2025-03-07	0.5B/1B/3B	未指定（記事内はベンチ高水準と記述）	未指定	公開モデル（配布）	同サイズ帯で日本語最高水準と記述（具体スコアは当該抜粋範囲では未指定） [32]
Sarashina2-Vision（8B/14B）	SB Intuitions	2025-03-17	8B/14B	MIT	商用利用可能	公開モデル（配布）	VLM評価の実施方針を公開（詳細スコアは別記事） [33]
OpenCALM（最大6.8B）	サイバーエージェント [34]	2023-05-17	6.8B	CC BY-SA 4.0	商用利用可能	公開モデル（配布）	日本語LLMとして最大級、学習データ（Wikipedia/Common Crawl）明記 [35]
Panasonic-LLM-100b（開発）	パナソニックホールディングス [36] ＋ストックマーク [37]	2024-07-02（開発開始）	100B（予定）	未指定（社内専用を前提）	未指定	自社活用前提（開発）	Stockmark-LLM-100bベースで日本語性能向上と記述（指標値は未指定） [38]
tanuki-8x8b（GENIAC成果例）	東京大学 [39]（プロジェクト言及）	2024年成果公開（最終更新 2024-12-20）	8x8B（MoE）	商用利用可能な形で公開（詳細ライセンスはリンク先参照）	商用可	公開モデル（配布）	JGLUE：76.0、Nejumi LB3総合0.57、Japanese MT-Bench 6.8〜7.0等 [40]

系列/モデル（代表）	開発主体	初回公開/提供開始	モデルサイズ（公表）	ライセンス形態	商用可否	API/オンプレ提供	日本語性能評価指標（公開範囲）
PLaMo-100B（GENIAC成果例）	Preferred Elements [41]	2024年成果公開（同上）	100B	未指定（技術ブログ・成果公開あり）	未指定	未指定	医師国試スコア（例：2022年317）、金融評価など比較値を提示 [40]
RakutenAI-7B	楽天グループ [42]	2024-03（論文公開）	7B	Apache 2.0	商用可	公開モデル（配布）	Japanese LM Harnessで7Bオープンモデル上位と記述 [43]

日本語性能評価指標の「事実上の標準化」状況

企業・行政がモデル選定を行う際の日本語評価は、単一指標ではなく“多面的”に見る方向に進んでいる。代表例として、entity"company","Weights & Biases Japan","ml tooling company japan"]は、日本語LLMの多面的評価サイト「Nejumi　LLMリーダーボード4」を公開し、高難度推論・知識、アプリ開発能力（関数呼び出し等）、安全性評価の拡充を明示している。 [44]
同時に、政府側でも経済産業省 [45] のGENIAC成果公開で、JGLUE（Nejumi　　v1相当）やNejumi　　LB3、Japanese　MT-Bench等を組み合わせた評価・比較が行われており、**「ベンチマーク体系（複数）」を前提にした選定**が実務になりつつある。 [40]

採用状況と導入事例

2026年時点で見える採用パターン

公開事例を俯瞰すると、国産LLM採用の導線は概ね次の3類型に分かれる。

第一に、**自治体・公共領域でのPoC→業務適用**である。公用文・例規・住民向け文書などは日本語特有の規範性が強く、また個人情報や行政内部情報の扱いから「閉域運用」志向が強い。実際、山口県では、職員不足を見据えた業務効率化の文脈で、tsuzumiを用いた生成AI実証を1年間の計画で行う旨を公表している。 [46]　また富士通は中央省庁でのPoC成果（12万字→10分、条文対応80%超）を公表しており、政策形成・立法支援への横展開（FY2026提供目標）まで言及している。 [9]

第二に、**金融・医療・製造など機微情報を扱う民間領域の「専有環境」前提導入**である。NECはcotomiを「低遅延でセキュアなLLM環境」やマネージドAPIで提供し、業種特化モデルを中核に据える戦略を示している。 [47] 富士通もTakaneをプライベート環境で展開できる形（Nutanix Enterprise AIでの認定LLM等）を整備している。 [48]

第三に、**大企業が内製・準内製で“自社専用LLM”を開発する動き**である。パナソニックHDはストックマークと、100Bの社内専用モデル（Panasonic-LLM-100b）開発を発表しており、「業務・データに合わせて作る」方向の投資が顕在化している。 [38]

