ARCHITECTURAL BLUEPRINT
v2026.11

CIO / AIアーキテクト /
戦略推進リーダー
向け診断レポート



次世代AIエージェントの設計図

国産LLMの「18カ月遅れ」仮説の解体と、ハイブリッド共存への青写真

Gemini Notebook

DIAGNOSTIC CONCLUSION（診断結論）



【Layer 1 & 2: 推論・基盤層】

「12〜18カ月遅れ」

長時間耐久実行や自律コーディングにおいて、海外フロンティア（OpenAI o1, DeepSeek-R1等）との間に依然として壁が存在。

【Layer 3: 実装・ハーネス層】

「0〜6カ月差（並走）」

CUA (ブラウザ自動操作) やマルチエージェント・オーケストレーション領域において、国産技術が世界標準と同時に商用展開を達成。

国産AIは「負けている」のではない。単一モデルの巨大化競争から、業務特化の「実装・制御層」へ主戦場をシフトさせている。

Gemini Notebook

断層型タイムライン



データ出典: Preferred Networks, SB Intuitions, NEC, カラクリ (PR TIMES), ビジネス+IT (Sakana Marlin), 技術評論社 (Sakana Fugu), Agent Lab (Sakana Aquarium), IT Leaders (富士通) , AI Revolution (ELYZA) , Spike Studio, WEEL, NII LLM-jp-4.1解説, エージェント実装動向調査

