国産生成AI基盤モデル「PLaMo 3.1 Prime」の技術的解剖と知的財産業務における戦略的実装

Gemini 4 Argon

PLaMo 3.1 Primeのアーキテクチャ革新と技術的到達点

2026年10月27日、株式会社Preferred Networks（PFN）は、フルスクラッチで開発する国産生成AI基盤モデル「PLaMo」の最新フラッグシップモデルである「PLaMo 3.1 Prime」をリリースした[1]。PFNは2025年7月1日付で、大規模言語モデル（LLM）の開発を担っていた完全子会社の株式会社Preferred Elements（PFE）を吸収合併し、同社の事業および開発した知的財産を全面的に承継することで、基盤モデルの研究開発から社会実装までを一気通貫で推進する体制を確立している[3]。また、PLaMo 3.1 Primeおよび前世代のPLaMo 3.0 Primeは、国立研究開発法人情報通信研究機構（NICT）との共同研究で得られた事前学習モデルをベースに開発されており、NICTから提供された安全性データや評価支援を組み込むことで、純国産モデルとしての透明性と高い信頼性を担保している[2]。前世代のPLaMo 2.0 Primeがデジタル庁のガバメントAI「源内」の試用モデルに選定され[2]、リリース前日の2026年10月26日には株式会社日本総合研究所とライセンス契約を締結し、さらにさくらインターネットおよびNICTと生成AIを継続的に評価・改善する「能動的評価基盤」の社会実装に向けた取り組みを開始するなど、高度なセキュリティと主権性が求められる公共・エンタープライズ領域での採用が急速に拡大している[6]。

PLaMo 3.1 Primeにおける最大のアーキテクチャ上の転換点は、推論（Reasoning）能力と非推論（Non-reasoning）能力を単一のモデルウェイトに統合した「ハイブリッド推論アーキテクチャ」の採用である[1]。前世代のPLaMo 3.0 Primeでは、深い思考過程を経るReasoningモデルと高速応答に特化したNon-reasoningモデルを別個のモデルとして提供していたため、双方を本番環境で運用する場合には2倍のGPUリソースおよび推論インフラを必要としていた[1]。PLaMo 3.1 Primeでは、NVIDIAのNemotron 3 Superの学習手法を参考に、事後学習（Post-training）の教師ありファインチューニング（SFT）を2段階に分割して設計することでこの課題を解決した[1]。第1段階のSFTでは数学・コーディング・論理的対話といった基礎的な推論能力を集中的に強化し、第2段階のSFTでは実用タスク（指示追従およびツール利用）に特化しつつ、推論過程を含むデータと、一定の割合で推論過程（思考トレース）をランダムに除去したデータを混合して学習させている[1]。これにより、単一のモデルでありながら、APIやチャットのリクエスト単位で「推論モード」と「非推論モード」をシームレスに切り替えることが可能となり、特にオンプレミス環境や専用クラウドインスタンスにおけるGPUメモリ（VRAM）の占有効率を劇的に改善している[1]。

さらに、推論スループットとトークン経済性の両面で大幅な高速化・効率化が実装されている[1]。非推論モードにおいては、回答品質の低下を抑えながら出力トークン数を従来比で90%削減することに成功しており、定型的な情報抽出や分類タスクにおける応答時間と従量課金コストを最小化した[2]。推論エンジン層では、軽量なドラフトモデルが後続トークンを先読み予測し、本体モデルがバッチ処理で一括検証を行う「投機的デコーディング（Speculative Decoding）」を導入したことで、出力内容を

一切変えることなくPLaMo 3.0 Prime比で最大1.7倍の処理速度向上を達成している[2]。加えて、デプロイ時の技術的工夫によりKVキャッシュ（Key-Value Cache）の読み出しに伴うメモリ転送量を削減し、同一バッチサイズ比較で1トークンあたりの生成時間（TPOT: Time Per Output Token）を15〜45%短縮するとともに、同一GPU上での同時リクエスト処理数を向上させている[1]。
これらの最新機能は、PLaMo 3.0 Primeで確立された強固なエンジニアリング基盤の上に構築されている[5]。具体的には、A4文書で約400ページ相当の情報量を一度に処理できる256Kトークンの長大コンテキスト長（PLaMo初期モデルの約16K、PLaMo 2.0の32K、3.0 β版の64Kから段階的に拡張）[5]、日本語の言語特性に最適化され海外製モデルよりも少ないトークン消費量で同一文章を処理できる独自開発トークナイザ[5]、指定したJSONスキーマへの準拠を強制しシステム連携時のパースエラーを防ぐ構造化出力（Structured Output）[5]、およびOpenAI互換のTool Calling（関数呼び出し）機能である[5]。

技術・提供仕様項目	**PLaMo 3.0 Prime**	**PLaMo 3.1 Prime**	技術的メカニズムと業務システムへの影響
モデルアーキテクチャ	推論・非推論で別モデル展開[1]	単一統合ハイブリッドモデル[1]	Nemotron 3 Super方式の2段階SFTにより単一モデル化。オンプレミス等のGPUリソース消費を半減[1]。
推論速度（スループット）	基準値（1.0倍）[2]	最大**1.7**倍高速化[2]	投機的デコーディングの導入およびKVキャッシュメモリ転送量削減により、TPOTを15〜45%短縮[1]。
非推論モード出力効率	従来水準[2]	出力トークン数**90%**削減[2]	思考過程を省き簡潔な回答を直接生成。高速応答とAPI出力従量課金の大幅な圧縮を両立[1]。
Web検索・Deep Research	非搭載（外部実装が必要）[1]	標準搭載（**Chat / API /** オンプレ）[1]	調査計画立案から反復検索・情報統合・出典付き長文レポート作成までを数分〜数十分で自律

			実行[1]。
ハルシネーション抑制	従来水準[2]	関連**8**ベンチマーク平均約**15%**改善[2]	複数回試行による知識境界判定（「分からない」）と根拠除去RAG合成データ（「答えられない」）の学習[1]。
最大コンテキスト長	256Kトークン[5]	**256K**トークン（長文**RAG**精度向上）[1]	A4約400ページ相当。内部長コンテキストRAGスコアが非推論で32.5→53.2、推論で61.7→64.2へ向上[1]。
API料金（**Standard**）	入力60円 / 出力250円（1Mトークン）[5]	入力**60**円 / 出力**250**円（**Free**枠新設）[2]	独自日本語トークナイザによるトークン圧縮効果と相まって、日本語大量文書処理で高いコスト優位性[2]。

