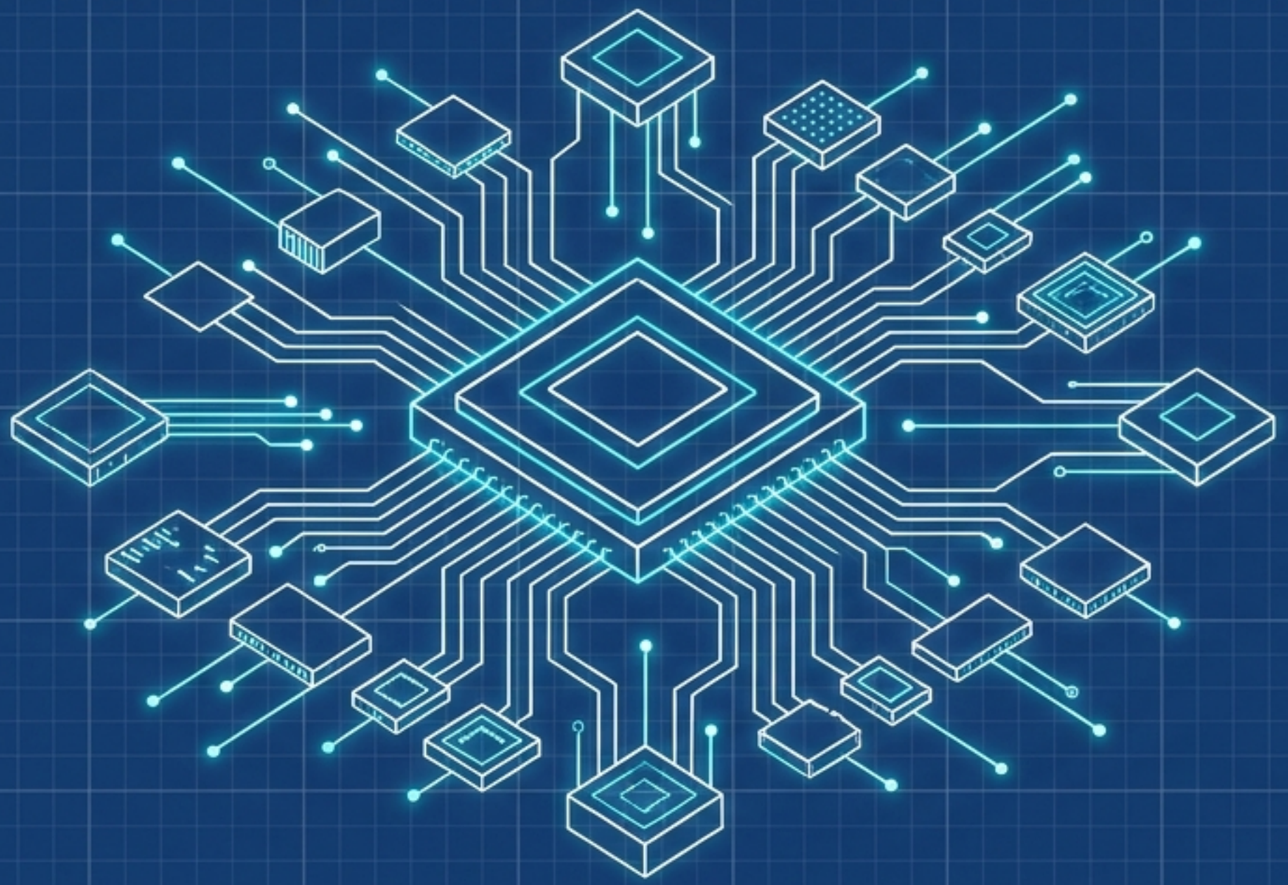


PLaMoとPreferred Networksの進化

「巨大単一モデル」から「垂直統合型AIインフラ」への戦略的転換

[Document ID: PLaMo-STRATEGY-B1] | [Audience: Executive & IT Decision Makers]

