国産生成AI「Sarashinaサービス」無償ベータ版の徹底解剖と知財実務への戦略的実装評価

Gemini 4 Argon

2026年10月13日、ソフトバンク株式会社の子会社であるSB Intuitions株式会社（本社：東京都港区、代表取締役社長 兼 CEO：井尻 善久）は、同社がフルスクラッチで研究開発を進める国産大規模言語モデル（LLM）「Sarashina（さらしな）」を無料で体験・検証できる「Sarashinaサービス」無償ベータ版の一般公開を開始した[1]。本発表は、単なる対話型AIチャットの無料開放にとどまらず、文章生成、高速軽量処理、テキスト埋め込み（Embedding）、指示付きリランキング（Rerank）、安全性判定（Guard）、および日英・英日翻訳（Translate）という実務システム構築に不可欠な6種類の特化型AIモデルをAPI経由で一挙に無償提供した点に最大の技術的・戦略的意義がある[1]。

知的財産（知財）業務の観点から本サービスを評価すると、日本語の専門文献に対する高い文脈理解力やトークン圧縮効率、さらにはRAG（検索拡張生成）構築用コンポーネントの充実度において極めて高いポテンシャルを有している[2]。一方で、無償ベータ版特有の利用規約（入力データのモデル学習への利用およびオプトアウト不可、法律等の専門的助言への利用禁止、1分あたりのリクエスト回数制限など）は、特許法上の「新規性喪失」や弁理士法・秘密保持契約（NDA）上の「守秘義務違反」と直結する重大な法的境界線を形成している[2]。

本レポートでは、公開されたプレスリリース、公式技術ブログ（Tech Blog）、API仕様書、および利用規約（2026年9月28日制定）の条文を徹底的に解剖した上で、知財部門および特許事務所がいかなる業務プロセスにおいて、どのようなアーキテクチャとガバナンスの下で「Sarashinaサービス」を活用すべきかを体系的に論じる[1]。

1. 「Sarashinaサービス」無償ベータ版の全貌と提供背景

1.1 プレスリリースの核心とソブリンAIエコシステムにおける位置づけ

SB Intuitionsが提供する「Sarashina」という名称は、同社の本社所在地である東京都港区「竹芝」の地名が歴史上初めて登場する平安時代中期の回想録『更級日記』に由来し、日本の地から世界で活用されるAIモデルを発信するという理念が込められている[4]。同社は国内最大規模の計算基盤と国内データセンターにおける厳格なデータ管理を強みとし、国産LLMとして日本語性能No.1を目指して開発を推進してきた[1]。

今回の「Sarashinaサービス」無償ベータ版公開の背景には、日本企業における生成AI活用のフェーズが「単発の実証実験（PoC）」から「自社業務システムや商用サービスへのAPI組み込み（本実装）」へと急速に移行している市場環境がある[1]。実務システムへの統合においては、汎用的なチャット応答だけでなく、機密・有害表現のフィルタリング、社内文書や専門データベースを対象とした高精度な検索（RAG）、検索結果の再順位付け、自然な多言語翻訳といった多様なモジュール群をパイプラインとして組み合わせる必要がある[1]。

時系列で俯瞰すると、SB Intuitionsは2025年11月に法人向け「Sarashina API」の提供を開始し[2]、2026年6月30日にはソフトバンクのソブリンクラウド基盤「Cloud PF Type A」（Oracle Alloy採用）上

にて最新世代の「Sarashina3 シリーズ」を有償提供開始していた[1]。また、2026年2月には国立研究開発法人情報通信研究機構（NICT）と高性能LLMの安全性技術・ガードレール技術に関する共同研究を開始し[11]、同年8月には順天堂大学等と「3省2ガイドライン」準拠の医療データ基盤上で国産医療特化型LLM「Med-Sarashina」の共同研究を開始したほか[15]、デジタル庁が推進する「ガバメントAI」の試用モデルとしても採択されるなど、高度な安全性と信頼性が求められる領域での実績を固めてきた[6]。今回の無償ベータ版は、これまで法人契約や「Cloud PF Type A」環境が前提であった「Sarashina3 シリーズ」の最新モデル群を、事前申し込み不要・Googleアカウントによるソーシャルログインのみで個人開発者や企業の企画・知財・DX担当者へ開放し、実装検証のハードルを劇的に引き下げると同時に、実利用フィードバックを収集して製品開発へ還元する戦略的施策である[1]。

1.2 提供される機能体系：ブラウザチャットと6種の特化型APIモデル

無償ベータ版（サービスサイト：https://sarashina.ai）は日本国内に居住する個人、および日本国内に本店を有する法人を対象としており（18歳未満は親権者等の同意が必要）、大きく「ブラウザ上のチャット機能」と「6つのAIモデルを呼び出せるAPI機能」の2軸で構成される[1]。

ブラウザ向けのチャット機能（chat.sarashina.ai）は、プログラミング知識を持たない担当者でも即座に利用できるインターフェースであり、「対話モード」（ウェブ検索を併用した調査や文章作成・要約）と「翻訳モード」（日英・英日の双方向翻訳）を切り替えて実行できる[1]。会話履歴の一覧表示機能を備える一方、バックエンドで稼働している具体的なモデル種別や1日あたりの利用回数上限は非公表であり、専用スマートフォンアプリは提供されずウェブブラウザからの利用に限定される[2]。

一方、開発者や知財DX推進者が自社ツールやスクリプトから呼び出すAPI機能（エンドポイント：https://api.sarashina.ai/v1、ドキュメント：docs.sarashina.ai）では、役割の異なる6つのモデルが提供されている[2]。各モデルの技術仕様、コンテキスト長、レート制限（RPM：1分あたりのリクエスト上限）、および性能指標は以下の通りである[2]。

モデル名（**APIモデルID**）	機能カテゴリ	最大入力長 / 出力上限	ペイロード上限	無償版レート制限 **(RPM)**	主な技術的特徴・ベンチマーク性能
Sarashina3 mini (sarashina3-mini)	標準文章生成・対話・エージェント	入力: 128kトークン 出力: 8,000トークン	4 MiB	**6**回 / 分	30Tトークン超の事前学習＋蒸留・多段階強化学習・OPSD[6]。日本文化・法律等を含むJamC-QAで**0.82**を記録し、GPT-4.1やGPT-5.4 miniを上回

					る[2]。ツール呼び出し・JSON構造化出力に対応[2]。
Sarashina3 nano (sarashina3-nano)	高速・軽量文章生成	入力: 128kトークン 出力: 8,000トークン	4 MiB	**60**回 / 分	スループット重視の軽量モデル[6]。大量テキストの分類・情報抽出・感情スコア推定等の定型バッチ処理に最適[6]。JamC-QAスコアは**0.61**で、小型OSS（Qwen3-4B）を凌駕[2]。
Sarashina3 embedding (sarashina3-embedding)	テキスト埋め込み（ベクトル化）	入力: 8,000トークン	4 MiB	**60**回 / 分	7種類のタスクタイプ切替に対応[5]。DARE-TIESによるモデルマージとMRL（マトリョーシカ表現学習）を採用[5]。256次元への圧縮時でもJMTEB平均**75.72**を維持し高効率な検索を実現[5]。

