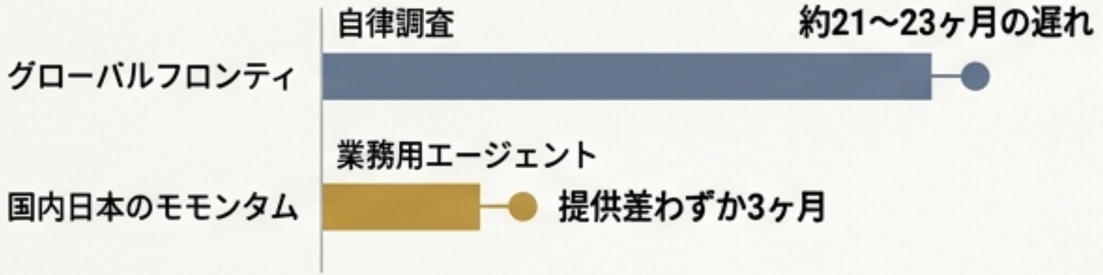
国産LLMの現在地とAIエージェントの国際比較

2026年後半における技術格差の実態とエンタープライズ実装戦略

Gemini Notebook

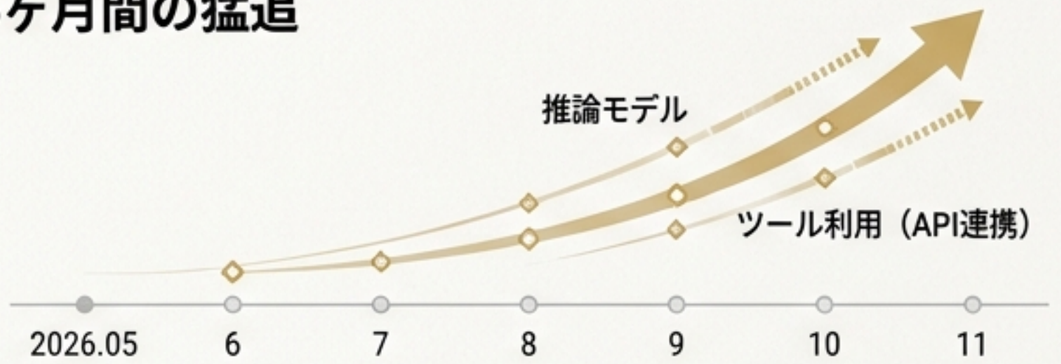
単一の「遅れ」は存在しない：ギャップの解体と実装の現実

技術の非対称性



「自律調査」の遅れは約21〜23ヶ月。「業務用エージェント」の提供差はわずか3ヶ月。

6ヶ月間の猛追



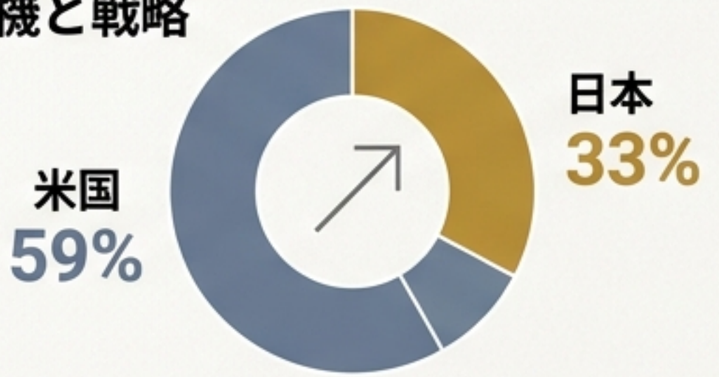
2026年5月〜11月の間で、国内の研究機関・企業が推論モデルとツール利用（API連携）の実装を急速に推進。

