2026年 日本の国産LLM採用状況：包括的調査報告

実験的プロジェクトから「国家の重要インフラ」への転換と最前線

Research Foundation provided by: perplexity

2026年は国産LLMの「社会実装元年」

過去 -
実験的プロジェクト

- PoC（概念実証）中心の運用
- 汎用タスクでの精度検証
- 海外APIへの過度な依存

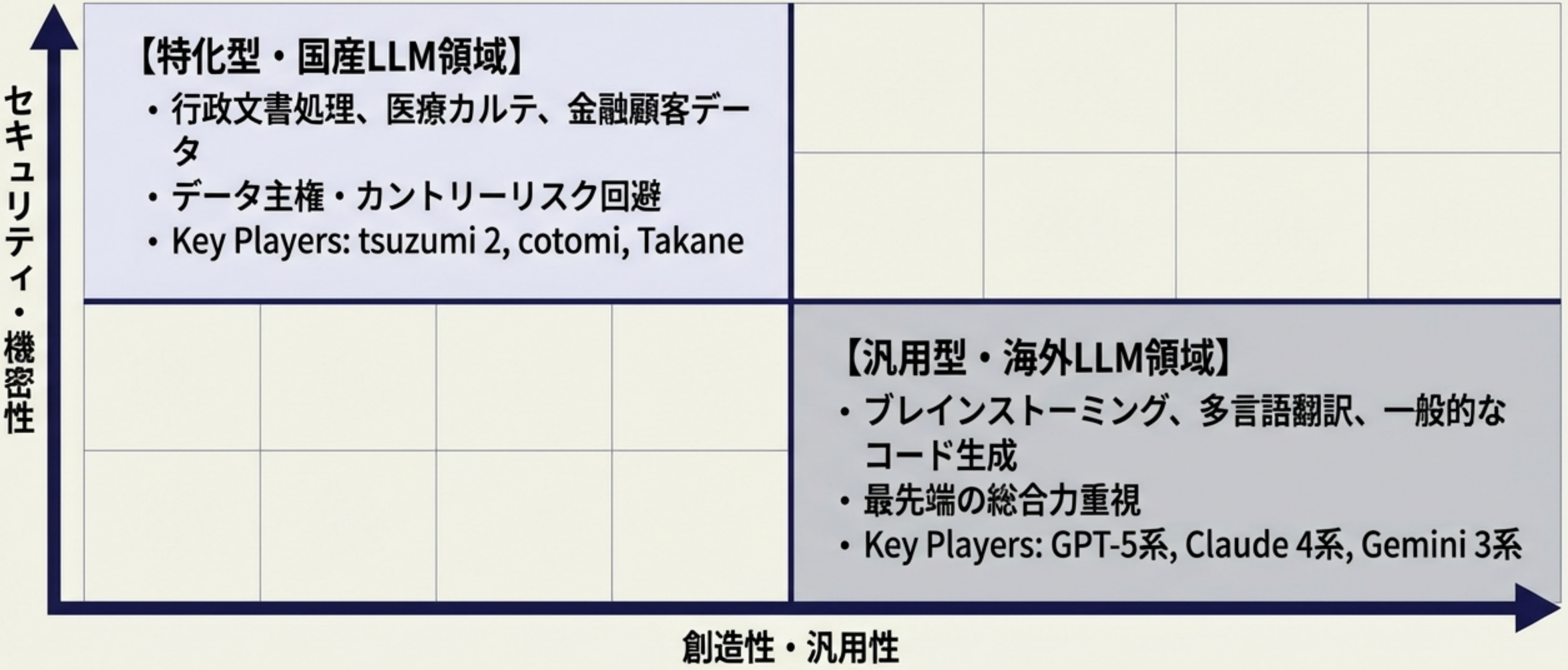
現在（2026）-
ミッションクリティカルな実装

- 金融・医療・自治体・製造業での本格稼働
- **3兆円**規模の官民連携インフラ投資
- 海外製モデルとの明確な「使い分け」フェーズへの移行

未来（2030）-
AI主権（経済安全保障）

- 完全国産エコシステムの確立
- 機密データの国内完結処理
- 自律型エージェント・フィジカルAIの社会基盤化

ハイブリッド戦略の定着：国産と海外モデルの明確な棲み分け



ハイブリッド運用によるコスト最適化：円安環境下での海外APIコスト増を回避し、
機密領域を国内オンプレミス・エッジ環境へ移行する企業が急増。

官民総額3兆円：かつてないスケールのインフラ投資

経済産業省：5年間で

約**1兆円**の

支援（2026年度予算案に3,000億円超計上）

ソフトバンク：6年間でデータセンターに

2兆円

円を投資

物理インフラ

- 国内最大級AI計算基盤の整備
- 苫小牧市・堺市のデータセンター稼働（2026年度）

出力・目標

- 新会社設立: ソフトバンク、PFNなど十数社が出資。約100人規模の技術者集団。
- 国家目標：

1兆パラメータ

規模の世界最先端基盤モデル開発。

- 最終形態：ロボットに搭載可能な「フィジカルAI」の実装。

ガバメントAI「源内」：市場を牽引する巨大な初期需要

最大の狙い：海外モデルへの依存から脱却し、日本語・行政文書・公的業務に最適化された完全国産AIの育成エコシステムを政府主導で構築する。

技術的アーキテクチャの分岐：規模追求型 vs 効率特化型

規模追求型（Scale-Seeking）

- アーキテクチャ: MoE（Mixture of Experts）
- パラメータ規模: 数千億〜7,000億クラス
- 目的: AGI的振る舞い、クラウドでの大規模サービス、汎用能力の極大化
