「Sarashina サービス」無償ベータ版の知財業務への活用評価

— SB Intuitions プレスリリース（2026 年 10 月 13 日）の調査と業務別評価 —

2026 年 10 月 14 日

Claude Opus 5.5

凡例：本文中の右肩の番号は、文末の参考文献の番号を示す。見出しの〔事実〕は出典で確認した事項、〔分析〕および本文の「【分析】」は公開情報に基づく推論・評価を示す。

要旨

- **未公開の発明情報や営業秘密は、無償ベータ版に入力しない。**公式発表は、品質向上・改善のために利用データを活用する場合があるとしている[1]。第三者による利用規約の要約では、学習利用のオプトアウトは有償プランだけとされる[2]。
- **有望な用途は、公開情報を扱う 3 つの領域に限られる。**公開特許の日本語セマンティック検索（embedding・rerank）、nano による公開公報の大量分類・要約、公開文献の下訳である。
- **法的判断を伴う業務は補助にとどめる。**生成モデル（mini・nano）は推論モデルではなく[3]、第三者の検証では架空の文献を示す例が報告されている[4]。クレーム解釈、FTO、中間応答の論理構成は任せられない。
- **機密情報を扱う本番利用は、法人向け環境で評価し直す。**ソフトバンクの「Cloud PF Type A」は国内データセンターで運用される[5]。ただし 2026 年 6 月 30 日の発表時点で、同環境の提供モデルに translate は含まれていない[5]。

1　発表内容の概要〔事実〕

項目	内容
提供開始	2026 年 10 月 13 日（火）[1]
提供機能	API 機能（Sarashina3 mini・nano・embedding・rerank・guard・translate の 6 モデル）と、チャット機能（ウェブ検索や文章作成などの対話、日英・英日翻訳）[1]
対象	開発者・IT エンジニアに加え、生成 AI を試したい個人も利用できる[1]
利用方法	Google アカウントでログインする。事前の申し込みは不要[1]
料金	無料。一定時間あたりのリクエスト数に上限がある[1]

項目	内容
提供期間	無償ベータ版の終了時期は未定[1]
データ	品質向上・改善のため、利用データを活用する場合がある。詳細は利用規約とプライバシーポリシーに記載[1,6]
正式版・価格	発表なし[1]

API で使える 6 モデル

モデル	用途	最大コンテキスト	RPM
Sarashina3 mini	標準の対話・文章生成。ツール呼び出しに対応	128k トークン	6
Sarashina3 nano	軽量・高速。抽出・要約・分類の大量処理	128k トークン	60
Sarashina3 guard	入力テキストの安全性判定（T1〜T12）	64k トークン	6
Sarashina3 embedding	テキストのベクトル化（1536 次元）	8,000 トークン	60
Sarashina3 rerank	検索候補の並べ替え（1 回 20 件まで）	8,000 トークン	60
Sarashina3 translate	日英・英日翻訳	5,000 文字	6

注：最大コンテキストと RPM（1 分あたりのリクエスト数）は、公式ドキュメント[7]の記載として第三者の検証記事[2]が引用した値である。

2　モデルの特徴〔事実〕

- Sarashina3 mini・nano はフルスクラッチで開発され、日本語・英語を中心とする 30 兆トークンで事前学習された。長い思考を行わない Instruct モデルで、推論（Reasoning）モデルではない[3]。パラメータ数は公表されていない[8]。
- 開発元の評価では、推論モードを使わない Qwen3-235B-A22B-Instruct、GPT-4.1、GPT-5.4 mini（non-reasoning）と全体的に同程度で、日本文化 QA（JamC-QA）ではこれらを上回る。ただし自社評価で、比較対象はいずれも非推論モデルである[3]。
- 開発元自身が、推論モデルが普及した現在の水準には、エージェント能力などでまだ届いていない点があると認めている[3]。
- embedding は、日本語埋め込みベンチマーク JMTEB v1.4.0 で平均 78.22（Retrieval 79.3、Reranking 88.4）。前世代（平均 78.69）より次元を 40%減らしつつ、検索系の 2 項目では上回る[6]。

- translate は、WMT2024 を使った LLM による一対比較で Google 翻訳と plamo-2-translate に勝ち越した。一方、人手評価では英日で Google に勝ち越したが（236 対 213）、日英では僅差で負けた（199 対 214）。原文への忠実さを重視する評価者は Google を支持する傾向があり、開発元は Sarashina の意訳傾向を指摘している[9]。
- 提供される 6 モデルはいずれもテキストを処理するもので[2]、画像や PDF を直接入力できるかは確認できなかった（非対応とは断定しない）。