Sarashina3 rerank (sarashina3-rerank)	検索結果の再順位付け（リランク）	入力: 8,000 トークン （設計上限 8,192）	4 MiB	**60**回 / 分	ModernBERT-Ja-1.4B ベース[4]。1クエリに対し最大20件の候補文書を精読し関連度スコア順に再配列[4]。5種のタスクタイプに加え、独自の評価軸を自然言語で指定できる **CUSTOMIZE指示機能**を搭載[4]。
Sarashina3 guard (sarashina3-guard)	入出力テキストの安全性・有害性判定	入力: 64k トークン （文章生成なし）	4 MiB	**6**回 / 分	テキストの「安全」「有害」を高速判定し、情報漏えい・違法行為のほう助・過激描写などの有害カテゴリコードを返却[13]。NICTとの共同研究等に基づく日本語特化ガードレール[1]。
Sarashina3 translate (sarashina3-translate)	日英・英日特化型機械翻訳	入力: 5,000 文字	1 MiB	**6**回 / 分	合成データ・独自訳質評価モデルによる自己蒸留・SFT+DPO

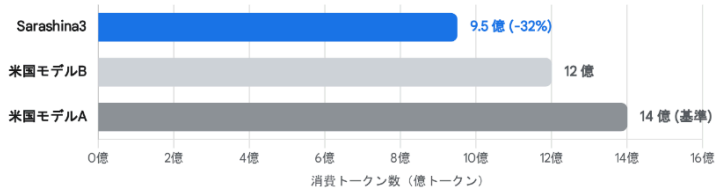
					で学習[1]。無生物主語の解消や文脈に即した流暢な意訳に強み[1]。WMT2024自動評価（英日・日英）および人手評価（英日）でGoogle翻訳に勝ち越し[1]。

国産生成AI「Sarashina3」技術仕様 & 業務適用評価

Sarashina3 Achieves 30% Higher Japanese Token Efficiency While Enforcing Strict Tiered API Rate Limits in Beta

Panel 1: 日本語トークナイズ消費量（全Wikipedia 約19億文字 処理時）

※トークン数が少ないほど高圧縮・長文処理コスト削減



例文検証 「山手線」等を含む同一日本語文のトークン消費: Sarashina3: 40 tokens vs 米国モデルA: 80 tokens (-50% 圧縮)

Panel 2: API 6モデルの性能特性・入力上限・レート制限 (RPM) 構造

利用規約: 入力データの学習利用同意必須 / 60秒毎リセット

高速・大量バッチ / RAG層 — 60 RPM

sarashina3-nano — 128k tokens
高速軽量生成・抽出 (JamC-QA: **0.61**)
適合: 特許公報の大量スクリーニング・要旨要約

sarashina3-embedding — 8,000 tokens
テキスト埋め込み (7タスク対応・MRL 256d: **75.72**)
適合: 社内特許DBの高速ベクトル化・第1段階検索

sarashina3-rerank — 8,000 tokens
指示付き再順位付け (最大20文書・5タスク対応)
適合: 検索上位特許の請求項適合度リランキング

緻密生成・翻訳・監査層 — 6 RPM（制限注意）

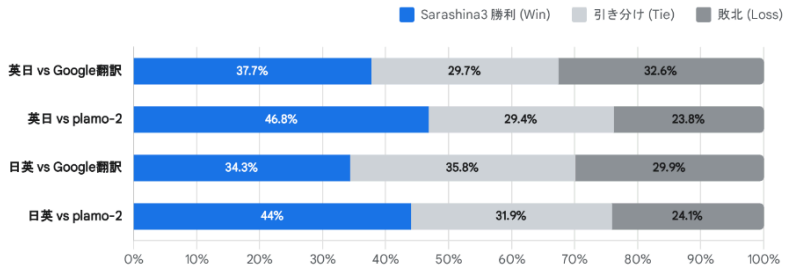
sarashina3-mini — 128k tokens
標準文章生成・ツール呼出 (JamC-QA: **0.82**)
留意: 100件処理に約17分所要。緻密な拒絶理由分析向け

sarashina3-translate — 5,000 文字
日英・英日特化翻訳 (SFT + DPO 2段階調整)
適合: 外内・内外特許アブストラクトの流暢な対訳化

sarashina3-guard — 64k tokens
安全性・情報漏洩コード判定 (生成なし)
留意: 入出力ガードレール。有償版移行前の安全性検査

Panel 3: WMT2024 機械翻訳ペアワイズ自動評価（GPT-5-mini 判定 / 997文）

提示順入れ替えによるバイアス排除評価



Top: Token count required to process the entire Japanese Wikipedia corpus (~1.9 billion characters), demonstrating Sarashina3's compact tokenization compared to major US models. Middle: Rate limits (Requests Per Minute) and input capacities across the 6 beta API models, highlighting the 60 RPM high-throughput tier for RAG and bulk classification versus the 6 RPM limit on generation and translation. Bottom: Pairwise evaluation results on WMT2024 (997 sentences, judged by GPT-5-mini with order-swapping) showing **Sarashina3 translate** outperforming both Google Translate and plamo-2-translate in win rates.

Data sources: AI-Revolution Media (Sarashina無料β版解説), SB Intuitions プレスリリース (2026-10-13), SB Intuitions TECH BLOG (翻訳モデル詳細), SB Intuitions TECH BLOG (Sarashina3 LLM評価)

1.3 技術的優位性：30兆トークン学習・日本語トークナイザ効率・OpenAI互換差分

知財システムへの実装を見据えた際、特筆すべき技術的特徴は3点存在する。第一に、基盤となるsarashina3-miniおよびsarashina3-nanoは、既存OSSモデルの継続事前学習（ファインチューニング）ではなく、新設計の学習パイプラインによって日本語・英語を中心に約30兆（30T）トークン以上の高品質データからフルスクラッチで事前学習されている[6]。事後学習においても、最新の蒸留技術、多段階の強化学習、およびモデル自身の生成結果を基に自己改善を図るオンポリシー自己蒸留（OPSD：On-Policy Self-Distillation）を組み合わせ、フィードバックループを複数回回すことで、日本の法律・歴史・文化・商習慣に関する選択式ベンチマーク「JamC-QA」においてsarashina3-miniが0.82という高スコアを達成し、GPT-4.1やGPT-5.4 mini (non-reasoning)、Qwen3-235B-A22B-Instruct-2507を上回る日本語知識性能を実証している[2]。なお、sarashina3-miniは長い思考過程を経る推論（Reasoning）特化モデルではなく標準的なInstructモデルとして設計されており、数学（AIME 2025）、Pythonコード生成（MBPP+）、指示追従（IFBench）、関数呼び出し（BFCL v3）、通信ドメインのマルチステップ対話（τ2-bench Telecom）、日本語対話（ELYZA-tasks-100）、および日本語出力安全性（Answer Carefully）の8ベンチマーク全般で前世代（Sarashina2 mini）から飛躍的な向上を果たし、ビジネス文書に求められる過剰な装飾を排した簡潔で正確な出力スタイルへとアライメントされている[2]。

第二に、日本語向けに専用設計されたトークナイザによる圧倒的な「トークン効率」である[2]。海外製LLMでは日本語の複合語や固有名詞が細かく分割されトークン数が膨張しやすいのに対し、Sarashina3のトークナイザは「山手線」のような固有名詞や複合語をひとまとまりの単位として処理できる[6]。その結果、日本語Wikipedia全体（約19億文字）をトークン化した場合、米国モデルAが14億トークン、米国モデルBが12億トークンを要するのに対し、Sarashina3は9.5億トークン（GPT-5等と比較して約3割少ないトークン数）で表現可能であり、特定の日本語例文では米国モデルAの80トークンに対して半分の40トークンで処理できる[2]。数万文字に及ぶ特許明細書や審決文を取り扱う知財業務において、トークン消費量が約3割削減されることは、同一の128kコンテキストウィンドウ内に実質約1.4倍の文字数の特許文献を同時投入できることを意味し、将来的な有償従量課金環境への移行時にも推論速度の向上とAPIコストの大幅削減に直結する[2]。