- 代表プレイヤー: 楽天、ソフトバンク

効率特化型（Efficiency-Focused）

- アーキテクチャ: 軽量ベースモデル＋タスク特化アダプタ
- パラメータ規模: 7B〜30Bクラス
- 目的: オンプレミス、エッジデバイス、閉域環境での低コスト運用
- 代表プレイヤー: NTT、NEC、ELYZA

計算資源の制約から生まれた日本の「軽量化技術」は、グローバル市場でも通用する独自の競争力へ進化。

規模追求型プレイヤー：グローバル水準への挑戦

楽天

モデル名：Rakuten AI 3.0

スペック：約**7,000億**パラメータ（MoE）/ アクティブ40B

実績：日本語MT-Benchスコア**8.88**（GPT-4oの8.67を凌駕）

戦略：2026年春に「オープンウェイト」として無償公開予定。日本のAI研究エコシステム全体の底上げを狙う。

ソフトバンク

モデル名：Sarashina 2 MoE

スペック：8 × 70B パラメータ

実績：「Large Telecom Model」として社内運用済み（通信品質予測精度90%超）

戦略：経済安全保障を担保し、学習から運用まで完全に国内で完結する生成AI基盤の構築。

効率特化型プレイヤー：ハードウェア制約を打破する技術力

NTT

モデル名：tsuzumi 2

スペック：30B（7B / 0.6Bモデルも提供）

強み：**1枚のGPU**で動作可能な圧倒的軽量・省電力設計

実績：グループAI受注額約1,478億円。tsuzumi単体で金融・医療を中心に**2,000件**の引き合い。

未来：光通信網で複数モデルを結ぶ「AIコンステレーション」構想。

NEC

モデル名：cotomi

スペック：130億パラメータ / 128kロングコンテキスト対応

強み：GPT-4oと比較して**2.2倍**の推論速度

実績：生成AI関連事業で売上500億円目標。

未来：2026年1月より自律型AI「cotomi Act」を展開。

特化型・主権確保プレイヤー：
完全なる国内エコシステムの構築

富士通 - Takane

- 130億パラメータ。Cohere社との共同開発。
- **行政特化:** 中央省庁でのパブリックコメント業務実証に成功。約12万文字の賛否分類・要約を10分で完了（正答率8割超）。2026年度サービス化。

