

NVIDIA RTX Spark（128GB）による 知財業務向けローカル LLM 構築

— 検討すべき国産 LLM と注意事項（2026 年 10 月時点） —

Claude Opus 5.

2026 年 10 月 10 日

凡例：本文中の上付き番号 [n] は文末「参考文献」の番号に対応する。【確認済み】は一次情報または信頼できる情報源で確認した事実、【分析】【推定】は本稿の分析・推計、【未確認】は確認できなかった事項を示す。参考文献のうち URL を確認できなかったものには「※URL 未確認」と付記した。

要旨

最初に試すべきは、国立情報学研究所（NII）の「LLM-jp-4.1-33b-thinking」と、これを土台にした ELYZA「ELYZA-Thinking-1.0-llm-jp-4-33b」である。いずれも国内フルスクラッチの基盤モデルで Apache 2.0 ライセンス、128GB 機では FP8/Q8 で余裕をもって動作する^[1,2]。大量文書の高速処理には同系列の 32B-A3B（MoE）、難度の高い推論のクロスチェックには「GPT-OSS Swallow 120B」（海外ベースの継続学習モデル）を組み合わせる構成を推奨する^[3]。

- ・ **ハードウェア**：「RTX Spark」は 2026 年 5 月 31 日に NVIDIA が発表した Windows（Arm）PC 向けのスーパーチップ／PC プラットフォームで、20 コア Grace CPU と 6,144 CUDA コアの Blackwell RTX GPU を統合し、最大 128GB 統合メモリ、FP4 で 1PFLOP を備える。2026 年秋に OEM 各社から発売される^[4]。Linux（DGX OS）で動く開発機「DGX Spark」（GB10、273GB/s）とは別製品である^[5,6]。メモリ帯域は NVIDIA のポータリングガイドで 300GB/s と記載され、生成速度はこの帯域で決まる^[7,8]。
- ・ **モデル**：主力は LLM-jp-4.1／ELYZA-Thinking（33B Dense・32B-A3B MoE、Apache 2.0、64K コンテキスト）。補助として Qwen3／GPT-OSS Swallow（Apache 2.0、海外ベース）、Stockmark-2-100B（MIT、フルスクラッチ。旧世代の Stockmark モデルは特許庁の特許文書で学習済み）。PLaMo 3 の公開重みはベースモデルのみで独自ライセンス、tsuzumi 2 と PLaMo 3.0 Prime は重み非公開、Rakuten AI 3.0（671B）は 128GB に載らない^[9,10]。
- ・ **注意**：国産 LLM の特許タスクでの公開評価はほとんどない。採否は自社・自事務所の事件データでの評価で決め、出力はすべて弁理士・専門家のレビューを前提とする。完全オフライン化（推

論ソフトの通信遮断、safetensors 限定、ログ暗号化）と、ライセンス・顧客契約の確認を導入の前提条件とする。

主要な知見

- ・【確認済み】 RTX Spark は Blackwell RTX GPU（6,144 CUDA コア、第 5 世代 Tensor コア、FP4）と 20 コア Grace CPU（MediaTek と共同設計）を NVLink-C2C で接続し、最大 128GB 統合メモリを備えた Windows PC として発売される。NVIDIA は、120B パラメータの LLM を最大 100 万トークンのコンテキストでローカル実行できると説明している^[4,11-13]。
- ・【確認済み】 LLM-jp-4.1（2026 年 9 月 28 日公開）は 8B／32B-A3B／33B の thinking モデルで、全モデル Apache 2.0。改良点は STEM 性能の向上、ツール呼び出しへの新規対応、回答の簡潔化であり、後継の LLM-jp-4.2 も開発中である^[1,14]。
- ・【確認済み】 特許文書を事前学習に明示的に用いたと一次情報で確認できた国産オープンモデルは、Stockmark（stockmark-13b：特許庁の特許文書 348 億トークン、stockmark-100b：375 億トークン）のみであった^[15,16]。
- ・【分析】 128GB 機のスイートスポットは、30B 前後の Dense（FP8/Q8）とアクティブ 3B 程度の MoE である。100B 級 Dense は載るが遅く（理論上限でも毎秒数トークン）、日常業務には向かない^[8,17]。
- ・【分析】 国産モデルの最大の利点は、知識性能そのものより「学習データの透明性・ライセンスの明快さ・説明責任」にある。純粋な性能では、同規模の海外オープンモデル（Qwen3.5-27B、Gemma 4 31B）が JMMLU や GPQA で上回る（ELYZA モデルカードの比較表）^[2]。

1. ハードウェアの確認

1-1. 製品像（RTX Spark と DGX Spark の違い）

表 1 RTX Spark と DGX Spark の比較

項目	RTX Spark	DGX Spark
位置づけ	Windows PC 向けスーパーチップ。ノート・小型デスクトップを OEM 各社が販売（ASUS、Dell、HP、Lenovo、Microsoft Surface、MSI、Acer、GIGABYTE が後続） ^[4]	NVIDIA 製の小型 AI 開発機 ^[5]
チップ	「RTX Spark superchip」（型番 N1X は販売店表記・二次情報 ^[18] 。GB10 と同一ダイかは公式には未確認）	GB10 Grace Blackwell Superchip ^[5]
CPU	20 コア Grace（Arm v9.2、Cortex-X925+A725 各 5 コア×2 クラスター）=ARM64 ^[7,13]	20 コア Arm（X925×10+A725×10） ^[19]