Gemini Notebook

MID-SIZED MODEL TRADE-OFFS & STRATEGIC SHIFT
（中規模モデルのトレードオフと戦略的シフト）

TELEMETRY DATA

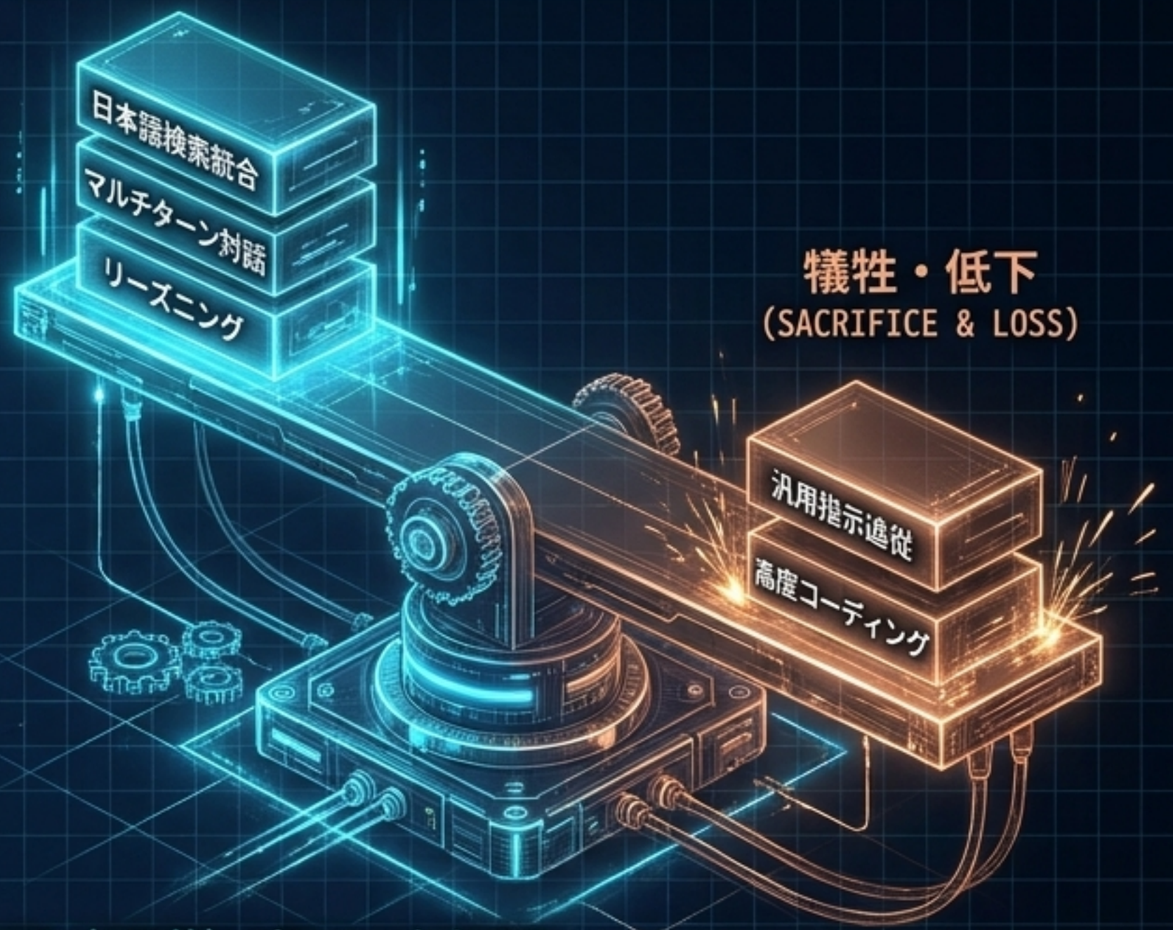


PLaMo 3.0 vs 3.1 Prime（推論モード）パフォーマンスシフト

DATA

Nejumi Leaderboard 4において、国産LLM-jp-4.1（33B）が国産オープン首位を獲得。日本語Web検索レポートや関数呼び出し（BFCL-nejumi）では劇的な向上を記録。

特化・上昇
(SPECIALIZATION & GAIN)



犠牲・低下
(SACRIFICE & LOSS)

30B（300億）パラメータクラス

TECHNICAL ARCHITECT NOTE

限られた計算資源の中で、汎用性を捨てて「実務における推論・検索」へリソースを全振りする戦略的選択。

Gemini Notebook

暗黙知抽出と構造化プロセス（Implicit Knowledge Extraction & Structuring Process）
ハーネス・CUA層が切り拓く日本企業の優位性

ハーネス・CUA層（暗黙知抽出メカニズム）
- NEC cotomi Agent: 人間の操作ログから暗黙知を学習（WebArena成功率80.4%）
- KARAKURI VL2: 8B軽量モデルで日本語PCアプリ横断操作特化



```
{
  "action": "click",
  "target": "button_A",
  "parameters": {...}
},
{
  "action": "input",
  "target": "field_B",
  "value": "data_C"
}
```

Ⓐ
【入力】現場の操作ログ / レガシーSaaS / 複雑な非定型帳票

Ⓑ
ハーネス・CUA層（暗黙知抽出メカニズム）

Ⓒ
【出力】構造化された自動実行手順

TECHNICAL ARCHITECT NOTE

勝負の鍵は「クローリングされたWebデータ量」ではなく、「現場のドメイン知識・操作ログ（暗黙知）の構造化」にあるため、日本企業が優位性を保ちやすい。

現場特有の複雑で非定型なプロセスを、軽量かつ特化型のAIモデルで効率的に構造化するアプローチ。

Gemini Notebook

国産AIエコシステムの戦略マトリクス



Gemini Notebook

国産AIの国産ゲーム





【見せかけの国産（The Wrapped Model）】
コアエンジンは海外オープンウェイトの流用（例：楽天 AI 3.0 におけるDeepSeek-V3リブランドの自主開示事例）
リスク：法的透明性、ライセンス条件、推論特性のブラックボックス化

【真のソブリン基盤（The Sovereign Scratch）】
事前学習データから自社管理されたフルスクラッチ（LLM-jp, PLaMo, Sarashina等）
価値：真のソブリン要件。自社データ保護と完全なガバナンス