3　データの取り扱い〔事実〕

論点	確認できた内容	確度
無償ベータの利用データ	品質向上・改善のため活用する場合がある[1]	一次情報
学習利用とオプトアウト	無償プランの入出力はモデルの学習や新規サービスの開発に使われ、学習利用を拒否できるのは有償プランだけ（利用規約、2026 年 9 月 28 日制定）[2]	二次情報（規約原文は未確認）
利用資格	日本国内に居住する個人と、国内に本店を置く法人[2]	二次情報
禁止事項	特定の個人を識別できる情報の入力、レート制限を回避するための複数アカウント作成など[2]	二次情報
保管場所・権利帰属・商用利用・免責	確認できなかった	未確認
参考：旧 Sarashina API（法人向け）	よくある質問で、入出力データをモデル学習に利用しないと明記。2026 年 6 月 16 日に新規受付を終了し、Cloud PF Type A へ順次統合[10]	一次情報
参考：Cloud PF Type A	Oracle Alloy を使った基盤を、ソフトバンクが国内データセンターで管理・運用する。データ主権と運用主権をうたう。東日本は 2026 年 6 月 30 日開始、西日本は 10 月予定。共有型（従量課金）と専有型（月額固定）がある[5]	一次情報

【分析】日本弁理士会のガイドラインは、入力情報が学習に利用される場合は営業秘密や非公開情報の入力を避けるべきだとしている（第 4 章 1）[11]。また、秘密保持規約のない第三者のプラットフォームに情報を載せると新規性を失う可能性があるとし（第 4 章 2(2)）、クライアントの秘密情報は学習の有無にかかわらず合意を取るべきだとする（第 4 章 4）[11]。無償ベータ版は「学習に利用される場合」に当たるとみて扱うのが安全である。

【分析】同ガイドラインは商用利用の可否の確認も求めている（第 4 章 1）[11]。無償ベータ版の商用

利用条件は本調査で確認できていないため、業務で使う前に規約原文を確認する必要がある。なお、AI への入力と新規性喪失・営業秘密の秘密管理性との関係には確立した解釈がなく、保守的に運用するのが妥当である。

4 「国産」の意味と政策上の位置づけ

- **技術：**モデルはフルスクラッチ開発である[3]。旧 Sarashina API の紹介ページは、他国への技術依存による規制変更のリスクを排除できると説明している[10]。
- **計算基盤：**「NVIDIA GB200 NVL72」を搭載した AI 計算基盤が 2025 年 12 月 22 日に稼働した。1,224 基の Blackwell GPU で構成され、将来は 4,000 基超・10.6EFLOPS 規模に拡張する予定で、Sarashina の商用サービス開発などに使われる[12]。
- **政府調達：**デジタル庁の「ガバメント AI 源内」で試用する国産 LLM5 モデルの 1 つとして契約された（他は tsuzumi 2、cotomi v3、Takane 32B、PLaMo 2.0 Prime）[13,14]。2026 年 8 月頃から 2027 年 3 月にかけて行政実務への適合性を評価し、優れたモデルは 2027 年度に有償で調達する予定とされる[13]。
- **GENIAC：**2026 年 9 月 9 日、SB Intuitions はロボット基盤モデル研究開発事業の採択 13 社（グループ）の 1 つに選ばれた[15]。Sarashina LLM 本体への GENIAC の支援は、本調査では一次情報で確認していない。

【分析】「国産」が指すのは、主にモデルの開発と、運用・データ管理の主体である。クラウド基盤（Oracle Alloy）や GPU（NVIDIA）を含め、技術スタックのすべてが国産という意味ではない。とはいえ、経済安全保障上センシティブな技術分野（防衛、半導体、重要インフラなど）では、国内運用とガバメント AI での試用実績が社内承認の説明材料になりうる。

5 性能と外部評価〔事実〕

- **Nejumi LLM リーダーボード 4：**Qualiteg による 2026 年 9 月 1 日時点の分析では、総合上位 50 に Sarashina3 は入っていない。首位は Claude Opus 5（0.8720）で、国内勢では GPT-OSS-Swallow-120B-RL-v0.1（0.6914）や llm-jp-4-33b-thinking（0.6760）が挙げられている[16]。Sarashina3 が評価対象に含まれているかは確認できていない。
- **ハルシネーション（個人による検証）：**難問 20 問を各 3 回出題した総合指数は、Sarashina3 mini が 21.7、Claude Opus 5.5 が 95.0、Claude Sonnet 5.5 が 70.0 だった。論文の出典を求めると、実