項目	RTX Spark	DGX Spark
GPU	Blackwell RTX、6,144 CUDA コア、第 5 世代 Tensor コア (FP4) ^[4]	Blackwell、第 5 世代 Tensor コア (FP4) ^[5]
メモリ	最大 128GB 統合 LPDDR5X、256bit、 300GB/s ^[7]	128GB LPDDR5X、256bit、 273GB/s ^[5,8]
AI 性能	FP4 で最大 1PFLOP ^[4]	FP4 で最大 1PFLOP (スパース時) ^[5]
OS	Windows 11 (Arm)	DGX OS (Linux ベース)
発売	2026 年 5 月 31 日発表、2026 年秋発売 ^[4]	2025 年発売済み

- ・ **名称の注意**：「RTX Spark」は実在する正式名称である。ただし「DGX Spark の Windows 版」という説明は不正確である（OS・帯域・販売形態が異なる）。「600GB/s」という報道値は CPU-GPU 間の NVLink-C2C 帯域でありメモリ帯域ではない。273GB/s は DGX Spark の値を流用したものである^[6,20]。
- ・ **【未確認】** RTX Spark で動く Linux（ネイティブ／WSL2）での CUDA 動作、各 OEM 機の電力・冷却設計（Surface RTX Spark Dev Box は 100W 枠とする二次情報あり^[13]）。

1-2. 128GB で動かせる規模の目安【分析・推定】

重みの必要メモリは、おおむね「パラメータ数×1 パラメータあたりのバイト数」（BF16=2、FP8/Q8=約 1、NVFP4/Q4_K_M=約 0.55~0.6）で見積もれる。OS と常駐アプリで 10~15GB 程度を差し引き、実際に使えるのは約 100~110GB と見込む^[8,17]。

表 2 量子化別の必要メモリと 128GB 機での判断（推定）

モデル規模	BF16	FP8/Q8	NVFP4/Q4	128GB 機での判断
8B	約 17GB	約 9GB	約 5GB	余裕。複数同時起動も可
32B-A3B (MoE)	約 64GB	約 34GB	約 19GB	余裕。高速
33B Dense	約 66GB	約 35GB	約 20GB	FP8/Q8 推奨 （長文 KV の余地あり）
120B (gpt-oss 系、MXFP4)	—	—	約 63GB 前後	実用可（MoE のため速い）
100B Dense (Stockmark-2)	約 200GB（不可）	約 100GB（ほぼ限界）	約 55~60GB	Q4 なら可。ただし遅い
671B MoE (Rakuten AI 3.0)	不可	不可	約 380GB 超	不可

- ・ **KV キャッシュ（推定）**：LLM-jp-4 33B（64 層）で GQA（KV ヘッド 8、ヘッド次元 128）と仮定すると、FP16 で約 0.26MB/トークン、64K トークンで約 17GB となる。FP8/Q8 の重み（約 35GB）と合わせても 60GB 以内に収まる（KV ヘッド数はモデルカードで要確認）。

- ・ **生成速度（推定）**：デコード速度の上限は「帯域÷1 トークンあたりに読み出す重みのバイト数」で概算できる。300GB/s の場合、33B Dense Q8（約 35GB）で約 8 トークン/秒、Q4（約 20GB）で約 15 トークン/秒、32B-A3B MoE で数十トークン/秒が上限となる。
- ・ **参考実測（DGX Spark、273GB/s）**：gpt-oss-120b（MXFP4、llama.cpp）について、llama.cpp の DGX Spark 性能スレッド（2025 年 10 月 14 日開始）内の利用者報告（2025 年 11 月 2 日、カーネル更新後の Fedora 環境）では、生成 60.6 トークン/秒、32K コンテキスト時 40.6 トークン/秒、プリフィル 1,956 トークン/秒が示されている^[21]（NVIDIA 自身の公表値は 55.4 トークン/秒とされるが、出典 URL は未確認）。Llama 3.1 70B FP8 では、デコード 2.7 トークン/秒、プリフィル 803 トークン/秒（SGLang 使用）が報告されている^[19]。いずれも RTX Spark（Windows）での実測ではない。
- ・ **長文入力（プリフィル）**は計算量で決まるため、Blackwell の演算性能が効く。それでも明細書全文（数万トークン）を投入すると、最初の応答まで数十秒～分単位かかる可能性がある（推定）。