パフォーマンスの現実



日本語の関数呼び出しや業務対話（BFCL-nejumi, J-tau）では国産モデルが海外モデルを凌駕。一方、長時間にわたる自律作業には課題が残る。

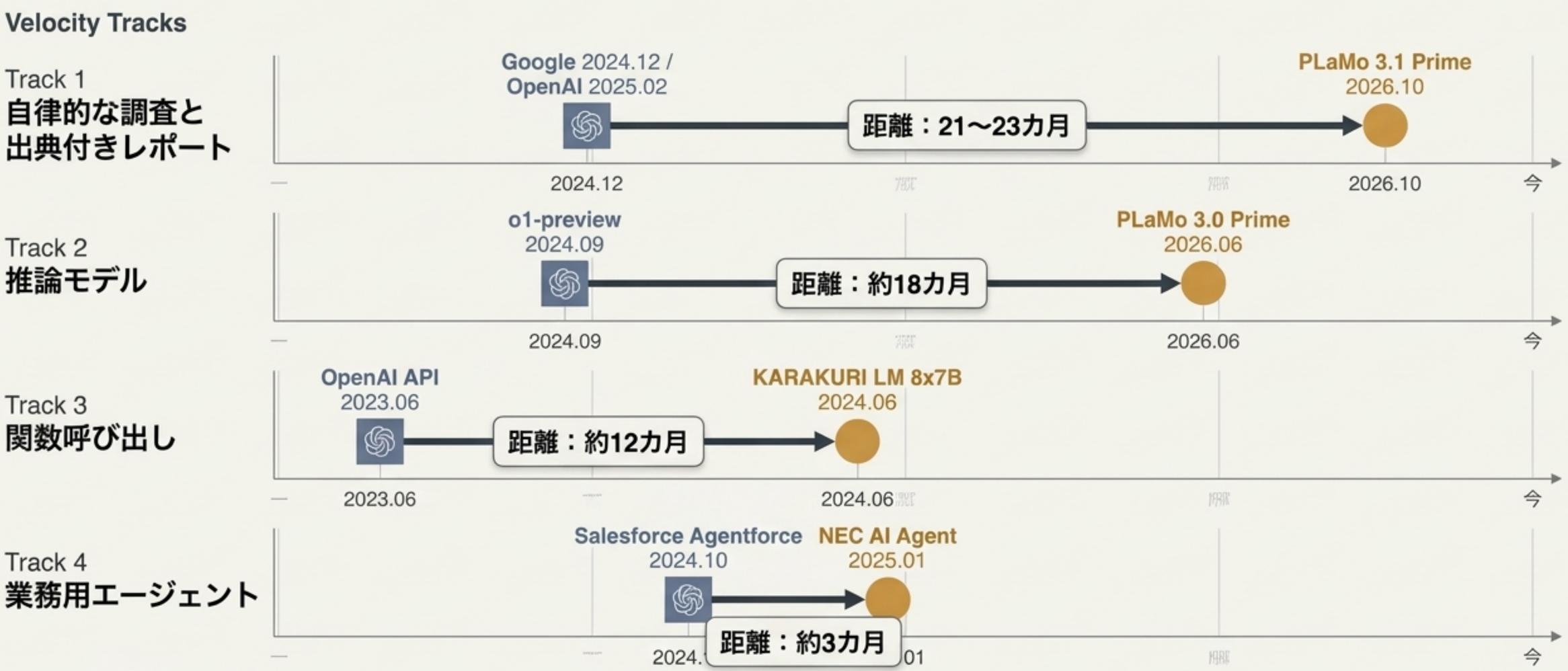
採用の危機と戦略



日本企業の導入率は33%（米国59%）。基礎モデルのキャッチアップを待たず、今すぐ「特化型エージェント」を実装することが至上命題。

Gemini Notebook

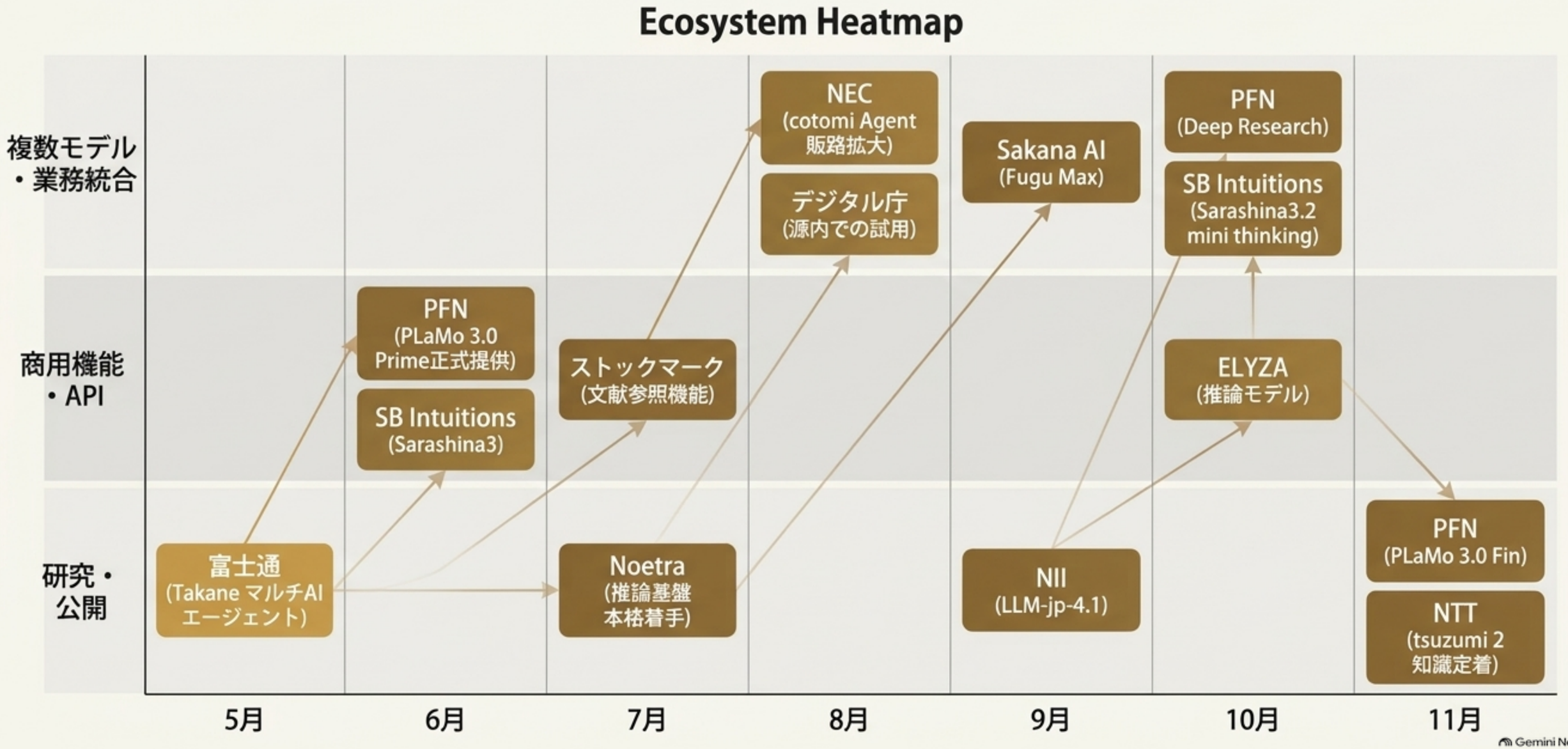