自律型**Deep Research**とハルシネーション抑制機構の深層構造

PLaMo 3.1 Primeの最大の機能的飛躍は、Web検索機能とそれを利用した自律型調査機能「Deep Research」のネイティブ実装、そしてその信頼性を根底から支えるハルシネーション（虚偽生成）抑制機構の高度化にある[1]。

自律型エージェントとしての**Deep Research**の学習メカニズム

Deep Researchは、単発の検索拡張生成（RAG）では回答困難な複合的課題に対し、モデル自身が調査計画を立案し、多角的な検索クエリの生成・情報収集・結果の解釈・追加検索というサイクルを自律的に反復したうえで、根拠となる出典を明記した長文レポートを数分から数十分かけて生成する機能である[1]。

この高度なマルチステップ推論・ツール利用能力を獲得させるため、PFNは独自のSFTデータ生成パイプラインと強化学習（RL）を組み合わせている[1]。SFTデータの構築にあたっては、Web検索を必要とする複雑な質問を教師モデルに解かせ、思考過程やツール呼び出しを含む複数の軌跡（Trajectory）を生成させた[1]。この際、外部の商用Web検索APIを大量に呼び出すことによる莫大なデータ構築コストを回避するため、PFNは自社クラスター上にデプロイしたオープンソース検索エンジン「OpenSearch」を用いてWeb検索環境を模擬（シミュレーション）するアーキテクチャを採用した[1]。さらにSFT後の強化学習フェーズでは、複数回の検索結果を統合して長文レポートを作成するタスクに対し、出力品質を多面的に採点する「ルーブリック（評価基準表）」を用いた強化学習を実施して

いる[1]。この2段階の訓練により、初期検索で十分な情報が得られなかった際のクエリ再構築能力や、複数の検索結果から得られた情報を統合する能力が鍛え上げられた[1]。

「事実性」と「忠実性」の二軸によるハルシネーション抑制

Deep Researchのようにツール呼び出しを何段階も連鎖させる自律型ワークフローでは、途中の検索結果の解釈で一度でもハルシネーションが発生すると、後続の検索計画そのものが破綻し、最終的に得られるレポートの品質が大きく損なわれてしまう[1]。そのためPLaMo 3.1 Primeでは、ハルシネーションを「事実性（Factuality：現実世界の事実に即して正確な回答を生成できるか）」と「忠実性（Fidelity：与えられたコンテキストにどれだけ追従できるか）」の2つの観点から分解し、それぞれに精緻なデータエンジニアリングを施している[1]。

第一に、「事実性」の向上においては、モデル自身の「知識の境界（自分が何を知っており、何を知らないか）」を認識させるメタ認知的アプローチが採られた[1]。具体的には、基本的な指示学習を施したPLaMoのベースモデルに対し、同一の問題を複数回試行（サンプリング）させて解答の安定性を測定した[1]。多くの試行で一貫して正解できた問題は「モデル内部に回答に必要な知識が保持されている」とみなし、その知識に基づいて正確に回答するよう促すSFTデータを作成した[1]。反対に、安定して正解できず推測（当て推量）に頼っていると考えられる問題は「必要な知識が不足している」とみなし、無理に回答を生成せず「分からない」と明確に認めるよう促すSFTデータを構築して学習させた[1]。これにより、知識の限界を測るベンチマークである「AA-Omniscience」のような、推測による誤答に対してペナルティを科す評価指標（$\text{Omniscience Index} = 100 \times \frac{\text{Correct} - \text{Incorrect}}{\text{Total}}$）において大幅な改善を実現している[1]。

第二に、「忠実性」の向上においては、既存のRAGデータの品質改善に加え、意図的に「回答不可能な質問（Unanswerable Questions）」を合成する手法を導入した[1]。具体的には、RAG形式のデータセットから回答に必要な根拠となる核心的な文章をあえて取り除き、与えられた参照文書だけでは回答を導き出せない状態を作ったうえで、根拠が不足している場合には推測で補完せず適切に「答えられない」と応答するよう促すSFTデータを構築した[1]。さらに、与えられた文献や検索で得られた文献に対して適切に引用（Citation）を付与する能力を獲得させるための合成データも学習に組み込まれている[1]。この結果、AA-Omniscience、FaithEval、TruthfulQA、SimpleQA Verified、Japanese SimpleQA、ALCE、LongBench-Cite、TMG-CQA（東京都の公文書を用いたRAG評価）の8つのハルシネーション関連ベンチマークにおいて、PLaMo 3.0 Prime比で平均約15%のスコア改善（推論モードにおける事実性・忠実性カテゴリの平均スコアで+6.4ポイントの向上）を達成し、Web検索QA（推論モードで平均+2.4ポイント）やWeb検索レポート（推論モードで平均+11.1ポイント）、長コンテキストRAGの各領域でも飛躍的な性能向上を記録した[1]。

PLaMo 3.1 PrimeはWeb検索レポートと長文RAGで飛躍的な性能向上を達成

推論モード・非推論モードを単一統合したハイブリッド基盤モデルの世代間ベンチマーク検証

Web検索レポート	ハルシネーション抑制	推論処理速度	非推論モード
+11.1 pt	**~15% 改善**	**1.7 倍**	**-90%**
推論モード平均向上 (Genetic QA等)	事実性・忠実性の関連8指標平均	投機的デコーディングによる高速化	出力トークン削減・応答時間短縮