1-3. 推論ソフトの対応状況

表 3 推論ソフトの対応状況

ソフト	状況	区分
llama.cpp（GGUF）	NVIDIA が RTX/DGX 向けの最適化を公表。Windows ARM64 版は Clang（clang-cl）でのビルドが必要 ^[18,22,23]	確認済み
Ollama／LM Studio	NVIDIA が llama.cpp バックエンドの「最適化済みの簡単なアプリ」として推奨。最適化は LM Studio・Ollama 経由でも提供 ^[18,24,25]	確認済み（Windows ARM64 ネイティブ版の提供状況は導入時に要確認）
vLLM	NVIDIA の案内では Linux が前提。Windows（RTX Spark）での直接利用は未確認 ^[24]	要注意
CUDA	CUDA Toolkit 13.4 developer preview が Windows 11 on ARM64 を初めて対象化 ^[23,26]	二次情報
TensorRT-LLM／NVIDIA NIM	NVIDIA は RTX Spark に TensorRT を含むと説明 ^[11] 。TensorRT-LLM・NIM の Windows ARM64 対応は未確認	未確認

2. 国産 LLM 候補の比較

2-1. 重み公開モデル

ローカル実行の前提となる「重み公開」モデルを中心に整理した。表中の「知財業務での適性」は本稿の分析である。

表 4 国産 LLM 候補（重み公開モデル）の比較

モデル（最新版）	開発元	規模・構造	ベース出自	ライセンス	128GB での実行目安	知財業務での適性（分析）
llm-jp-4.1-33b-thinking (2026/9/28) ^[1,27]	NII LLMC	約 332 億・Dense、64K	フルスクラッチ (約 11.7 兆トークン)	Apache 2.0	FP8/Q8 (約 35GB) 推奨	◎ 主力候補。学習データ公開で説明責任を果たしやすい
llm-jp-4.1-32b-a3b-thinking ^[1]	NII LLMC	約 320 億・MoE（アクティブ約 3B）	フルスクラッチ	Apache 2.0	Q8/BF16 も可、高速	○ 大量文書の一次要約・分類向け
llm-jp-4.1-8b-thinking／LLM-jp-4-VL 9B ^[1,28]	NII LLMC	約 86 億・Dense／VLM 約 9B	フルスクラッチ	Apache 2.0	余裕	△ 補助（図面・画像入り文書の試行は VL）
ELYZA-Thinking-1.0-llm-jp-4-33b (2026/10/2) ^[2,29]	ELYZA	33B Dense、推論モデル	llm-jp-4-33b-base に中間学習・事後学習	Apache 2.0	FP8/Q8 推奨。コミュニティ版 GGUF あり ^[30]	◎ 日本語の指示追従・知識想起が強い（M-IFEval-ja 85.84、JFBench 39.02、JEMHopQA 73.85）
ELYZA-Thinking-1.0-llm-jp-4-32b-a3b ^[29]	ELYZA	32B MoE	llm-jp-4-32b-a3b-base	Apache 2.0	高速	○
GPT-OSS Swallow 120B／20B (2026/2/20) ^[3]	東京科学大学・産総研	120B／20B MoE、推論型	OpenAI gpt-oss の継続学習	Apache 2.0（ベースの規約も要確認）	120B は約 60GB 台で実用可	○ 難しい推論・クレーム対比のセカンドオピニオン。英語タスク平均 0.804
Qwen3 Swallow 8B／30B-A3B／32B ^[31]	同上	Dense／MoE、推論型	Alibaba Qwen3 の継続学習	Apache 2.0	余裕（GPTQ 版は繰り返し生成の問題で公開停止）	○
Stockmark-2-100B-Instruct ^[32]	ストックマーク	約 960 億・Dense	フルスクラッチ (約 2.0 兆トークン：英 60%・日 30%・コード 10%)	MIT	Q4 (約 55～60GB) で可、遅い	○ ビジネス・技術文書。特許データ使用の有無は公式記載なし
stockmark-13b／stockmark-100b (旧世代) ^[15,16]	ストックマーク	13B／100B	フルスクラッチ。特許庁の特許文書で事前学習（13b：348 億、100b：375 億トークン）	商用可（13b）／要確認（100b）	可	△ 特許用語への親和性はあるが世代が古い

モデル（最新版）	開発元	規模・構造	ベース出自	ライセンス	128GB での実行目安	知財業務での適性（分析）
Stockmark-DocReasoner-Qwen2.5-VL-32B (2026/4) ^[33]	ストックマーク	32B VLM	Qwen2.5-VL ベース	要確認	FP8 版あり	○ 図表入り文書の読解（公報 PDF・図面の試行）
PLaMo 3 NICT 2604 31B Base (2026/6) ^[34-37]	PFN×NICT	31B（120B から蒸留）、SWA 混合	フルスクラッチ	PLaMo Community License（商用は PFN への登録・連絡が必要。PFN の責任上限 500 円、業務利用時は免責）	可（非公式 GGUF あり）	△ ベースモデル（指示調整なし）。チャット用途にはそのまま使えない
plamo-2-translate ^[38,39]	PFN	翻訳特化	フルスクラッチ	PLaMo Community License	可	○ 日英翻訳（商用は申請が必要）。plamo-3-translate は HF 公開予定（未公開）
Sarashina2.2-0.5B/1B/3B-Instruct ^[40,41]	SB Intuitions	小型	フルスクラッチ	MIT	余裕	△ 前処理・分類など軽作業
Rakuten AI 3.0 ^[10,42]	楽天	671B MoE（アクティブ 37B）、128K	DeepSeek V3 ベース（経緯は 4-3 参照）	Apache 2.0（+DeepSeek の MIT 表示）	不可	× 本機では対象外

