ChatGPT

国産生成AI基盤モデル PLaMo Prime／PLaMo 3.1 Primeの知財実務への適用可能性に関する調査報告書

エグゼクティブサマリ

第一に、PLaMo 3.1 Primeは「日本語に強い国産LLM」という位置づけから、企業向けの調査エージェント基盤へ明確に進化した。 2026年10月27日に提供開始されたPLaMo 3.1 Primeでは、PLaMo Chat/API双方にWeb検索が加わり、Deep Researchでは調査計画の作成、反復検索、情報整理・分析、出典付きレポート生成までを自律的に実行する。さらにDeep Researchをオンプレミス環境でも利用でき、社内検索システムと接続して閉域内文書だけを調査させる構成もPFN自身が明示している。これは未公開発明、FTO、無効資料、契約書、訴訟準備資料などを扱う知財部門にとって特に重要な差別化要素である。 [1]

第二に、知財実務では「調査・整理・草案作成」に非常に適する一方、「法的結論」や「証拠評価」を人間から置き換えるべきではない。 PLaMo 3.1 Primeは事実性・忠実性を重点的に事後学習しており、PLaMo 3.0 Prime比でハルシネーション関連8ベンチマークのスコアを平均約15%改善したとPFNは公表している。しかしPFN自身の比較でも、STEM/CodeやWeb検索QAでは一部競合モデルに劣るとしており、Web検索結果の誤読がDeep Research全体を損なう可能性も開発ブログで明示している。したがって、特許性・侵害性・無効理由・契約解釈などの最終判断では「一次資料確認＋弁理士・弁護士レビュー」を必須とする運用が妥当である。 [2]

第三に、知財利用では「通常API」と「Web検索／Deep Research」の情報管理を分けて考える必要がある。 Standard/Providerの通常APIはZDR（入力・出力を保持しない）の対象で、学習利用もされない。一方、現時点のWeb検索とDeep ResearchはZDR対象外であり、検索クエリ等が国内リージョン外の外部検索サービスへ送信され、Deep Research結果は完了後24時間保持される。Freeプランでは入力・出力自体が保存・学習利用される。このため、未公開発明名、秘密の製品仕様、相手方との交渉内容、個人情報等をそのままWeb検索に流す運用は避けるべきである。 [3]

第四に、オンプレミスはPLaMoの知財用途で最も戦略的価値が高い提供形態である。 PFNはモデル本体、Dockerイメージ、モデル＋GPUサーバー一式の提供に加え、Snowflake、さくらのAI Engine、Amazon SageMaker、Microsoft Foundry等のプライベートな展開形態を用意している。モデル本体ではファインチューニング・追加学習も可能とされる。ただしPLaMo 3.1 Primeそのものの総パラメータ数、必要GPU数、最小VRAM、推奨GPU型番、具体的オンプレミス価格は現時点で公開資料上**未指定**であり、導入費を精密に見積もるにはPFNへの個別見積りが必要である。 [4]

第五に、知財部門への導入は「検索エンジンの代替」ではなく、「調査オーケストレータ＋証拠整理器」として始めるのが適切である。 特許番号・公開番号・出願日・優先日・引用文献・請求項対応表などを構造化させ、J-PlatPat、特許庁公報、Espacenet等の一次資料へ人間が戻れる設計にする。PLaMoはOpenAI形式に近いChat Completions/Responses API、JSON SchemaによるStructured Outputs、Function Callingを備えるため、このような知財ワークフローへの組込みが比較的容易である。なおWeb検索とJSON Schemaを同時利用した場合は引用が返らないというAPI上の制約があるため、「引用取得フェーズ」と「構造化フェーズ」を分離する実装が望ましい。 [5]

技術仕様、Deep Research、信頼性

PLaMo 3.1 Primeについて公開資料から確認できる仕様と、確認できない事項を以下に整理する。「未指定」は、PLaMo 3シリーズ一般の研究情報から推測できる場合でも、PLaMo 3.1 PrimeについてPFNが明示していないものを意味する。