PFN × さくらインターネット × NICT - PLaMo 2.0

- 310億パラメータ。完全フルスクラッチの純国産。
- **データ優位性:** NICTが保有する700億ページ超の日本語Webデータを独占的活用。クラウドからアプリまで完全国内完結。

ELYZA（KDDI）

- 1,300億パラメータ（Llamaベース）。
- **商用展開:** ローカル実行に最適化され、KDDIのWAKONX基盤を通じて商用提供中。

採用が加速するミッションクリティカル分野と投資対効果

行政・公共

導入モデル: ガバメントAI「源内」, Takane

ユースケース: パブリックコメント処理、法令整合性チェック

KPI: 12万文字を **10分** で処理（富士通実証）

金融・医療

導入モデル: tsuzumi 2

ユースケース: 顧客データ分析、医療カルテ処理（閉域網処理）

KPI: FP技能試験2級レベルの専門知識を少量の追加学習で習得

製造業

導入モデル: cotomi, Stockmark

ユースケース: 設計図面・部品表の構造理解、熟練ノウハウの継承

KPI: パナソニック等、大手製造業での採用実績拡大

通信インフラ

導入モデル: Sarashina (Large Telecom Model)

ユースケース: 通信品質の予測とインフラ最適化

KPI: 予測精度 **90%以上** を達成

国産LLMの現在地：独自の強みと直面する構造的課題

【強み】

データ主権と安全性: 機密データの国内処理、カントリーリスクの完全排除

日本語処理の最適化: 敬語・謙譲語・業界専門用語の正確な理解

エッジAIへの適合性: 世界トップクラスの軽量化技術、低コストなオンプレ運用

コスト効率: 円安による海外APIのコスト高騰に対する強力なヘッジ

【課題】

汎用性能のギャップ: 総合力では依然としてGPT-5/Claude 4系に劣後。1兆パラメータ級「怪物モデル」の不在。