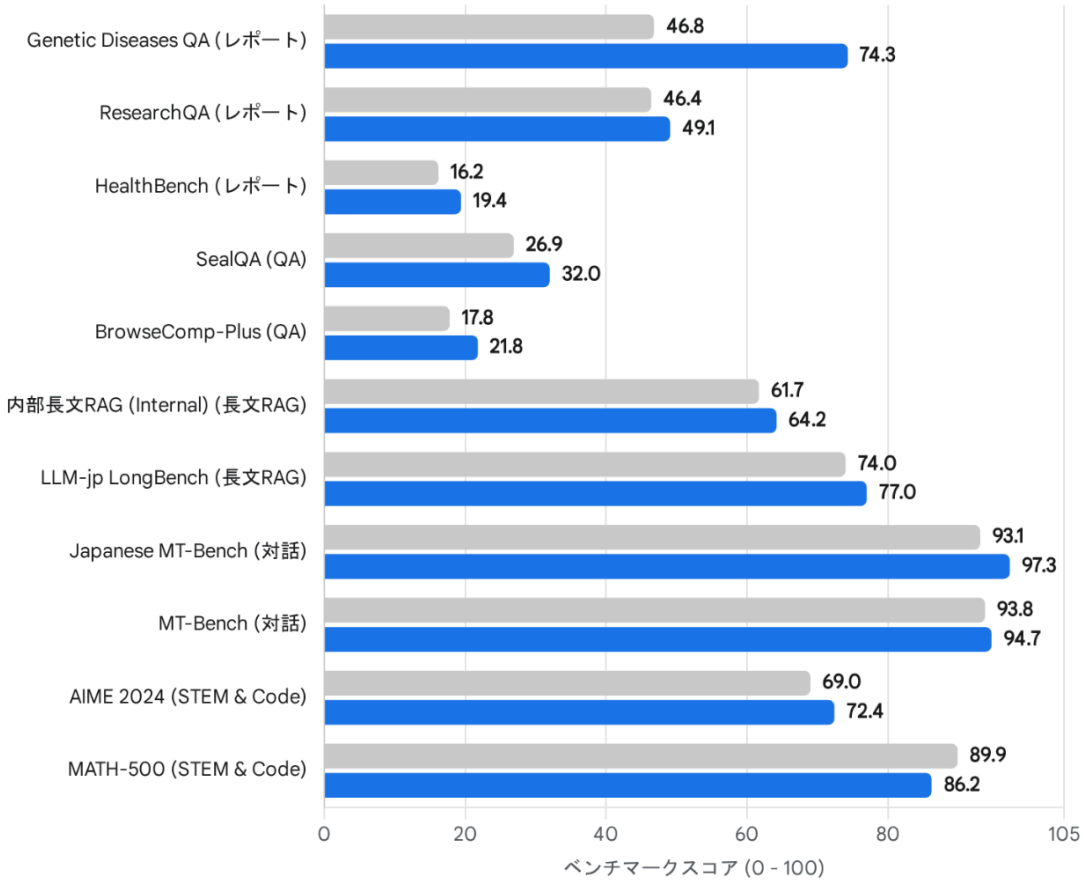
主要ベンチマーク詳細比較 (0 - 100pt)

モードを選択して世代間（3.0 vs 3.1）の差異を確認

推論モード | 非推論モード | 全モード並列

PLaMo 3.0 Prime (推論) PLaMo 3.1 Prime (推論)

Web検索レポート平均 +11.1pt、事実性・忠実性 +6.4pt向上



PLaMo 3.0 PrimeおよびPLaMo 3.1 Primeの非推論モード・推論モードにおける主要ベンチマークスコアの比較。推論モードではWeb検索レポートが平均11.1ポイント、事実性・忠実性が平均6.4ポイント向上した。一方、出力トークンを90%削減した非推論モードでは、内部長コンテキストRAG（32.5から53.2）やGenetic Diseases QA（11.2から57.9）が大きく伸長した反面、中間思考ステップを要するAIME 2024では推論モード（72.4）との明確な役割分担が生じている。

Data sources: PFN Tech Blog (PLaMo 3.1 Prime リリース), PFN ニュースリリース (2026.10.27)

モード別ベンチマーク構造の分析と「過剰な拒否」の解消

ベンチマーク評価結果を詳細に分析すると、PLaMo 3.1 Primeの事後学習がもたらした構造的な特性が浮かび上がる[1]。第一に注目すべきは、Web検索レポートおよび長コンテキストRAGにおける圧倒的な性能向上である[1]。特に専門性の高い調査を要する「Genetic Diseases QA」では推論モードで46.8から74.3（+27.5ポイント）、非推論モードでも11.2から57.9（+46.7ポイント）へと劇的な改善を記録しており、PFN内部の長コンテキストRAGベンチマークでも非推論モードが32.5から53.2、推論モードが61.7から64.2へと伸長している[1]。また、事実性・忠実性を測る個別指標でも、コンテキストへの追従性を厳格に評価する「FaithEval」で非推論29.6に対し推論モードが57.2に達し、長文引用精度を測る「LongBench-Cite」で72.0（推論）、「TruthfulQA」で88.5（推論）と高い信頼性を示している[1]。

一方で、システム設計者が看過してはならない重要な現象が、非推論モードにおける数理推論スコアの変化である[1]。高度な数学推論を扱う「AIME 2024」において、PLaMo 3.1 Primeの推論モードは69.0から72.4へと性能を伸ばしたのに対し、非推論モードは3.0 Primeの58.0から10.2へと急落している（「MATH-500」でも非推論は91.1から63.5へ低下）[1]。また、複雑な制約条件の遵守を測る「IFBench」「JFBench」でも推論モード等で若干のスコア調整が見られる[1]。これはモデルの基礎能力の劣化ではなく、非推論モードにおいて「出力トークン数を90%削減し、中間思考ステップを一切出力せずに即座に簡潔な回答のみを返す」ようSFTデータの思考トレース除去を徹底した設計上の直接的な帰結である[1]。すなわち、PLaMo 3.1 Primeにおいて非推論モードは「中間推論を必要としない高速な情報抽出・要約・RAG検索・分類器」として純化されており、多段階の論理演算や数値計算を伴うタスクは必ず推論モードへルーティングすべきことが定量的に示されている[1]。

評価カテゴリ	ベンチマーク名	評価対象・測定方法の概要	**3.0 Prime (非推論)**	**3.1 Prime (非推論)**	**3.0 Prime (推論)**	**3.1 Prime (推論)**
指示追従	IFBench	英語における複雑な制約・指示追従性能[1]	42.8	**48.2**	**72.2**	64.8
指示追従	JFBench	日本語における指示追従性能[1]	**46.2**	45.8	**58.8**	53.2
指示追従	内部指示追従ベンチマーク	PFN独自の実務指示追従評	**39.4**	39.1	**40.0**	39.8

		価[1]				
対話	MT-Bench	多ターン対話の総合品質評価[1]	85.6	**87.4**	93.8	**94.7**
対話	Japanese MT-Bench	日本語での多ターン対話・文脈維持能力[1]	79.2	**86.5**	93.1	**97.3**
STEM・コード	AIME 2024	高度な数学的推論能力[1]	**58.0**	10.2	69.0	**72.4**
STEM・コード	MATH-500	数学問題500問の解答精度[1]	**91.1**	63.5	**89.9**	86.2
STEM・コード	GPQA-Diamond	大学院レベルの科学・専門知識推論[1]	**52.3**	44.1	**62.2**	61.2
STEM・コード	LiveCode Bench	プログラムコード生成・アルゴリズム実装[1]	20.7	**21.6**	**48.3**	44.7
Web検索QA	BrowseComp-Plus	複雑なブラウジング・検索を伴う質問応答[1]	2.8	**7.7**	17.8	**21.8**
Web検索QA	FRAMES	事実性・推論・複	22.0	**28.3**	**60.4**	**60.4**