2-2. 重み非公開（オンプレ／API 提供のみ）

- ・ **tsuzumi 2（NTT）**：30B、フルスクラッチ、1GPU で動作し、オンプレミス／プライベートクラウド／Azure で提供される。重みの一般公開はなく、RTX Spark への自力導入はできない（NTT との個別契約による提供形態は要相談）^[43-45]。
- ・ **PLaMo 3.0 Prime（PFN）**：2026 年 6 月 22 日に正式リリースされ、API・チャット経由で提供されている。重みは非公開である^[9]。
- ・ **cotomi（NEC）、Takane（富士通）、CyberAgent CALM、KARAKURI、ABEJA（ABEJA-QwQ32b-Reasoning 等）、Kotoba、Sakana AI**：今回の調査では最新版の一次情報を確認できなかった【未確認】。候補とする場合は、各社公式ページで重み公開の有無・ベースモデル・ライセンスを個別に確認する。

2-3. 特許・知財分野での評価実績（確認できた範囲）

- ・ Katsuki Chousa・Ryosuke Sugiura による「JAPAS: A Benchmark and Neural Approach for Japanese Patent Support Relation Extraction」（LREC 2026、pp.4188–4198）は、サポート要件（請求項と明細書記載の対応関係）の検証を対象に、日本の特許について 2,000 件超の事例を人手で注釈した初の公開ベンチマークである。最高値はファインチューニングした Qwen3-14B の F1 0.50 で、国産モデルの結果は示されていない^[46]。
- ・ 特許庁委託の調査報告書「特許情報の機械翻訳における大規模言語モデルの利用可能性に関する調査報告書」（日本特許情報機構、2026 年 3 月 19 日）がある（数値は本稿では未抽出）^[47]。
- ・ 弁理士試験を用いた学術的ベンチマークは見つからなかった（個人ブログの試行のみで、参考程度）。

3. 知財業務向けの推奨構成

3-1. 用途別のおすすめ【分析】

表 5 用途別のおすすめモデルと運用のポイント

用途	第一候補	補助・比較用	運用のポイント
長文明細書の要約	llm-jp-4.1-33b-thinking (Q8)	32B-A3B（大量処理）	64K を超える文書は章ごとに分割して要約→統合。「段落番号付き引用」を必須にする
クレーム対比・構成要件分説	ELYZA-Thinking-1.0-llm-jp-4-33b	GPT-OSS Swallow 120B	2 モデルの結果が食い違う箇所を人間が重点確認する「クロスチェック運用」
拒絶理由・意見書の検討補助	llm-jp-4.1-33b-thinking	GPT-OSS Swallow 120B	審査基準・判例は RAG で最新版を与える（モデルの知識に頼らない）
日英翻訳（明細書・引用文献）	plamo-2-translate（商用申請後）	LLM-jp-4.1/Swallow	訳語統一のための用語集をプロンプトに常時注入
先行技術調査・FTO・無効資料の補助	RAG + 32B-A3B	33B で最終判定補助	検索式・分類（FI/CPC）の提案と、スクリーニングの一次判定に限定する
IP ランドスケープ	32B-A3B（大量分類）	33B（考察）	集計そのものは表計算・BI で行い、LLM はラベリングと解釈に使う
発明提案書・社内文書の整理	ELYZA-Thinking 33B	8B（定型処理）	社内テンプレートを few-shot で与える

3-2. RAG・埋め込み（国産）

- ・ **Ruri v3（名古屋大学）**：30M～310M、最大入力 8,192 トークン、JMTEB 平均 75.85（310m）。軽量で CPU でも動くため第一候補である（ライセンスは HF で要確認）^[48–50]。

- ・ **sarashina-embedding-v2-1b (SB Intuitions)** : JMTEB 平均 76.38 と高性能だが、「Sarashina Model NonCommercial License」のため顧客業務には使えない。Sarashina3 embedding (78.22) は API 提供のため、オフライン用途の対象外である^[48,51]。
- ・ **構成例【分析】** : Ollama/LM Studio (または llama-server) をローカル API として立て、Ruri v3 で公報・社内 DB をベクトル化する。ローカルのベクトル DB で検索し、33B モデルで回答を生成する。回答には必ず「公報番号・段落番号」を返させる。

3-3. 国産と海外オープンモデルの位置づけ【分析】

- ・ **国産を選ぶ理由** : 学習データとコーパスの公開 (LLM-jp)、Apache 2.0/MIT のシンプルな条件、日本語の指示追従、顧客・社内への説明のしやすさ (データ主権・経済安保の観点)。
- ・ **国産を選ばない (または併用する) 理由** : 知識・数学・コードでは同規模の海外モデルが上回る。ELYZA モデルカードの比較表では、JMMLU が ELYZA-Thinking 33B の 84.69 に対し Qwen3.5-27B が 90.11、Gemma 4 31B が 90.17、GPQA (ja) が 60.97 に対し 79.07/79.69 である^[2]。
- ・ **結論** : 「国産フルスクラッチ (LLM-jp 系) を主力とし、海外ベースの国産継続学習モデル (Swallow) または海外モデルを検証用に併用する」構成が、説明責任と精度を両立する現実的な選択である。