導入実績一覧（公表ベース）

導入組織（公表名）	業界	導入目的	導入規模（公表値）	導入時期	導入形態	実行環境	使用LLM/方式	出典
山口県	自治体	生成AI活用実証（業務効率化、職員不足対応）	未指定（1年計画）	2024-10（開始）	PoC	未指定	tsuzumi	46
神戸市	自治体	行政業務向け生成AIの実証	未指定	2025-01（開始）	PoC	未指定	cotomi	49
相模原市	自治体	生成AIの行政業務活用（例：文書作成/要約等）	未指定	2025-01（公開）	未指定	未指定	cotomi	50
北九州市	自治体	行政業務での生成AI活用	未指定	2025-01（公開）	未指定	未指定	cotomi	51
大塚商会	IT/商社	生成AIサーバセットで業務活用	未指定	2025-04-23（発表）	本番/提供開始（文脈上）	サーバ提供（オンプレ想定）	cotomi（NEC Generative AI Server）	
中国電力	エネルギー	問い合わせ対応（RAGを含む社内業務）	トライアル期間：2026/1〜3	2026-01〜03	PoC/トライアル	未指定（閉域前提の文脈）	tsuzumi 2	52
東京通信大学	教育（大学）	学内LLM基盤整備（授業Q&A、教材/試験作成支援等）	未指定	2025-10（導入決定の記載）	本番準備（導入決定）	学内ローカル運用（要件）	tsuzumi 2	53

導入組織（公表名）	業界	導入目的	導入規模（公表値）	導入時期	導入形態	実行環境	使用LLM/方式	出典
中央省庁（名称非公表）	政府	パブリックコメント業務の分類・要約・条文紐付け	12万字→約10分、条文特定80%超	2025（PoC実施）	PoC	未指定（業務実証）	Takane	9
パナソニックHD（社内専用）	製造	社内専用LLM（100B）開発・業務日本語性能向上	100B（開発）	2024-07-02（開発開始）	開発（内製/共同）	社内利用前提	Stockmark-LLM-100bベース	38
ソフトバンク（社内）	通信	Sarashina mini（API提供前の社内トライアル）	約2万人対象	2025-06〜（実施）	トライアル	社内利用	Sarashina mini	30

上表から、**自治体はPoCが多く、民間は「サーバ/専有環境」または「問い合わせ対応（RAG）」系の入口が多い**ことが読み取れる。さらに、国産LLM採用の明確な動機として「クラウド依存回避」「データを組織内に留める」「ガバナンス要件充足」が繰り返し現れる。 54

時系列の変化（主要イベント）

```
timeline
  title 国産LLMの主要リリース/採用イベント（公開情報ベース）
  2023-05 : CyberAgentが6.8B日本語LLMを公開（CC BY-SA 4.0）
  2023-11 : tsuzumi発表
  2023-12 : NECがcotomi強化と生成AI事業戦略を発表（130億級方針）
  2024-03 : tsuzumi商用開始（資料上のマイルストーン）
  2024-04 : cotomi Pro/Light発表（ELYZA Tasks 100/MT-Benchスコア公表）
  2024-07 : Panasonic×Stockmarkが100B社内専用LLM開発を発表
  2024-09 : Takane提供開始（JGLUE詳細スコア公表）
  2025-03 : Sarashina2.2・Sarashina2-Vision公開（MIT等）
  2025-05 : 行政向け生成AIの調達・利活用ガイドライン（デジタル庁）
  2025-10 : tsuzumi 2提供開始（30B、1GPU推論前提）
  2025-11 : Sarashina API法人提供開始（Sarashina mini）
  2025-12 : cotomi v3が推論API基盤で提供開始（さくらのAI Engine）
  2026-02 : 中央省庁でTakaneのパブコメPoC成果を公表（12万字→10分）
  2026-01 : 中国電力でtsuzumi 2トライアル（2026/1〜3）
```

上記年表の根拠は、各社公式発表および公的ガイドライン等である。 55

技術評価と運用要件

性能ベンチマークの整理（日本語）

国産LLMの性能議論は、（a）日本語言語理解（JGLUE等）、（b）対話・作文（Japanese MT-Bench等）、（c）安全性（有害出力・漏えい等）、（d）実務適用（RAGでの正答率等）に分解されることが多い。