企業は「国産」というラベルを鵜呑みにせず、事前・事後学習の「系譜（Lineage）」を検証する評価プロセスが不可欠。

Gemini Notebook

国産AIの国産タイブル進化

Phase 1: テキスト・行政実務

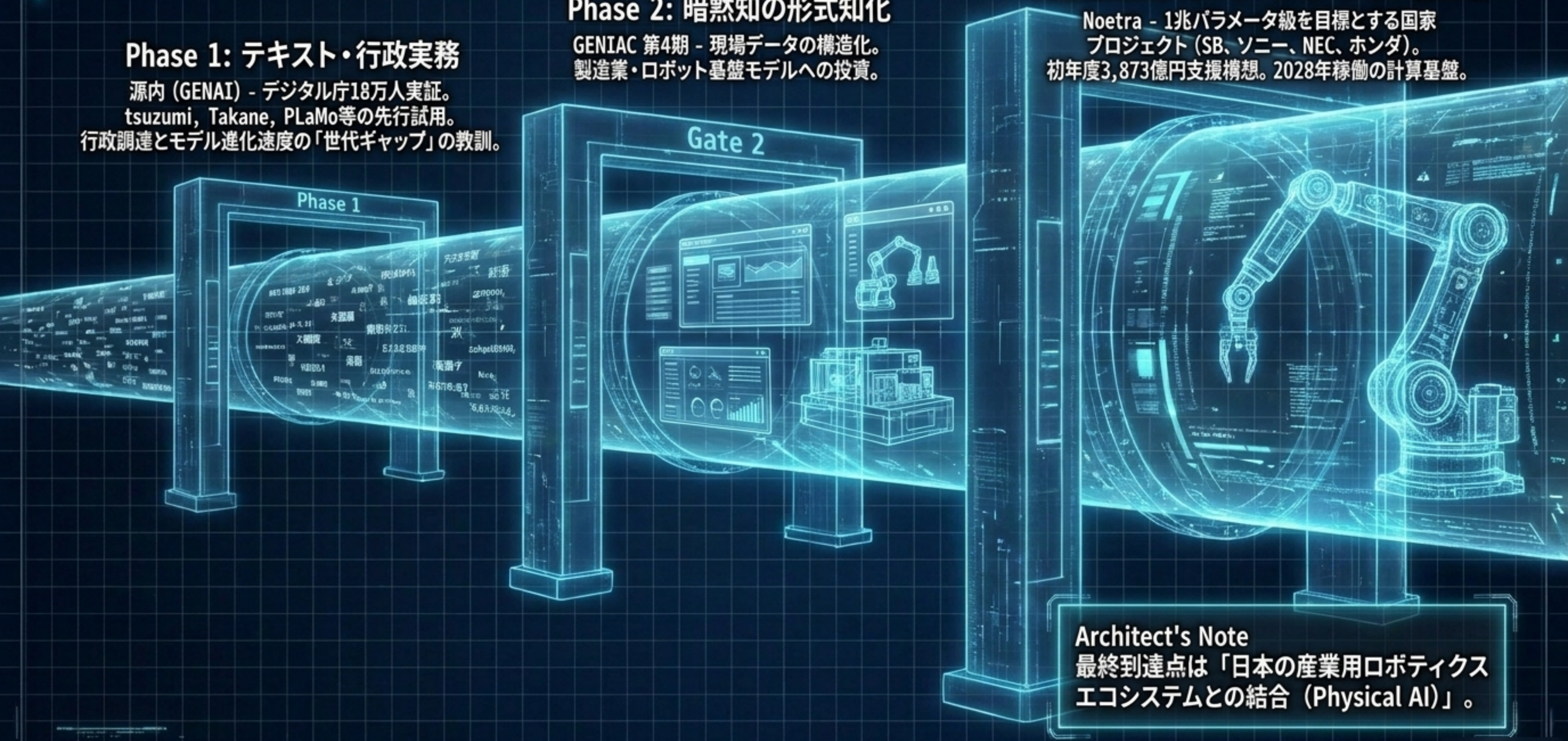
源内（GENAI）- デジタル庁18万人実証。
tsuzumi, Takane, PLaMo等の先行試用。
行政調達とモデル進化速度の「世代ギャップ」の教訓。

Phase 2: 暗黙知の形式知化

GENIAC 第4期 - 現場データの構造化。
製造業・ロボット基盤モデルへの投資。

Phase 3: オムニモーダル・フィジカルAI

Noetra - 1兆パラメータ級を目標とする国家プロジェクト（SB、ソニー、NEC、ホンダ）。
初年度3,873億円支援構想。2028年稼働の計算基盤。



Architect's Note
最終到達点は「日本の産業用ロボティクス
エコシステムとの結合（Physical AI）」。

Gemini Notebook

単一の最強モデルは存在しない。「業務適合性」による選択が急務。

	海外フロンティアAPI (Claude 5, GPT-6系)	国産中規模フルスクラッチ (PLaMo, Sarashina系)	国産特化型CUA・ハーネス (cotomi, Kozuchi等)
長時間自律コーディング	最高	低	中
マルチステップ汎用推論	最高	中	低
日本語固有の複雑・非定型帳票	中	高	最高
社内レガシーシステム(GUI)自動操作	低	中	最高 - 成功率80%+
オンプレ(1GPU)での機密データ処理	不可	最高	最高