戦略的パラダイムシフト：単一の巨大モデルから、高効率な専門特化モデル群へ

2024年初頭

汎用100Bモデルの追求

2兆トークン（英1.3T / 日0.7T）を用いた
「PLaMo-100B」による大規模分散学習の確立。

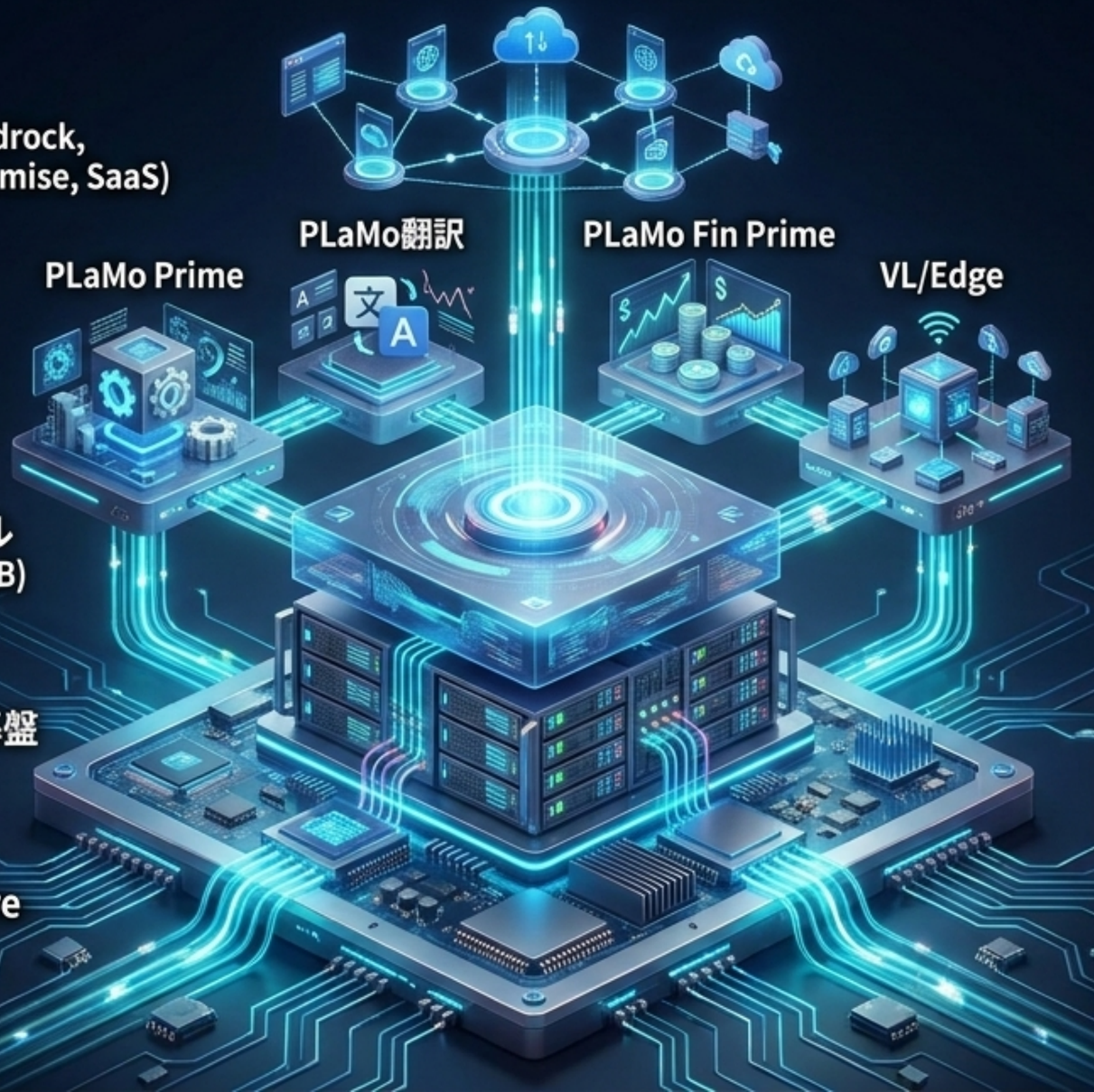
PLaMo 2.x & 3.0以降

用途別ハイブリッド・アーキテクチャ

「高品質データ」「軽量化（Pruning/蒸留）」
「Samba系ハイブリッド構造」を組み合わせた、
1B/2B/8B/31Bの軽量高効率モデルへの転換。