項目	PLaMo 3.1 Primeの確認結果	知財実装上の意味
リリース	2026年10月27日。最新フラッグシップモデル。 [6]	3.0以前とはWeb検索・DR能力が異なるため評価を分けるべき
モデルサイズ	**未指定**	GPU容量・オンプレ費用を公開情報だけでは算定不可
PLaMo 3系列研究モデル	31B、120B等をPFN/NICTが研究・公開。120Bベースは3兆トークン規模の事前学習。これは3.1 Primeのサイズを意味しない。 [7]	Primeのパラメータ数を120Bと仮定してはいけない
基本アーキテクチャ	PLaMo 3系列ではAttention＋Sliding Window Attentionを採用。PLaMo 3.1 Prime固有の層数等は**未指定**。 [8]	オープンモデルほど詳細な容量計画はできない
コンテキスト	262,144 tokens	長大な明細書群・契約群・証拠資料の一括分析に有利。 [9]
最大出力	20,480 tokens	長文クレームチャート等は分割生成推奨。 [9]
推論方式	1モデル内で推論／非推論を切替えるハイブリッド方式	調査・侵害分析はreasoning、分類・要約はnon-reasoningという使い分けが可能。 [10]
SFT	2段階SFT。第1段階は推論中心、第2段階は推論あり／なしを混合	複雑分析と高速処理を同一基盤で扱える。 [11]
強化学習	汎用RL＋Web検索特化RLを実施	検索タスクへの事後最適化が明示されている。 [11]
DPO	非推論タスクを中心にDPO実施	応答形式・簡潔性調整に利用。 [11]
RLHF	「RLHF」という方式名は**未指定**。強化学習・DPOの実施は明示。	RL＝RLHFと断定しないこと
学習データ	日本語重視の独自データ、NICT関連データ等をPLaMo 3開発で活用。3.1では利用可能性を個別確認した社内データも学習データ生成に利用。全コーパス一覧・比率は**未指定**。 [12]	特定公報が学習済みか等は判定不能
Web検索	Chat/API対応、回答にURL citation付与	最新判例・法改正・企業ニュース補完に有効。 [13]

項目	PLaMo 3.1 Primeの確認結果	知財実装上の意味
Deep Research	`plamo-3.1-prime-deep-research`、Responses API、非同期ジョブ	複数ソース型の知財調査に適する。[9]
Structured Output	JSON Schema対応	特許番号、請求項要素、証拠日付等の抽出に有効。[14]
Function Calling	通常Chat Completionsで対応	社内特許DB、契約DB、J-PlatPat連携用ゲートウェイ等を接続可能。[14]
Web検索＋JSON	JSON Schema指定時、引用を返さない	「証拠探索→引用固定→JSON化」の2段階処理が必要。[15]
オンプレ	モデル本体、Docker、サーバー一式等を提供。Deep Researchもオンプレ可。[16]	機密知財への適合性が高い
GPU要件	**未指定**	PFNによるサイジング必須
監査ログ	PLaMo標準サービスとして完全なSIEM向け監査ログ仕様は**未指定**	自社側API Gatewayでログを補完すべき
利用管理	テナント／プロジェクト、ロール、APIキー、日次利用量、予算上限、アラート、2要素認証あり	部門・案件単位の費用統制が可能。[17]

特に重要なのは、Deep Researchが単に検索APIを付加したものではなく、検索を必要とする質問を教師モデルに解かせたSFT軌跡、検索クエリ生成、ツール呼出し、複数文献統合、長文レポートに対するルーブリックベースの強化学習を行っている点である。PLaMo 3.0 Prime比で、推論モードのWeb検索QAは平均2.4ポイント、Web検索レポートは平均11.1ポイント改善したとPFNは報告している。[18]

一方、Deep Researchを「出典があるから正しい」と評価するのは危険である。PFN自身が、検索結果を途中で一度でも誤解釈すれば最終レポート品質が大きく損なわれ得ると説明している。3.1では「知らない場合は知らないと答える」SFT、根拠を除去した回答不能RAGデータ、引用付与データ等を追加し、事実性・忠実性カテゴリを3.0比6.4ポイント改善しているが、これは誤りがなくなったことを意味しない。[18]

知財向けリスク評価は「中」。 一般的な調査補助としては十分に実用候補だが、先行技術調査では「検索漏れ」、侵害分析では「請求項解釈の誤り」、無効資料調査では「基準日前後の日付混同」が致命的になり得る。したがって、出典URLだけでなく「資料名・公報番号・段落番号・請求項番号・公開日・優先日・取得日時」を別フィールドで保存し、一次資料との一致を機械的に検査するのが推奨される。これは、PFNが引用能力を明示的に学習している一方、競合比較でWeb検索QAや長コンテキストRAGにはなお改善余地があると自ら評価しているためである。[19]

知財業務別ユースケースと推奨運用

PLaMo 3.1 Primeの最大の価値は、専門家を置き換えることではなく、「検索→抽出→比較→論点化→証拠リンク付与」の工数を減らすことにある。特許庁自身もAIを用いた先行技術調査高度化を導入フェーズへ進め、生成AIの特許審査業務への適用を技術実証対象としているため、知財実務におけるAI利用そのものは制度運用側でも現実的な検討対象になっている。[20]