レイヤーによって劇的に変化する「グローバルとの時間差」



「基礎能力」のギャップは依然として大きいが、「業務適用（エージェント）」のギャップは急速に縮小している。

Gemini Notebook

エコシステムの猛追：2026年後半の6ヶ月間における爆発的進化

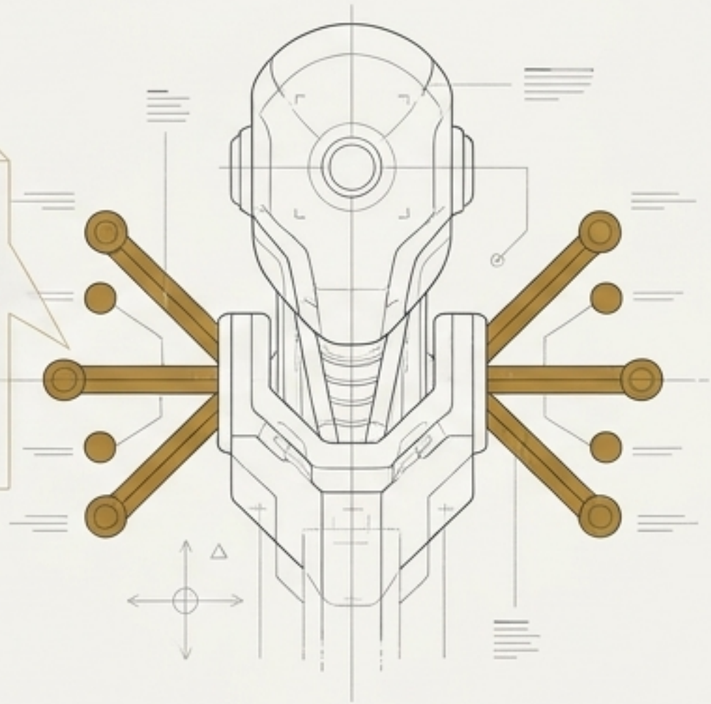


Gemini Notebook

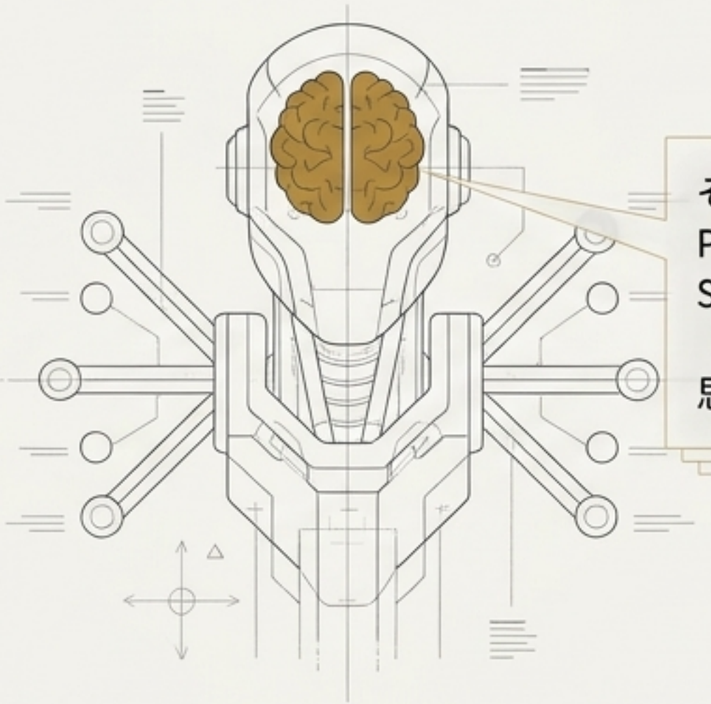
