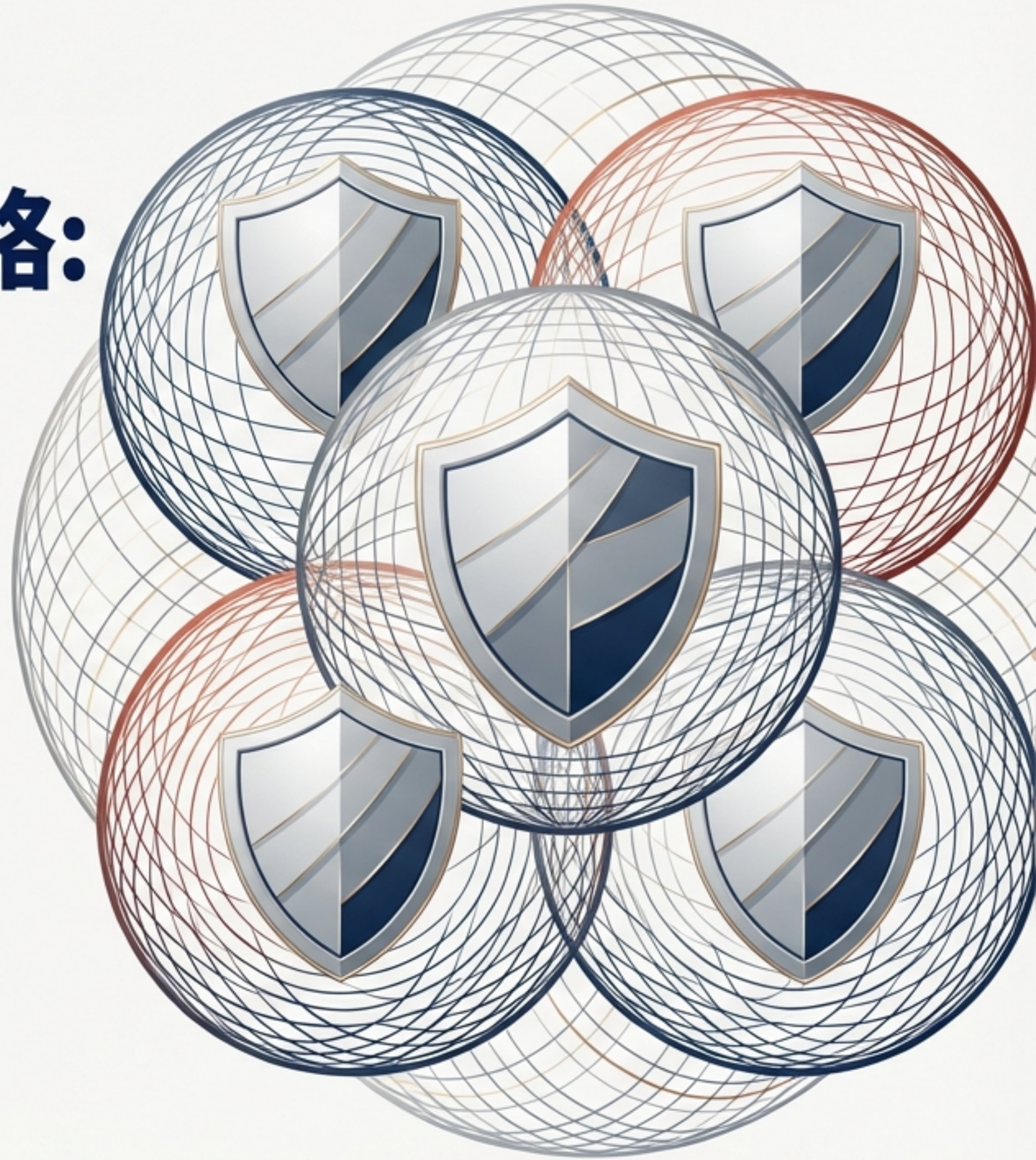
2026年の生成AI導入戦略: 国産LLMが切り拓く「エンタープライズ本番運用」への道

セキュリティ・コスト・データ主権の壁を越えるための実践的プレイブック

経営層・CIO / DX推進リーダー 各位



生成AI利用は加速。しかし「本番運用の壁」が立ちはだかる

言語系生成AI導入率
41.2%（準備中含む）

導入企業の**約7割**が
効果を実感

出典：JUAS「企業IT動向調査2025」速報

データ主権・機密性
（社外に出せないデータ）
データ主権を機密するデータなど
社外に出せないと考える

ガバナンス
（監査・脆弱性リスク）
ガバナンスに、監査に論理するる、
監査や脆弱を検証する

コストと効果測定
（推論コスト高騰・効果測定の難しさ）
推論コストの高騰により、コストと
効果測定・効果測定の難しさ

国産LLMの急浮上

公共・金融・重要インフラ
での採用加速

国産LLM市場の急速な進化（2023〜2026）

2023 黎明期
国産LLM開発の第一波

- 5月: CyberAgent 6.8B公開
- 11月: NTT「tsuzumi」発表
- 12月: NEC「cotomi」130億級方針

2024 実装準備
特化型モデルと商用展開

- 4月: cotomi Pro/Light発表
- 7月: Panasonic×Stockmark 100B専用モデル開発
- 9月: 富士通「Takane」提供開始