業務	PLaMoの利用方法	利点	主な制約・リスク	推奨運用・対策
先行技術調査	発明を要素分解→日本語/英語検索語・IPC/CPC候補生成→Web/社内DB検索→候補文献ランキング	多段検索・比較・引用に適合	網羅性保証なし、検索対象DBの差	「AI候補探索＋専門検索者による最終検索」。基準日以前資料だけを別検証
無効資料調査	請求項要素ごとに単独文献／組合せ文献を探索、公開日・優先日・記載位置を整理	Deep Researchと相性が良い	日付錯誤が致命的	公開日、優先日、出願日を別フィールド化し原公報で確認
特許明細書作成補助	発明者メモ→課題・構成・作用効果・実施形態の草案	日本語文章生成、256K context	新規事項、発明者性、過度な一般化	発明者ヒアリング結果のみをground truthにし、人間が最終校閲
クレーム作成支援	独立請求項案、従属項バリエーション、サポート箇所マッピング	多案生成が高速	権利範囲の不適切な拡縮	「原明細書サポートなしの文言を禁止」し、段落番号を必須出力
OA・拒絶理由対応	引用発明と請求項差分、審査官ロジック、補正候補、意見書骨子を整理	複数文献比較に有効	進歩性判断を過信しやすい	最終論理・補正方針は担当弁理士承認
侵害/FTO分析	claim chart作成、製品資料との要素対応、非充足要素・不明要素を抽出	長文資料を横断比較	クレーム解釈・均等論を誤る可能性	「充足／非充足／不明」の3値＋証拠引用で出力し、法律判断を分離
商標調査	類似候補、称呼・外観・観念、指定商品・役務候補を整理	文字・概念的比較が高速	類否判断・DB網羅性は保証不可	J-PlatPat等の正式検索と併用
意匠	物品・用途・形態特徴のテキスト分類、先行意匠候補の説明	テキスト整理には有効	PLaMo 3.1 APIは画像＋Web検索併用不可	画像検索系を別システム化しPLaMoはメタデータ整理に限定。[15]
契約レビュー	ライセンス、NDA、共同研究、職務発明条項等の抽出・差分・リスク整理	長文契約処理に適する	法律意見を自動確定させるリスク	社内プレイブックをRAG化、逸脱条項のみ人へエスカレーション
著作権調査	類似性・依拠性に関連する資料整理、ライセンス条件比較	情報整理に有効	法的類似性判断をLLM単独で決定不可	原作品・権利情報・利用態様を専門家が確認
証拠収集・時系列整理	メール、契約書、設計資料、Git履歴等からイベント抽出	256K context、JSON出力が有用	AI要約が原資料と異なる可能性	原本ハッシュ、ファイルID、ページ、日時を全イベントへ付与

特に**先行技術・無効調査**では、Deep Researchに「答え」を求めるより、「次に検索すべき語・分類・出願人・引用ネットワーク」を反復生成させる方が安全である。特許庁もAI利用の先行技術検索高度化を進めて

おり、PLaMo側もBrowseComp-PlusやBRIGHTのような検索・情報収集系ベンチマークで能力を評価しているため、この使い方はモデルの設計思想にも合う。 21

侵害調査では、結論ではなくClaim Chart生成を主用途にするべきである。 推奨出力は「Claim element」「解釈候補」「対象製品の該当記載」「証拠URL/文書ID」「充足候補」「反対解釈」「不足情報」であり、PLaMoには「判断不能なら不明とする」「引用のない充足認定は禁止」と指示する。PLaMo 3.1 Primeは根拠不足時に回答不能と判断するデータで訓練されているため、この設計との親和性は高い。 18

特許明細書・クレームでは発明者性の記録を残す必要がある。 日本の特許庁は発明者を自然人に限ると明示しており、AIを発明者として記載することは認めていない。したがって、生成AIを発明創出過程で使う場合は、誰が技術的課題を設定し、誰が具体的構成を着想し、AI提案のどこを採否・修正したかを「発明者寄与ログ」として残す運用が望ましい。 22

データプライバシー、法的・倫理的リスク、証拠管理

クラウドPLaMoの情報管理は機能ごとに大きく異なる。

利用形態	学習利用	入出力保持	外部送信	知財機密への推奨度
Standard/Provider API、Web検索なし	なし	ZDR対象	原則外部送信なし	**高** 3
PLaMo Chat	なし	削除まで履歴保存	Webなしなら外部送信なし	中 3
Web検索	なし	現時点ZDR対象外	検索クエリ等を外部検索サービスへ送信	**機密案件は低** 3
Deep Research	なし	完了後24時間保持	検索クエリ等を外部検索サービスへ送信	**秘密情報を匿名化できる案件のみ** 3
Free	入出力を学習利用	保存	Web利用時は外部送信あり	**知財機密には不適** 3
オンプレ／閉域	PFN学習利用なし	自社設計次第	社内文書限定DRなら外部Web遮断可能	**最高** 16