日本独自の技術進化：「思考」に先行した「ツール操作」の獲得

ステップ1（手：ツール連携・外部API）

PFN PLaMo 2.1 Prime（2025年10月）およびSarashina3（2026年6月）において、外部APIの自動選択やツール利用が先に実装。



ステップ2（脳：推論・深考）



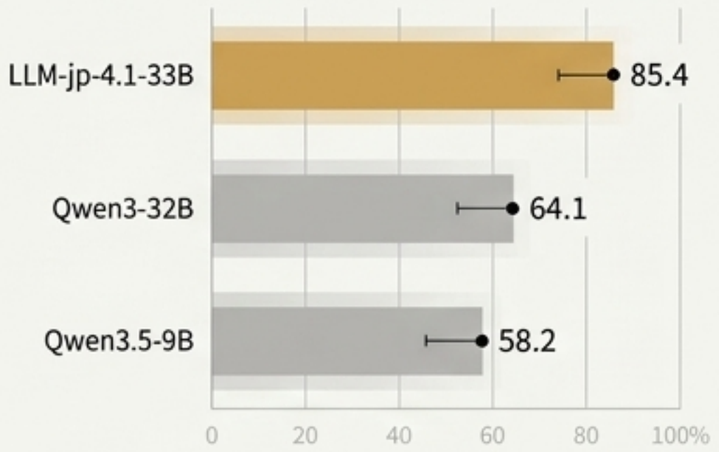
その後、PLaMo 3.0 Prime（2026年6月）およびSarashina3.2 thinking（2026年10月）において思考モデルが登場。