第三に、システム実装上の注意点として、Sarashina APIはOpenAIのChat Completions APIに類似した設計を採用しているものの、いくつかの重要な仕様上の差分が存在する[2]。知財部門のエンジニアが既存のOpenAI用スクリプトを流用する際は、出力トークン上限パラメータとしてmax_tokensではなくmax_completion_tokens（初期値・最大値ともに8,000トークン）を指定しなければならず、公式ドキュメントに記載のないパラメータ（logprobsやstream_optionsなど）を送信すると無視されずにHTTP 400エラーが返却される仕様を吸収するラッパー設計が必須となる[2]。また、ツール呼び出し（Function Calling）においてtoolsパラメータを指定した際、tool_choiceを省略するとOpenAIのauto（モデルが呼び出し要否を自律判断）とは異なりrequired（必ずいずれかのツールを呼び出す）として動作するため、自律判断を行わせるには明示的にtool_choice: "auto"を指定せねばならず、特定関数を一本釣りするオブジェクト指定には非対応（none / required / autoのみ対応）である点にも留意が必要である[2]。一方で、構造化出力（Structured Outputs）に関してはresponse_formatでjson_objectおよびstrict指定付きのjson_schema、さらにはstream: trueによるSSE形式のストリーミングに完全対応しているため、特許公報からの構成要件抽出や書誌事項のデータベース格納において型安全なJSON出力を強制できる[2]。

API仕様項目	**OpenAI Chat Completions API** の一般的挙動	**Sarashina API**（無償ベータ版）の仕様・挙動	知財システム実装時の対応策
出力トークン上限指定	max_tokens または max_completion_tokens	max_completion_tokens のみ対応（初期値・上限ともに **8,000**トークン）[2]	既存コードの max_tokens 指定を置換し、長文生成時は8,000トークン以内に収まるようプロンプトまたは分割出力を設計する[2]。
未対応パラメータ送信時	一部パラメータは無視される場合あり	ドキュメント記載外のパラメータ（logprobs, stream_options等）送信で **400**エラー を返却[2]	LangChainやLlamaIndex等のフレームワーク利用時は、暗黙的に付与される未対応パラメータを除外するアダプタを介在させる[2]。
tool_choice 省略時の挙動	デフォルトで "auto" として動作	tools 指定時に tool_choice を省略すると **"required"**（必須呼び出し）扱い。関数名指定オブジェクトは非対応[2]	エージェント実装時は必ず tool_choice: "auto" を明示的に宣言し、none / required / auto の3値で制御する[2]。
構造化出力・ストリーミング	json_object, json_schema (strict), stream: true	json_object および json_schema（strict 指定あり）、SSE形式の stream: true に対応[2]	公報からのクレーム構成要件抽出や課題・解決手段のマトリクス化に json_schema (strict: true) を積極活用する[2]。
APIキー管理・通信制約	サーバーサイド管理推奨	ブラウザやスマホアプリ等のクライアント側コードへのAPIキー記述は規約で禁止（.envを.gitignoreへ登録必	必ずバックエンドサーバー経由でAPIを呼び出し、上限判定単位である「契約者×モデル」ごとの60秒間リクエスト数をス

		須）[2]	ロットリング制御する[2]。

2. 利用規約・ガバナンス制約の徹底検証と知財法務リスク評価

知財業務において生成AIを評価する際、モデルの言語性能以上に決定的な意味を持つのが「利用規約（データガバナンス）」である[7]。2026年9月28日に制定された「Sarashinaサービス」無償ベータ版の利用規約を精査すると、知財実務家が絶対に踏み越えてはならない法的リスクの境界線が浮かび上がる[2]。

2.1 データ学習利用（オプトアウト不可）と特許法上の新規性喪失・守秘義務リスク

無償ベータ版の利用規約第11条第2項および第9条第3項により、ユーザーが入力したプロンプトおよびAIが出力した回答データは、サービスの提供・維持のみならず、「AIモデルの学習および新サービスの開発」のためにSB Intuitionsが無償かつ無制限に利用（複製・改変・解析・学習等）できる権利を許諾する構造となっている[2]。さらに決定的な点として、第11条第3項により「学習への利用を拒否（オプトアウト）できるのは今後提供予定の有償プランのみであり、無償ベータ版では一切オプトアウトできない」と明記されている[2]。加えて、一度学習に利用されたデータは後から有償プランへ移行しても遡って削除されず、学習後のモデルから当該データを削除する義務も提供元は負わない（第9条第4項）[2]。また、特記事項第2条により提供元は会話データの保存・削除義務を負わず、1年以上未使用のアカウントは事前通知なく削除される場合があるため、履歴管理は利用者自身のバックアップに委ねられている[2]。

この規約構造を知財法務の観点に照らすと、特許法第29条第1項における「新規性喪失」および不正競争防止法上の「営業秘密の喪失」という第一の重大な法的帰結が導かれる[7]。出願前の発明提案書、未公開の実験データ、あるいは作成途中の特許明細書ドラフトを無償ベータ版に入力した場合、その技術情報は提供元の学習パイプラインに取り込まれ、外部からアクセス可能な状態または将来的に他者への出力として再現されるリスクを生じるため、公然知られた発明（新規性喪失）と認定されるおそれ、ならびに営業秘密としての「秘密管理性」を喪失するおそれがある[7]。

第二の法的帰結は、弁理士法第30条・第77条が定める「守秘義務」および企業間の秘密保持契約（NDA）に対する違反リスクである[8]。日本弁理士会が2025年4月に公表した「弁理士業務AI利活用ガイドライン」（ならびに欧州特許弁理士会epiの2024年ガイドライン等）では、AIツール選定の安全基準として「通信の暗号化」と「入力データがAIモデルの再学習に利用されないこと（およびその規約上の明示的保証）」を中核要件として掲げている[7]。再学習への利用が規約上明記されオプトアウトが認められない無償ベータ版に対し、クライアントから預かった未公開の発明情報や秘密の契約書案を入力する行為は、クライアントの同意の有無にかかわらず技術的安全性基準を欠くものであり、弁理士の法定守秘義務およびNDAに対する重大なコンプライアンス違反を構成する[7]。

2.2 専門的助言の禁止条項（第10条）・免責規定と日本弁理士会ガイドラインとの整合性

さらに見落とされがちな重要条項として、利用規約第10条における「医学・法律・金融などの専門的な助言としてAIの回答を使用することの禁止」が挙げられる[2]。知的財産業務は特許法、実用新案

法、意匠法、商標法、著作権法、不正競争防止法に立脚する高度な法務専門領域である[18]。したがって、無償ベータ版の出力をそのまま「特許有効性鑑定」「侵害成否鑑定（FTO意見書）」「契約書上の法的助言」として利用・提供することは利用規約第10条に直接抵触する[2]。
また、入出力の知的財産権自体は利用者に帰属する（第9条第2項）ものの、第25条により出力の正確性や第三者権利非侵害の保証は一切なされておらず、第24条により無料プラン利用者に損害が発生してもSB Intuitionsは全面的に免責される[2]。さらに第10条では、個人情報（氏名・住所・連絡先等）や他人の著作物（書籍の全文など）の入力禁止、出力を用いた競合サービスの開発禁止、出力を自身の独自作成物やSB Intuitionsの公式見解であるかのように装う行為の禁止、および事前問い合わせなしにプログラム等で出力データを自動的に大量抽出する行為の禁止が定められている[2]。日本弁理士会のガイドラインにおいても、AIが実在しない公報番号や判例を捏造するハルシネーションへの対策として一次情報による裏取りを義務付け、かつAI生成物が他者の著作権（複製権等）を侵害していないかの確認責任はすべて弁理士・利用者側にあると強調されており、無償ベータ版の活用においてはこれらの規約・倫理基準を厳格に遵守した運用設計が求められる[7]。