4. 注意事項

4-1. ライセンス・規約

- ・ 継続学習モデルは、ベースモデルの規約も引き継ぐ前提で確認する (Qwen3→Apache 2.0、gpt-oss→Apache 2.0+OpenAI の利用ポリシー、DeepSeek V3→MIT 表示)^[3,10,31]。Apache 2.0 では、再配布時にライセンス文・NOTICE の同梱が必要である。
- ・ PLaMo Community License は、商用利用に PFN への登録・連絡が必要で、業務利用時の PFN 免責規定がある。社内利用も商用扱いになるとの指摘 (二次情報) もあるため、顧客業務で使う前に PFN への確認を推奨する^[35-37,39]。
- ・ Sarashina の埋め込みモデル・TTS は非商用ライセンスである。同じシリーズでもモデルごとにライセンスが分かれている (MIT と非商用)^[41,52]。
- ・ 出力を別モデルの学習に再利用する場合、学習データ生成に使った上流モデルの規約が問題となり得る (法的評価は定まっていない)。

- ・ GGUF の多くは非公式のコミュニティ変換である（mmnga-o、nori-sdc 等）。元のライセンスがそのまま適用されるうえ、変換品質の保証はない^[30,36,53]。

4-2. 情報セキュリティ

- ・ **完全オフライン化**：モデル取得後はネットワークを物理的または論理的に遮断する。Ollama、LM Studio、OpenShell、Windows のテレメトリ・自動更新・モデル自動取得を無効化し、ファイアウォールで外向き通信を拒否する。
- ・ **真正性**：取得元は Hugging Face の公式組織（llm-jp、elyza、tokyotech-llm、stockmark、pfnet、sbintuitions 等）に限り、ハッシュを記録する。手元で量子化できる場合は公式の safetensors から自分で変換する。
- ・ **ファイル形式**：pickle 形式（.bin／.pt）は任意コード実行のリスクがあるため使わず、safetensors または GGUF に限定する。trust_remote_code を要求するモデル（PLaMo、ELYZA の vLLM プラグイン等）は、コードを確認してから実行する。
- ・ **ログ管理**：プロンプトと出力には未公開発明が含まれる。チャット履歴・キャッシュの保存場所を把握し、BitLocker 等で暗号化したうえで、保存期間と削除手順を定める。

4-3. 「国産」の定義

- ・ 「国産」には、フルスクラッチ（LLM-jp、PLaMo、Stockmark-2、tsuzumi 2、Sarashina）と、海外ベースの継続学習（Swallow、Rakuten AI 3.0、Stockmark-DocReasoner）が混在する。Rakuten AI 3.0 は、2026 年 3 月 19 日の ITmedia 取材にはベースモデルを「非開示」としたが、3 月 27 日に楽天広報が DeepSeek のオープンモデルを採用したと回答し、当初明示しなかったことが問題視された^[42]。
- ・ ELYZA についても、ITmedia AI+が 2026 年 10 月 2 日の記事で用いた「完全国産」の見出し・本文を同日中に修正しており、ELYZA は日本語訳のニュアンスが想定と異なると説明している^[54]。「完全国産基盤」という表現が指す範囲については解説記事もある^[29]。顧客に「国産」と説明する際は、「事前学習の出自」まで明示する。
- ・ 経済安保・データ主権の観点で本当に重要なのは「重みを手元で動かし、データが外に出ないこと」であり、ローカル運用ならベースモデルの国籍によるリスクは限定的である（分析）。ただし、取引先の調達規程が「国産」を要件にしている場合は定義の確認が必要である。

4-4. 精度・品質

- ・ ハルシネーション（存在しない公報番号・段落・判例の捏造）、クレーム解釈や均等論・進歩性判断の誤りが起こり得る。モデルは最新の法改正・審査基準・判例を知らない（知識は学習時点で止まっている）。
- ・ 長文では性能が落ちる。64K コンテキストに収まっていても、中盤の記載を見落とす可能性がある。分割処理と原文の引用を義務付ける。
- ・ 量子化（特に Q4 以下・GPTQ）で品質が落ちる場合がある。Swallow では、GPTQ 版が繰り返し生成の問題で公開停止となった^[31]。重要な判断には FP8/Q8 以上を使う。
- ・ 最終判断は必ず弁理士・専門家がレビューし、LLM の出力は「下書き・論点抽出」と位置付ける。

4-5. 弁理士法・守秘義務・契約（推測を含む）

- ・ 弁理士には秘密保持義務がある（弁理士法 30 条）。ローカル運用は外部送信リスクを下げるが、端末の紛失・共有やログ管理の不備は依然として漏えいの原因になる（本稿では条文本文を再確認していない）。
- ・ 顧客との秘密保持契約・業務委託契約に「生成 AI の利用」「第三者ツールの利用」の制限条項がないかを確認し、必要なら利用の明示と同意を得る（推測）。
- ・ 弁理士資格を持たない者が、他人の求めに応じ報酬を得て特許の鑑定・意見書作成等を業として行くと、弁理士法上の業務制限（75 条）に抵触するおそれがある。LLM を使っても線引きは変わらないと考えるのが安全である（法的評価は個別判断。推測）。