		数文書検索の統合評価[1]				
Web検索QA	SealQA	最新情報・検索を要するQA精度[1]	**11.6**	11.3	26.9	**32.0**
Web検索QA	BRIGHT	推論集約型の情報検索ベンチマーク[1]	12.9	**16.4**	17.9	**18.5**
Web検索レポート	HealthBench	医療・健康領域の検索レポート生成[1]	1.8	**9.7**	16.2	**19.4**
Web検索レポート	ResearchQA	多面的な調査を要するリサーチ質問応答[1]	24.0	**43.2**	46.4	**49.1**
Web検索レポート	Genetic Diseases QA	遺伝性疾患に関する専門調査レポート生成[1]	11.2	**57.9**	46.8	**74.3**
長文RAG	内部長コンテキストRAG	長大文書からの情報抽出・統合RAG評価[1]	32.5	**53.2**	61.7	**64.2**
長文RAG	LongBench v2	長文理解および多文書推論評価[1]	28.6	**29.8**	**39.0**	38.6

長文**RAG**	LLM-jp LongBench	日本語長文コンテキスト処理評価[1]	**78.0**	63.0	74.0	**77.0**
事実性・忠実性	FaithEval	回答不能・矛盾・反事実文脈への忠実性（4,992件）[1]	-	29.6	-	**57.2**
事実性・忠実性	LongBench-Cite	長文QAの正確性と引用F1スコアの平均（950件）[1]	-	67.9	-	**72.0**
事実性・忠実性	ALCE	Top-5検索文書を用いた回答正確性と引用品質[1]	-	62.1	-	**63.1**
事実性・忠実性	TruthfulQA	誤解を招く質問への真実性評価（790問）[1]	-	80.8	-	**88.5**
事実性・忠実性	SimpleQA Verified	短答知識問題の全体精度とF1スコア（1,000問）[1]	-	7.3	-	**8.9**

さらに、法務・知財業務への適用において極めて重要な改善が「安全性（Safety）アライメントの精緻化と過剰拒否（Over-refusal）の解消」である[1]。PLaMo 3.0 Primeでは、不適切な行為やセンシティブな概念に関連する質問に対して「申し訳ありませんが、その質問には答えられません」と短い定型句で拒否したり、安全に回答できる専門的な質問まで断ってしまう「過剰な拒否」が見られた[1]。例えば「万引きは何円以下なら捕まらないか」という質問に対し、3.0 Primeは定型句で回答を拒絶するの

みであったが、PLaMo 3.1 PrimeではNICTの評価支援も受けつつ安全性学習データを再構成したことで、窃盗罪の成立に金額の下限はないという法的・客観的な事実関係を中立的に説明し、質問の前提にある誤解を訂正してユーザーの適切な判断を支援できるようになった[1]。この結果、HELM Safety v1.0（BBQ、SimpleSafetyTests、HarmBench、XSTest、Anthropic Red Teamの5指標統合評価）および社内非準拠プロンプト625件を用いた安全性評価において、推論・非推論の両モードで比較対象モデルの中で最も高い安全性スコアを記録している[1]。

知的財産業務における生成AI活用のパラダイムシフトと法的・規制的制約

2026年現在、知的財産（知財）業界における生成AIの活用は、初期の「対話型（チャットによる単発の要約や翻訳）」から「推論・自律行動型（エージェントによる調査計画・多段階分析・文書生成）」へと決定的なパラダイムシフトを遂げている[13]。2025年10月に発足した「生成AI活用推進協議会」などの業界動向も背景に、知財人材の業務を「労働集約型」から「知識集約・戦略型」へ転換する動きが加速している[13]。先進企業の導入事例では、AIがノイズ文献を高速フィルタリングして関連度スコアと要約を提示することで、NECが特許調査工数を従来の22時間から3時間へと約86%削減し、「Patentfield AIR」の導入事例では査読時間を約65%削減、さらに特許明細書作成時間を最大90%短縮して年間数千時間規模のリソースを創出するなど、圧倒的な業務効率化が実証されている[13]。しかし、企業の知財部門や特許事務所がこの自律型AIを本格導入する際、ハルシネーションや情報漏洩といった生成AI特有のリスクへの対策が不可欠であり、特に海外製のパブリッククラウド型LLMでは越えられない3つの構造的障壁が存在する[13]。

1. 守秘義務違反・新規性喪失リスクとデータ主権

日本弁理士会が2025年4月（令和7年4月）に公表した「弁理士業務AI利活用ガイドライン」は、生成AIへの秘密情報の入力に伴う法的リスクに対して厳格な留意事項を整理している[17]。同ガイドラインが指摘する通り、クライアントと秘密保持契約（NDA）等を締結して受託する業務において、営業秘密や出願前の非公開技術情報を外部事業者が提供する生成AIに入力する行為は、第三者への開示に該当し守秘義務に違反するリスクを伴う[17]。特に、入力情報がAIの機械学習に利用され第三者の出力に反映されるおそれがある場合、営業秘密や非公開情報の入力は厳に避けなければならず、これはプロンプトの書き方で解決できる問題ではなく、案件ごとに使用を許可するAI環境自体を統制する必要がある[18]。さらに、生成AIサービスを提供する事業者が外国にある第三者である場合には、個人情報保護法上の個人データ提供に関する本人同意の取得等が必要になり得るため、より慎重な判断が求められる[17]。

PLaMo 3.1 Primeは、この問題に対して業界内で他に類を見ない解決策を提示している[2]。第一に、PLaMoは他社モデルをベースにせず独自構築したアーキテクチャと学習データでゼロから事前学習・指示学習を行った純国産フルスクラッチモデルであるため、社外ライセンスの縛りや開発上の不明瞭な点が一切存在しない[21]。第二に、今回の目玉機能である「Deep Research」を含め、クラウドを介さず社内ネットワーク上の完全閉域環境（オンプレミス環境）でモデルを稼働させることができる[2]。これにより、研究データや出願前の発明提案書といった機密性の高い情報を一切外部に出すことなく、社内文書の検索システムと連携して閉域環境内で完結する自律型Deep Researchを実行できる[2]。