**経営の意思決定：単なる「モデルベンダー」ではなく、
日本語実務に最適化された「エコシステム」の構築へ。**

PFNの真の競争優位：半導体からSaaSまでを貫く「垂直統合モデル」



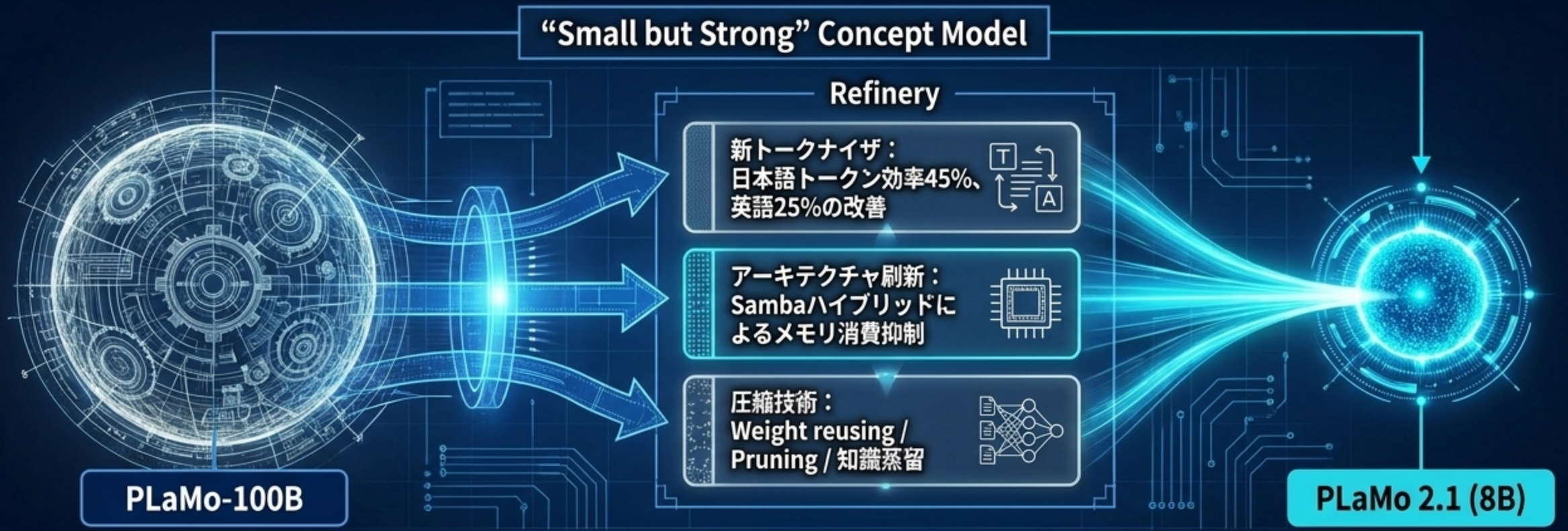
Key Insight

2024年12月の190億円調達、開発子会社Preferred Elementsの本体吸収合併（2025年5月）は、このスタック全体への投資を加速し、モデル研究から「本格的な社会実装」へ事業の重心を移したことを証明している。