4-6. ハードウェア面

- ・ 300GB/s のメモリ帯域が速度の上限となる。100B 級 Dense や長文の多並列処理は遅くなる^[7,8,55]。
- ・ Windows on Arm では、x86 前提の Python ライブラリ・CUDA 拡張が動かない場合がある。vLLM は Linux 前提である^[12,24]。初期のドライバは成熟度が低いとの報告もある（二次情報）^[6]。
- ・ 薄型ノートでは、長時間の推論で発熱・サーマルスロットリングが起き、速度が低下するおそれがある。常時稼働させるなら小型デスクトップ型（例：ASUS ProArt GR1X）が有利と考える（推定）。

5. 未確認事項・留意点

- ・ RTX Spark の最終仕様（OEM 機ごとのメモリ速度・電力）、および搭載チップと GB10 の同一性は未確認である。300GB/s は NVIDIA Porting Guide v0.1.0 の記載値である^[7]。
- ・ LLM-jp-4.1 の各ベンチマーク値は ELYZA モデルカードの比較表に基づく（LLMC 自身のスコア一覧は今回取得していない）^[2]。GPT-OSS Swallow 120B の日本語タスク平均 0.642、日本語 MT-Bench 0.916（Swallow チームがこれまで評価した全 LLM で最高タイ。GPT-5.1 Thinking は 0.897）、英語平均 0.804 は、Swallow LLM 公式ページで確認済みである^[3]。
- ・ Ruri v3 のライセンス、Stockmark-2 における特許データ使用の有無、stockmark-100b のライセンス、Sarashina3 mini/nano の重み公開状況、cotomi・Takane・CALM・KARAKURI・Kotoba・Sakana AI の最新版は未確認である。
- ・ RTX Spark 上の推論速度は実測値ではなく、帯域からの推定と DGX Spark での第三者実測値に基づく。導入前に、実機で自社・自事務所の事件データを使ったベンチマーク（要約の正確性、引用の正確性、処理時間）を行うことを強く推奨する。

参考文献

番号は本文での初出順。閲覧日はいずれも 2026 年 10 月 10 日。

- [1] 国立情報学研究所 大規模言語モデル研究開発センター（LLMC）「大規模言語モデル『LLM-jp-4.1』シリーズを公開」、2026 年 9 月 28 日。 <https://llmc.nii.ac.jp/topics/llm-jp-4-1/>
- [2] ELYZA「elyza/ELYZA-Thinking-1.0-llm-jp-4-33b（モデルカード）」、Hugging Face。
<https://huggingface.co/elyza/ELYZA-Thinking-1.0-llm-jp-4-33b>
- [3] Swallow LLM（東京科学大学・産業技術総合研究所）「GPT-OSS Swallow」、2026 年 2 月 20 日。
<https://swallow-llm.github.io/gptoss-swallow.ja.html>
- [4] NVIDIA「NVIDIA and Microsoft Reinvent Windows PCs for the Age of Personal AI」、NVIDIA Newsroom、2026 年 5 月 31 日。 <https://nvidianews.nvidia.com/news/nvidia-microsoft-windows-pcs-agents-rtx-spark>
- [5] NVIDIA「Personal AI Supercomputer Powered by Blackwell（DGX Spark 製品ページ）」、NVIDIA。
<https://www.nvidia.com/en-us/products/workstations/dgx-spark/>
- [6] LeCompute「RTX Spark vs DGX Spark: specs, memory, price (2026)」。
<https://lecompute.fr/en/silicon/rtx-spark-vs-dgx-spark/>
- [7] NVIDIA「System Overview — NVIDIA Windows on Arm Porting Guide（v0.1.0）」、NVIDIA Docs。
<https://docs.nvidia.com/rtx-spark/rtx-spark-porting-guide/0.1.0/overview.html>
- [8] GPUWerk「DGX Spark specs: GB10, 128 GB memory, 273 GB/s explained」。
<https://gpuwerk.com/dgx-spark/specs/>
- [9] Oflight「PLaMo 3.0 Prime Deep Dive — Preferred Networks' Flagship Japanese LLM Officially Released June 22, 2026 …」。
<https://www.oflight.co.jp/en/columns/plamo-3-0-prime-pfn-japanese-llm-2026-06>