個人情報についても注意が必要である。個人情報保護委員会は、生成AI提供者による学習利用等を伴う場合、個人データの第三者提供規制等が問題になり得ると注意喚起している。またPLaMo自身の現行仕様では、Web検索・Deep Researchへの個人情報入力は不可とされ、要配慮個人情報はプランを問わず入力できない。 23

著作権リスクは「学習」と「生成・利用」を分ける必要がある。 日本法では著作権法30条の4により、思想・感情の享受を目的としない情報解析等には一定の権利制限があるが、著作権者の利益を不当に害する場合等は除外される。文化庁も、特定作品の表現を出力させることを目的とした追加学習等では同条が適用されない場合があるとの整理を示している。 24

生成物については、AIだから一律に著作物になる／ならないわけではない。文化庁は、人の創作意図と創作的寄与があればAI利用者の著作物となり得る一方、単純な指示だけでAIが自律生成したものは著作物に該当しない可能性を示している。また既存作品との著作権侵害判断では、AI生成か否かにかかわらず「類似性」と「依拠性」が主要論点となる。 25

したがって、明細書、調査報告書、契約文案等をPLaMoで作成する場合には、**人間の編集履歴を残すこと**が知財管理上有益である。プロンプト、生成日時、生成物、修正差分、採用理由、最終承認者を案件IDに紐付けて保存すれば、著作物性だけでなく、発明者性、品質管理、監査の説明可能性にも寄与するというのが本報告の推奨である。文化庁が創作的寄与を個別具体的に判断するとしている点、特許庁が発明者を自然人に限る点から導かれる実務的措置である。 [26]

弁護士・弁理士業務との関係では、人間の専門家による最終判断を外さないことが重要である。 弁護士法72条は、非弁護士が報酬目的で法律事件に関する鑑定・代理等を業として扱うことを原則禁止している。弁理士法75条も、弁理士等でない者による、一定の特許庁手続の代理、鑑定、書類作成等を業として行うことを制限している。企業内部での補助ツール利用と、外部顧客にAI法律判断をサービスとして提供する場合はリスク構造が異なるため、後者は個別に弁護士・弁理士レビューを要する。 [27]

証拠能力については、「AIレポート」そのものより原データの保全が重要である。 日本の民事訴訟法には電磁的記録の証拠調べの規定があり、電磁的記録について書証関連規定が準用される。一方、生成AIが作った時系列表は原資料そのものではないため、AI出力だけを保存するのではなく、元メール、元PDF、取得Webページ、ハッシュ、取得日時、ファイルID、PLaMoへの入力、モデルバージョンを保全する方が証明力・再現性の観点で合理的である。 [28]

リスクを総合すると、次のように評価できる。

リスク	評価	主要対策
ハルシネーション	中〜高	引用必須、一次資料検証、回答不能を許容
検索漏れ	高	Patent DBによる専門検索を別工程化
機密漏洩	Web検索時は高	オンプレ／検索語匿名化／Webオフ
個人情報	高	検索機能への入力禁止ルール
著作権	中	原作品照合、人間編集、ライセンス確認
発明者性	中	人的寄与ログ
非弁・非弁理士	外部サービスでは中〜高	AIを補助者に限定、専門家承認
証拠改変・出所不明	高	原本＋hash＋timestamp＋モデルログ保存
プロンプトインジェクション	中〜高	Web/RAG文書を命令ではなくデータとして扱い、ツール権限を最小化
モデル更新による再現性低下	中	model ID、prompt、temperature相当設定、引用資料を保存

AIガバナンス全般については、経済産業省・IPA/AISIの最新版「AI事業者ガイドライン第1.2版」が2026年3月に公表されており、知財AIについてもリスクベースのガバナンス、モニタリング、利用者側統制をこの枠組みに合わせるのが妥当である。 [29]

導入コスト、実装ロードマップ、QA

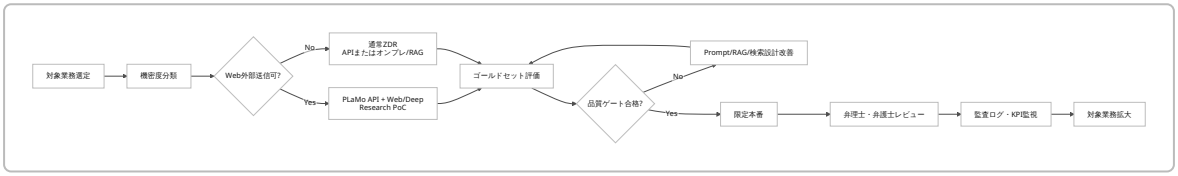
公開情報だけで確定できるPLaMo 3.1 Primeの費用は限定的である。Web検索・Deep Researchについては Standardで検索1回2円のキャンペーン価格が2027年1月31日まで適用され、通常Web回答では1〜3回、Deep