利用規約の該当条項・仕様制限	規定内容の要旨	知財業務における具体的リスク	必須となる運用統制・回避策
第11条2項・3項 / 第9条3項・4項 （データの学習利用・オプトアウト不可）	入出力データをモデル学習等に無償・無制限利用[2]。無料版は拒否不可で、学習後のモデルからの削除義務なし[2]。	未出願発明の新規性喪失（特許法29条）、営業秘密の秘密管理性喪失、弁理士守秘義務（弁理士法30条）・NDA違反[7]。	無償ベータ版への入力は**「発行済みの公開特許公報・審決公報」「公開Web情報」「架空のダミー発明データ」に厳格限定**する[2]。
第10条（禁止事項：専門的助言）	医学・法律・金融などの専門的な助言として回答を使用することを禁止[2]。	侵害予防調査（FTO）の法的結論や無効鑑定書として出力を用いると規約違反および善管注意義務違反となる[2]。	AIの役割を「公開文献の一次スクリーニング・要約・技術要素の構造化（素材生成）」に限定し、法的評価と最終判断は必ず弁理士・知財担当者が行う[8]。
第10条（禁止事項：個人情報・他者著作物）	個人情報（氏名・連絡先等）や他人の著作物（書籍全文等）の入力禁止[2]。	公報に記載された発明者氏名・住所や、有料学術論文・市販技術書籍の全文入力による規約違反・著作権侵害（複製権侵害）[2]。	公報テキスト投入時に書誌事項の発明者氏名・住所フィールドを前処理スクリプトでマスキング（除外）し、商業論文の無断投入を禁止する

			[2]。
第**10**条（競合開発・自動大量抽出の制限）	出力を用いた競合サービス開発禁止、プログラムによる自動大量抽出は事前問合せが必要[2]。	自社独自の知財特化型LLMの蒸留データ作成や、無断での大量スクレイピング的利用が規約違反となる[2]。	レート制限（6 RPM / 60 RPM）の範囲内で適正にAPI検証を行い、複数アカウント作成による制限回避（規約禁止）を厳禁とする[2]。
第**16**条（再許諾の事前同意） 第**24**条・**25**条（免責・非保証）	自社サービスの一部として第三者に提供（再許諾）するにはSB Intuitionsの事前同意が必要[2]。損害賠償免責[2]。	特許事務所や知財ベンダーが、無償APIを裏側で組み込んだクライアント向けWebツールを無断提供すると規約違反となる[2]。	無償ベータ版は「所内・社内の技術検証（PoC）」に留め、クライアント提供や機密処理の本番運用は有償の「Cloud PF Type A」等へ移行する[2]。

3. 知財業務領域別の徹底評価：どの業務にどう使えるか

上記の法務・規約上の制約と、6つのモデルの技術的特性（トークン効率、コンテキスト長、RPM制限、ベンチマーク特性）を交差分析すると、「Sarashinaサービス」無償ベータ版が真価を発揮する知財業務と、有償ソブリン環境（Cloud PF Type A）へ移行して初めて解禁される業務が明確に区分される[2]。

Public Data Workflows Thrive in the **Sarashina Free Beta**, While Pre-Filing Tasks Require Sovereign Cloud Migration

知財実務6大領域における適合性・実装マトリクス

SB Intuitions 国産LLM「Sarashina3」業務適用評価

✓ **公開情報の活用領域（無料β推奨）**：
特許公開公報・登録公報・公開IR・学術論文などは機密漏洩リスクがなく、無償枠（特に60 RPMモデル）で即時PoC・検索基盤構築が可能。

! **出願前・法的鑑定の制約（ソブリン移行必須）**：
未公開発明の入力は新規性喪失リスク及び規約11条（学習利用）に直結。また規約10条（法律・専門助言禁止）より侵害鑑定・意見書作成はCloud PF Type Aへ移行。

適性判定凡例： ◎ 即時活用 / 本番運用　○ 条件付き有効 / PoC　△ 補助利用（厳格校閲要）　× 利用不可・規約厳禁

知財業務領域 (IP Workflow)	推奨モデル構成 & RPM	無償ベータ版	Cloud PF Type A	具体的実装メソッド・実務留意点
1. 公開公報の先行技術調査・SDI・RAG構築 類似特許抽出、無効資料調査、社内知財DB検索	`embedding` (60 RPM) `rerank` (60 RPM)	◎ 即時活用 PoC推進	◎ 本番運用 大規模インフラ	**MRL 256次元圧縮で公報を省メモリ索引化** 公開公報全文をベクトル化し、`rerank`の`CUSTOMIZE`指示（「構成要件・課題の一致度順に並べ替え」）で上位20件を精読。公開情報のため規約11条の学習リスクが完全無害化され即時実用可能。
2. IPランドスケープ・大量公報分類・ノイズ除去 競合ポートフォリオ分析、課題/解決手段タグ付け	`nano` (60 RPM) `mini` (6 RPM)	◎ 即時活用 公開情報限定	◎ 戦略統合 社内R&D連携	**strict JSON Schemaによる高速構造化** 軽量・高速な`nano`（60 RPM）を用い、数千件の要約・請求項から「技術課題・解決手段・効果」をJSON抽出。抽出結果を`mini`で集約し、経営・R&D向けのHTMLダッシュボードや共創資料を生成。
3. 外国公開公報・海外知財ニュースの和訳査読 USPTO/EPO公開文献、訴訟動向、技術記事の抄訳	`translate` (6 RPM) Webチャット翻訳モード	○ 有効 長文分割要	◎ 高速運用 バッチ処理	**無生物主語を解消した高自然度和訳** 技術英語特有の構文を自然な日本語に変換し、技術者の一次スクリーニング時間を短縮。無料版は1リクエスト5,000文字上限及び6 RPMのため、明細書は段落・章単位にチャンク分割して投入するパイプライン設計が必須。
4. 外国出願用の正式明細書・クレーム日英翻訳 PCT出願、米国特許庁（USPTO）向け直接出願書類	`translate` (6 RPM) `mini` (6 RPM)	× 利用不可 機密・漏出懸念	△ 補助利用 弁理士校閲必須	**未公開機密の投入厳禁 & 訳抜け・漏き出しリスク** 出願前の未公開原稿は無料版規約11条（再学習・無償利用）により絶対入力禁止。また特化モデルは流暢さを優先して限定詞脱落（訳抜け）や幻覚が生じる懸念があり、クレーム直訳には`mini`の厳格プロンプトと専門弁理士の全件検閲が必須。
5. 拒絶理由通知（OA）分析・引例対比・意見書案 特許庁審査官の拒絶理由精査、引用文献との構成対比	`mini` (6 RPM) 128kコンテキスト	○ PoCのみ 公開経過情報	◎ 本番運用 高速下書き作成	**日本語3割省トークンで128k枠に公報全文投入** 本願公報と複数引用文献の全文を128kコンテキストに一括入力し、構成要件対比表と反論論点の下書きを生成。無料版ではJ-PlatPat既公開の拒絶経過案件による検証に限定（規約10条の法律助言禁止に留意し、最終判断は弁理士が実施）。
6. 発明発掘・出願前明細書ドラフト・FTO鑑定 発明提案書ドラフト、実施例作成、侵害予防クリアランス	`guard` (6 RPM) `mini` (6 RPM)	× 厳禁 新規性喪失・規約	○ 補助ツール 国内データ主権	**未発表発明の無料版入力は「公知化」致命リスク** 出願前の実験データやノウハウ入力は新規性喪失及び規約11条（SB Intuitionsへの無償利用権付与）により致命傷となるため無料版では絶対不可。`guard`フィルタと閉域環境を備えた再学習なしの有償ソブリン基盤（Cloud PF Type A）でのみ運用可能。