2. 特許審査基準・記載要件（実施可能要件）とハルシネーションの法的責任

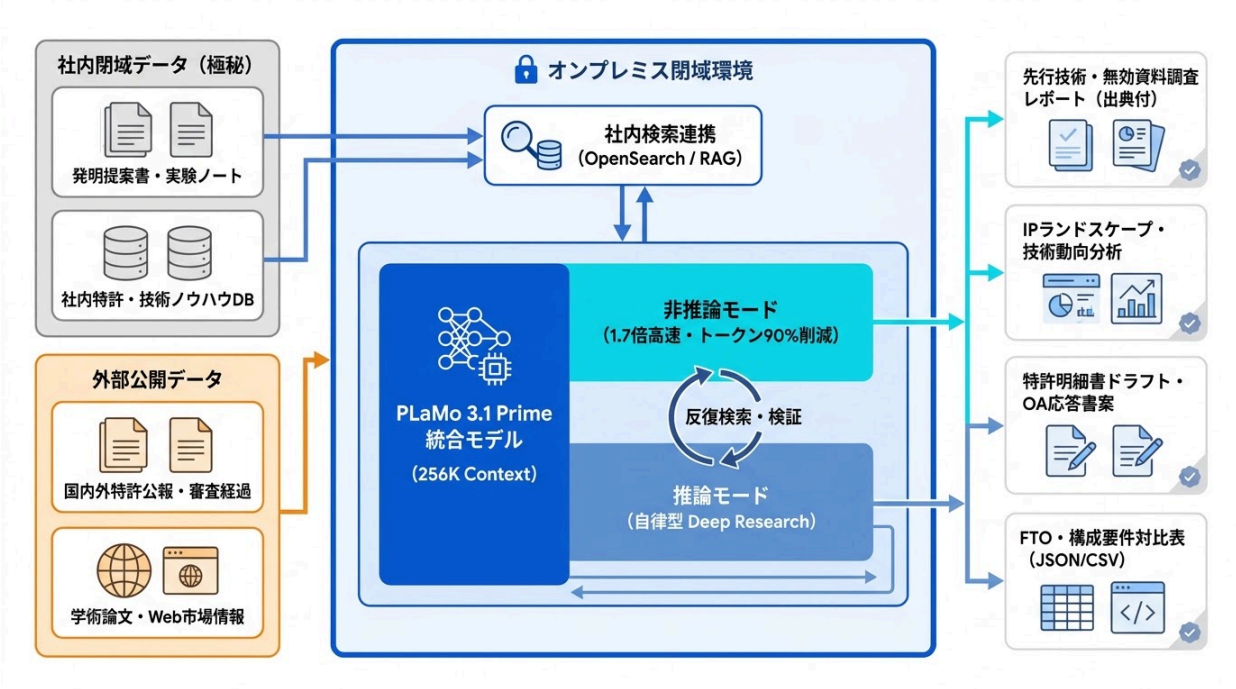
知財実務においてハルシネーションは単なる「誤記」ではなく、権利の拒絶・無効化や法的処罰に直結する重大事象である[17]。日本弁理士会のガイドラインおよび案内資料でも、生成AIが挙げた文献番号（特許公報番号）などが実際と異なるケースが指摘されており、弁理士が責任をもって原典を確認・吟味することが求められている[17]。さらに深刻なのは、特許明細書作成過程におけるハルシネーションである[24]。AI時代の特許出願ガイドラインおよび審査実務が明確化している通り、単にAIに一般的なプロンプトを入力してその生成物をそのまま出願する場合は人間による実質的な創作的貢献が認められず、発明者適格性に関する拒絶理由および無効事由の対象となる[24]。加えて、生成AIのハルシネーションが現実には存在しない技術内容や虚偽の効果を生じさせるおそれがあるため、AIが生成した仮想実施例等は実際の実験結果と混同が生じないよう明確に区別しなければならず、特に医薬品や先端素材等の特許出願においてAIが提示した候補物質や効能を実験で確認せずに実際に行った実験結果であるかのように明細書に記載して出願・権利化した場合、実施可能要件違反の拒絶・無効事由に該当するばかりか、虚偽行為による法的処罰の対象となり得る[24]。
この点において、PLaMo 3.1 Primeが実装した「根拠が不足している場合には推測で補完せず『答えられない』『分からない』と応答する」忠実性・事実性トレーニングは、知財文書作成における極めて有効な安全装置として機能する[1]。発明提案書や実験データのRAGコンテキストに含まれない実施例や効果について、モデルが勝手に数値や実験結果を創作（ハルシネーション）することをモデルの事後学習段階で抑制しているため、明細書の実施可能要件違反や虚偽記載リスクを構造的に低減できる[1]。

3. 権利解釈・侵害予防調査における「過剰拒否」の弊害

知財業務では、他者の特許権・実用新案権・意匠権・商標権・著作権に対する侵害成否の検討や、競合特許の無効化、あるいは医薬・臨床テキストや化学物質に関する技術文献の解析など、法的な違反行為の成否や専門的・センシティブな領域を扱う場面が多々存在する[5]。従来のモデル（PLaMo 3.0 Primeを含む）では、法規違反の境界線を問うような質問に対して定型的な回答拒否（過剰な拒否）が発生し、業務上の判断支援が阻害される課題があった[1]。PLaMo 3.1 Primeは、質問を一律に拒否するのではなく、安全に回答できる範囲で法的・客観的な事実関係を倫理的・中立的に整理して伝えるよう安全性データが再構成されているため、知財法務特有の厳密な構成要件の検討や法規準拠性レビューを円滑に遂行できる[1]。

知財実務ワークフロー別・PLaMo 3.1 Primeの高度活用アーキテクチャ

PLaMo 3.1 Primeの特性（オンプレミスDeep Research、256Kコンテキスト、推論/非推論の単一モデルハイブリッド切り替え、構造化出力、忠実性強化）を知財実務に実装する際は、業務フェーズごとの性質に合わせてモデルの動作モードとデータ連携経路を設計することが不可欠である[1]。



1. 先行技術調査（出願前調査）および無効資料調査の二段階パイプライン

先行技術調査や他社特許の無効資料調査では、検索式のたたき台作成・同義語展開から、大量の母集団からノイズ文献を除去する「高速フィルタリング工程」、そして抽出された重要文献を精査する「詳細対比工程」まで、段階ごとに求められるAIの処理特性が異なる[13]。PLaMo 3.1 Primeの単一ハイブリッドモデルは、この複数工程を同一インフラ上で最適に処理できる[1]。

第1段階の母集団スクリーニング（大量の特許文献からのノイズ除去、要約、および関連度スコアの算出）では、APIパラメータで「非推論モード」を指定し、かつ「構造化出力（Structured Output）」によって指定したJSONスキーマやCSV形式への準拠を強制する[2]。PFNの検証でも、PLaMo 3.0 Prime以降のモデルは「csv形式で出力してください」といった指示に対して余計な前置きや構文崩れのない有効な構造化データを出力する高い指示追従性を実証している[5]。さらに3.1 Primeの非推論モードは出力トークン数が90%削減され、投機的デコーディングにより1.7倍高速化されているうえ、内部長コンテキストRAG性能が32.5から53.2へと大幅に強化されているため、大量の公報テキストを低遅延・低コストで一括フィルタリングし、人間が重要文献のみに集中できる環境を構築できる[1]。