Researchでは15〜25回程度検索するのが目安とされる。したがって**Deep Researchの検索部分だけなら1件約30〜50円**であり、これにモデルの入出力トークン料金が加算される。 3

費用項目	クラウド/API	オンプレミス
初期導入費	APIは初期費用不要とPFNが案内。 30	**個別見積／未指定**
PLaMo 3.1 Prime トークン単価	本調査で取得できた3.1固有の確定表は**未指定**。公式サイトは「約百円で百万文字」規模の処理と案内。 31	ライセンス・サーバ費とも**個別見積**
Web検索	2円/検索（〜2027-01-31）、以降価格変更予定。 32	Web接続設計による
Deep Research	約15〜25検索＋token課金。 3	ライセンス・検索基盤費とも**未指定**
GPU台数	不要	**未指定**
VRAM	不要	**未指定**
GPU型番	不要	**未指定**
Docker	不要	PFN提供可能、価格**未指定**。 4
モデル＋サーバ	不要	提供あり、価格**個別見積**。 4
運用人員	API連携、IAM、評価担当	左記＋GPU/Kubernetes/MLOps担当
監査基盤	自社ログ基盤を推奨	自社SIEMと直接連携可能

したがって、**オンプレのTCOを公開情報だけからGPU台数×単価で計算するのは適切でない。** PLaMo 3.1 Primeのパラメータ数と推奨VRAMが公開されていないためである。PLaMo 3系列には120Bモデルが存在するが、それを3.1 Primeと同一モデルとみなす根拠はない。 7

実装は次のゲート方式を推奨する。PLaMoの通常APIがZDR対応、Deep Researchが24時間保持、Web検索が外部検索サービスへクエリを送るという差異を踏まえ、機密度判定を最初に置く。 3



PoCは6種類のゴールドセットを作るのが実務的である。 例えば各50〜100件程度の「特許要約」「請求項要素分解」「先行技術候補」「OA比較」「契約条項抽出」「侵害Claim Chart」を用意し、社内専門家の正解と比較する。件数は本報告の推奨値でありPFN仕様ではない。

評価指標は単なる「回答の良さ」ではなく、以下のように分解する。

KPI	測定方法	本番ゲート例
Citation correctness	引用文献が当該主張を実際に支持する割合	≥98%を目標
Citation completeness	検証可能な主張のうち引用付き割合	≥95%
Patent-ID accuracy	公開番号・登録番号等の正確率	100%を目標
Date accuracy	優先日・公開日・出願日の正確率	100%を目標
Claim-element recall	人手作成要素の抽出再現率	≥98%目標
Unsupported assertion rate	根拠なし断定率	<1%目標
Prior-art recall@k	人間が重要とした文献を上位k件に含む率	業務別に設定
Human acceptance	修正なし／軽微修正で採用できた率	段階的改善
Time saved	人手のみとの工数差	PoCで測定
Cost/task	API・検索・人レビュー費用総額	現行業務と比較

閾値はPFNが保証する値ではなく、知財部門向けの**推奨管理値**である。とりわけ公報番号や日付は「ほぼ正しい」では不十分であるため、自然言語評価と別に決定論的バリデーションを行うべきである。

社内体制は、知財責任者、弁理士／弁護士、検索担当、AI/ML担当、情報セキュリティ、IT基盤、個人情報・コンプライアンス担当からなる小規模な横断チームが望ましい。PLaMo APIはテナント／プロジェクト単位のロール管理、APIキー、利用上限・アラートを備えるため、案件別プロジェクトまたは部門別プロジェクトに分離するとコスト・権限管理を行いやすい。 17

競合モデルとの比較

下表は、2026年10月29日時点で確認できる主要モデルを、**知財実務で重要な「日本語・長文・検索・オンプレ・機密性」**の観点から比較したものである。性能はベンチマーク条件に大きく依存するため、単純なランキングとはしていない。

モデル	形態・規模	Context	Reasoning / Web	オンプレ・自社運用	知財用途での位置づけ
PLaMo 3.1 Prime	商用国産モデル。サイズ**未指定**	262K	推論/非推論、Web、Deep Research	**対応。Deep Researchもオンプレ可**	日本語＋機密調査＋国内運用の組合せが強み。 33
LLM-jp-4.1 33B Thinking	国産・約33B、公開モデル	詳細はモデルカード参照	Thinking	**可能**	自社改変・完全ローカル性を重視する組織向け。NIIが4.1シリーズを公開。 34

