国産LLMの現在地：推論モデル到達と「エージェント化」の新たな壁

18カ月の遅延構造と、エンタープライズに向けたハイブリッド設計の指針

2026年11月20日時点 最新調査レポート
出典：Claude Fable 5.2 調査報告

Gemini Notebook

Executive Summary: The Narrative Arc



1. The Reality: 18カ月の遅延構造

国産LLMに推論（思考）モデルが出揃った一方で、高度なAIエージェント化においても海外フロンティアモデルと間に約18〜22カ月の構造的な遅れが継続している。



2. The Pivot: 戦略的ピボットと「独自の堀」

国産ベンダーの主戦場は、フロンティア性能の追従から、「軽量・ソブリン・オンプレミス環境での機密データ処理」へと明確にシフトしている。



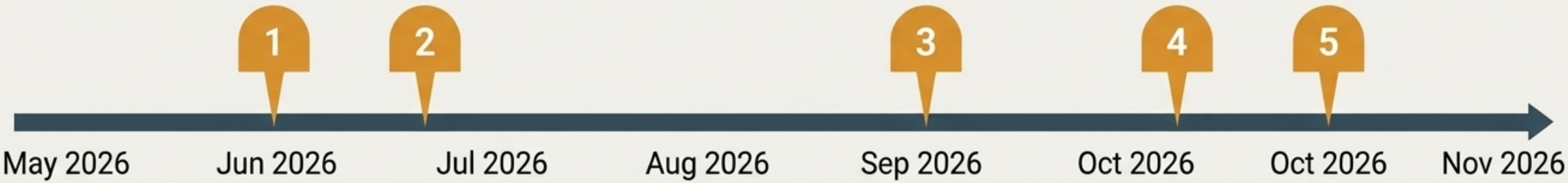
3. The Playbook: ハイブリッド設計への移行

「国産か海外か」の二元論を脱却し、タスクの自律度とデータの機密度に応じた「適材適所のエコシステム」を構築することが今後の最適解となる。

Gemini Notebook

Milestone Reached: 国産フルスクラッチにおける「推論モデル」の到達

推論・思考モデルは
「あって当然」の段階へ



2026年6月（PFN）
「PLaMo 3.0 Prime」**正式版。推論/非推論モデルを使い分け、**256K**長文脈に対応。**

2026年6月（SB Intuitions)
「Sarashina3」**提供開始。複数ツールの同時呼び出しに対応。**

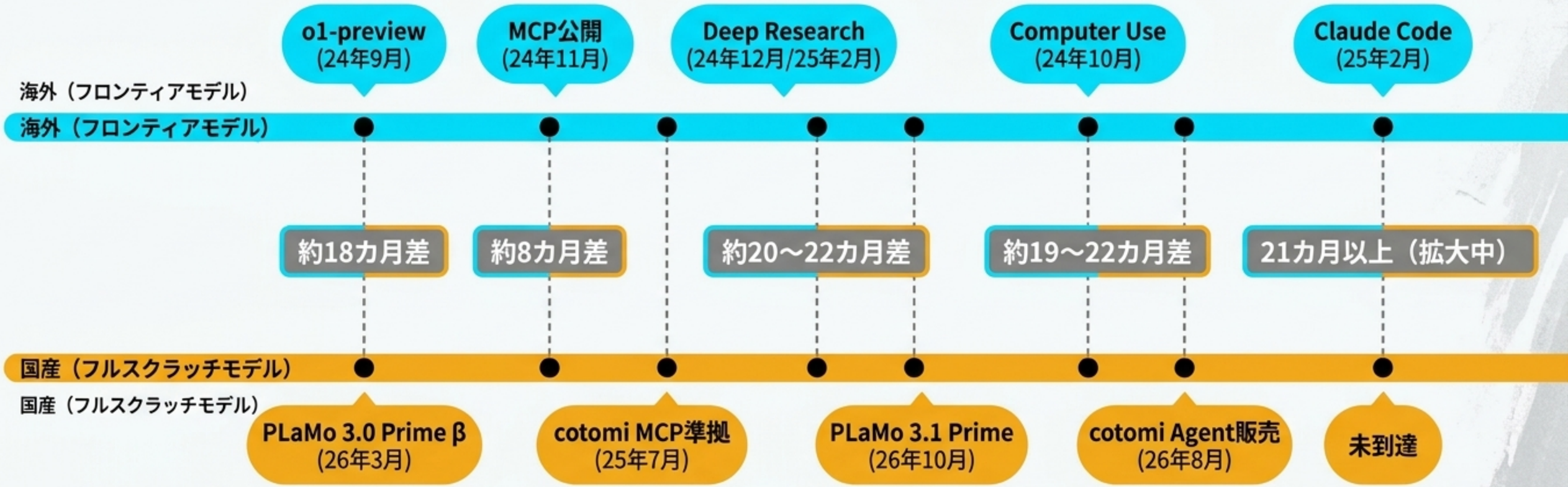
2026年9月（NII）
「**LLM-jp-4.1**」**公開。ツール呼び出しに新規対応。**