日本のモデルは、高度な基礎推論力を獲得する前に、業務システムと接続するための「手（関数呼び出し）」を先に鍛え上げた。これが現在実務で機能している理由である。

Gemini Notebook

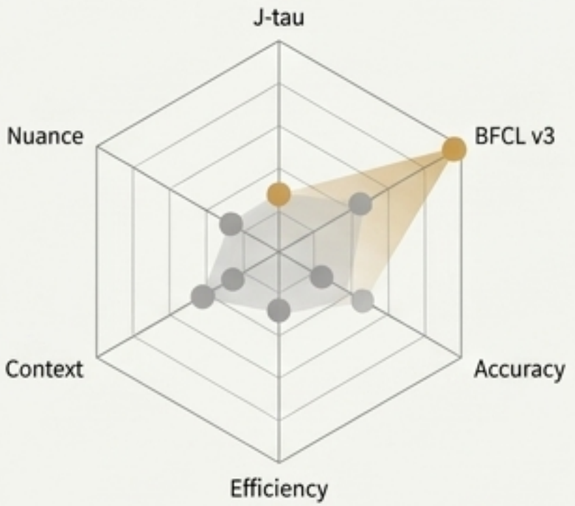
パフォーマンスの実態 #1：日本語コンテキストにおける明確な優位性

関数呼び出しと複数ターン



日本語版関数呼び出し評価「BFCL-nejumi」の複数ターン課題において海外モデルを上回る。

業務対話と最新スコア



「Sarashina3.2 mini thinking」が、日本語業務対話評価「J-tau」および「BFCL v3」において、GPT-5.4 miniを含む比較対象中で最高スコアを記録。