2025 規格化と展開
公共・法人向けサービス拡大

- 5月: デジタル庁「生成AIの調達・利活用ガイドライン」策定
- 10月: tsuzumi 2 (30B/1GPU推論)
- 11月: SoftBank「Sarashina API」法人提供

2026 本番横展開
自律運用と社会実装

- 2月: 中央省庁でのTakaneパブコメPoC成果公表
- 専有環境での自律運用・モデル改善サイクルが本格化

用途で分かれる、国産LLM「3つの戦略的類型」

軽量・オンプレ運用型

推論コスト圧縮 × 閉域運用

比較的小規模で、1〜2GPUでの高速推論・オンプレミス稼働を前提とする。

代表モデル：
tsuzumi（7B, 30B），
cotomi（130億級）

エンタープライズ・プライベート特化型

高精度 × 監査・統制

プライベート環境での展開を前提とし、エンタープライズの厳格なセキュリティ要件に統合される中規模〜大規模モデル。

代表モデル：
Takane（富士通）

オープンウェイト・素材型

内製RAG × 追加学習の基盤

商用利用可能な形で公開され、企業や政府が自社専用モデルを構築するためのベースとなる素材。

代表モデル：
Sarashina2.2, tanuki-8x8b,
OpenCALM

主要国産LLM・最新カタログ（2026年版）

tsuzumi 2（NTT）

30B　プロプライエタリ　オンプレ/プライベート

強み：1GPU推論前提。圧倒的なコストパフォーマンス。

cotomi Pro / Light / v3（NEC）

非公開/130億級　プロプライエタリ　API/セキュア環境

強み：GPU1〜2枚での高速推論、業種特化モデル・マネージドAPI展開。

Takane（富士通）

中規模　独占提供　DI PaaS統合

強み：JGLUE等で高スコア。セキュアなプライベート環境と監査機能の統合。

Takane（富士通）

中規模　独占提供　DI PaaS統合

強み：日本語最高水準を目指す。社内2万人トライアルを経た法人向けAPI提供。

Sarashina 系 (SoftBank)

最大460B　API / 公開モデル

強み：日本語最高水準を目指す。社内2万人トライアルを経た法人向けAPI提供。

GENIAC発・独自開発

100B級 / MoE　商用可/社内専用

強み：tanuki-8x8b, PLaMo-100Bなど。自社活用や内製基盤として活躍。

どこでどう使われているか？「3つの典型的な導入パターン」

自治体・公共領域
(PoCから横展開へ)

公用文・例規・住民向け文書。
閉域運用志向。

事例：山口県(tsuzumi実証),
神戸市・相模原市(cotomi)

金融・医療・製造
(機微情報を扱う専有環境)

セキュアなLLM環境でのRAG、
社内問い合わせ。

事例：大塚商会(生成AIサーバ),
中国電力(tsuzumi 2 閉域トライアル)

大企業の内製・準内製
(自社専用モデル構築)

業務・データに合わせた専用LLMの
スクラッチ/追加学習開発。

事例：Panasonic HD×Stockmark
(100B社内専用モデル)