2026年10月 (PFN)
「PLaMo 3.1 Prime」**提供開始。**Web**検索・オンプレミスでの**Deep Research**搭載。**

2026年10月（SB Intuitions)
「Sarashina3.2 mini thinking」**発表。シリーズ初の思考モデル。**

Gemini Notebook

The 18-Month Gap: エージェント化で再生産・拡大する遅延



ELYZA-Thinking（25年5月）のような「海外モデル派生型」であれば推論への追従は約8カ月で可能だった。しかし、自前の事前学習モデルに推論・エージェント能力を実装するフルスクラッチ開発においては、後工程になるほど差が拡大している。

Gemini Notebook

Explanatory Visual: エージェント化の階層構造と「遅れの持ち越し」

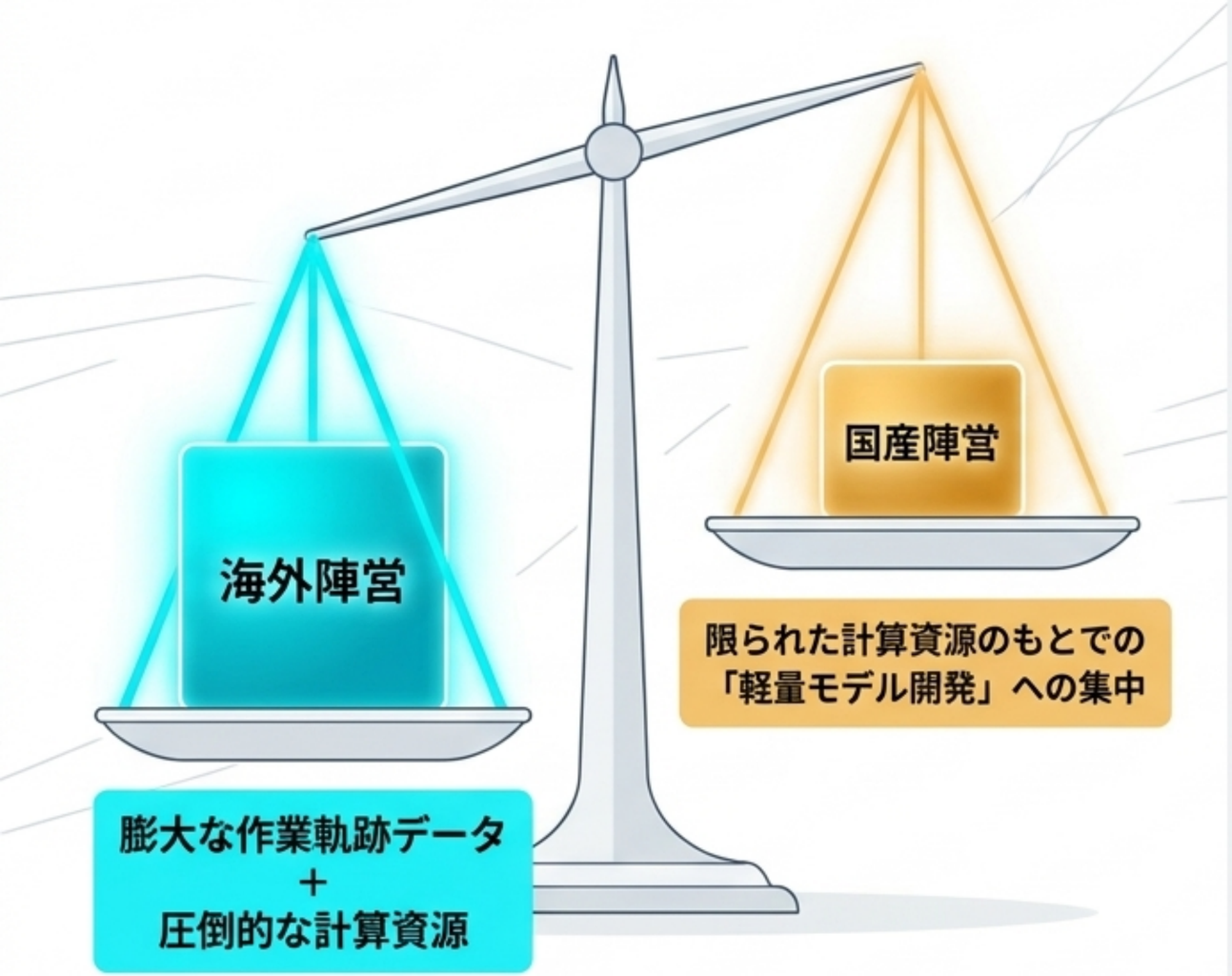


土台の構築遅延が、
そのまま上層の
着手遅延に直結する

インサイト: エージェント能力は、推論能力の上に積み上げる「後工程」である。推論モデル化の遅れ（Layer 2）は、そのまま長時間自律タスクの訓練（Layer 5）の出発点の遅れに直結する。