- TakaneはJGLUE平均0.92、JSQuAD 0.93などタスク別の数値を公開し、Nejumi Leaderboard3でもカテゴリ首位級のスコアを示したとしている。 56
- cotomi Pro/Lightは、ELYZA Tasks 100とJapanese MT-Benchでスコア（3.87/3.53、7.71/6.61）を明示し、日本語対話性能と総合力の両面での“測れる形”を提示している。 22
- 政府のGENIAC成果公開では、tanuki-8x8bのJGLUE 76.0、日本語MT-Bench 6.8〜7.0など、公開モデルについて政府サイト上で比較可能な数値が示されている。 40

他方、tsuzumi 2は比較評価の枠組み（llm-jp-evalやM-IFEval_Ja、AnswerCarefully等）を提示しているが、公開ページ上は図表中心であり、第三者が独立に再現できる形での“全スコア公開”とは限らない（少なくとも本調査の取得範囲では表形式スコアが限定的）。 57

推論コスト・速度・メモリ要件

推論要件は「パラメータ規模×量子化ビット幅×同時実行（バッチ）×コンテキスト長」に依存するが、国産LLMは“オンプレ前提”の説明により、GPU枚数や概算コストに踏み込む例が増えている。

- tsuzumi 2は30Bで、8bit量子化を前提に「必要GPUメモリ＝パラメータ数×量子化サイズ/8bit」の形で説明し、30Bなら約30GBと示している。 58
- 同資料の推論時GPUコスト比較では、tsuzumi 2（30B）をA100 40GB相当1基＝約500万円、比較として400B級（Llama-4想定）約5千万円、700B級（DeepSeek-v3.1想定）約1億円とし、大規模モデル比で推論コストを約10〜20分の1に低減可能と記載している。 59
- cotomi Pro/Lightは、標準的GPU2枚でGPT-4の約1/8〜1/15の時間で処理できるとし、RAG用途での高速性も具体例で述べている（評価条件としてA100×2・16bit等）。 22

これらの“ベンダー試算・条件付き比較”は、実際の運用（同時ユーザー数、プロンプト長、RAG構成、入出力比）で変動しうるため、採用検討では自社データでの再現検証が不可欠である（この点はJUAS調査の「効果測定が課題」とも整合的）。 60

セキュリティ・プライバシー機能（ベンダーの差別化点）

国産LLM採用の最大の差別化は「専有環境（閉域）＋統制機能」になりやすい。

- tsuzumi 2は「オンプレ／プライベートクラウドでの運用が可能で、機密情報を安全に扱える」ことを明確にし、学習データのコントロール（新聞等データの自主削除など）にも言及している。 61
- cotomiは「情報漏洩・脆弱性等のセキュリティ面」や、セキュアな環境での提供ニーズを明示し、業種特化・マネージドAPIの戦略と結びつけている。 62
- 富士通の「Kozuchi Enterprise AI Factory」発表では、独自定義を含む**7,700種超の脆弱性**に対応したスキャナーとガードレール、ルール自動生成・自動適用による運用自動化をうたい、非専門家でも安全性・信頼性を確保しやすい設計思想を示している。 63

さらに同社はTakaneの強化として、量子化技術を適用し**1ビット量子化でメモリ消費量最大94%削減、精度維持率89%、量子化前の3倍高速化**などの数値を明示している。 64

市場動向・競合と採用障壁

海外LLMとの比較軸（性能・価格・主権性）

国産LLMが海外フロンティアモデルと競合する際、意思決定軸は「最高性能」だけではなく「総所有コスト」と「統制可能性」にシフトしている。背景として、海外APIは価格が明確である一方、モデル更新・提供リージョン・データ取り扱いなどが契約体系に依存し、組織要件によってはオンプレ／専有環境が必要になる。 [65]

2026-03-01時点に近い公式価格例（API、1M tokensあたり）は次の通り（代表例、為替影響は別）。

- OpenAI [66] ：GPT-5.2 入力$1.750／出力$14.000（キャッシュ入力$0.175） [67]
- Anthropic [68] ：Claude Opus 4.5 入力$2.50／出力$12.50（ドキュメント上の価格表） [69]
- Google [70] ：gemini-3.1-pro-preview 入力$2／出力$12（<200k tokens）等（Gemini 3ガイド記載） [71]

これに対し国産LLMは、価格の公開度が低い（個別見積もり、承認ユーザーに表示等）ケースが多い。たとえば「さくらのAI Engine」でのcotomi v3は価格が承認ユーザーのみに表示されると明記されている。 [24]
一方で、tsuzumi 2のようにオンプレ想定のGPU構成と概算費用を提示し、海外の超大規模モデルの推論コストと比較する（ベンダー試算）ことで、導入の経済合理性を示そうとする動きもある。 [59]