01. レート制限（RPM）の実務適合性
60 RPMの**nano / embedding / rerank**はバッチ調査・分類パイプラインに最適。一方、対話・翻訳標準の**mini / translate**は6 RPMのため、大量公報処理ではキュー制御と分割待機設計が必須。

02. 知財ガバナンスと規約境界線
無償版規約第10条（法的アドバイス禁止）と第11条（学習利用オプトアウト不可）が最大の分水嶺。「公開特許の調査・分析・可視化」は無償版で加速し、「出願前ドラフト・侵害鑑定」は必ず有償ソブリン基盤に隔離する二段構えが正解。

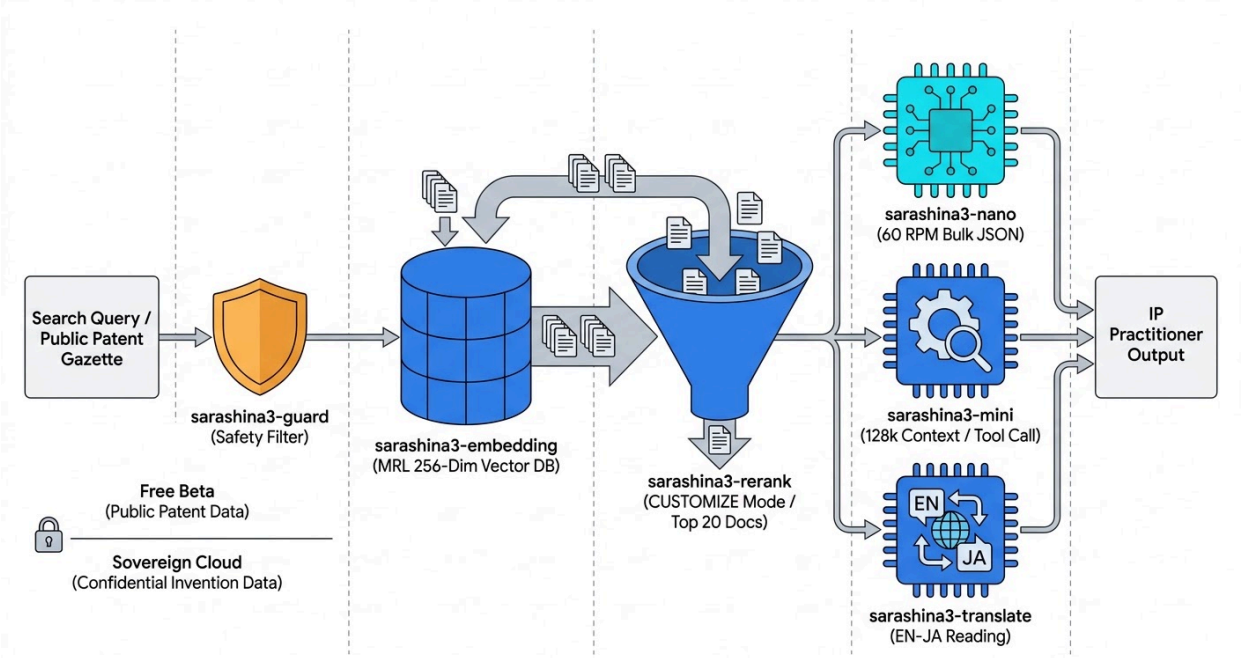
Evaluation of six core Intellectual Property (IP) workflows across the **Sarashina Service Free Beta** and **Cloud PF Type A**. Workflows relying strictly on published patent gazettes (J-PlatPat) and public corporate data can be immediately prototyped and executed in the Free Beta using the 60 RPM models (**embedding**, **rerank**, **nano**), whereas workflows involving unpublished inventions or formal legal opinions are blocked in the Free Beta by Terms of Service Articles 10 and 11.

Data sources: SB Intuitions プレスリリース, SB Intuitions Tech Blog, AI Revolution, The Invention 2026, Science & Technology, note (知財実務・生成AI)

3.1 先行技術調査・無効資料調査・SDI（定期監視）におけるRAGパイプライン構築

無償ベータ版において最も導入効果が高く、かつ法的リスクを完全にクリアして即日実践できる領域が、公開特許公報（J-PlatPat等から取得した発行済み公報テキスト）を対象とした先行技術調査・無効資料調査・SDI（定期監視）の高度化システム（知財RAG）の開発・検証である[2]。公開公報はすでに世の中に公表された公知文献であるため、学習利用される無償ベータ版に入力しても新規性喪失や営業秘密漏えいの問題は生じない（ただし、書誌事項に含まれる発明者の個人情報は前処理で除去し、かつ調査クエリ自体に自社の未公開製品スペックをそのまま書き込むことは避け、抽象化した技術要素や公知の課題表現に置き換える統制が必要である）[2]。

従来の特許ベクトル検索では、「キーワードや表面的な語彙は似ているが、解決しようとする技術課題や発明の文脈が全く異なるノイズ文献」が大量に混入する課題があった[4]。これに対し、Sarashinaサービスが提供するsarashina3-embeddingとsarashina3-rerankは、いずれも無償ベータ版でありながら1分あたり60回（60 RPM）という高いレート制限が付与されており、実用的な二段階検索パイプラインを構築できる[2]。



この二段階検索パイプラインの第一段階では、sarashina3-embeddingのMRL（マトリョーシカ表現学習：Matryoshka Representation Learning）を活用した公開公報の大規模・省メモリベクトル化を行う[5]。本モデルは、Retrieval・Reranking・STSタスクに強い「検索特化モデル」とClassification・Clusteringタスクに強い「分類特化モデル」を別々に学習した上でDARE-TIES手法により均衡統合し、さらにMRLを組み込むことで、前世代（Sarashina2 embedding）比でベクトル次元を40%削減しながら出力ベクトルの次元数を切り詰めても検索性能が劣化しにくい設計を実現している[5]。日本語テキスト埋め込みベンチマーク（JMTEB）において、ベクトルを256次元まで圧縮した場合でも平均スコア75.72を記録しており、これは768次元を要する高性能オープンソースモデルruri-v3-310m（スコア75.85）に匹敵する[5]。知財部門が数十万～数百万件の公開特許公報（要約・請求の範囲・課題解決手段）をベクトルデータベースに格納する際、256次元への圧縮を適用することで、メモリ効率やス

トレージサイズを大幅に節約しつつ高速な一次候補抽出（Retrieval）が可能となり、さらに7種類のタスクタイプを用途（公報検索、技術クラスタリング、類似度算出など）に応じて動的に切り替えることができる[5]。

続く第二段階では、一次検索（ベクトル検索およびIPC/FI・キーワード検索）で絞り込んだ上位20件の候補公報を、約4.4兆トークンを学習したSB Intuitions独自のModernBERT-Ja-1.4Bをベースとするsarashina3-rerankへ投入する（1リクエストあたり最大20文書、クエリと候補文書中で最も長い文書のトークン数の合計が8,192トークン以内となるよう設計されており、JMTEBのESCI・JaCWIR・JQaRA・MIRACL・MLDRの5タスクにおいて既存のSOTA日本語リランカーcl-nagoya/ruri-v3-reranker-310mと同水準の性能を実証している）[4]。本モデルの最大の武器は、一般検索（GENERAL_RETRIEVAL）、QA、ニュース検索、意味的類似度評価などの5種の定型指示だけでなく、CUSTOMIZEモードによって「何を基準に関連度が高いと判定するか」を自然言語の独自指示文として動的に注入できる点にある[4]。