Gemini Notebook

The Resource Bottleneck: 埋めがたいリソースの非対称性



Noetraの大型投資とタイムラグ

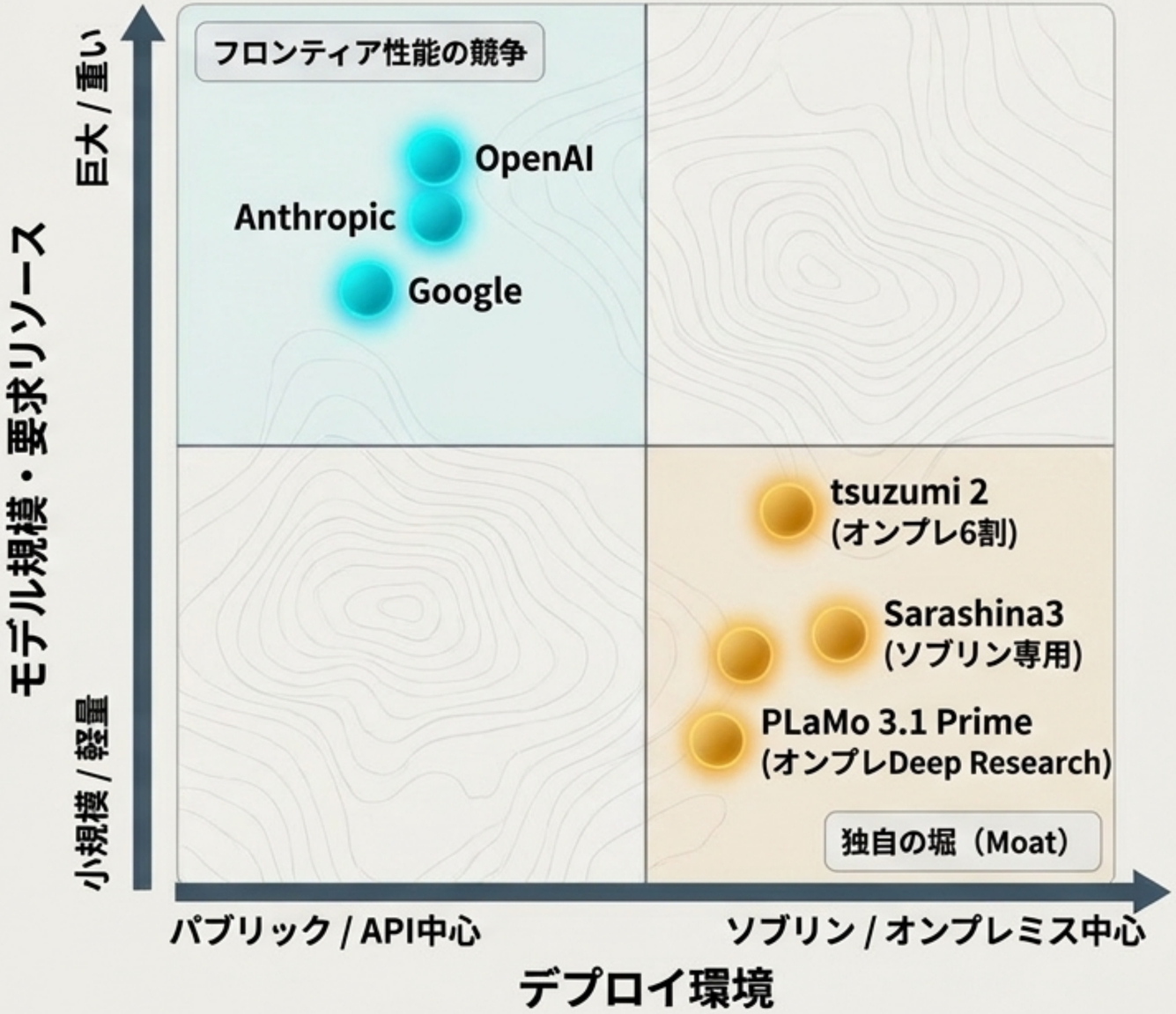
- **事実:** 2026年度、経済産業省がAI開発会社「Noetra」に3,873億円を拠出
- **計画:** オムニモーダル基盤モデルの開発目標は2028年度
- **計算基盤:** 2万7,500基のGPUを搭載した計算基盤の本格稼働は2028年6月予定

結論：計算資源の圧倒的な不足が解消され、フロンティア級に挑戦できる環境が整うのは、最速でも「2028年半ば以降」となる。

Gemini Notebook

Strategic Shift: フロンティアのレッドオーシャンを回避する

競争軸はパラメータ数から「機密データを国内・社内で扱えること」へ完全にシフトした



Gemini Notebook

The Three Layers of Agentic Capability: エコシステムによる補完

[Layer 3] 業務適用層（競争力：高）
富士通「自己進化マルチAIエージェント」（27年3月専有型AI基盤）、NEC「cotomi Act」（WebArenaタスク成功率80.4%記録）。製品化で価値を創出。

[Layer 2] オーケストレーション層（競争力：高）
Sakana AI「Fugu Max / Ultra v2」。他社のフロンティアモデルやオープンウェイトモデルを作業役として自動選択・指揮するルーター技術。海外最前線と互角。

[Layer 1] 基盤モデル層（競争力：遅延20カ月超）
フルスクラッチの国産LLM単体でのエージェント能力。

インサイト：基盤モデル単体（Layer 1）での遅れを、外側のオーケストレーション（Layer 2）や特定の業務システムへの組み込み（Layer 3）で補完するのが現在の日本勢の勝ち筋である。