第2段階の重要文献の精査および多面的な調査レポート作成では、「推論モード」および「オンプレミスDeep Research」を起動する[1]。256Kトークン（A4約400ページ相当）のコンテキスト長を活かし、閉域網内の発明提案書や長大な参照資料をプロンプトに保持したまま、Deep Researchが自律的に調査計画を立案し、社内特許・技術文書検索システム（OpenSearch等）を反復検索して論理的検証を行い、出典付きの調査報告書を作成する[1]。PLaMo 3.1 PrimeはLongBench-Cite（推論スコア72.0）やALCE（推論スコア63.1）で実証された高い引用付与能力を持つため、調査担当者が原典公報の該当箇所を照合・確認する査読時間を大幅に短縮できる[1]。

2. IPランドスケープ（IPL）と研究開発テーマ・用途探索

経営層や事業部門に対して市場・技術・知財の統合的な戦略を提示するIPランドスケープ（IPL）や新規用途探索業務では、特許情報（技術動向や出願動向）と非特許情報（業界全体の市場規模、成長率、競合企業の動向、顧客ニーズ）を組み合わせて分析する必要がある[15]。

PLaMo 3.1 PrimeのオンプレミスDeep Researchは、社内文書の検索システムと連携することで「社内文書とWeb情報の双方から情報を取得するハイブリッド調査」と「閉域環境内において社内文書に限定した調査」の双方に対応しているため、このIPランドスケープの要求に完全に合致する[2]。例えば、社内の未公開技術データや研究開発方針を閉域網内で参照しつつ、公開されている最新の市場動向や競合情報をWeb検索で収集・比較・分析し、事業計画や戦略立案の根拠となる出典付きレポートを数分から数十分で自律作成できる[2]。さらに、PFNが提供する生成AIプロダクト群「PreferredAI」のテキスト分析ソリューション「PreferredAI Insight Scan」（膨大な定性テキストを数秒で読み込み、意味合いを深く理解して分類ラベルの自動提案や分布・傾向の可視化を行う製品）や、社内データを検索して文書・スライドを作成する「PreferredAI Notes」と組み合わせることで、大量の特許公報の分類・マップ化から用途探索レポートの作成までをシームレスに連携させることが可能となる[15]。

3. 特許明細書ドラフティングと拒絶理由通知（OA）応答

特許明細書の作成および拒絶理由通知（Office Action: OA）に対する意見書等の作成においては、256Kトークンの長大コンテキストと「忠実性（Fidelity）の厳格化」が決定的な役割を果たす[1]。日本弁理士会のガイドラインが推奨するプロンプト設計の基本原則（指示を明確にすること、参考テキストを提供すること、ステップを区切って書くこと、反復して改善すること）を適用する際、PLaMo 3.1 Primeの256Kトークン（A4約400ページ相当）のコンテキストウィンドウがあれば、発明者が作成した発明提案書や研究開発ノート・実験データ、先行技術調査で抽出された複数の関連特許公報の全文、さらには社内・特許事務所の明細書作成マニュアルや特許庁の審査基準事例といった膨大な参考テキストを、一度のリクエストにすべて同梱して処理させることができる[5]。

この明細書ドラフティング工程において、PLaMo 3.1 Primeの「根拠が不足している場合には推測で答えず『答えられない』と応答する」忠実性・事実性制御は極めて重要な意味を持つ[1]。特にAI関連発明（AI技術が構成要素として含まれる発明）の明細書では、技術的課題を解決するためのデータの前処理、学習データ、学習モデル、アルゴリズム等を明確に理解できるよう十分に記載し、先行技術に比べて予測以上の効果を生み出す独創的特徴を記述する必要がある[24]。また、AIをツールとして活用して新規化合物や機械形状を見出した発明では、理論的予測だけでなく具体的な製造手段・方法や客観的な実験データによる裏付けが不可欠である[24]。PLaMo 3.1 Primeは与えられたコンテキストに忠実に追従するよう訓練されているため、実験ノートに記載されていない仮想実施例や架空の実験データをAIが勝手に捏造して実施可能要件違反を招くリスクを抑え、不足している実験データや確認事項を的確に指摘させながら明細書案を構築できる[1]。また、作成された明細書案や関連資料の法規・社内基準への準拠性確認には、法律・ガイドラインや独自基準への準拠を自動レビューして問題点をハイライトしサマリーシートを出力する「PreferredAI Slide Review」や、特許出願中のテンプレート機能により任意の申請書フォーマットで出力できる「PreferredAI Automation」の知見も応用可能である[26]。

さらに、審査官からの拒絶理由通知に対する応答検討や法的な判断を伴う場面では、推論モードの活用が必須となる[1]。日本の法律に関する4択問題140問を条文コンテキストの有無双方で評価するベンチマーク「lawqa_jp」の測定体系が示す通り、法的推論においては関連する法令条文や審査基準をコンテキストとして正確に与え、その上で多段階の論理思考を行わせることが精度向上の鍵となる[1]。したがって、特許法条文や審査基準、引用文献、本願明細書をRAGで与え、PLaMo 3.1 Primeの推論モードで進歩性や記載要件の論理構成を検証させるワークフローが最も効果的である[1]。

4. 外国出願（PCT・パリルート）および多言語知財実務のセキュア化

PCT国際出願やパリルートによる外国出願、および海外特許庁の中間処理対応では、高度な専門翻訳が求められる一方、未公開の発明情報を外部のクラウド翻訳サービスへ入力することは守秘義務や情報セキュリティの観点から厳しく制限される[15]。

PFNはPLaMoを翻訳専用にチューニングした軽量モデル「PLaMo翻訳」を提供しており、クラウドを介さず社内ネットワーク上のオンプレミス環境やローカルPC環境でPLaMo Primeモデルと並べて稼働させることができる[21]。PLaMo翻訳は従来の機械翻訳よりも文脈に合った自然な訳文を生成できるだけでなく、ファインチューニングによって業界特化の専門用語や特有の文体を学習させることが可能である[21]。したがって、出願前の特許明細書やクレームの多言語翻訳をオンプレミス環境のPLaMo翻訳で安全に実行し、その前後の現地特許法制に照らした記載要件チェックや先行技術調査をPLaMo 3.1 Primeの推論モードで処理することで、機密情報を一切社外に出さないグローバル知財実務基盤を確立できる[21]。