SB Intuitionsの公式検証でも、同一のクエリ（「ログデータを外部サービスに送信して分析したい」）と候補文書群に対し、「エンジニア向けの実装仕様文書を優先せよ」と指示した場合にはAPIキーやJSON形式を解説した文書がスコア0.9999995で最上位となり、「法務担当者向けの文書を優先せよ」と指示した場合には個人情報の保存期間や委託先管理・匿名化を解説した文書がスコア0.9999679で最上位へと劇的に入れ替わることが実証されている[4]。これを知財調査に応用すると、同一の先行技術候補群に対して、無効資料調査では「請求項1の構成要件である特定機構の機械的接続構造または数値限定範囲が具体的に開示されている公報を最優先せよ」と指示して概念的な言及にとどまる公報を除外し、課題・効果一致調査では「単に同一の材料を使用しているだけでなく、高温環境下での剥離防止という技術課題を解決することを主目的とした文献を優先せよ」と指示して異分野のノイズ文献を沈め、さらにライセンス標的調査では「製造方法ではなく最終製品の装置クレームとして権利化されており侵害発見が容易な文献を優先せよ」と指示するなど、調査官の熟練のスクリーニング眼をAPIのパラメータとして直接コード化できる[4]。

3.2 IPランドスケープ・大量公報スクリーニングと他部門共創レポートの自動生成

近年の知財実務においては、特許庁の「VC-IPASガイドブック」等が提唱するように、特定企業のコア技術・用途の特定、特許ポートフォリオの分析、競合他社の事業・技術との比較という3ステップを通じて、経営層や事業部・研究開発部門と戦略を練り上げる「IPランドスケープ」および「共創型知財分析」の重要性が急拡大している[21]。また、先進企業では生成AIによるノイズ文献フィルタリングと要約・関連度スコアリングにより、特許調査工数を22時間から3時間へ約86%削減（NEC事例）、あるいは査読時間を約65%削減（Patentfield AIR事例）する成果が報告されている[22]。

この領域において「Sarashinaサービス」無償ベータ版を活用する際の核心的な設計戦略は、sarashina3-nano（60 RPM）とsarashina3-mini（6 RPM）の非対称なレート制限を活かした役割分担である[2]。標準モデルのsarashina3-miniは1分間に6回しか呼び出せないため、100件の公開公報を処理するだけで最短でも約17分を要し、大量文献の一次処理には向かない[2]。一方、軽量高速モデルのsarashina3-nanoは1分間に60回（1時間で最大3,600件）のリクエストが可能であり、128kトークンの広大なコンテキストとresponse_formatにおけるstrict指定付きjson_schemaによる厳格なJSON構造化出力をサポートしている[2]。公式ユースケースでもECサイトの大量口コミの並列感情・スコア分析や表データのJSON変換が挙げられている通り[6]、競合他社の公開特許公報群や公開プレスリリース・製品Webページをsarashina3-nanoに連続投入し、公報番号、技術課題、解決手段の核心、

従来技術に対する効果、想定アプリケーション、および重要度スコアを定型JSONスキーマで高速に構造化抽出させる[2]。
その上で、抽出・集計されたJSONデータセットを最後に1回～数回だけsarashina3-mini（高度な数理理解・コーディング能力を保持）[6]に渡し、根拠となる公開公報へのリンクを埋め込んだインタラクティブなHTMLレポートや図解ダッシュボードのコードを生成させる[21]。これにより、知財部員が一方的にテキスト報告書を読ませるのではなく、研究開発者や事業部メンバーが自らの関心軸（競合企業別・技術課題別・用途別）でクリックして深掘りし、必要に応じて根拠となる特許文献の原文まで遡れる「触れる知財資料（共創の場）」を無償ベータ版の枠内で迅速に構築できる[21]。

3.3 外国特許公報の査読・技術理解と外国出願翻訳における「意訳特性」の使い分け

知財業務における翻訳ニーズは、大きく「外国特許公報（US/EP/WO等）や海外知財訴訟ニュースを日本語に訳して素早く技術内容・法的争点を把握する業務（インバウンド査読）」と、「日本語の特許明細書・クレームを外国出願（パリルート・PCT国内移行）や中間対応のために英語へ翻訳する業務（アウトバウンド権利化）」の2つに分かれる[23]。
SB Intuitionsの松野智紀氏らがTech Blogで公開したsarashina3-translate（2026年4月訓練完了）の詳細な学習パイプライン（汎用LLMによる大規模データ合成と品質フィルタリング、独自の訳質評価モデルを用いた自己蒸留、およびMetricXやMQMを組み合わせたSFT＋DPOの2段階アライメント）ならびに評価実験データは、この2つの知財翻訳業務に対する明確な適性差を示している[1]。

評価指標・比較観点	**Sarashina3 translate** の数値・特性	**Google翻訳 (Cloud Translation API Advanced)** の数値・特性	知財実務における解釈と業務適性への示唆
翻訳アプローチの基本設計	専門用語を伝わりやすい表現に置き換え、無生物主語のねじれを解消し、文脈を汲み取った流暢な意訳を重視[1]。	原文の構文構造に忠実な直訳的翻訳を行う傾向[1]。	外国公報の素早い通読・R&Dへの共有にはSarashina3が圧倒的に読みやすいが、一言一句の厳密な対応が求められるクレーム翻訳では意訳がリスクとなる[1]。
LLMペアワイズ自動評価 （WMT2024 / GPT-5-mini判定）	英日: 勝率 **37.7%** vs 敗北 32.6%（引分 29.7%） 日英: 勝率 **34.3%** vs 敗北 29.9%（引分	Sarashina3に対して英日・日英ともに負け越し（なおplamo-2-translateに対してもSarashina3が英日46.8%対	正確性と流暢性を総合判定した場合、英日・日英の双方で既存の商用翻訳モデルを上回る高い総合翻訳品質を持つ[1]。

	35.8%）[1]	23.8%で圧勝）[1]。	
文字列一致度（**BLEU / chrF**）	英日: BLEU **24.37** / chrF **34.04** 日英: BLEU **21.27** / chrF **49.44** [1]	英日: BLEU **28.11** / chrF **37.58** 日英: BLEU **25.43** / chrF **52.38** [1]	語順変更や意訳を行うため、参照訳との表層的な文字列一致度（BLEU等）では直訳型のGoogle翻訳を下回る[1]。
意味的類似度・誤訳評価 （COMET / GEMBA-MQM）	英日: COMET **0.8821** / MQM **-2.91** 日英: COMET **0.8318** / MQM **-1.63** [1]	英日: COMET **0.8888** / MQM **-2.86** 日英: COMET **0.8404** / MQM **-1.59** [1]	文脈の意味的保存度を測るCOMETや重み付き誤訳スコア（GEMBA-MQM）ではplamo-2-translate（英日MQM -3.80）を大きく凌駕し、Googleと同等水準を維持[1]。
社内アノテーター**5**名によるブラインド人手評価	英日: **236**勝 213敗（5名中3名が支持し勝利） 日英: 199勝 **214**敗（僅差で敗北）[1] ※「自然」「ニュアンスを捉えている」と高評価[17]。	英日で敗北、日英で僅差勝利[1]。 ※「湧き出し（原文にない情報の付加）」や「意味の乖離」が少ない点が厳密派の評価者から支持[17]。	重要知見: 厳密性を最重視する評価者からは、Sarashina3の表現の工夫が「湧き出し」等の懸念として評価が分かれた[17]。特許クレーム翻訳ではこの特性の制御が必須となる[17]。