- [10] 楽天グループ「Rakuten/RakutenAI-3.0（モデルカード）」、Hugging Face。
<https://huggingface.co/Rakuten/RakutenAI-3.0>
- [11] AI Made Tools「NVIDIA RTX Spark: Complete Guide to the AI-First Windows PC (2026)」。
<https://www.aimadetools.com/blog/nvidia-rtx-spark-complete-guide/>
- [12] OfflineCreator「NVIDIA RTX Spark Local AI: 128GB, Ollama & ComfyUI」。<https://offlinecreator.com/rtx-spark-local-ai>
- [13] Wikipedia「Nvidia RTX Spark」、Wikipedia（英語版）。https://en.wikipedia.org/wiki/Nvidia_RTX_Spark
- [14] PC Watch（インプレス）「NII、ツール呼び出し対応・STEM 強化の『LLM-jp-4.1』公開」。
<https://pc.watch.impress.co.jp/docs/news/2144150.html>
- [15] ストックマーク「stockmark/stockmark-100b（モデルカード）」、Hugging Face。
<https://huggingface.co/stockmark/stockmark-100b>
- [16] Kosuke Takahashi, Takahiro Omi, Kosuke Arima, Tatsuya Ishigaki「Pretraining and Updates of Domain-Specific LLM: A Case Study in the Japanese Business Domain」、arXiv:2404.08262、2024 年 4 月 12 日投稿（v3：2024 年 11 月 6 日）。<https://arxiv.org/abs/2404.08262>
- [17] Kubesimplify「Day 3: DGX Spark Unpacked. GB10, Unified Memory, sm_121, and NVFP4」。
<https://blog.kubesimplify.com/day-3-the-dgx-spark-unpacked-gb10-unified-memory-sm-121-and-the-one-reason-this-hardware-exists>
- [18] igor'sLAB「NVIDIA RTX Spark N1X: ARM Chip with 128 GB Memory」。<https://www.igorslab.de/en/nvidia-rtx-spark-n1x-october-arm-chip-blackwell-graphics-128-gb-unified-memory/>
- [19] Jerry Zhou, Richard Chen（LMSYS Org）「NVIDIA DGX Spark In-Depth Review: A New Standard for Local AI Inference」、LMSYS Blog、2025 年 10 月 13 日。<https://www.lmsys.org/blog/2025-10-13-nvidia-dgx-spark/>
- [20] DataHardware.ai「RTX Spark vs Strix Halo: Buy Now or Wait for October?」。<https://datahardware.ai/blog/rtx-spark-vs-strix-halo-2026>
- [21] ggml-org/llama.cpp「Performance of llama.cpp on NVIDIA DGX Spark（Discussion #16578）」、GitHub、2025 年 10 月 14 日開始。本文の数値は同スレッド内の 2025 年 11 月 2 日の利用者投稿。
<https://github.com/ggml-org/llama.cpp/discussions/16578>
- [22] NVIDIA Technical Blog「Accelerating LLMs with llama.cpp on NVIDIA RTX Systems」。
<https://developer.nvidia.com/blog/accelerating-llms-with-llama-cpp-on-nvidia-rtx-systems/>
- [23] Conan Blog「CUDA meets Windows on ARM64: getting your applications ready for NVIDIA RTX Spark with Conan」、2026 年 7 月 30 日。
<https://blog.conan.io/cuda/gpu/armv8/arm64/windows/conan/nvidia/jetson/jetpack/2026/07/30/CUDA-Meets-Windows-On-Arm64-NVIDIA-RTX-Spark.html>
- [24] NVIDIA「How to Get Started With Large Language Models on …（RTX 向け LLM 導入ガイド）」、NVIDIA Build。<https://build.nvidia.com/rtx/llms>
- [25] Runtime Wire「NVIDIA says new optimizations make local agents up to 1.9x faster」。
<https://runtimewire.com/article/nvidia-local-agent-inference-llama-cpp-vllm-blackwell>
- [26] NVIDIA Developer「Port Windows Apps to ARM: RTX Spark Guide」。
<https://developer.nvidia.com/topics/ai/local-ai/port-apps>
- [27] Claypier「国産 LLM『LLM-jp-4 33B』公開 332 億パラメータの Dense 型、Apache 2.0 で商用利用も可」。
<https://claypier.com/llm-jp-4-33b-open-model/>
- [28] LLM-jp「LLM-jp-4-VL に関する公式ブログ記事」。※URL 未確認

