純国産生成AI「PLaMo 3.1 Prime」と知財業務の革新

高度な推論力と完全閉域オンプレミスが実現する、次世代IP戦略の設計図



Gemini Notebook

PLaMo 3.1 Primeが知財実務にもたらす3つの構造的変革



1. アーキテクチャの革新

推論（Reasoning）と非推論（Non-reasoning）を単一モデルに統合。インフラ要件を半減しつつ、タスクに応じた圧倒的な処理効率を実現。



2. 知財特化の安全性と主権確保

完全閉域オンプレミスでの自律型「Deep Research」稼働。機密情報の漏洩リスク（新規性喪失など）を完全に遮断。



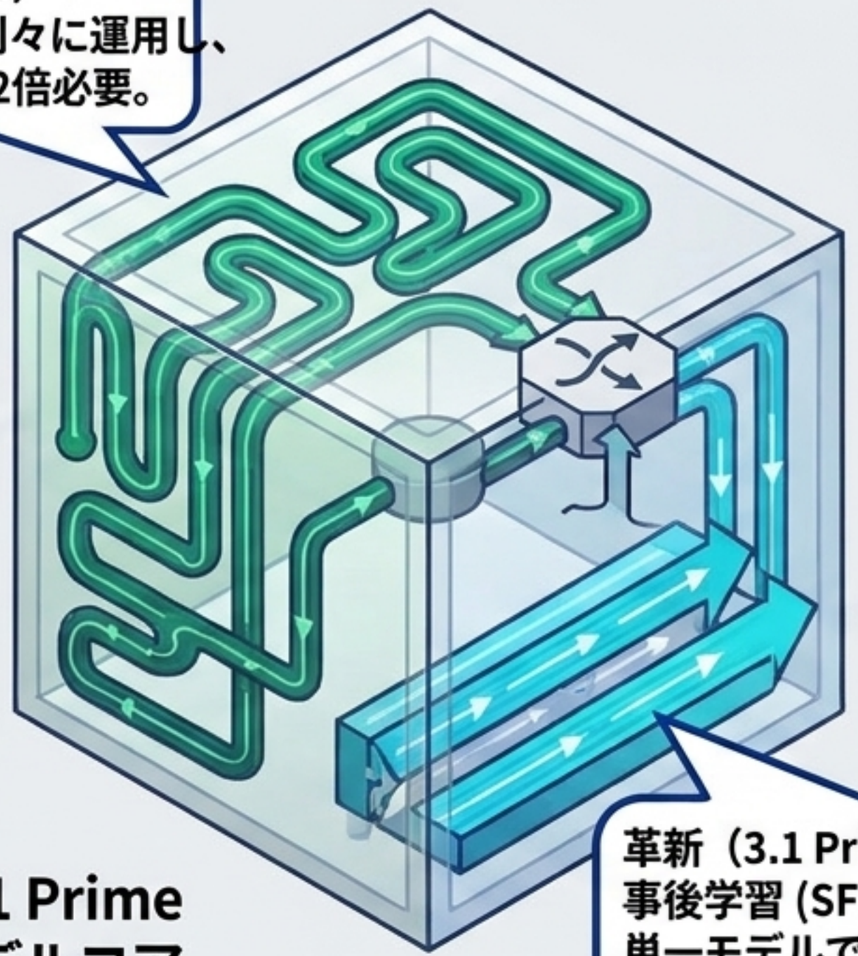
3. 圧倒的なトークン経済性と精度

非推論モードでの出力トークン90%削減と、日本語独自トークナイザによるコスト優位性。ハルシネーションを15%抑制。

Gemini Notebook

推論と非推論を単一ウェイトに統合した「ハイブリッドアーキテクチャ」

従来（3.0 Prime）：
2つのモデルを別々に運用し、
GPUリソースが2倍必要。



推論エンジン
(Deep Research)