ベンダーのビジネスモデル（国産LLMの“売り方”）

公開情報から整理すると、国産LLMの収益化は「モデル単体」よりも、以下の束ね方が主流である。

- **AIプラットフォーム統合型**：TakaneをKozuchi/DI PaaSに統合し、RAG・監査・コンサル（Wayfinders等）と一体提供（富士通）。 [72]
- **マネージドAPI＋業種特化型**：cotomiを軸に業種・業務特化モデルを整備し、マネージドAPIとして提供（NEC）。 [21]
- **ソブリン/オンプレ推進型**：tsuzumi 2の「1GPU」「オンプレ/プライベートクラウド」を前面に、自治体・企業の閉域需要に寄せる（NTT）。 [73]
- **API提供（法人向け）＋周辺API（埋め込み等）**：Sarashina miniをChat Completion/Embeddingsとして提供し、社内利用→外販へ展開（ソフトバンク/SB Intuitions）。 [30]

採用障壁（法規制・コスト・人材・データ・信頼性）

国産LLMの採用障壁は、海外LLMと共通する部分と、国産特有の部分がある。

行政領域では、生成AIの利活用促進とリスク管理を表裏一体で進めるため、デジタル庁が「行政の進化と革新のための生成AIの調達・利活用に係るガイドライン」を策定し、調達・運用・リスク整理の枠組みを提示した。 [74]　これは、ベンダー選定時に「ログ、監査、説明責任、情報管理」などがより厳密に問われることを意味し、PoCから本番への移行に時間がかかりやすい。

企業側の障壁としては、JUAS調査が示すように「効果は出たが測定が難しい」という運用課題があり、これが本番化の遅れにつながり得る。 [3]　また、生成AIの業務利用率は国際比較で日本が相対的に低いとの言及もあり、社内教育・業務設計・ガバナンス整備が普及の前提条件となる。 [75]

コスト面では、推論コストだけでなく、RAGのための文書整備（権限設計、メタデータ、更新運用）、安全対策、AI運用体制（MLOps/LLMOps）が総コストを押し上げる。ベンダーが「ガードレール」「監査」「脆弱性対応」など運用機能をパッケージ化するのは、こうした障壁を下げる意図とも整合する。 [76]

展望と推奨事項

2026年以降の見通し（公開情報からの帰結）

公開情報に基づけば、2026年は「PoCの横展開」と「専有環境での自律運用（モデル改善サイクル）」が一段進む年になり得る。富士通は中央省庁PoCを踏まえ、政策形成・立法支援の生成AIサービスをFY2026に提供する目標を掲げている。[9] 同時に、同社はモデル開発・運用・追加学習を企業が自律的に回すプラットフォーム提供（2026年7月正式提供予定）を打ち出しており、**「一回作って終わり」ではなく、継続改善を前提にした採用**が増える可能性が高い。[63]
NTT側もtsuzumi 2で「事前実証が多数」「導入実績の創出」を明記し、公共・金融・医療への知識強化を前面に出している。[77]

また、政府（経産省）のGENIAC成果公開は、公開モデルの性能・開発ノウハウが“商用利用可能な形で公開”されるケースを含み、内製・準内製の裾野を下支えする。[40] これにより、国産LLM採用は「ベンダー製モデルを買う」だけでなく「公開モデルを基盤に自社RAG/微調整で作る」という二極化が進むと推測される（推測である点は後述）。[78]

推奨事項（採用判断を“失敗しにくくする”実務提案）

国産LLM採用を検討する組織に対し、公開事例・ガイドライン・ベンチマーク動向から、以下を推奨する。

第一に、**要件を「主権性・機密性」「性能」「コスト」「統制（監査・安全）」に分解し、優先順位を明示**することである。国産LLMは「オンプレ/専有」など主権性で優位を作る一方、最高性能だけを見れば海外フロンティアが上回る局面があるため、要件分解がないと選定が迷走する。[79]

第二に、評価は**単一ベンチマーク禁止**を原則にし、（a）Nejumi等の多面的リーダーボード、（b）自社業務ベンチ（RAG正答率、監査要件、ヒヤリハット率）、（c）推論要件（GPU/遅延/同時実行）をセットで設計する。政府事業（GENIAC）も複数指標で評価しており、ベンダー各社も「RAG」「安全性」「エージェント」など能力分解を前提にしている。[80]