この評価データから導かれる知財業務への具体的な適用指針は極めて明確である。第一に、「外国特許公報（英語）の日本語査読・R&D技術者への情報共有」には、sarashina3-translate（およびWebチャットの翻訳モード）が極めて有効である[1]。米国特許（USPTO）や欧州特許（EPO）の英文明細書は、関係代名詞の多重入れ子や無生物主語（例えば「it's most interested in...」のような構文）が多用されており、従来の直訳型機械翻訳を通すと日本語として破綻した悪文になりやすく、査読者の認知負荷を著しく増大させていた[1]。sarashina3-translateは無生物主語のねじれを「〜に重点が置かれている」といった自然な日本語構文へ再構築し、専門用語や多義語を文脈に応じて的確に訳し分けるため（人手評価の英日翻訳でGoogle翻訳に236対213で勝利）[1]、大量の外国公報の発明概要や実施例を技術者が素早く理解する用途に最適である。ただし、APIの入力上限が1リクエストあたり5,000文字（1 MiB）、レート制限が6 RPMであるため[2]、英文明細書全文を一括投入するのではなく、Abstract、Background、Summary、各Embodimentのセクション単位（5,000文字以内）にチャンク分割して順次リクエストを送る設計が必要となる[2]。

第二に、「外国出願用の特許請求の範囲（クレーム）および明細書の正式な日英翻訳」においては、sarashina3-translateのそのままの適用は避けるべきである[2]。そもそも出願前の明細書や未公開の中間処理方針は機密情報であり、学習オプトアウトができない無償ベータ版には入力できない[2]。さらに、将来的に有償のソブリン環境へ移行した場合であっても、特許翻訳においては「原文にない情報の付加（湧き出し）」や「構成要件の勝手な言い換え・省略（訳抜け）」は、特許法上の新規事項の追加（New Matter）による拒絶・無効理由や、権利範囲の意図せぬ減縮・禁反言（エストッペル）を招く致命傷となる[9]。社内アノテーターによる全960件の評価コメント分析（Error Span Annotationフレームワーク準拠）でも、原文への忠実さ・厳密性を最重視する評価者からは「湧き出し」の少ないGoogle翻訳が支持されたことが報告されている[1]。したがって、公開された日本語公報をベースに外国出願翻訳のPoCを行う際や、後に有償環境で実務利用する際には、意訳傾向のあるsarashina3-translateではなく、指示追従性の高いsarashina3-miniに対して特許用語対訳辞書（Glossary）を与え、「意訳や主語の補完を禁じ、構成要件を一対一で直訳せよ」という厳格な制約プロンプトを課すアプローチが適している[2]。

3.4 拒絶理由通知（OA）分析・明細書作成支援・形式チェックにおける検証アプローチ

権利化業務（発明発掘、明細書作成、拒絶理由通知への対応）は、知財実務の中で最も工数削減インパクトが大きい領域であり、島津製作所における「プロンプトドリブン改革」（発明提案書からクレームドラフトや実施例構成案を生成）や国内企業における出願依頼書作成時間の約15時間から1〜2時間への短縮（約90%削減）など、劇的な効率化事例が報告されている[22]。無償ベータ版においては、秘密保持の制約から「未公開の実案件」を投入することは厳禁であるが[2]、J-PlatPat上で公開されている過去の審査経過情報（公開公報、拒絶理由通知書、意見書、手続補正書、審決文）や架空のダミー発明提案書を用いた「プロンプトエンジニアリングおよびエージェント挙動の検証環境」として、sarashina3-miniおよびsarashina3-guardは極めて高い価値を持つ[2]。

まず、拒絶理由通知への対応（中間対応）の検証においては、sarashina3-miniが備える128kトークンの長文コンテキストと高い日本語トークン効率が威力を発揮する[2]。中間対応では、本願の出願時明細書・請求項、審査官が起案した拒絶理由通知書、および引用文献1〜3（主引例・副引例）の全文を同時に突き合わせ、審査官の認定の妥当性と技術的相違点を精査する必要がある[20]。Sarashina3は日本語トークン効率が米国製モデルより約3割高いため[2]、128kトークンのコンテキストウィンドウ内に本願と複数の長大な引用文献を欠落なく同時格納できる[2]。公開済みの審査経過データを用いて、「審査官が引用文献1の該当段落から認定した構成と、本願請求項1の構成との文言上の差異および技術的意義の差異を抽出し、進歩性主張（動機づけの欠如・阻害要因）の論理構成案を出力せよ」というプロンプトの精度を無償で徹底検証できる[6]。

また、検索式作成や分類特定の自動化に向けては、sarashina3-miniのツール呼び出し（Function Calling）機能を活用したエージェントの試作が有効である[6]。本モデルは関数呼び出しベンチマーク（BFCL v3）でGPT-4.1と同等の性能を有し、複数ツールの自律的な選択・実行に対応している[6]。技術用語の類義語展開APIやIPC/FI分類コード検索関数をtoolsとして定義（必ずtool_choice: "auto"を明示指定）[2]することで、ユーザーが入力した技術テーマに対して自律的に分類コードと類義語を調査し、論理演算子を組み合わせた特許検索式を提案するAIエージェントを無償でプロトタイピングできる[6]。なお、日本語特許特有の言い回し（「〜からなる」と「〜を含む」の権利範囲の違いなど）や旧公報の表記ゆれについてはAIの示唆をそのまま鵜呑みにせず、最終的に弁理士・調査担当者が検索

式をチューニングする分業体制が不可欠である[9]。
さらに、明細書の作成支援に関しては、生成AIにゼロからクレームを起案させると実施例の具体的構成に引きずられて権利範囲が過度に狭くなる「AIドラフトの罠」や、若手のクレーム設計力が育たない教育上の問題（特許庁の弁理士制度小委員会でも指摘されている論点）が存在するため、「ゼロ稿や観点チェックはAI、クレーム設計と最終判断は人間」という分業が鉄則となる[9]。とりわけ人間が設計したクレームと明細書案に対する「品質チェックリストの自動監査」には生成AIが極めて強い威力を発揮するため[20]、公開公報やダミー明細書をsarashina3-miniに入力し、発明が解決しようとする課題と発明の効果の対応関係、用語定義の一貫性と表記ゆれ、図面符号の記載漏れ・不整合、およびサポート要件・実施可能要件の観点からの記載不足箇所を列挙させるプロンプトテンプレートを無償ベータ版で磨き込んでおくことが推奨される[2]。加えて、社内R&D部門向けの知財AIポータルを設計する際には、フロントエンド直後に64kトークン対応のsarashina3-guardを配置し、入力テキストの「安全／有害」および有害カテゴリコード（情報漏えい・違法行為のほう助等）を高速判定してフィルタリングするゲートキーパー機構の動作検証を、無償APIを用いて事前に確立しておくことができる[2]。

4. 知財部門・特許事務所における段階的導入ロードマップと結論

4.1 無償ベータ版（PoC）から有償ソブリン環境（Cloud PF Type A）への二段階移行設計

以上の分析を総合すると、「Sarashinaサービス」無償ベータ版は、それ単体で機密情報を扱う本番知財業務を完結させるためのものではなく、ソフトバンクのソブリンクラウド「Cloud PF Type A」上の商用「Sarashina3 シリーズ」へシームレスに移行するための「無償開発・評価サンドボックス」として位置づけるのが最も合理的かつ安全なアーキテクチャ戦略である[2]。
無償ベータ版は提供終了時期が未定であり、損害賠償免責や再許諾制限、6 RPMのレート制限が存在するため、本番業務の継続性を担保する基盤としては利用できない[2]。しかし、同じモデル群が稼働する「Cloud PF Type A」へ移行すれば、入力データの再学習非利用が明示的に保証され、国内データセンター完結によるデータ主権（海外法域リスクの排除）とアクセス制御が担保されるため、知財業務におけるセキュリティおよびガバナンス上のボトルネックがすべて解消される[6]。