- [29] AI Revolution 「ELYZA-Thinking-1.0 とは？ LLM-jp-4 ベースの国産推論モデル・33B/32B-A3B の性能・使い方・LLM-jp-4.1 比較【2026 年 10 月】」。 <https://ai-revolution.co.jp/media/what-is-elyza-thinking/>
- [30] mmnga-o 「ELYZA-Thinking-1.0-llm-jp-4-33b-gguf（コミュニティ変換版）」、Hugging Face。
<https://huggingface.co/mmnga-o/ELYZA-Thinking-1.0-llm-jp-4-33b-gguf>
- [31] Swallow LLM（東京科学大学・産業技術総合研究所）「Qwen3 Swallow」。 <https://swallow-llm.github.io/qwen3-swallow.ja.html> <https://swallow-llm.github.io/qwen3-swallow.en.html>
- [32] ストックマーク「stockmark/Stockmark-2-100B-Instruct（モデルカード）」、Hugging Face。
<https://huggingface.co/stockmark/Stockmark-2-100B-Instruct>
- [33] ストックマーク「Stockmark-DocReasoner-Qwen2.5-VL-32B：ペルソナに基づくデータ合成を用いた日本語ドキュメント読解のための Chain-of-Thought 視覚言語モデル」、Stockmark Tech Blog、2026 年 4 月 28 日。
<https://stockmark-tech.hatenablog.com/entry/2026/04/28/100048>
- [34] Preferred Networks「pfnet/plamo-3-nict-2604-31b-base（モデルカード）」、Hugging Face。
<https://huggingface.co/pfnet/plamo-3-nict-2604-31b-base>
- [35] Preferred Networks「pfnet/plamo-3-nict-8b-base（PLaMo Community License 掲載）」、Hugging Face。
<https://huggingface.co/pfnet/plamo-3-nict-8b-base>
- [36] nori-sdc「PLaMo-3-NICT-2604-31B-Base-GGUF（非公式変換版）」、Hugging Face。
<https://huggingface.co/nori-sdc/PLaMo-3-NICT-2604-31B-Base-GGUF>
- [37] coplus-one.jp「ローカル LLM 日本語 | 国産モデルは選ぶべき？一覧と判断表」。 <https://coplus-one.jp/local-llm-nihongo/>
- [38] Preferred Networks「plamo-3-translate: フロンティア級の翻訳特化 LLM」、PFN Blog。
<https://www.preferred.jp/ja/blog/tech/plamo-3-translate-product-release>
- [39] XEXEQ「Preferred Networks が PLaMo 翻訳モデルを公開、多様なモデルと商用利用に関する注意点」。
<https://xexeq.jp/blogs/media/topics48611>
- [40] Spike Studio「【2026 年版】国産 LLM15 選 | 日本企業の日本語モデルを商用利用の可否で比較」。
<https://spikestudio.jp/blog/japan-llm-companies-2026>
- [41] ISSOH「Sarashina とは？モデル一覧・使い方・ライセンス | TTS/OCR/API 提供形態」。
<https://www.issoh.co.jp/tech/details/11856/>
- [42] AI Revolution「楽天 Rakuten AI 3.0=DeepSeek V3 リブランド炎上のその後 | GENIAC 補助金・国産 LLM 定義問題の現在地【2026 年 7 月続報】」。 <https://ai-revolution.co.jp/media/rakuten-ai-3-deepseek-scandal/>
- [43] AI Revolution「NTT tsuzumi 2 とは？政府 Gennai 基盤採用の日本語最高性能 LLM を徹底解説」。 <https://ai-revolution.co.jp/media/what-is-tsuzumi-2/>
- [44] Impress Watch「NTT『tsuzumi 2』提供開始 30B で強力な金融・医療知識」。
<https://www.watch.impress.co.jp/docs/news/2056477.html>
- [45] 日経クロステック「NTT が純国産 LLM の新版『tsuzumi 2』提供開始、30B パラメーターで日本語性能を強化」。 <https://xtech.nikkei.com/atcl/nxt/news/24/02892/>
- [46] Katsuki Chousa, Ryosuke Sugiura「JAPAS: A Benchmark and Neural Approach for Japanese Patent Support Relation Extraction」、Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pp.4188–4198、2026 年 5 月。 <https://aclanthology.org/2026.lrec-1.330/> DOI: 10.63317/4yxepyo7gna
- [47] 一般財団法人日本特許情報機構「令和 7 年度 特許情報の機械翻訳における大規模言語モデルの利用可能性に関する調査事業 調査報告書」、特許庁ウェブサイト、令和 8 年（2026 年）3 月 19 日。
https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/document/kikai_honyaku/2025_01.pdf

- [48] SB Intuitions 「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。 <https://www.sbintuitions.co.jp/blog/entry/2026/07/02/101054>
- [49] 塚越駿・笹野遼平 「Ruri: 日本語に特化した汎用テキスト埋め込みモデル」、自然言語処理 33 巻 2 号 (J-STAGE) 。 https://www.jstage.jst.go.jp/article/jnlp/33/2/33_591/_pdf/-char/ja
- [50] gonnta (note) 「『OpenWebUI』埋め込みモデル ruri-v3 を使えるようにしてみる」。 https://note.com/gonnta_poyoko/n/n8d5166f48afd
- [51] SB Intuitions 「sbintuitions/sarashina-embedding-v2-1b (モデルカード)」、Hugging Face。 <https://huggingface.co/sbintuitions/sarashina-embedding-v2-1b>
- [52] tt-tsukumochi (note) 「AI ニュースで英語学習 SB Intuitions、日本語特化の大規模言語モデルベース音声合成システム『sarashina2.2-tts』を公開」。 https://note.com/tt_tsukumochi/n/n5c6a99213859
- [53] nori-sdc 「PLaMo-3-NICT-2604-31B-JP-YUA1-instruct (非公式派生版)」、Hugging Face。 <https://huggingface.co/nori-sdc/PLaMo-3-NICT-2604-31B-JP-YUA1-instruct>
- [54] ITmedia AI+ 「ELYZA-Thinking-1.0 (llm-jp-4 版) の公開に関する記事 (見出し・本文の「完全国産」表記を同日修正)」、2026 年 10 月 2 日。 ※URL 未確認 (記事の特定に至らず)
- [55] iFactory 「NVIDIA DGX Spark Review: Personal AI Supercomputer Verdict」。 <https://ifactoryapp.com/sap-integration/on-prem-ai/nvidia-dgx-spark-review-enterprise>