モデル	形態・規模	Context	Reasoning / Web	オンプレ・自社運用	知財用途での位置づけ
ELYZA-Thinking-1.0-llm-jp-4-33b	llm-jp-4をPost-trainingした約33B	基盤依存	Thinking	**公開モデルとして自社展開可**	日本語推論・カスタマイズ候補。ELYZAは商用利用可能として公開。[35]
gpt-oss-120b	117B総パラメータ、約5.1B active MoE	131K	Reasoning、open-weight	**可能。単一80GB GPU級で運用可能とOpenAIが案内**	ローカル運用と高度推論の比較基準。[36]
GPT-5.5	OpenAIクラウド旗艦	約1.05M	Reasoning、Web/File Search、各種Tools	モデル本体オンプレ不可	長文・高度ツール利用で強力。APIはinput $5/M、output $30/M。[37]
Claude Sonnet 5.5	Anthropicクラウド	モデル／提供面により長コンテキスト	Thinking、agent/tool利用	モデル本体オンプレ不可。AWS/Google/Microsoft経由あり	文書・エージェント型業務の比較対象。$2/M input、$10/M output。[38]
Gemini 3.1 Pro系	Googleクラウド	1M input、最大64K output	Thinking、Google Search、File Search、URL Context、Code Execution	モデル本体オンプレ不可	巨大文書セット・Google検索連携が強み。[39]

PLaMo 3.1 Primeは、PFN自身の35ベンチマーク比較では、安全性や非推論時の指示追従に強みを示した一方、STEM/CodeとWeb検索QAでは比較対象のgpt-oss-120b、Gemma 4 31B IT、Qwen3.8 27B等を下回り、長コンテキストRAGでも劣るケースがあった。この点は「国産だから最も高性能」と評価するのではなく、「国内・オンプレ・日本語・Deep Researchの総合条件」で選ぶべきことを示している。[40]

一方、知財部門では純粋なベンチマーク性能以外の差が大きい。GPT-5.5は1M超のコンテキストと多数の組込みツールを備え、Gemini 3系列も1M contextとGoogle Search/File Searchを備えるため、大規模公開資料分析には有力である。gpt-ossやLLM-jp/ELYZAは完全ローカル運用の自由度が高い。PLaMo 3.1 Primeの特徴は、その中間で、**商用サポート付き国産モデル、OpenAI互換性の高いAPI、Web検索、Deep Research、モデル本体／Docker／サーバ提供、閉域Deep Research**を一つの製品系列で選択できる点にある。[41]

したがって知財部門の選定基準は、「最高ベンチマークモデル1つ」を決めるより、**機密案件＝オンプレPLaMoまたはopen-weightモデル、公開情報Deep Research＝PLaMo/GPT/Claude/Geminiを社内ベンチマーク、超長文＝1M contextモデルも候補**というマルチモデル戦略の方が合理的である。これはPLaMo自身の比較結果にもモデルごとの得手不得手が存在することから導かれる運用上の推奨である。[40]

推奨プロンプト設計、評価方法、参考文献

知財プロンプトでは、一般的な「専門家として回答せよ」より、**事実・推論・法的判断を明示的に分離する**方が重要である。PLaMo 3.1 Primeが「根拠がなければ答えられない」とする忠実性学習を強化しているため、`不明`を正式な出力値として認める設計が特に適している。[18]

先行技術調査用テンプレート

```
目的：
以下の発明について、特許性検討のための先行技術候補を調査する。

基準日：
YYYY-MM-DD
基準日より後に公開された資料は「参考資料」とし、先行技術候補に含めない。

発明：
[発明説明]

手順：
1. 発明を必須構成要件A, B, C...へ分解する。
2. 各要件について同義語、日本語検索語、英語検索語、IPC/CPC候補を生成する。
3. 単一文献で全要件を満たす候補を優先探索する。
4. 次に複数文献の組合せ候補を探索する。
5. 各資料について公開番号、名称、公開日、優先日、出願人、
   該当箇所、対応する構成要件、URLを示す。
6. 原資料で確認できない情報は推測せず「未確認」とする。
7. 各主張には出典を付す。
8. 最後に「追加検索すべき検索語・分類・出願人」を提示する。

禁止：
・存在を確認できない特許番号を生成しない。
・公開日と優先日を混同しない。
・新規性・進歩性を最終的に断定しない。
```

評価は、重要文献Recall@10/20、存在しない公報番号率、日付正確率、claim-element coverage、引用支持率を計測する。Deep Researchは反復検索に適するが、専門特許DBを直接検索するものとは限らないため、最終先行技術調査を一般Webだけで完結させない。PLaMoがWeb検索レポート能力を強化した一方、Web検索QAでは競合比の改善余地を自ら認めている点もこの運用を支持する。[19]