在の研究者名と無関係な実在の arXiv 番号を組み合わせた架空の文献を示した。祝日法や改元日のようにルールを当てはめる問題で誤り、本題は正しいのに補足で作り話をする例もあった[4]。

- **速度（個人による検証）**：最初の 1 文字が出るまでの時間は mini が 0.11 秒、nano が 0.10 秒で、Claude Haiku 5.5（0.71 秒）の約 7 分の 1 だった。書き出した後の速度では mini は Haiku に及ばず、nano は毎秒 585〜680 字と突出していた。日本語の文章を約 2.1 字/トークンで表し、Claude の約半分のトークン数で済むが、英語でも同程度に効率的なため日本語特化とは言えないと投稿者は指摘している[17]。別の検証では、mini の初回応答は 0.50 秒だった[2]。
- **API 仕様（個人による検証）**：OpenAI 互換を前提に max_tokens を送ると 400 エラーになる。構造化出力は型は守るが、日付の形式指定は守らなかった。翻訳では「見送らせていただく」が postpone（延期）と訳された。embedding は 1 リクエストに 1 件だけで、rerank のスコアは確率として校正されておらず、閾値による足切りに向かない（公式ドキュメントの記載）。translate には用語集や文体を指定するパラメータがない[2]。

注：第三者の検証はいずれも公開初日の個人によるもので、サンプル数も少なく、統計的な評価ではない[2,4]。

6　知財業務別の評価〔分析〕

有用度は、◎＝ベータ版でも PoC の価値が高い、○＝条件付きで有用、△＝補助的、×＝不適、とする。ベータ版では公開情報と架空データだけを扱い、機密情報は法人向け環境で扱うことを前提とする。

業務	活用方法	有用度	留意点
先行技術調査	embedding と rerank で公開公報のセマンティック検索・RAG を試作する。mini で技術要素の分解や同義語の展開を行う	◎（検索） ○（生成）	embedding は 1 件/回・60RPM で、1 時間に約 3,600 件（試算）。rerank は 1 回 20 件まで。スコアで足切りせず順位だけを使う[2]。検索式の最終設計と網羅性の担保は人が行う
特許文献の要約・読解	nano で公開公報を大量に要約し、課題・解決手段・効果を抽出する	○	128k トークンは日本語で 20 万字超に相当するという試算があり[2]、明細書 1 件は収まる見込み。補足部分での作り話の例[4]があるため、原文との突合が必須
クレーム解析・クレームチャート	構成要件の分節や対比表の下書き	△	推論モデルではないため、文言の解釈や均等論の判断は任せられない。構造化出力の形式も保証されない[2]
明細書のドラフト	公開技術を題材にした文章校正や	△（ベータ）	未公開の発明は入力しない。記載要件の判

業務	活用方法	有用度	留意点
支援	言い換え	○（法人環境）	断は人が行う
中間処理（拒絶理由への応答）	引用文献（公開済み）の要約や相違点候補の列挙	△	本願の補正案は未公開情報になりうる。論理の組み立ては推論モデルに分があると見込まれる
侵害予防調査（FTO）	公開特許の母集団の一次スクリーニング（nano による分類）	△	自社製品の仕様は機密。見落としのリスクが大きく、AI の判定だけで除外しない
無効資料調査	公開文献の類似度による並べ替え、英文文献の下訳	○	書誌を生成させると架空の文献が出るおそれがある[4]。書誌は必ずデータベースで照合する
IP ランドスケープ・特許分析	nano で大量の公報を分類・タグ付けし、embedding でクラスタリングする	◎	nano は 60RPM で大量処理に向く。mini は 6RPM で、100 件の処理に約 17 分かかる[2]
発明発掘・アイデア創出	公開技術のトレンドを相手にした壁打ち	△	自社の研究開発情報は入力しない。発想の幅は海外フロンティアモデルが優位と推定
契約書レビュー	公開ひな形を使った条項説明や論点の洗い出しの練習	△	実際の契約書は機密。法的判断に誤りが出るリスクがある
翻訳（外国出願・外国文献）	translate で公開文献を下訳する	○（公開文献） △（出願書類）	意訳傾向があり、人手評価では日英で Google がわずかに優勢[9]。用語集機能がなく、1 回 5,000 字まで[2]。逐語性が求められる特許翻訳とは相性が悪い可能性。未公開の明細書は入力しない
社内知財教育	教材や Q&A の作成、制度説明の平易化	○	公開情報で完結しやすく、ベータ版に向く。ルールの当てはめで誤る例[4]があるため、法令・判例の記述は確認する
知財情報開示（CG コード対応）	公開済みの IR 資料や統合報告書の要約・推敲	○	開示前の文案はインサイダー情報になりうるため、ベータ版では扱わない