非推論エンジン
（高速・軽量）

革新（3.1 Prime）：
事後学習 (SFT) を2段階に分割。
単一モデルでAPIリクエストごとにモードを動的切り替え。

PLaMo 3.1 Prime
モデルコア

処理速度（スループット）

投機的デコーディングにより最大1.7倍の高速化（TPOT 15〜45%短縮）。

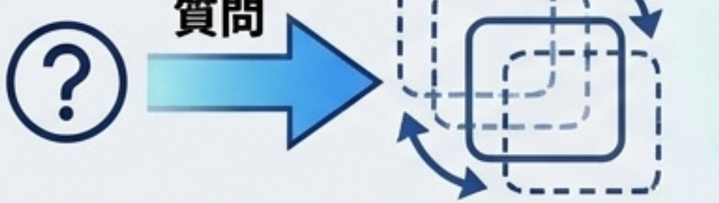
出力効率

非推論モードにおいて出力トークン数を90%削減。

Gemini Notebook

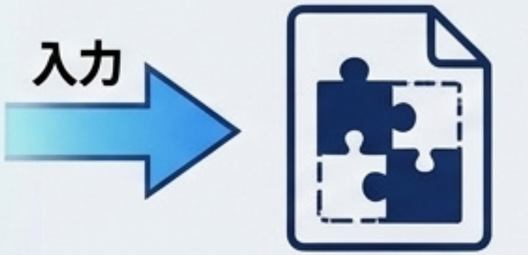
「事実性」と「忠実性」の分離学習によるハルシネーションの構造的抑制

シールド1：事実性のメタ認知プロセス（知識境界の判定）



推測を排除し、明確に『分からない』と認めるようSFTで強制

シールド2：忠実性の強化（合成データの意図的投入）



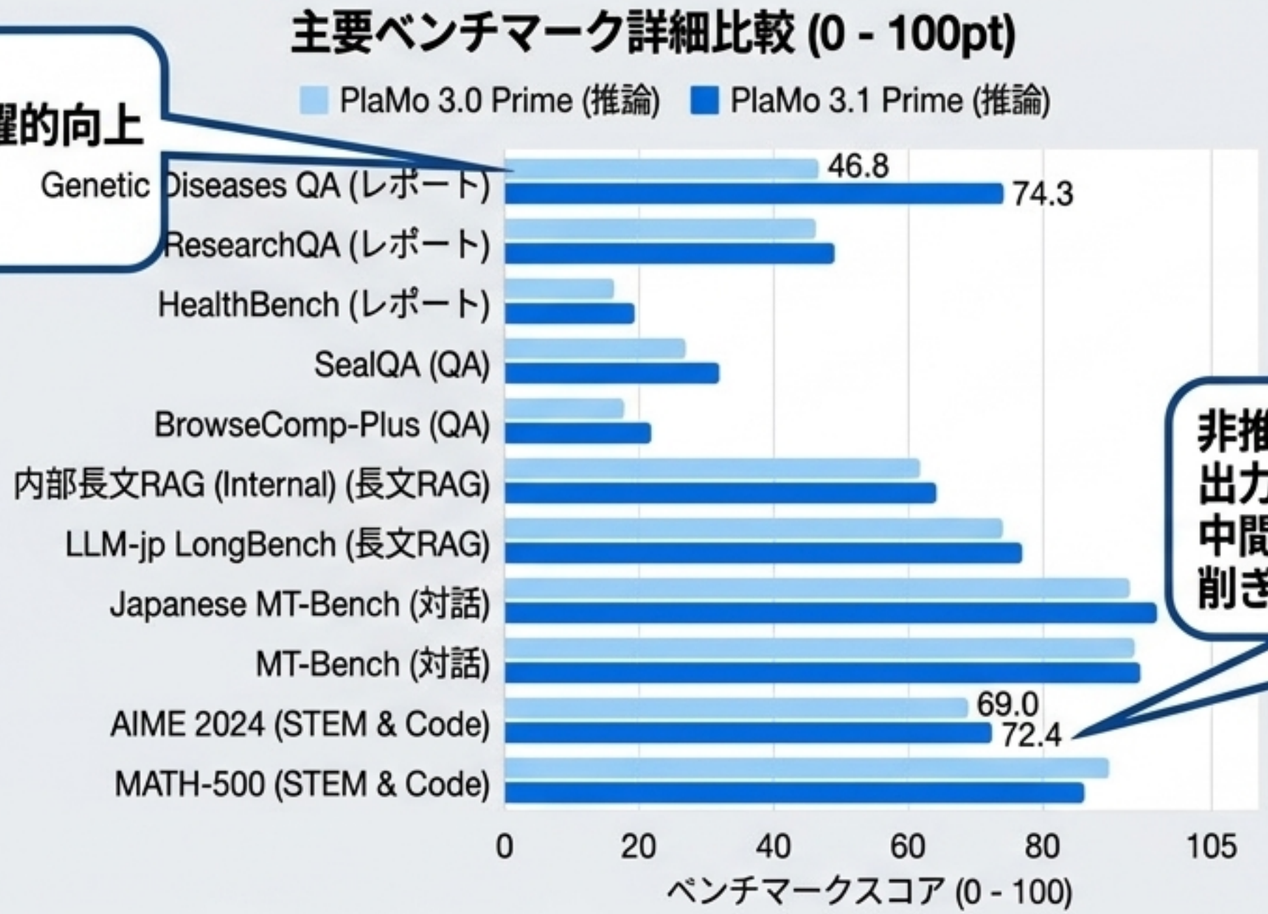
根拠不足の際は『答えられない』と応答させる意図的学習

実績：関連8ベンチマーク（TruthfulQA等）で平均約15%の改善、事実性・忠実性カテゴリで+6.4ポイント向上

Gemini Notebook

データが証明する「モード純化」と役割分担の明確化

推論モードの圧倒的進化：
Web検索レポート領域で飛躍的向上（+11.1pt）。46.8から74.3へ急伸。



非推論モードの戦略的スコア低下：
出力トークン90%削減のため、中間思考ステップを徹底的に削ぎ落とした結果。

結論：情報抽出・分類は「非推論」へ、複雑な論理演算は「推論」へルーティングすべき き定量的な証明。

Gemini Notebook

生成AIの知財実務適用を阻む「3つの構造的障壁」