特許要約テンプレート

```
以下の特許文献を、記載内容だけに基づき分析する。
外部知識で補完してはならない。

出力：
1. 書誌事項
2. 解決しようとする課題
3. 独立請求項ごとの必須構成
4. 各構成に対応する明細書段落番号
```

```
5．発明の作用・効果
6．従来技術との差異
7．用語定義
8．実施形態の限定事項
9．広く解釈できる点
10．不明点・記載不足

各項目には請求項番号または段落番号を付す。
根拠が無い場合は「記載なし」とする。
```

評価は「原文忠実性」「請求項要素の漏れ」「段落citation正確率」「原文にない因果関係を付加した割合」で行う。PLaMo 3.1 PrimeはRAG忠実性と回答不能データを重点的に学習しているため、この種の文書拘束型タスクが有力なPoC候補となる。 [18]

侵害分析／FTO補助テンプレート

```
役割：
法的結論を確定する者ではなく、Claim Chart作成補助者として分析する。

入力：
[対象請求項]
[対象製品資料]
[裁判所・審判等で既に確定している解釈があればその資料]

各claim elementについて以下を出力：
- Element ID
- 請求項文言
- 解釈候補
- 対象製品に対応する記載
- 引用元文書・ページ
- 状態：充足候補 / 非充足候補 / 不明
- 充足を支持する理由
- 非充足を支持する理由
- 追加で必要な証拠

ルール：
証拠がない場合は必ず「不明」とする。
均等論・間接侵害等の法的結論を自動確定しない。
引用箇所以外の情報を対象製品の事実として扱わない。
```

評価では、各elementのEvidence Precision、専門家とのclassification agreement、unsupported fact rate、反対論点の抽出率を測る。「充足」と断定させるより両論併記させることで、LLMのもっともらしい断定を抑える運用が適している。PLaMo 3.1 Primeが事実性・忠実性改善を明示している一方、ハルシネーションを完全には排除していないためである。 [18]

最終的な推奨アーキテクチャは、**PLaMoを直接ユーザーに開放するのではなく、「案件DMS／特許DB → 検索・RAG → PLaMo → Citation Validator → Rule Engine → 専門家承認 → 監査ログ」のパイプラインに組み込む形**である。通常の機密文書処理にはZDR APIまたはオンプレ、公開情報調査だけをWeb検索／Deep Researchへ送る「二経路構成」が、安全性と利便性のバランスがよい。PLaMoがモデル本体・Docker・サー

バ・クラウドAPIを横断して提供され、Deep Researchもオンプレ利用可能であることが、この構成を現実的にしている。 42

主要参考文献・一次資料。 本調査では、PFN「PLaMo 3.1 Primeをリリースしました」（2026年10月27日）を技術仕様・学習法・評価の中心資料とした。 43 PFN「国産生成AI基盤モデルPLaMo PrimeにWeb検索・Deep Research搭載」（同日）を提供形態・オンプレDeep Researchの一次資料とした。 6 PLaMo公式API Referenceおよび料金・データ取扱い資料を、API、262K context、Responses API、ZDR、検索時外部送信、24時間保持等の根拠とした。 44 PLaMo 3系列の基礎技術についてはPFN/NICTの120B・31Bモデル研究資料を参照した。 7

法制度については、文化庁「AIと著作権について」および「AIと著作権に関する考え方について」を著作権分析の中心資料とした。 45 個人情報は個人情報保護委員会の生成AI利用注意喚起を参照した。 46 特許・発明者性については特許庁の「発明者等の表示について」およびAI関連技術に関する調査研究・審査事例を参照した。 47 弁護士・弁理士業務については弁護士法72条、弁理士法75条、電子証拠については民事訴訟法231条の2・231条の3等を参照した。 48

総括すると、**PLaMo 3.1 Primeは、知財部門において「日本語生成モデル」としてよりも、「国内運用可能な調査・文書分析エージェント基盤」として評価すべきモデルである。** 特にオンプレミスDeep Research、社内文書との閉域連携、Web検索と引用、256K級コンテキスト、推論／非推論切替えは、企業知財との適合性が高い。一方、モデルサイズ・オンプレGPU要件・本体価格等は**未指定**であり、検索漏れ・引用誤り・法的判断のハルシネーションも残存リスクである。したがって、本番導入の成立条件は「PLaMoそのものの性能」より、①情報分類、②一次資料のCitation Validation、③専門家承認、④案件単位監査ログ、⑤知財専用ゴールドセットによる継続評価、の五つを周辺システムとして実装できるかにある。 49

---

1 6 国産生成AI基盤モデルPLaMo PrimeにWeb検索・Deep Research搭載 調査計画の立案から出典付きレポートまで自律的に実行、オンプレミスにも対応 - 株式会社Preferred Networks
https://www.preferred.jp/ja/news/pr20261027