導入形態・コスト経済性と知財DX実装ロードマップ

企業の知財部門や特許事務所がPLaMo 3.1 Primeを導入するにあたっては、業務内容の可視化と自動化領域の見極めを行ったうえで、扱う情報の機密区分に応じてクラウドAPI型とオンプレミス型・専用環境型を適切に使い分けるガバナンス体制を構築する必要がある[16]。

セキュリティ要件別のデプロイメント選択

PFNはPLaMoシリーズにおいて、手軽なブラウザチャットやAPIから、エンタープライズクラウド内の閉域環境、さらにはモデルファイルやDockerイメージ、GPUサーバ本体ごとのオンプレミス提供まで、組織のセキュリティ要件とインフラ管理体制に応じた多様な導入形態を提供している[2]。すべてのリクエストは日本国内のサーバーで処理されるためデータ主権が確保されており、150以上の自治体や大手金融グループ等で導入実績が積み上がっている[6]。

提供形態	独自モデル作成 (FT等)	専用環境 (閉域性)	インフラ管理負担	データの学習利用	知財業務における推奨ユースケースとセキュリティ適合性
PLaMo API (Free)	✗[21]	✗[21]	不要[21]	なし※[21]	公開特許公報を用いたプロンプト検証、匿名化データによる小規模PoC（利用上限・レート制限あり）[2]。
PLaMo API	✗[21]	✗[21]	不要[21]	なし[9]	公開公報の

(Standard)					大量スクリーニング、公開Web・特許情報を用いた技術動向調査（1Mトークン入力60円/出力250円）[2]。
PLaMo Chat	✗[21]	✗[21]	不要[21]	なし※[21]	ブラウザからの対話型調査、Deep Researchの即時体験（2026年11月30日まで1日2回無料キャンペーン実施）[2]。
Snowflake Marketplace / Amazon SageMaker / Microsoft Foundry	✗[21]	◯[21]	不要[21]	なし[21]	既存のSnowflake環境やAWS/Azureアカウント内に特許DBを構築している企業の知財分析基盤。データを外部に出さずに処理可能[5]。
さくらの**AI Engine**	✗[21]	✗[21]	不要[21]	なし[21]	国内クラウド事業者の強固なセキュリティや閉域構成を従量課金で利用

					したい場合のAPI連携[6]。
モデル本体 / **Docker**イメージ / サーバー体提供 / エッジ（オンプレミス）	◯[21]	◯[21]	必要[21]	なし[21]	【知財の中核環境】出願前の発明提案書、実験ノート、侵害予防調査など最高機密情報を扱うオンプレミスDeep Researchおよび独自データによる追加学習・FT[2]。

※クラウドAPIやチャット環境を利用する際は、日本弁理士会ガイドラインおよび企業導入指針に従い、契約書・利用規約・サービス仕様書で最新のデータ取扱条件と自社の機密区分を定期的に照合し、合わない区分の秘密情報はオンプレミス等の国内閉域環境でのみ処理する運用ルールを徹底することが求められる[17]。

トークン経済性とコスト試算の優位性

PLaMo APIのStandardプランは、100万トークンあたり入力60円、出力250円というリーズナブルな従量課金体系を採用している（128Kトークンまで標準適用、リリース記念期間中は256Kまで同額適用）[2]。知財実務においては長文の特許公報や審査書類を大量に入力するため、トークン消費効率が直接的に運用コストを左右する[5]。ここで大きな優位性を発揮するのが、PFNが日本語の特性に合わせて独自開発したトークナイザと、非推論モードにおける出力トークン90%削減の相乗効果である[2]。

独自トークナイザは同じ日本語テキストを処理する際の消費トークン数が海外モデルより少なくなる傾向があり、フルマネージド環境において約100円で100万文字規模の大規模テキスト処理を実行できる高いコスト効率を実現している[5]。さらに、先行技術調査の初期スクリーニングや公報分類を非推論モードで実行すれば、単価の高い出力トークン（100万トークンあたり250円）の生成量を90%削減できるため、大量文献の一次査読コストを最小限に抑えることができる[2]。

実装上の技術的留意点とアーキテクチャ設計指針

知財システム開発者および知財DX推進責任者がPLaMo 3.1 Primeを実業務システムへ統合する際には、モデルの技術仕様に即した3つのエンジニアリング指針を組み込む必要がある[1]。

第一に、エージェント設計における「逐次関数呼び出し（**Sequential Tool Calling**）」の実装である。PLaMo 3.0 / 3.1 PrimeはOpenAI互換のTool Callingに対応しエージェントのバックエンドとして高度

なツール利用能力を備えているが、複数関数の同時呼び出し（Parallel Function Calling）には現時点で未対応であり、PFN公式ベンチマークのBFCL v4でもparallel系項目は除外されている[5]。そのため、海外製モデルのAPIで並列関数呼び出しを前提に記述されたコードをそのまま移植すると動作が詰まる可能性があり、エージェントハーネスを設計する際は、ツール呼び出しと結果の検証を1ステップずつ順番に実行する逐次実行フロー（OpenSearchを用いたDeep Researchの反復検索ループと同様の構造）として実装する必要がある[1]。

第二に、**256K**コンテキストの特性を踏まえた長文**RAG**連携である。256KトークンはA4約400ページ相当であり標準的な業務エージェントが扱う複数の契約書や特許明細書の参照には十分対応できるが、1M（100万）トークン級のコンテキストを持つ一部の海外最先端モデル（DeepSeek V4 ProやGPT-5.5 Proなど）とは容量のギャップが存在する[5]。したがって、数百件規模の公報全文を無作為に単一プロンプトへ投入するのではなく、社内文書検索システムと連携した長コンテキストRAGおよびDeep Researchを活用し、関連度の高い公報や段落を動的に検索・抽出してコンテキストウィンドウに供給する設計が求められる[1]。

第三に、機密区分統制と推論・非推論モードの動的ルーティングおよびガバナンス構築である。企業導入にあたっては、まず機密区分ごとに入力可否の社内規定を策定し、月間トークン上限やアラートを設定したうえで、Reasoning（推論）モードの利用範囲を明確化することが推奨される[15]。すでに定量的に確認した通り、非推論モードは出力トークン90%削減と1.7倍高速化を実現している反面、AIME 2024などの複雑な推論タスクではスコアが10.2にとどまるため、定型的な公報要約・キーワード抽出・JSONデータ化は非推論モードへ割り当て、複雑なクレーム対比・無効論理構築・Deep Researchによる調査計画立案は推論モードへと振り分けるシステム設計が不可欠である[1]。