1. 守秘義務違反とデータ主権の喪失

出願前の非公開技術や営業秘密を外部AIに入力することは、第三者開示に該当。個人情報保護法やNDA違反の致命的リスク。



2. 実施可能要件違反と仮想実施例の捏造

ハルシネーションによる架空の実験データや文献番号の虚偽記載。そのまま出願すれば実施可能要件違反による拒絶・無効事由、最悪は法的処罰の対象。

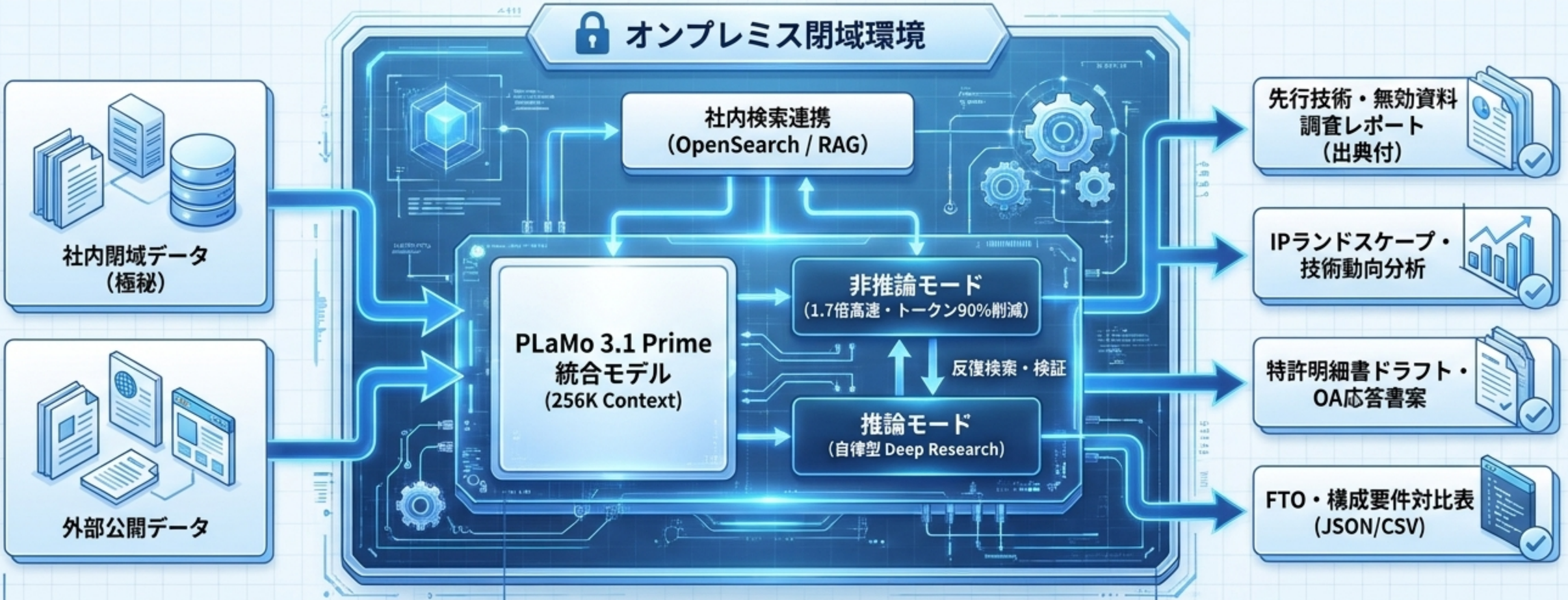


3. 権利解釈における「過剰拒否 (Over-refusal)」

法規違反の境界線を問う専門的質問に対し、従来のAIは「答えられない」と一律に回答を拒絶し、侵害予防調査等の業務支援を阻害。

Gemini Notebook

完全閉域・オンプレミス稼働による「自律型AI調査空間」の確立



入力：社内の極秘データと
外部の公開データを統合。
外部にデータは一切漏れない。

処理：オンプレミス環境内で
「OpenSearch」と「PLaMo統合モデル」が
自律的に反復検索 (Deep Research) を実行。

出力：先行技術調査、明細書
ドラフト、FTO対比表などの
安全な知財アウトプット。

Gemini Notebook

Workflow 1｜先行技術・無効資料調査の「二段階パイプライン」



Step 1: 高速フィルタリング工程

- 使用モード：非推論モード（1.7倍高速・トークン90%削減）＋構造化出力（JSON/CSV強制）
- 処理内容：数千件の公報からノイズを除去し、関連度スコアと要約を低コストで一括抽出。

Step 2: 詳細対比・精査工程

- 使用モード：指論2：推論モード＋自律型Deep Research
- 処理内容：処理内容：抽出された重要文献に対し、自律的な反復検索・論理的検証を実施し、出典付きの多面的な調査報告書を生成。