2 10 11 18 19 40 49 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

3 料金とデータの取り扱い | リファレンス
https://docs.plamo.preferredai.jp/ja/pricing

4 16 30 41 42 PLaMo
https://plamo.preferredai.jp/?utm_source=chatgpt.com

5 9 13 33 44 API リファレンス | リファレンス
https://docs.plamo.preferredai.jp/ja/api?utm_source=chatgpt.com

7 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事前学習モデル、31B蒸留モデルの評価｜株式会社Preferred Networks
https://www.preferred.jp/ja/blog/tech/plamo_3_120b_distillation_cpt

8 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8B, 31Bの小規模モデルによる事前学習の検証｜株式会社Preferred Networks
https://www.preferred.jp/ja/blog/tech/plamo_3_8b_31b?utm_source=chatgpt.com

12 生成AI基盤モデルPLaMo 3.0 Primeβ版のモニター企業募集 - 株式会社Preferred Networks
https://www.preferred.jp/ja/news/pr20260319?utm_source=chatgpt.com

14 API Reference | Reference
https://docs.plamo.preferredai.jp/en/api

15 API Reference | Reference
https://docs.plamo.preferredai.jp/en/api?utm_source=chatgpt.com

17 PLaMo APIコンソール マニュアル | リファレンス
https://docs.plamo.preferredai.jp/ja/console?utm_source=chatgpt.com

20 21 特許庁における人工知能（AI）技術の活用に向けたアクション・プランの令和7年度改定版について | 経済産業省 特許庁
https://www.jpo.go.jp/system/laws/sesaku/ai_action_plan/ai_action_plan-fy2025.html?utm_source=chatgpt.com

22 47 発明者等の表示について | 経済産業省 特許庁
https://www.jpo.go.jp/system/process/shutugan/hatsumei.html?utm_source=chatgpt.com

23 46 生成AIサービスの利用に関する注意喚起等について ｜個人情報保護委員会
https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/?utm_source=chatgpt.com

24 著作権法 | e-Gov 法令検索
https://laws.e-gov.go.jp/law/345AC0000000048/20201001_502AC0000000048?utm_source=chatgpt.com

25 26 AI生成物の場合の「創作的寄与」
https://www.bunka.go.jp/seisaku/chosakuken/pdf/94097701_02.pdf?utm_source=chatgpt.com

27 48 弁理士法 | e-Gov 法令検索
https://laws.e-gov.go.jp/law/412AC0000000049/20260521_504AC0000000048?utm_source=chatgpt.com

28 民事訴訟法 | e-Gov 法令検索
https://laws.e-gov.go.jp/law/408AC0000000109/20260521_504AC0000000048?utm_source=chatgpt.com

29 AI事業者ガイドライン（METI/経済産業省）
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20260331_report.html?utm_source=chatgpt.com

31 PLaMo
https://plamo.preferredai.jp/en/?utm_source=chatgpt.com

32 料金とデータの取り扱い | リファレンス
https://docs.plamo.preferredai.jp/ja/pricing?utm_source=chatgpt.com

34 大規模言語モデル「LLM-jp-4.1」シリーズを公開 ｜ 国立情報学研究所 大規模言語モデル研究開発センター
https://llmc.nii.ac.jp/topics/llm-jp-4-1/?utm_source=chatgpt.com

35 ELYZA、「AIによるAI開発」の研究を推進する「ELYZA RSI Research」を設立。基盤モデル・評価・自律改善に関する研究成果を公開　～研究成果を活用した「バーティカルAI事業」も開始、第一弾として「ELYZA Voice Agent」を開発～ | 株式会社ELYZA | 株式会社ELYZA
https://elyza.ai/news/2026/10/02/rsiresearch?utm_source=chatgpt.com

36 gpt-oss-120b Model | OpenAI API
https://developers.openai.com/api/docs/models/gpt-oss-120b?trk=article-ssr-frontend-pulse_little-text-block&utm_source=chatgpt.com

37 GPT-5.5 Model | OpenAI API
https://developers.openai.com/api/docs/models/gpt-5.5?utm_source=chatgpt.com

38 Introducing Claude Sonnet 5.5 \ Anthropic
https://www.anthropic.com/claude-sonnet-5-5?utm_source=chatgpt.com

39 Gemini 3 developer guide ｜ Gemini API ｜ Google AI for Developers
https://ai.google.dev/gemini-api/docs/gemini-3?thinking=explicit&utm_source=chatgpt.com

43 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?utm_source=chatgpt.com

45 AIと著作権について | 文化庁
https://www.bunka.go.jp/seisaku/chosakuken/aiandcopyright.html?afsrc=1&utm_source=chatgpt.com