PLaMo世代進化マトリクス：アーキテクチャの劇的な進化

	PLaMo-100B	PLaMo 2.x Series	PLaMo 3.0 Prime β
アーキテクチャ	Transformer系 (QK Norm, z-loss, Zero Bubble)	Samba系ハイブリッド (Mamba2 + Sliding Window)	ゼロベース再構築・ Reasoning（推論思考）対応
データ戦略	日本語Webコーパスの 自前構築 (Books3不使用)	教育価値フィルタ・ 世界最大規模(100B級)の 合成データ	NICTデータ + 医療・実務特化データ追加
効率・推論特性	学習最適化 (540 TFLOP/s/GPU)	日本語トークン効率45% ● 改善、KVキャッシュ削減 による低コスト化	長文(YaRN 64K)、 ● 推論能力向上と引き換えの 計算コスト増
最適な ユースケース	大規模基礎研究・ ベンチマーク	エッジ・オンプレ・ 低コスト実務API	複雑な多段思考・ エージェント・高度専門業務

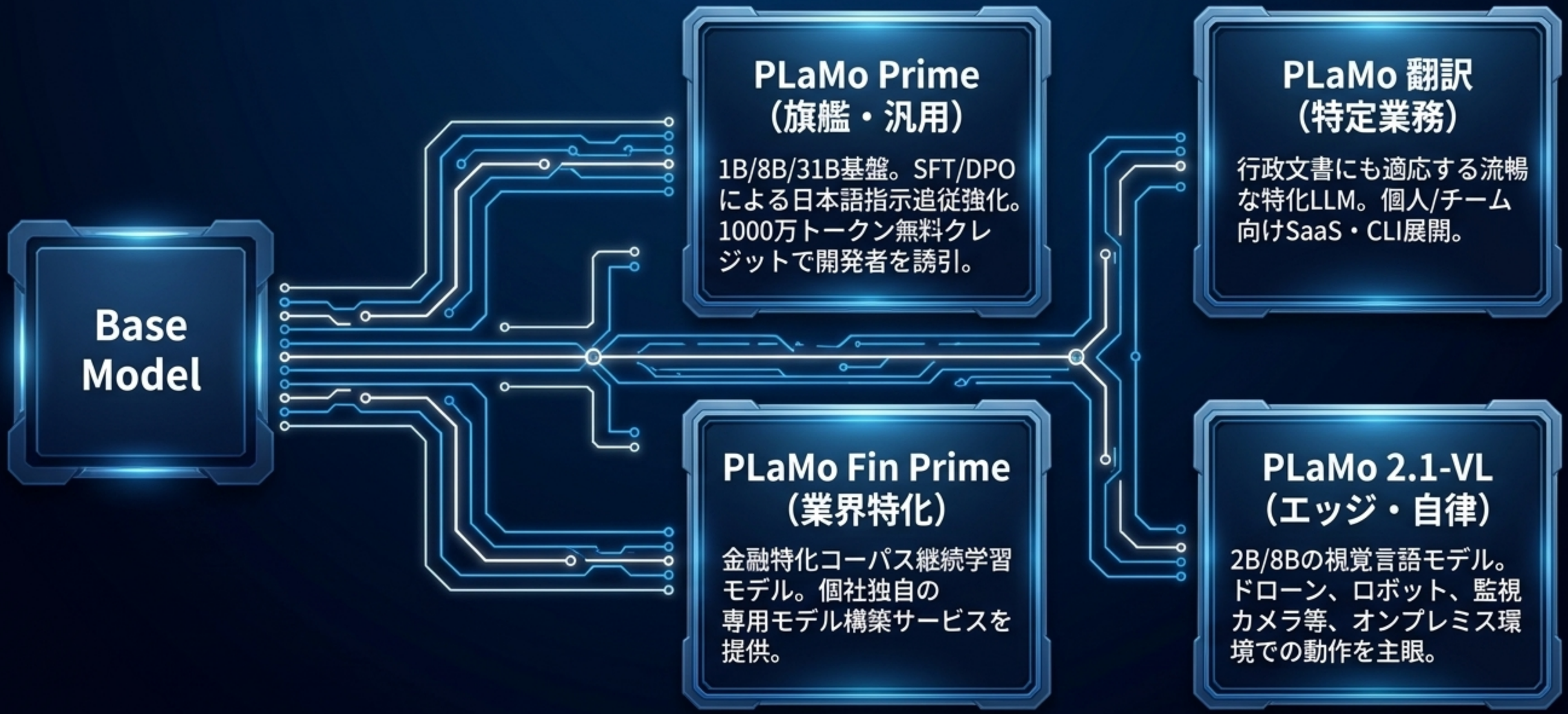
「小さくても強い」：PLaMo 2.1が実現した驚異的な高効率化メカニズム



「100Bモデルの 1/12 のサイズで、日本語・コーディング性能が同等以上」

ビジネスへの影響：この圧縮技術により、PC上のローカル動作や、オンプレミス環境・エッジデバイスでの自律稼働（VL）が現実的な選択肢となった。

PLaMo進化の系統樹：専門分化した派生ラインナップ



データ品質戦略とガバナンス：量から「高品質・高純度」へのシフト

厳格なフィルタリング

厳格なフィルタリング

CommonCrawlから自前でパイプラインを構築。著作権侵害懸念のある「Books3」を使用しない方針を明言（リスク回避とコンプライアンスの優先）。