Gemini Notebook

制御・オーケストレーション層

マルチLLM振り分け (Gateway)
Multi-LLM Routing & Privacy

自己進化ハーネス
Self-Evolving Harness

暗黙知抽出CUA
Tacit Knowledge Extraction & Analyze

海外API（難易度高・非機密）

汎用推論・自律コーディング
General Reasoning & Autonomous Coding

国産ソブリンLLM - 閉密実行

1GPUオンプレ・ガバメントAI溉内
1GPU On-Premise & Government AI Gennai

日本語複雑帳票・金融・自治体特化
Japanese Complex Forms, Finance, Local Gov

次世代国策・フィジカルAI基盤

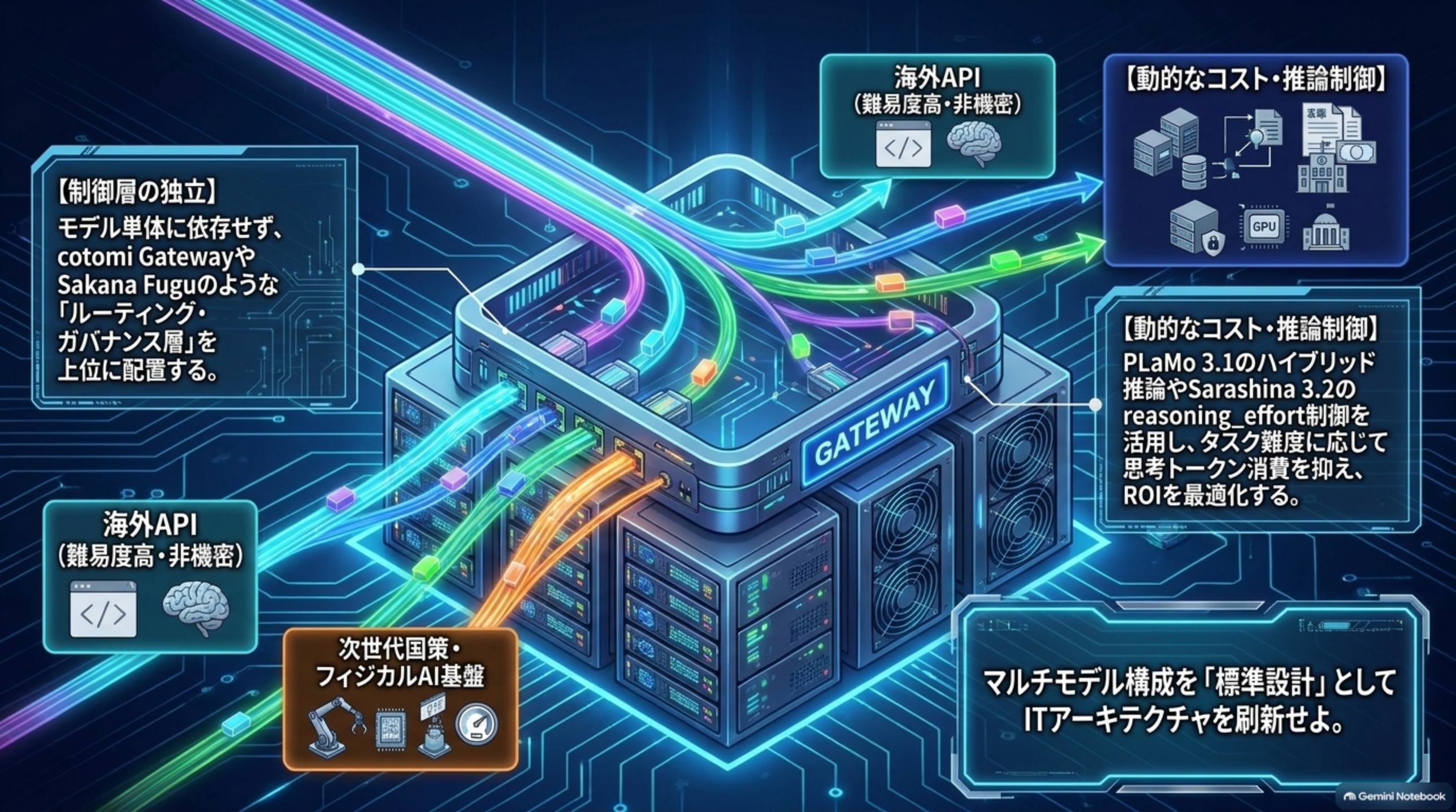
Noetra（1兆パラメータ目標）

GENIAC第4期・ロボティクス連携

エージェント・ネイティブ事前学習

「海外か国産か」の二者択一は誤り。オーケストレーターを介した「共存アーキテクチャ」こそが勝者の設計図

Gemini Notebook


【制御層の独立】
モデル単体に依存せず、
cotomi Gatewayや
Sakana Fuguのような
「ルーティング・
ガバナンス層」を
上位に配置する。
海外API
(難易度高・非機密)
【動的なコスト・推論制御】
GPU
【動的なコスト・推論制御】
PLaMo 3.1のハイブリッド
推論やSarashina 3.2の
reasoning_effort制御を
活用し、タスク難度に応じて
思考トークン消費を抑え、
ROIを最適化する。
GATEWAY
海外API
(難易度高・非機密)
次世代国策・
フィジカルAI基盤
マルチモデル構成を「標準設計」として
ITアーキテクチャを刷新せよ。
Gemini Notebook



Evals（評価基盤）を制する企業が、AIエージェント時代の実装を制する

Gemini Notebook



AIエージェント時代における真の勝負は、基盤モデルのパラメータ競争から、「自社ドメインの暗黙知をいかに組み込むか」というアーキテクチャの設計力へと完全に移行した

Gemini Notebook