第三に、PoC設計では「本番障壁」を先に潰す。具体的には、①データ棚卸し（権限・更新・機密区分）、②ログ設計（説明責任・監査）、③プロンプト/ガードレールの運用手順、④効果測定（工数・品質・再作業率）をPoC段階から組み込む。JUAS調査が示す“効果測定”の弱さは本番化を遅らせるため、ここを仕様化することが投資回収に直結する。[81]

未確認情報・推定と仮定（明確化）

本レポートでは未公表情報を推測しない方針だが、分析上、以下は「推定・仮定」を含む（明確に区別する）。

- **推定**：国産LLM採用が「専有環境需要」と結びつきやすいのは、各社の提供形態（オンプレ/プライベート）や公共PoC事例から合理的に推測できるが、国産LLM全体の採用率を示す統計がないため、定量的シェアとしては断定できない。[82]
- **仮定**：価格が非公開（例：cotomi v3の推論API料金）な場合、比較表には「未指定」とし、海外API価格との単純比較は行わない。[83]
- **未確認**：一部モデル（例：cotomi v3のパラメータ数、tsuzumi 2の外部再現可能な総合スコア等）は、公式が数値を全面公開していないため、第三者検証可能な形での確定ができない。[84]

主要参考ソース

- NTT「tsuzumi 2」提供開始（2025-10-20） 85
- NTT「tsuzumi 2 進化のポイント」（推論コスト、30B、評価条件など） 6
- NTT「AI For Quality Growthの実現へ tsuzumi 2のリリース」（受注件数・相談件数等） 86
- NTTデータ「tsuzumi on Azure MaaS ユーザーガイド」（7B、Rakuda言及等） 18
- NEC「cotomi事業戦略」（130億級、Rakuda、30万字等） 21
- NEC「cotomi Pro/Light」（ELYZA Tasks 100、Japanese MT-Bench、速度比較） 22
- さくらインターネット「さくらのAI Engineでcotomi v3提供開始」（2025-12-02） 24
- 富士通「Takane」提供開始（JGLUE詳細スコア、Nejumiカテゴリスコア等、2024-09-30） 56
- 富士通「中央省庁でのTakane活用PoC（パブコメ）」成果（12万字→10分、80%超等、2026-02-03） 9
- 富士通「生成AI再構成技術」（1bit量子化、最大94%削減、精度維持率89%等、2025-09-08） 64
- ソフトバンク/SB Intuitions「Sarashina API」（Sarashina 4600億、法人提供、社内2万人トライアル） 30
- SB Intuitions「Sarashina2-Vision」（MIT、8B/14B公開） 33
- サイバーエージェント「6.8B日本語LLM公開」（CC BY-SA 4.0） 35
- 経済産業省 GENIAC「性能評価結果詳細」（JGLUE/Nejumi/MT-Bench等の成果公開） 40
- デジタル庁「生成AIの調達・利活用ガイドライン」（2025-06最終更新） 11
- JUAS「企業IT動向調査2025」速報（言語系生成AI導入41.2%） 3
- OpenAI API Pricing（価格レンジ） 67
- Anthropic Claude API Pricing（価格レンジ） 69
- Gemini API（Gemini 3.1 Pro Previewの価格・コンテキスト等） 87

---

1 2 11 74 79 https://www.digital.go.jp/news/3579c42d-b11c-4756-b66e-3d3e35175623
https://www.digital.go.jp/news/3579c42d-b11c-4756-b66e-3d3e35175623

3 16 29 37 60 81 https://juas.or.jp/cms/media/2025/02/it25_2.pdf
https://juas.or.jp/cms/media/2025/02/it25_2.pdf

4 18 https://www.nttdata.com/jp/ja/-/media/nttdatajapan/files/lineup/tsuzumi/tsuzumionazuremaasv10-2.pdf
https://www.nttdata.com/jp/ja/-/media/nttdatajapan/files/lineup/tsuzumi/tsuzumionazuremaasv10-2.pdf

5 27 https://www.fsastech.com/ja-jp/resources/topics/2025/1010.html
https://www.fsastech.com/ja-jp/resources/topics/2025/1010.html