評価軸	「**Sarashina**サービス」無償ベータ版（現在公開中）	「**Cloud PF Type A**」有償環境（**On-Demand**共有型 / **Dedicated**専有型）
主な利用目的・フェーズ	公開特許公報・ダミーデータを用いたPoC、RAG精度検証、プロンプト資産の構築[2]	出願前発明・FTO調査・中間対応・契約審査を含む全知財業務の本番運用[2]
入力データの再学習利用	学習に利用される（オプトアウト不可）[2]	再学習には一切利用されない（規約・仕様により明示保証）

		[6]
データ主権・セキュリティ	一般パブリックWeb/API経由（Googleアカウント認証）[2]	国内データセンター完結（海外移転なし）、ソフトバンク及びOracleの非アクセス設計、Private Endpoint / IAM制御対応[6]
利用可能モデルと提供形態	6モデルすべて無料（mini/guard/translateは6 RPM、nano/embedding/rerankは60 RPM）[2]	**On-Demand**（従量課金）: 5モデル（mini, nano, guard, embedding, rerank）対応 **Dedicated**（月額固定・**NVIDIA HGX B200**専有）: mini, nanoの超高速・安定大量処理に対応[6]
日本弁理士会ガイドライン適合性	秘密情報の入力は不適合（公開情報処理のみ適合）[2]	暗号化・再学習禁止・国内法域完結を満たし、機密情報の取扱基準に完全適合[6]

4.2 総括：知財実務家が採るべきアクションプラン

2026年現在の知財実務において、知財人材に求められる役割は、単なる「労働集約的な調査・書類作成の代行者」から、AIシステムの設計思想を理解し業務目的に合わせて最適な文脈とパイプラインを構築する「コンテキスト・アーキテクト（文脈の設計者）」および「事業部・R&Dとの共創の場を設計するハブ」へと構造変化を遂げている[21]。

企業の知財部門および特許事務所は、今回の「Sarashinaサービス」無償ベータ版の公開を受け、直ちに3段階のアクションを順次実行に移すべきである。第一段階（ガバナンスルールの即時周知）として、所内・部内において「Sarashina無償ベータ版は入力データが学習に利用されオプトアウトできないため、未公開の発明提案書、出願前明細書、クライアント機密情報、および発明者個人情報の入力を厳禁とし、法的助言として出力を直接使用しない」旨のガイドラインを明文化して通達する[2]。

続く第二段階（公開公報を用いた知財RAGとプロンプト資産の無償構築）では、公開特許公報（J-PlatPatデータ）と公開企業情報を題材として、sarashina3-embedding（MRL 256次元圧縮）とsarashina3-rerank（CUSTOMIZE指示モード）を組み合わせた高精度な先行技術スクリーニング基盤を試作する[4]。同時に、sarashina3-nano（60 RPM＋JSON Schema）による公報群の高速構造化およびsarashina3-translateによる外国公報の自然な和訳査読ワークフローを確立し、公開審査経過データを用いてsarashina3-mini（128kコンテキスト）による拒絶理由通知の引例対比・明細書チェックリストのプロンプトテンプレートを完成させる[2]。

そして第三段階（ソブリンクラウド環境への本番移行とコスト最適化）において、無償ベータ版で検証した日本語性能（JamC-QA 0.82）と約3割のトークン削減効果をもとに投資対効果（ROI）を算出し[2]、

機密性の高い出願業務・FTO調査・契約業務へ適用範囲を拡大するタイミングで、再学習非利用と国内データ主権が保証された「Cloud PF Type A」のOn-Demand型（従量課金）またはDedicated型（NVIDIA B200専有・月額固定）へとAPI接続先およびOracle Database等とのデータ連携基盤を切り替える[2]。この「公開情報による無償ベータ版での徹底したパイプライン検証」と「ソブリンクラウドでの安全な機密情報本実装」を峻別した二段階アプローチこそが、新規性喪失や守秘義務違反のリスクをゼロに抑えつつ、国産最高峰の生成AI「Sarashina3」の恩恵を知財業務で最大化する最適解である[2]。

引用文献

1. 国産生成AI「Sarashina」を利用できる「Sarashinaサービス」の, https://www.sbintuitions.co.jp/news/press/20261013_01/
2. Sarashina無料β版とは？始め方・6モデルAPI・制限・利用規約の, https://ai-revolution.co.jp/media/sarashina-free-beta/
3. 国産生成AI「Sarashina」の無償ベータ版を提供開始, https://businessnetwork.jp/article/37791/
4. Sarashina3 rerank: 日本語検索に指示付きリランキングを導入する, https://www.sbintuitions.co.jp/blog/entry/2026/07/03/100056
5. Sarashina3 embedding: 日本語に強い最新のテキスト埋め込みモデル, https://www.sbintuitions.co.jp/blog/entry/2026/07/02/101054
6. Sarashina3 mini / Sarashina3 nano: フルスクラッチで開発した最新, https://www.sbintuitions.co.jp/blog/entry/2026/06/30/100000
7. 弁理士業務AI利活用ガイドライン（2025年4月公表）の評価と反響, https://yorozuipsc.com/uploads/1/3/2/5/132566344/2ad1da223bf80d38ed2c.pdf
8. 日本の弁理士 向け生成 AI ガイドラインに関する詳細な 分析と戦, https://yorozuipsc.com/uploads/1/3/2/5/132566344/1cc6220fa1e5998eb9d0.pdf
9. 弁理士のAI活用術 | 業務効率化と集客の実践ガイド - 志師塾, https://44jyuku.com/blog/benrishi-ai-katsuyo/
10. SB Intuitions、独自の日本語LLMを構築 アカデミアや産業界の研究, https://www.sbintuitions.co.jp/news/press/20240614_01/
11. SB Intuitionsと情報通信研究機構、高性能LLMの安全性技術に関する, https://www.sbintuitions.co.jp/news/press/20260226_01/
12. 国産LLM「Sarashina mini」のAPI接続サービス ... - SB Intuitions, https://www.sbintuitions.co.jp/news/press/20251105_01/
13. ソフトバンクのクラウドサービス「Cloud PF Type A」で国産生成AI, https://www.sbintuitions.co.jp/news/press/20260630_01/
14. 「Oracle Alloy」を採用した「Cloud PF Type A」で - SB Intuitions, https://www.sbintuitions.co.jp/news/press/20260416_01/
15. 国産医療特化型LLM「Med-Sarashina」を活用した共同研究を開始, https://www.sbintuitions.co.jp/news/press/20260810_01/
16. 国産生成AI「Sarashina」を利用できる「Sarashinaサービス」の, https://news.nicovideo.jp/watch/nw19874026
17. Sarashina3 translate: 翻訳品質を追求した日英・英日翻訳モデル, https://www.sbintuitions.co.jp/blog/entry/2026/07/14/100002
18. 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
19. 士業のAI活用｜守秘義務の条文と、事務所で決める環境・ルール, https://coplus-one.jp/shigyo-ai/
20. 弁理士×生成AI【7事例】業務を30%削減する完全ガイド, https://nocode-sol.co.jp/blog/patent-attorneys-generative-ai-use-cases/
21. 生成AIが変えるのは、特許調査の速度ではなく「知財の伝わり方」, https://note.com/tsunobuchi/n/n0d4c217c9bef
22. 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
23. B260709: 知財業務における生成AIの賢い使い方, https://www.science-t.com/seminar/B260709.html
24. 記事一覧 - SB Intuitions TECH BLOG, https://www.sbintuitions.co.jp/blog/archive
25. 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
26. 知財業務における生成AIの活用と知財業界の将来展望, https://www.japia.or.jp/topics_detail/id=6479
27. ソフトバンクCloud PF Type Aとは？ソブリンクラウドの特徴・料金, https://www.issoh.co.jp/tech/details/11819/