7　他の選択肢との使い分け〔分析〕

選択肢	強み	弱み	推奨用途
Sarashina サービス（無償ベータ）	無料。日本語の検索部品、初回応答の速さ、トークン効率	学習利用あり、レート制限、推論は弱い	公開データでの PoC、教育、検索部品の評価
Cloud PF Type A 上の	国内データセンターでの運	単価は非公開[8]。6 月 30 日時	機密を扱う社内 RAG（社

選択肢	強み	弱み	推奨用途
Sarashina3（法人）	用、データ主権[5]、ガバメント AI での試用[13]	点で translate なし[5]。推論性能の差は残る	内規程、自社特許群、未公開の技術資料）
海外フロンティア LLM（法人契約）	推論力、長文の論理、第三者検証での低いハルシネーション率[4]	データの越境、経済安全保障上の説明負担	クレーム解釈の補助、中間応答の論点整理（契約条件の確認後）
国内の特許特化 AI サービス	特許データベースや業務画面との統合、セキュリティ資料の整備	内部で使う LLM はサービスごとに異なり、海外 LLM を使うものもある（各社の開示を個別に確認）	日常の調査・読解のワークフロー
ローカル実行の公開モデル（Sarashina2.2 系）	3B モデルや OCR モデルは MIT ライセンスで手元で動かせ、日本語文書の OCR で高い評価がある[8]	3B 級で能力は限られる。埋め込み v2 や音声合成は非商用ライセンス[8]	未公開資料の社内での OCR・前処理（ライセンスとモデルカードの確認が前提）

【示唆】Sarashina の差別化点は、汎用的な推論力ではなく、日本語の検索部品と、国内運用によるデータ主権にある。知財判断の質が問われる業務は、当面は海外の推論モデルや特化サービスに分がある。Sarashina は検索・分類・前処理の部品として組み込む使い方が現実的である。

8　ベータ版での PoC の進め方〔分析〕

(1) **ガバナンスを先に決める：**入力してよい情報（公開公報・公開文献・架空データ）を文書で定める。業務用の Google アカウントを用意し、私用と分ける。法務が利用規約とプライバシーポリシーの原文（保管場所、権利帰属、商用利用、免責）を確認する[11]。

(2) **評価セットを作る：**自社分野の公開公報 50〜100 件に、正解ラベル（関連・非関連、要約の要点、訳語集）を付ける。

(3) **検索部品を測る：**embedding と rerank を、既存の検索（キーワード・分類）と、上位 10 件に正解が入る率などで比べる。

(4) **生成を測る：**nano と mini で要約・抽出の正確性を確かめ、書誌・数値・請求項番号の捏造の頻度を人手で数える。

(5) **翻訳を測る：**請求項や明細書の抜粋で、逐語性、用語の一貫性、数値・単位の保持を既存の訳と比べる。

(6) **ほかの選択肢と比べる：**同じ評価セットを法人契約の海外 LLM や特化サービスにもかけ、品質・速度・コスト・データ条件を一覧にする。

(7) **移行を判断する：**成果があれば、Cloud PF Type A などで学習利用なし・国内保管・ログ保持を契約上確認したうえで、機密を扱う PoC に移る。ベータ版での大量評価は nano で行い、mini は件数を絞る[2]。

9　留意事項（未確認の事項）

- 利用規約（https://sarashina.ai/terms）とプライバシーポリシー（https://www.sbintuitions.co.jp/privacy-policy/）の所在は公式ブログで確認したが[6]、原文は本調査で読めていない。オプトアウト、利用資格、禁止事項は第三者による要約[2]に基づく。導入の判断前に原文を確認すること。
- 正式版の時期と価格、Cloud PF Type A の単価とデータ条件の詳細は未確認である[1,8]。
- 公式ベンチマークは自社評価で、比較対象は非推論モデルである[3]。第三者の検証は公開初日の少数サンプルによる。
- AI への入力と新規性喪失・営業秘密性との関係は、法的に確立した解釈がない。
- 第 6 章・第 7 章の評価は公開情報に基づく推論であり、自社データでの PoC による検証を前提とする。