出力：高度な調査報告書（出典付き）

Gemini Notebook

Workflow 2 | IPランドスケープと新規用途探索のハイブリッド分析

ストリームA（閉域・極秘）



社内の未公開技術データ、研究開発方針、特許・ノウハウDB。

ストリームB（公開・Web）



最新の市場規模、競合動向、顧客ニーズのWeb検索結果。

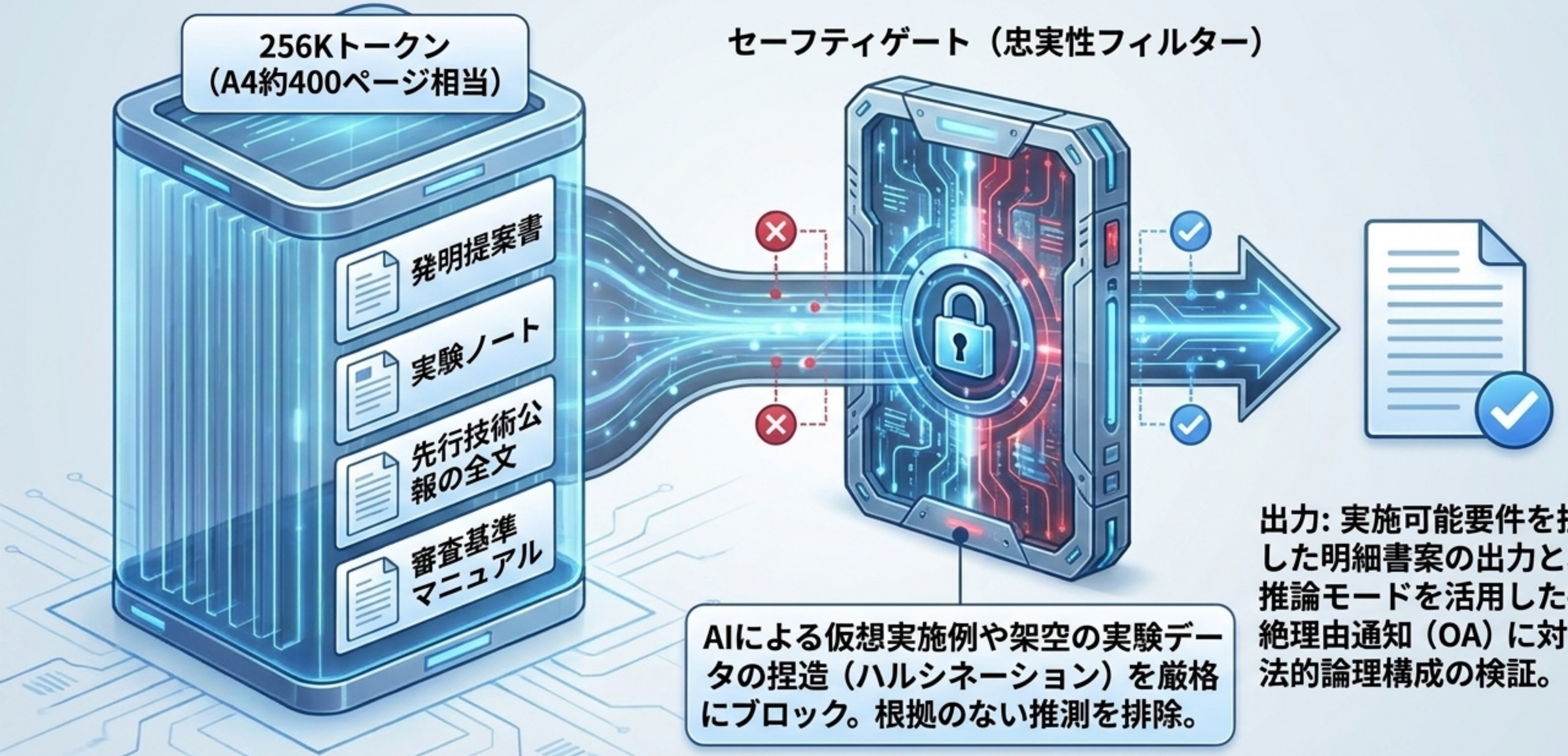
Deep Researchエンジン（PLaMo 3.1 Prime）



統合プロセス：膨大な定性テキストの意味合いを深く理解。外部環境と内部リソースを掛け合わせた戦略的な事業計画・用途探索レポートを数分～数十分で自律作成。

Gemini Notebook

Workflow 3 | 特許明細書ドラフトとOA応答におけるコンテキスト統合



Gemini Notebook

Workflow 4 | 多言語知財実務とグローバル対応のセキュア化

完全閉域オンプレミス環境



1. 入力

出願前の特許明細書（極秘日本語ドキュメント）。

2. ローカル翻訳

軽量モデル「PLaMo翻訳」を使用し、業界特化の専門用語を反映させた自然な現地語へセキュアに変換。

3. 法制チェック

推論モードが現地の特許法制や記載要件に照らしてレビュー。

結論：PCT・パリルートの外国出願において、機密情報を一切社外のクラウドサービスに出さないグローバル知財基盤の確立。

Gemini Notebook

データ機密性に基づくデプロイメント選択マトリクス

【PoC・初期検証】 PLaMo API (Free/Standard) & Chat	公開公報のスクリーニングやプロンプト検証。インフラ負担ゼロ。
【既存クラウド連携】 Snowflake / AWS SageMaker等	既に社内特許DBを構築している企業の分析基盤向け。データ外部流出なし。
【知財の中核環境】 オンプレミス（モデル本体 / サーバー一体提供）	最高機密（出願前発明、侵害予防調査）を扱うDeep Research環境。独自データによる追加学習（FT）も可能。日本国内処理による完全なデータ主権確保。

Gemini Notebook