GPU調達のボトルネック: NVIDIA製GPUの激しい争奪戦と円安による調達コスト増。

「2026年問題」: 高品質な日本語学習用データの枯渇と生成基盤の不足。

タレント不足: AI研究者・トップエンジニアの慢性的な不足。

次のパラダイムシフト：「チャット」から「自律型エージェント」へ

過去: 対話型AI

- プロンプトに対する受動的な応答
- テキスト生成と要約が中心

現在: RAG・外部知識連携

- 社内データと連携したハルシネーション抑制
- 特定業務の効率化・アシスタント機能

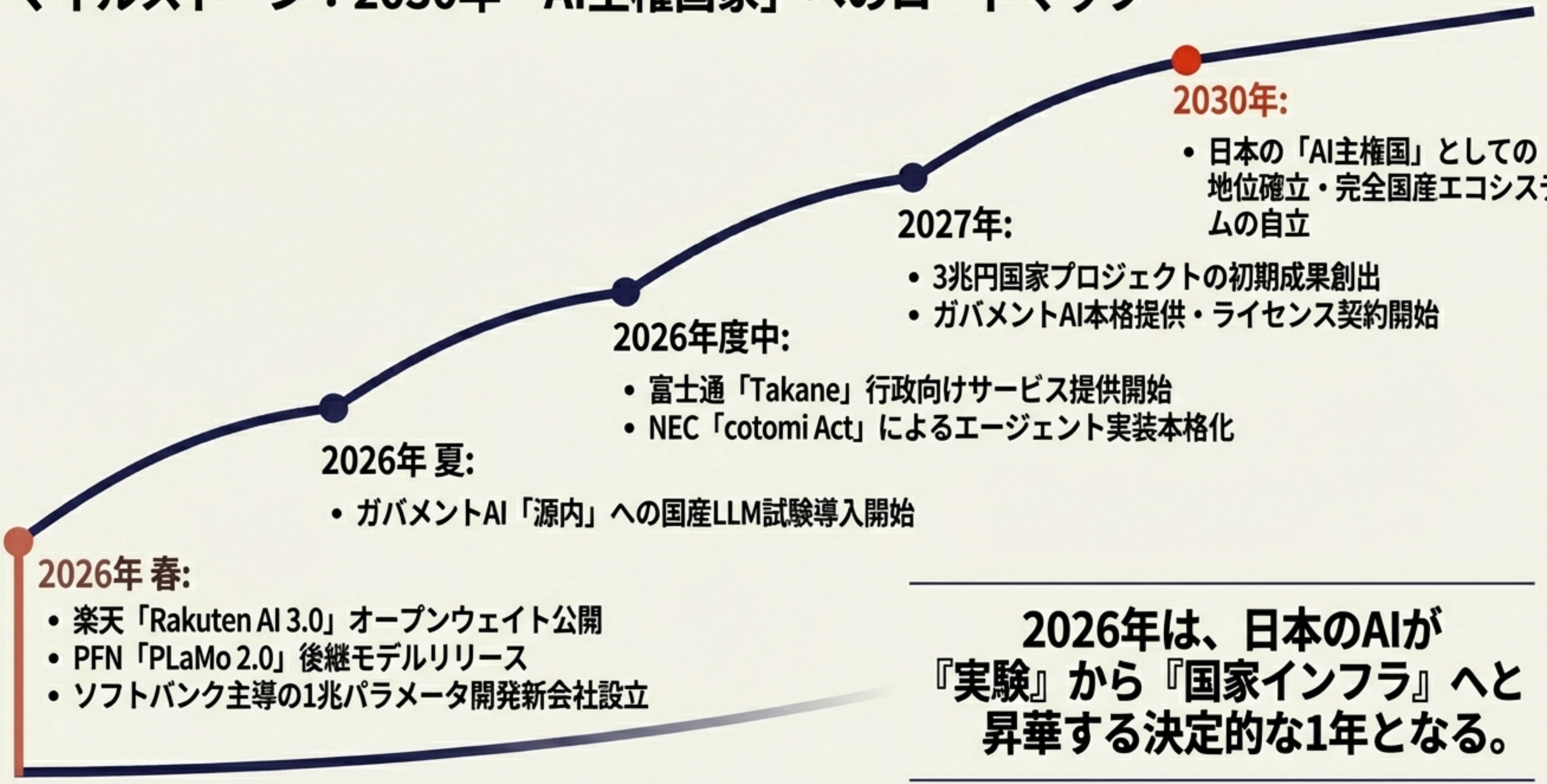
2026年以降: 自律型エージェント

- AIが計画立案・システム操作・外部API連携を自律的に実行

【国産の最前線】

- NEC：「cotomi Act」による業務ノウハウ自動抽出・組織資産化
- ソフトバンク：AI-RAN構想によるインフラの自律制御
- 新会社構想：ロボット制御を担う「フィジカルAI」への発展

マイルストーン：2030年「AI主権国家」へのロードマップ



2026年 春:

- 楽天「Rakuten AI 3.0」オープンウェイト公開
- PFN「PLaMo 2.0」後継モデルリリース
- ソフトバンク主導の1兆パラメータ開発新会社設立

2026年 夏:

- ガバメントAI「源内」への国産LLM試験導入開始

2026年度中:

- 富士通「Takane」行政向けサービス提供開始
- NEC「cotomi Act」によるエージェント実装本格化

2027年:

- 3兆円国家プロジェクトの初期成果創出
- ガバメントAI本格提供・ライセンス契約開始

2030年:

- 日本の「AI主権国」としての地位確立・完全国産エコシステムの自立

2026年は、日本のAIが『実験』から『国家インフラ』へと昇華する決定的な1年となる。