高品質合成データの生成

高品質合成データの生成

汎用LLMを活用し、世界最大規模の「100B token級」学習データを構築。

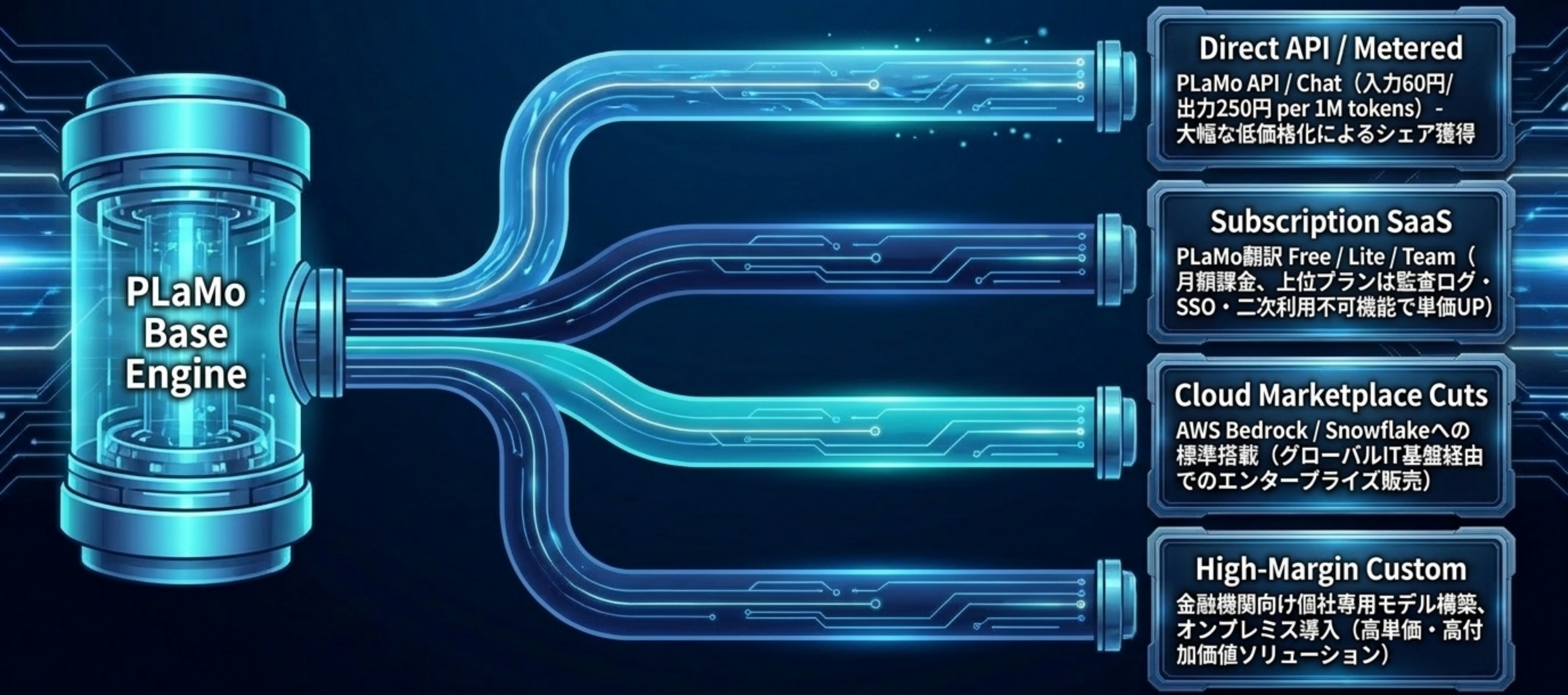
ドメイン特化データの統合

ドメイン特化データの統合

NICTデータ、Talent Scouter、MedRECT-ja、JMLE等の専門領域データを追加。

結論：教育的価値によるフィルタリングと合成データにより、「単純なスケールアップ」から「データ品質と日本語適合性」による競争優位の確立へ。

多層的マネタイズ構造：B2B流通網の完全マッピング



社会実装の証明：公共セクターと大企業における圧倒的トラクション

公共セクター (Public Sector)

- デジタル庁のガバメントAI「源内」にPLaMo翻訳を採用 (2025年12月開始)。
- 中央省庁向け試用モデルに「PLaMo 2.0 Prime」が選定 (2026年3月)。

大企業 & SaaSエコシステム (Enterprise)

- 約700の自治体が導入する「QommonsAI」に標準搭載。miibo、Tachyon生成AI等への組み込み。
- AWS Bedrock、Snowflakeでの提供開始による、既存B2Bワークフローへの摩擦なき統合。

戦略的意義：PFNは最終顧客と直接取引するだけでなく、既存のSaaSやクラウドの「国内日本語エンジン」として裏側に組み込まれることで、爆発的な販売効率を実現している。

国内競争環境：主要国産モデルにおけるPLaMoの独自ポジション

	戦略の診断	
PFN (PLaMo)	差別化：フルスクラッチ日本語・垂直統合・専門派生（翻訳/金融/エッジ）。	統合度：半導体からSaaSまで自前。