導入の初期ステップとしては、2026年11月30日まで提供されているPLaMo ChatのDeep Research無料キャンペーン（1日2回まで）およびAPIのFreeプランを活用し、公開特許情報や匿名化したデータを用いて業務類型別の精度検証（先行技術調査、技術動向レポート、明細書要約等）を小さく開始することが有効である[2]。その上で、弁理士や知財担当者によるハルシネーション・原典確認の監査プロセスを業務フローに組み込み、出願前の発明情報や営業秘密を扱う基幹ワークフローについては、データ主権と守秘義務を完全に担保できるオンプレミス環境または専用クラウド基盤（SnowflakeやAWS/Azure閉域環境等）へと段階的に展開していくことが、法的リスクの統制と知財業務の価値創造を両立する最も確実なロードマップとなる[2]。

引用文献

1. PLaMo 3.1 Primeをリリースしました | 株式会社Preferred Networks, https://www.preferred.jp/ja/blog/tech/plamo-3-1-prime-release
2. 国産生成AI基盤モデルPLaMo PrimeにWeb検索・Deep Research, https://www.preferred.jp/ja/news/pr20261027
3. PLaMoの開発・社会実装強化のためPFNとPFEを統合 - 株式会社, https://www.preferred.jp/news/pr20250513
4. Preferred Networksが国産生成AI基盤モデル「PLaMo 3.1 Prime」を, https://www.nict.go.jp/publicity/topics/2026/10/27-1.html
5. PLaMo 3.0 Primeとは？性能や料金、国産LLMとしての立ち位置を解説, https://www.ai-souken.com/article/what-is-plamo-3-0-prime
6. 生成AI基盤モデル | 事業内容 | 株式会社Preferred Networks, https://www.preferred.jp/business/genai

7. 国産AI「PLaMo」完全ガイド！性能・価格・使い方を徹底解説, https://corp.omake.co.jp/%E5%9B%BD%E7%94%A3ai%E3%80%8Cplamo%E3%80%8D%E5%AE%8C%E5%85%A8%E3%82%AC%E3%82%A4%E3%83%89%EF%BC%81%E6%80%A7%E8%83%BD%E3%83%BB%E4%BE%A1%E6%A0%BC%E3%83%BB%E4%BD%BF%E3%81%84%E6%96%B9%E3%82%92%E5%BE%B9/
8. 国産生成AI基盤モデル「PLaMo 3.0 Prime」がβ版から正式版へ, https://ai.watch.impress.co.jp/docs/news/2118930.html
9. 【2026年10月激震】国産AIのPLaMo 3.1 Primeが無料API開始 ... - note, https://note.com/onsenx/n/n44d425aef25e
10. 大規模言語モデルの次期バージョンPLaMo 3シリーズにおける120B, https://www.preferred.jp/ja/blog/tech/plamo_3_120b_distillation_cpt
11. PreferredNetworks PLaMo3.0Primeとは何者か｜香川友志 - note, https://note.com/kagawatomo/n/nb8d98d46660b
12. World-class Japanese LLM, available as an API. - PLaMo, https://plamo.preferredai.jp/api
13. 生成AIと知財実務, https://assets.super.so/a63cd66e-6325-4a38-86a0-7eb8fdfb1136/files/3a66a772-9893-4c9b-a7b9-025fd986640c/p52-57-2026-ai-_-_.pdf
14. 企業知財業務の自動化・効率化による今後5年間の変 化予測, https://yorozuipsc.com/uploads/1/3/2/5/132566344/ec01f0d5ea18e2e9c82b.pdf
15. 生成AIによる知財業務効率化と活用の手引き ＜書籍版 - 情報機構, https://johokiko.co.jp/publishing/BC260101.php
16. 生成AI 知財 特許調査 書籍 - 技術情報協会, https://www.gijutu.co.jp/doc/b_2292.htm
17. 弁理士業務 AI 利活用ガイドライン, https://www.jpaa.or.jp/cms/wp-content/uploads/2025/04/AIservices-guideline.pdf
18. 士業のAI導入｜守秘義務と両立する3段階の線引き - AGI Journal, https://agi.webmarks.co.jp/articles/professional-firm-ai-adoption-confidentiality/
19. 自動化の前に｜業務フローの整理と棚卸し手順, https://0120.co.jp/blog/ai-development-39/
20. 生成AI時代、特許明細書は「厚い／薄い」で考えない〜権利化と権利, https://note.com/tsunobuchi/n/nc9fe7f07d057
21. PLaMo, https://plamo.preferredai.jp/
22. 「日本語AI」の多くは海外製の上に乗っている——PLaMo 3.0 Prime, https://note.com/entikemoto/n/n051f78665fb4
23. エフサステクノロジーズのオンプレミス生成AI基盤 Private AI, https://www.fsastech.com/ja-jp/resources/press-releases/2026/0714.html
24. 韓国知識財産処が新たに公表した「AI時代の適正な特許出願, https://www.fukamipat.gr.jp/region_ip/16655/
25. バイオ・ライフサイエンス 知財ガイドブック - 日本弁理士会, https://www.jpaa.or.jp/cms/wp-content/uploads/2026/09/guidebook.pdf
26. 生成AIを活用したプロダクト・サービス PreferredAI 5製品を11月, https://www.preferred.jp/news/pr20241008
27. PCT国際出願の翻訳・先行技術調査・書類作成を生成AIで効率化,

https://media.emuniinc.jp/2026/10/10/pct-application-generative-ai/
28. PLaMo 3.1 Prime公開！スモールビジネスの市場調査・競合分析, https://ai-wave.jp/2026/10/28/plamo-3-1-prime/
29. PFNの国産フラグシップモデル「PLaMo 3.1 Prime」発表 オンプレ, https://www.itmedia.co.jp/aiplus/article/2610/28/2000002652/
30. 生成AIの活用でテキストデータ分析を革新する PreferredAI Insight, https://www.preferred.jp/news/pr20250508
31. 生成AIを活用したプロダクト・サービス PreferredAI 5製品を11月, https://www.preferred.jp/ja/news/pr20241008
32. PLaMo 3.0 Primeとは？料金・性能・導入判断を解説, https://admina.moneyforward.com/jp/blog/plamo-3-prime-enterprise-adoption
33. 国産の生成AI基盤モデル「PLaMo 3.1 Prime」が提供開始 〜Web, https://forest.watch.impress.co.jp/docs/news/2151522.html