導入事例ハイライト：実証されたROIと実効性

圧倒的な処理速度と精度（富士通 / 中央省庁）

12万字 → 約10分

パブリックコメント業務PoC。
分類・要約を実行。
法案条文と意見の対応付けで
80%超 の該当条文特定に成功。

大規模な導入・引き合い（NTT / tsuzumi）

受注 1,827件
相談 1,818件

FY25 1Q時点(生成AI関連全体)。
公共領域での高い関心。
事前実証を多数経て本番導入へ。

全社規模のトライアル（SoftBank / Sarashina）

約2万人

Sarashina mini API提供前の
社内大規模トライアルを
実施し、品質を検証。

国産LLM選定を左右する「4つの意思決定軸」

① セキュアな専有環境

オンプレ / プライベートクラウド

クラウド依存の回避と、データを組織内に留める物理的・論理的統制。

② 日本語品質と実務適用

敬語・公用文・専門語

RAG構成時の正答率、公的文書や業界特有の用語に対する安定した出力。

③ 推論コストと運用要件

GPU枚数・メモリ

ランニングコスト（TCO）の抜本的削減と、少ないハードウェアリソースでの自律稼働。

④ ガバナンス

ログ・監査・脆弱性対策

行政の調達ガイドライン等に適合するリスク管理と説明責任の担保。

[評価軸1・4] 最大の差別化要因：セキュリティとガバナンス

脆弱性対応とガードレール

富士通「Kozuchi」は独自定義を含む7,700種超 の脆弱性スキャナーと自動ルール適用を実装。

学習データの統制

NTT「tsuzumi 2」は新聞等のデータの自主削除など、データのコントロール性を明言。

行政水準のコンプライアンス

デジタル庁「調達・利活用ガイドライン」が求める「ログ、監査、説明責任、情報管理」の厳密な要件に適合するオンプレ・閉域設計。

Public APIs

[評価軸3]「推論コスト1/10〜1/20」という破壊的インパクト

国産最適化 (tsuzumi 2 / 30B) : 約 500万円
（A100 40GB相当 1基で推論可能）

海外400B級モデル : 約 5,000万円

海外700B級モデル : 約 1億円

富士通の1ビット量子化技術
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。

NEC cotomi Pro/Light
標準GPU2枚で海外メガモデルの約1/8〜1/15 の時間で高速処理。

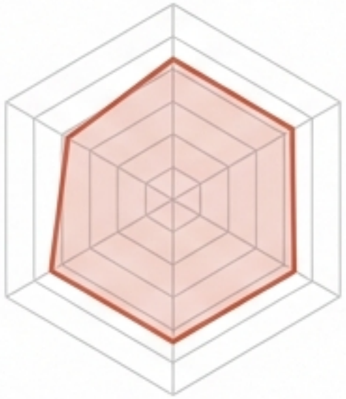
[評価軸2]「事実上の標準化」が進む多面的な日本語評価体系

単一のベンチマークではなく、「多面的評価」が実務のスタンダードへ。

Nejumi LLM リーダーボード 4

高難度推論、アプリ開発能力（関数呼び出し）、安全性を総合評価。

JGLUE / JSQuAD



日本語言語理解の基礎力
（例: TakaneはJGLUE平均0.92を記録）。

Japanese MT-Bench / ELYZA Tasks 100

対話・作文・指示追従能力の測定。

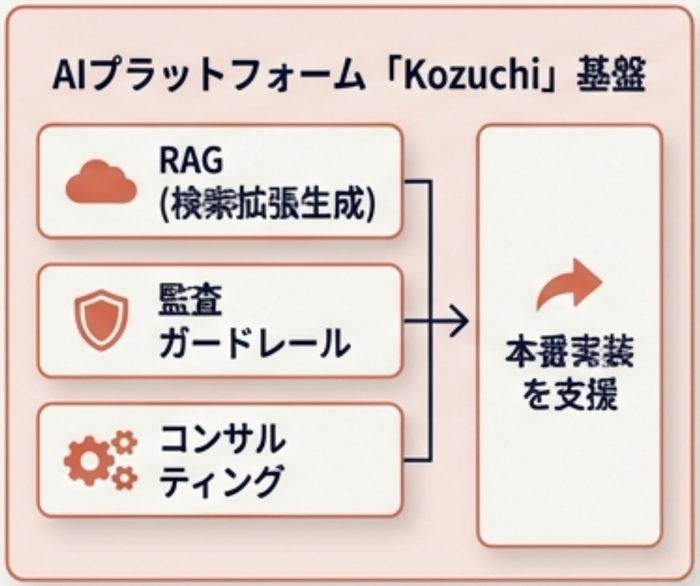
経産省「GENIAC」の成果公開でも複数指標の組み合わせが前提となっており、自社RAG環境での再現検証が必須に。

ベンダーのビジネスモデル：国産LLMはどう「売られて」いるか

各社の戦略と提供形態の多様化

AIプラットフォーム統合型

例：富士通 Takane



モデル単体ではなく、「Kozuchi」等の基盤に統合。RAG、監査ガードレール、コンサルティングまでを一体提供し、本番実装を支援。

マネージドAPI＋業種特化型

例：NEC cotomi

自治体・金融・医療などの実証知見を活かした「業種特化モデル」を整備し、セキュアなマネージドAPIとして展開。

ソブリン / オンプレ推進型

例：NTT tsuzumi 2

「1GPU推論可能」という軽さを前面に出し、自治体や企業の「どうしても外に出せないデータ」を扱う閉域・オンプレ需要を直撃。

「PoC死」を招く、エンタープライズ導入のリアルな障壁

ガバナンス要件の壁

デジタル庁ガイドライン等で求められる「ログ、監査、説明責任」の厳密化によるリードタイム長期化。

効果測定の壁（JUAS調査のジレンマ）

「効果は出たが測定が難しい」―定量的なROIが示せず、本番化の稟議が停滞する。

データ準備とMLOpsの壁

RAGのための文書整備（権限設計、メタデータ、更新運用）や、AI運用体制（MLOps/LLMOps）の構築工数が総コスト（TCO）を押し上げる。

成功へのプレイブック：2026年以降を見据えた「3つの推奨アクション」

1

1. 要件の厳密な分解と優先順位付け

「主権性・機密性」「性能」「コスト」「統制（監査）」のどれを最優先するかを定義する。
最高性能のみを追うと選定が迷走する。

2

2. 自社専用の「多面的ベンチマーク」設計

公開リーダーボードに依存せず、自社業務のRAG正答率、監査要件、ヒヤリハット率、
GPU遅延をセットにした評価軸を構築する。

3

3. PoC段階での「本番障壁」の事前排除

PoCの設計段階から、「データ棚卸し（機密区分)」「ログ設計」「効果測定（工数・再作業率)」を仕様に組み込み、PoC死の要因を先に潰す。

**「一回作って終わり」ではなく、継続改善（自律運用）を見据えた
プラットフォーム選定が、生成AI時代の競争力を決める。**