NTT (tsuzumi 2)	差別化：高セキュリティ・低コスト純国産。	ターゲット：通信インフラ・企業基盤。
Fujitsu (Takane)	差別化：政府機関実証・業界特化AIとの結合。	統合度：特定業界システムとの連携主導。
SB Intuitions (Sarashina)	差別化：日本文化/ビジネス慣行理解。	ターゲット：ソフトバンク/オラクル流通網。
ELYZA	差別化：大企業向けLLM実装支援とプロダクト伴走。	ターゲット：個別カスタマイズとPoC支援。

Strategic Conclusion

PFNの優位性は、モデル単体の性能ではなく「半導体計算資源・豊富な派生モデル（翻訳/VL）・AWS等のグローバル流通網」を独自に束ねている点にある。

AIガバナンスとコンプライアンス：エンタープライズの信頼を担保する防壁

EU AI Act & Global Regs

技術文書とリスク対応の整備を通じた
将来のグローバル展開への布石。

日本政府ガイドライン / APPI

プライバシーとサイバーセキュリティの徹底。

著作権・知財リスク

学習データからの「Books3」排除による
リーガルリスクの抜本的遮断。

SaaS Level Protections

PLaMo翻訳 Lite/Teamプランにおける「データ二次利用なし」「即時破棄」「完全削除」の保証。
性能だけでなく、セキュリティによって上位プランへ誘導する設計。

現実的な自己評価とSWOT診断：事業の強みと技術的課題

強み (Strengths)

- 日本語特化のフルスクラッチ開発と高品質データパイプライン。
- 半導体からソリューションに至る独自の垂直統合。
- 行政・大企業・AWS/Snowflakeにおける強固な流通チャネル。

弱み/課題 (Weaknesses)