参考文献

[1] SB Intuitions 株式会社「国産生成 AI『Sarashina』を利用できる『Sarashina サービス』の無償ベータ版を提供開始」プレスリリース、2026 年 10 月 13 日。https://www.sbintuitions.co.jp/news/press/20261013_01/

[2] skrtk98「国産 LLM『Sarashina3』の無償ベータ API を 6 モデルとも叩いた：速いが、OpenAI 互換を前提にすると 400 になる」Qiita、2026 年 10 月 13 日。https://qiita.com/skrtk98/items/9851aa76c64bfa06256f

[3] Yuyang Dong・廣淵亮太・平子潤「Sarashina3 mini / Sarashina3 nano: フルスクラッチで開発した最新の国産 LLM」SB Intuitions TECH BLOG、2026 年 6 月 30 日。https://www.sbintuitions.co.jp/blog/entry/2026/06/30/100000

[4] kotowazaman「Sarashina でハルシネーションチェックしてたら頼んでもないのにアニメの話を始めた」Qiita、2026 年 10 月 13 日。https://qiita.com/kotowazaman/items/45fb269ac8d6b324af91

[5] ソフトバンク株式会社「ソブリン性を備えたクラウドサービス『Cloud PF Type A』で国産生成 AI『Sarashina』を提供開始」2026 年 6 月 30 日。https://www.softbank.jp/business/news/2026/0630-01

[6] 小川隼斗ほか「Sarashina3 embedding: 日本語に強い最新のテキスト埋め込みモデル」SB Intuitions TECH BLOG、2026 年 7 月 2 日（2026 年 10 月 13 日追記あり）。

https://www.sbintuitions.co.jp/blog/entry/2026/07/02/101054

[7] SB Intuitions「Sarashina API Docs」公式 API ドキュメント。https://docs.sarashina.ai/ （概要ページは調査段階で参照。レート制限ページ https://docs.sarashina.ai/rate-limits は本報告では直接未確認で、数値は文献[2]の引用による）

[8] 株式会社一創「Sarashina とは？モデル一覧（TTS・OCR・Vision）と使い方・ライセンス・API 提供形態」テックブログ、2026 年 4 月 21 日公開・7 月 17 日更新。https://www.issoh.co.jp/tech/details/11856/

[9] 松野智紀「Sarashina3 translate: 翻訳品質を追求した日英・英日翻訳モデル」SB Intuitions TECH BLOG、2026 年 7 月 14 日。https://www.sbintuitions.co.jp/blog/entry/2026/07/14/100002

[10] ソフトバンク株式会社「Sarashina API」サービスページ（2026 年 6 月 16 日付お知らせ、よくあるご質問を含む）。https://www.softbank.jp/business/service/ai/sarashina-api/

[11] 日本弁理士会「弁理士業務 AI 利活用ガイドライン」令和 7 年 4 月。https://www.jpaa.or.jp/cms/wp-content/uploads/2025/04/AIservices-guideline.pdf

[12] ソフトバンク株式会社「『NVIDIA GB200 NVL72』を搭載した AI 計算基盤が稼働開始」プレスリリース、2025 年 12 月 25 日。https://www.softbank.jp/corp/news/press/sbkk/2025/20251225_01/

[13] DX マガジン「デジタル庁、国産 LLM5 社を選定 『ガバメント AI 源内』で行政利用へ」2026 年 6 月 30 日。https://dxmagazine.jp/news/ykmt262407/

[14] AIsmiley 編集部「【2026 最新】ガバメント AI『源内』とは？デジタル庁の生成 AI 基盤とできることを徹底解説」2026 年 6 月 30 日。https://aismiley.co.jp/ai_news/government-ai-gennai-2026/

[15] 朴尚洙「GENIAC の活動は AI×ロボティクスへ、ロボット基盤モデル開発の採択事業者を発表」MONOist、2026 年 9 月 10 日。https://monoist.itmedia.co.jp/mn/articles/2609/10/news053.html

[16] Qualiteg「Japanese LLM Rankings 2026 — Benchmark Analysis Report (September 1 Edition)」Qualiteg Journal、2026 年 9 月 1 日。https://journal.qualiteg.com/llm-ranking-2026/

[17] kotowazaman「ちょっぴりお馬鹿な国産 LLM Sarashina の速さの秘密」Qiita、2026 年 10 月 13 日。https://qiita.com/kotowazaman/items/5e2ec143012d661f1e0a