Gemini Notebook

Evaluation Shift: 動くゴールポストと見えなくなる実力差

旧来の評価軸（言語力・知識ベース）

- 指標: J-tau, BFCL v3 など（単発の推論やツール呼び出し）
- 状況: Sarashina3.2 mini thinkingなどが、同等クラスの軽量モデル群の中で最高点を獲得するなど、健闘を客観的に証明可能。

評価指標のシフト

新たな評価軸（長時間自律エージェント）

- 指標: SWE-bench Verified (多段のコーディング), Artificial Analysis v5 （エージェント型タスク）
- 状況: 国産モデルの公表値はほぼ存在しない。

警告: 評価の物差しが「知識」から「長時間自律タスク」へと移行する中、国産モデルは比較表（リーダーボード）の土俵にすら上がれなくなるリスクを孕んでいる。

Gemini Notebook

Short-term vs. Long-term Outlook: 2つのフェーズ

2027

Phase 1: 底上げと実運用期（2027年）

- NTT「tsuzumi 2」論理数学・コード対応高度化
- NII「LLM-jp-4.2」強化学習導入
- デジタル庁「源内」の2027年度調達において「tool use（外部ツール呼び出し）」機能の説明が要件化

2028+

Phase 2: フロンティアへの再挑戦（2028年以降）

- Noetraの計算基盤（2.7万基GPU）本格稼働による、大規模・長時間自律型学習の開始

インサイト：2027年中は、実用レベルの軽量エージェントの社内定着を急ぐフェーズとなる。フルスクラッチでのフロンティア級への挑戦は2028年以降を待つ必要がある。

Gemini Notebook

The Synthesis Arc: 「二者択一」から「ハイブリッド」へ

時代遅れの二元論

- 「機密データがあるからすべてオンプレミスの国産で統一する」
- 「性能が低いからすべて海外APIに依存する」

推奨されるハイブリッド・エコシステム



- 証明されたユースケース: 富士通「Uvance」は、自社IP由来の特化型AI（Takane）と海外フロンティアAIを適材適所で組み合わせて提供する方針を明確化。
- メッセージ: 企業の業AIの設計思想は、単一エンジンの選定から、複数モデルを束ねる複雑な工場のアーキテクチャ設計へと進化した。

Gemini Notebook

Decision Matrix for Enterprise AI: 適材適所の導入フレームワーク



Gemini Notebook

Summary & Key Takeaways

現実の直視（基盤レイヤーの遅れ）

推論モデル到達は評価すべきマイルストーンだが、長文脈・強化学習を要するエージェント化においては20カ月超の遅れが存在し、2028年までは埋まらない。

戦略のシフト（ソブリンという活路）

国産勢はパラメーター競争から降り、「軽量・オンプレミス・ソブリン」という機密データ処理の土俵へ見事にポジショニングのピボットを遂げた。

企業向け指針（オーケストレーション）

今後のエンタープライズAI戦略は「単一モデルの選定」ではない。データの機密性とタスクの自律度を軸に、国産・海外・ルーターを組み合わせるハイブリッド設計が唯一の解となる。

Gemini Notebook

Sources & Methodology

ソース情報:
本資料は、2026年11月20日時点の公開情報に基づく調査報告「国産LLM動向とAIエージェント化の遅れ（Claude Fable 5.2)」を元に再構成されています。

分析のアプローチ:

- **確認済み事実**: 各社のプレスリリース（PFN PLaMo 3.1 Prime, SB Intuitions Sarashina3, NTT tsuzumi 2, NII LLM-jp-4.1 等）、政府発表（Noetra投資3,873億円、源内調達要件）などの一次情報。
- **分析・推定**: 各種事象間の「遅れの月数」の算出、およびエージェント化の構造的要因に基づく今後の到達時期（2028年以降）に関する予測。

免責・留意点:
遅れの月数は初出の公表月の差であり、性能水準の完全な同等性を示すものではありません。また、国産モデルによる長時間タスク（SWE-bench 等）の客観的評価値は、調査時点で公開されていません。
本資料は「確認済み事実」と「分析・推定」を厳格に区別して記述しています。

Gemini Notebook