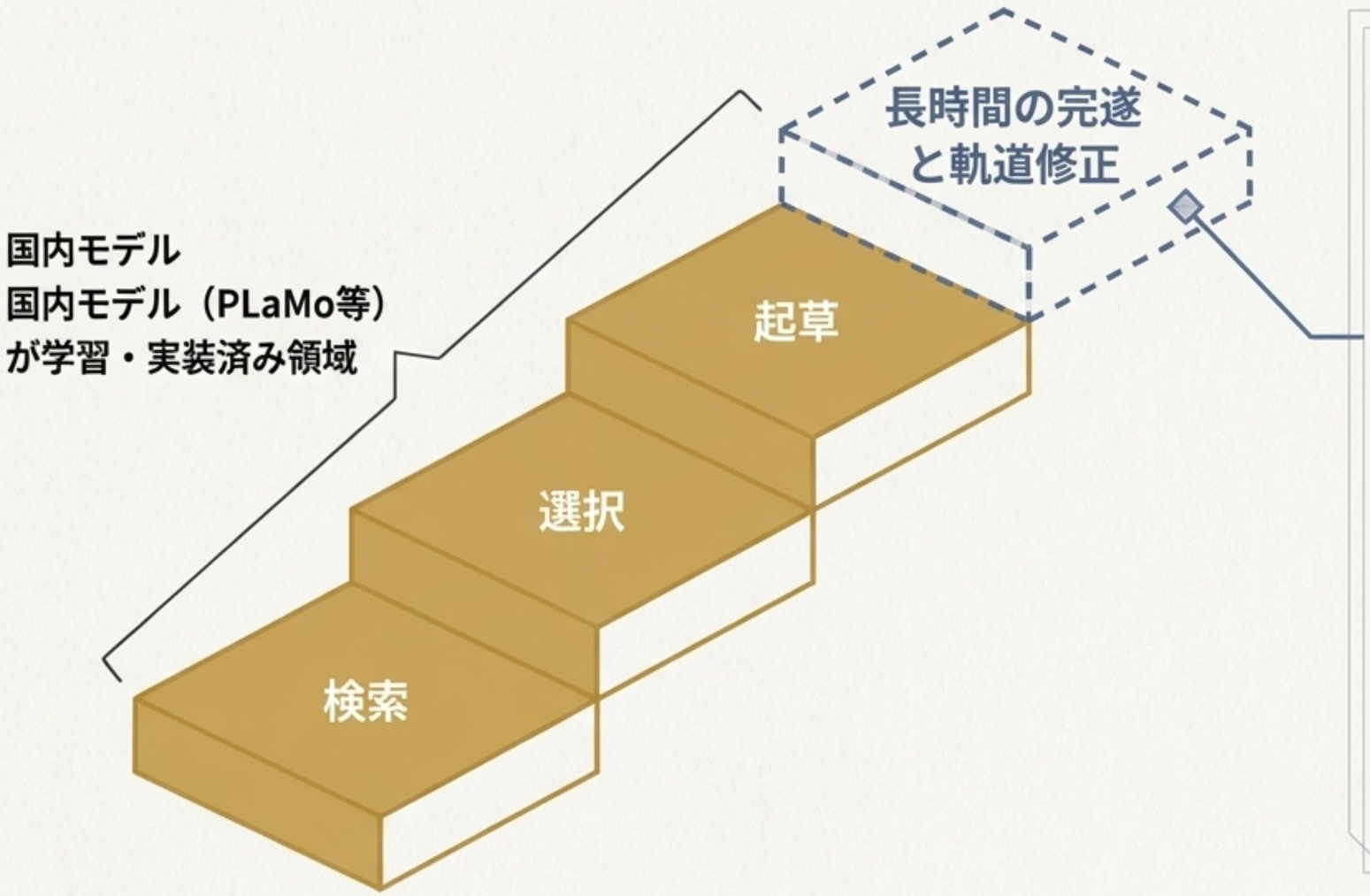
自律調査の進化



PLaMo 3.1 Primeの推論モードが、Web検索レポート作成において旧モデル比で**平均11.1ポイントの劇的な改善。**

Gemini Notebook

パフォーマンスの実態 #2：長時間の自律作業に残る「空白」



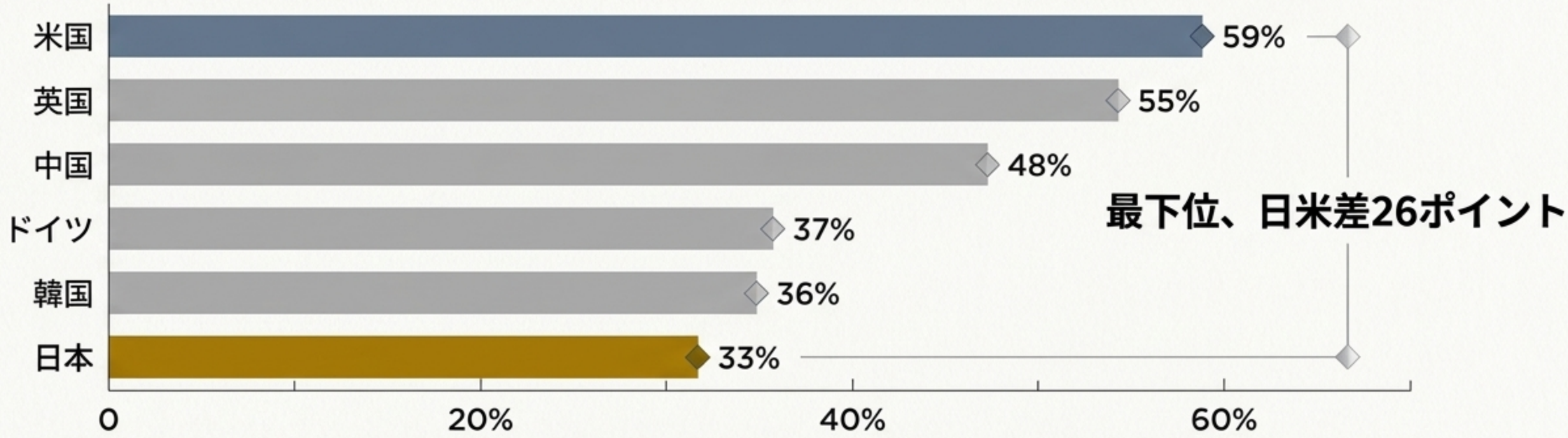
グローバルの評価基準（METRのソフトウェア課題完遂率、OpenAIの長時間の知識労働評価「GDPval v2」）に対し、同じツール・時間・費用上限で比較した国内データは不足している。英語の対話型業務「τ^2-bench」でも差が残存。

この「長時間にわたる自己補正と完遂」こそが、21〜23ヶ月の技術的ギャップの正体である。

Gemini Notebook

実装の危機：世界との間に広がる「認識と導入」のギャップ

AIエージェントについて理解し、導入済み／導入中の割合（売上500億円以上）



技術開発における3ヶ月の差ではなく、企業側の「実装に向けた決断」の遅れが、最も深刻なギャップを生み出している。

Gemini Notebook

実装のシグナル：「検証」から「大規模運用」への移行

政府・インフラ領域 / デジタル庁



- 「源内」での国産基盤モデル（tsuzumi 2, Takane 32B, PLaMo 2.0 Prime）をさくらのクラウド上で試用開始。
- 政府職員「約18万人」を対象とした順次導入の計画。