6 13 57 58 59 https://group.ntt/jp/newsrelease/2025/10/20/pdf/251020ac.pdf
https://group.ntt/jp/newsrelease/2025/10/20/pdf/251020ac.pdf

7 54 86 https://group.ntt/jp/newsrelease/2025/10/20/pdf/251020ab.pdf
https://group.ntt/jp/newsrelease/2025/10/20/pdf/251020ab.pdf

8 21 47 https://jpn.nec.com/press/202312/20231215_02.html
https://jpn.nec.com/press/202312/20231215_02.html

9 23 45 https://global.fujitsu/en-global/pr/news/2026/02/03-01
https://global.fujitsu/en-global/pr/news/2026/02/03-01

10 38 https://news.panasonic.com/jp/press/jn240702-3
https://news.panasonic.com/jp/press/jn240702-3

12 44 80 https://prtimes.jp/main/html/rd/p/000000024.000119963.html
https://prtimes.jp/main/html/rd/p/000000024.000119963.html

14 19 36 53 61 73 77 85 https://group.ntt/jp/newsrelease/2025/10/20/251020a.html
https://group.ntt/jp/newsrelease/2025/10/20/251020a.html

15 26 39 56 72 https://www.fujitsu.com/global/about/resources/news/press-releases/2024/0930-01.html
https://www.fujitsu.com/global/about/resources/news/press-releases/2024/0930-01.html

17 25 40 68 78 https://www.meti.go.jp/policy/mono_info_service/geniac/selection_1/result_1/result_details_1/index.html
https://www.meti.go.jp/policy/mono_info_service/geniac/selection_1/result_1/result_details_1/index.html

20 46 https://www.fujifilm.com/fb/company/technical/ai/related-info003
https://www.fujifilm.com/fb/company/technical/ai/related-info003

22 41 62 https://jpn.nec.com/press/202404/20240424_01.html
https://jpn.nec.com/press/202404/20240424_01.html

24 28 83 84 https://cloud.sakura.ad.jp/news/2025/12/02/ai-engine-models-cotomi3/
https://cloud.sakura.ad.jp/news/2025/12/02/ai-engine-models-cotomi3/

30 42 70 https://www.softbank.jp/corp/news/press/sbkk/2025/20251105_01/
https://www.softbank.jp/corp/news/press/sbkk/2025/20251105_01/

31 https://www.nict.go.jp/press/2026/02/26-1.html
https://www.nict.go.jp/press/2026/02/26-1.html

32 https://www.sbintuitions.co.jp/blog/entry/2025/03/07/093143
https://www.sbintuitions.co.jp/blog/entry/2025/03/07/093143

33 https://www.sbintuitions.co.jp/blog/entry/2025/03/17/111659
https://www.sbintuitions.co.jp/blog/entry/2025/03/17/111659

34 63 76 82 https://global.fujitsu/ja-jp/pr/news/2026/01/26-02
https://global.fujitsu/ja-jp/pr/news/2026/01/26-02

35 55 https://www.cyberagent.co.jp/news/detail/id%3D28817
https://www.cyberagent.co.jp/news/detail/id%3D28817

43 https://arxiv.org/abs/2403.15484
https://arxiv.org/abs/2403.15484

48 https://pr.fujitsu.com/jp/news/2025/04/16.html
https://pr.fujitsu.com/jp/news/2025/04/16.html

49 https://japan-telework.or.jp/teleworknext/ntt
https://japan-telework.or.jp/teleworknext/ntt

50 https://news.mynavi.jp/techplus/article/20251105-3628490/
https://news.mynavi.jp/techplus/article/20251105-3628490/

51 66 https://digitalpr.jp/r/120589
https://digitalpr.jp/r/120589

52 https://www.pref.yamaguchi.lg.jp/press/273242.html
https://www.pref.yamaguchi.lg.jp/press/273242.html

64 https://global.fujitsu/ja-jp/pr/news/2025/09/08-01
https://global.fujitsu/ja-jp/pr/news/2025/09/08-01

65 69 https://docs.anthropic.com/en/docs/about-claude/pricing
https://docs.anthropic.com/en/docs/about-claude/pricing

67 https://openai.com/api/pricing/
https://openai.com/api/pricing/

71 87 https://ai.google.dev/gemini-api/docs/gemini-3
https://ai.google.dev/gemini-api/docs/gemini-3

75 https://www.jri.co.jp/report/economistcolumn/detail/16445/
https://www.jri.co.jp/report/economistcolumn/detail/16445/