日本語処理における圧倒的なトークン経済性とコスト優位性

日本語独自
トークナイザによる
消費量圧縮

×

非推論モードによる
出力トークン
90%削減

=

**大量文献の
一次査読
コスト最小化**

API Standard料金：
100万トークンあたり
入力60円 / 出力250円。

出力コスト削減の比較



大量の公報分類やスクリーニングにおいて、単価の高い出力側トークンを劇的に削減することで、極めて高いROIを実現。

Gemini Notebook

知財DX実装のための3つのエンジニアリングアーキテクチャ指針



1. 逐次関数呼び出し（Sequential Tool Calling）の実装

並列呼び出しではなく、1ステップずつ実行する逐次実行フローでエージェントを設計し、Deep Researchの反復検索ループを確実にする。



2. 256Kコンテキストの長文RAG連携

全公報を無作為に投入せず、OpenSearch等の社内検索と連携し、関連度の高い公報・段落を動的に抽出してコンテキストウィンドウに供給する。



3. モードの動的ルーティングとガバナンス

定型要約・分類は「非推論」、クレーム対比・調査計画は「推論」へ振り分ける。社内規定による機密区分ごとの入力可否ルールをシステム側で統制。

Gemini Notebook

セキュアな知財DX実装に向けた段階的ロードマップ

Phase 1: API・ChatでのPoC検証（即時実行）

公開特許情報や匿名化データを用い、API Free枠やChatキャンペーンを活用して業務類型別の精度検証をスモールスタート。

Phase 2: 監査プロセスのシステム組み込み

非推論と推論のルーティングを実装。弁理士や知財担当者によるハルシネーション・原典確認（Human-in-the-Loop）を業務フローに統合。

Phase 3: オンプレミスでの基幹業務完全統合

出願前の発明情報や営業秘密を扱う業務を完全閉域のDeep Research環境へ移行。データ主権の確保と高度な知財戦略の自動化を両立。

結びのメッセージ：法的リスクを統制し、労働集約から知識集約へのパラダイムシフトを完成させる。

Gemini Notebook