- 商用Primeの内部仕様が非公開であり客観比較が困難。
- PLaMo 3.0 Prime βのReasoning採用による、計算コストと応答時間（レイテンシ）の増加。
- AIME、GPQA、BFCLの一部における遅れを自己開示。

機会 (Opportunities)

- 政府・自治体の国産AI需要とオンプレミス/閉域網ニーズの拡大。

脅威 (Threats)

- OpenAI/DeepSeek等グローバル勢の進化と、国内競合（tsuzumi等）によるシェア争奪。

PLaMo戦略ロードマップ：1〜5年の事業展望

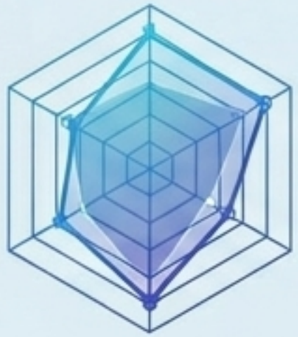
Horizon 1: 直近1年（最適化と実装）	Horizon 2: 今後3年（多層収益化とガバナンス）	Horizon 3: 今後5年（インフラとしての定着）
PLaMo 3 正式版の安定化（Reasoningレイテンシの改善）	Prime + 翻訳 + 金融 + 医療の多層収益化の確立	「日本語業務基盤モデル群」としての国内デファクトスタンダード化
BFCL / Agent / 長文性能の強化	Bedrock / Snowflake / オンプレミス販売の爆発的拡大	エッジVL（視覚言語モデル）と業界専用モデルの普及
行政・大企業でのPoC横展開、翻訳SaaSのチーム機能強化	EU AI Act / 著作権対応の監査体制整備	ハードウェア込みの差別化完成
	MN-Core（AI半導体）と組み合わせた強固な統合提案	

次期成長を牽引する重要業績評価指標（KPI）

製品性能 (Product Performance)

指標: JFBench, IFBench, BFCL, 長文推論(LongBench), MedRECT, JMLE。

狙い: 日本語実務、エージェント能力、 専門領域ごとの精度を厳密に追跡。



経済性・効率 (Economics)

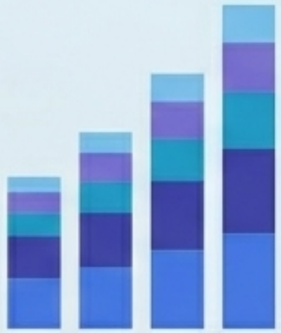
指標: 100万トークンあたりの粗利、平均応答時間(レイテンシ)、推論コスト/req、GPU稼働効率。

狙い: Reasoning導入による品質向上と収益性のトレードオフ監視。

市場浸透 (Market Penetration)

指標: 有償API顧客数、Bedrock/Snowflake経由売上、オンプレ案件数、自治体導入数。

狙い: SaaS・チャネルパートナー経由の流通面での信認シグナルを可視化。



派生事業と信頼性 (Derivatives & Trust)

指標: 翻訳SaaS継続率、個社カスタムモデルARR、法務レビュー/顧客監査通過率。

狙い: 高粗利な専門特化モデルの成長と、規制対応を営業上の「武器」に転換。

結論：「ベンチマーク競争」から「日本語実務インフラ」への昇華

**「勝負は、もはやベンチマークそのものではなく、
“どこで、どう売り、どう運用されるか” に移っている。」**

1. データからシリコンまでの掌握

MN-CoreからSaaSに至る垂直統合が、計算コストの連統合が、計算コストのショックを吸収し、競合の模倣を防ぐ最大の堀（Moat）となる。

2. 実務に刺さる「専門特化」

PLaMoは単一の万能モデルではなく、行政（翻訳）、金融（Fin Prime）、エッジ（VL)へと分岐する「実務基盤モデル群」である。

3. 社会インフラとしての定着

デジタル庁の採用やAWS/Snowflakeでの標準化化が示す通り、PLaMoはすでに「一企業のAI」を超え、日本のデジタルインフラの一部として稼働を始めている。