エンタープライズ領域 / 楽天



- 社内AI Workforceにおいて月間利用者「7,100人以上」のアクティブ運用実績。
- 消費者向け20サービスへのエージェント提供計画。

現場導入に向けたハイブリッド検証

ELYZA Voice Agent等、現場導入に向けた検証段階のプロジェクトも多数進行中。
基盤モデルの国籍を問わず、適材適所のハイブリッド構成が現実解となっている。

Gemini Notebook

評価のブループリント：知財調査におけるエージェントの実務評価フレームワーク

入力条件の統制

[同じ特許DB]
＋
[同じ工数上限]

自律プロセスの実行

[検索式の修正] → [文献選択] → [請求項と根拠段落の対応付け]

※未公開情報にはPFNのような閉域調査を活用

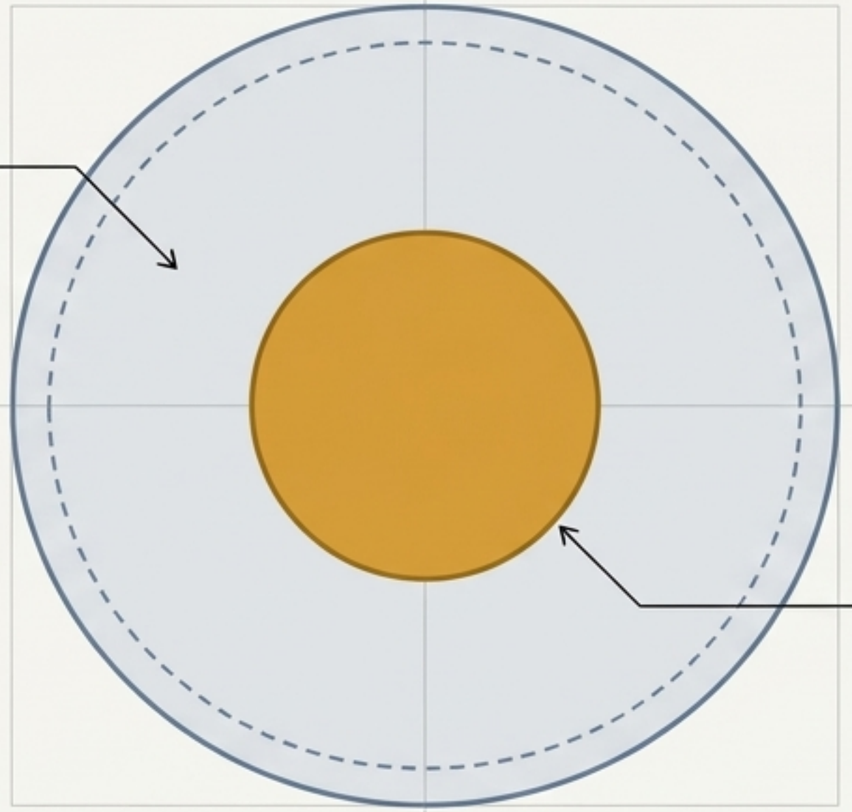
ROI測定マトリクス（4つの重要指標）

1. 既知文献の回収率 (Recall)	2. 根拠段落の正確さ (Accuracy)
3. 人間が修正に要した時間 (Human-Hours Saved)	4. 完遂1件当たりの費用 (Cost per Task)

カタログ上の性能ではなく、
「人間の専門家の時間をどれだけ削減できたか」を測定せよ。

Gemini Notebook

戦略的至上命題：基礎モデルの「完成」を待つな



「20ヶ月の基礎推論ギャップ」：
汎用的な長期間の自律作業能力。
ここは基礎モデルの開発側に任せる領域。

「3ヶ月の実装ギャップ」：
業務特化型エージェントの提供。
日本企業が今すぐ戦うべき領域。

日本のAIモデルは「日本語でのツール利用」という実務に直結する能力において、すでに戦える水準にある。長時間の汎用自律性における「20ヶ月の遅れ」を理由に導入を躊躇することは、致命的な経営判断の誤りである。「３ヶ月の遅れ」しかない業務用エージェントを今すぐ実装し、組織側のAI受容能力を高めること。窓は今、開いている。

Gemini Notebook