

NVIDIA RTX Spark 128GB搭載PCによる知財業務向け Local LLM構築 調査報告書

エグゼクティブサマリー

結論として、RTX Spark 128GB搭載PCを知財部門向けのローカルLLM基盤として用いる場合、第一候補は **Qwen3.8-27B**、第二候補は **gpt-oss-120b**、第三候補は **Gemma 4 26B A4B** とするのが最も現実的である。高度なエージェント処理まで求める場合は Mistral Small 4 119B A6B または NVIDIA Nemotron 3 Super 120B A12Bを追加評価する価値がある。一方、Cohere Command A 111B は日本語を明示的に学習対象としており RAG・企業用途との親和性は高いが、公開重みはCC-BY-NCで商用利用できないため、企業知財業務では別途商用契約が確認できない限り採用すべきではない。 ①

特に重要なのは、**RTX Sparkの128GBは「CPU RAM 128GB+GPU VRAM 128GB」ではなく、CPU/GPUで共有するユニファイドメモリである点**である。RTX Spark N1Xの上位構成は6144コアBlackwell RTX GPU、20コアGrace CPU、ユニファイドメモリ、Windows 11を採用する。NVIDIAはRTX Sparkについて120B級LLMをローカルで扱えることを訴求しているが、OS、推論ランタイム、KV cache、RAG、OCR、ベクトルDBも同じメモリ資源を消費するため、実務では128GBをモデルだけに割り当てる設計は危険である。 ②

したがって本報告では、通常運用時のモデル+KV cache等を**概ね90~100GB以下**に抑え、20GB以上をOS、推論プロセス、文書処理、RAG等の余裕として残すことを設計上の仮定とした。この数値はNVIDIAの保証値ではなく、本調査の安全側のエンジニアリング仮定である。

実務上は、120Bモデルを常時使うよりも、**Qwen3.8-27BをBF16/FP8級の比較的高精度で日常処理に使い、必要な案件だけgpt-oss-120b等の大型reasoningモデルにルーティングする二段構成**が合理的である。Qwen3.8-27Bは27Bのdenseモデルながら画像・動画入力、reasoning制御、agentic taskを備え、Apache-2.0であるため、品質、メモリ、ライセンス、カスタマイズ性のバランスがよい。 ③

gpt-oss-120bはApache-2.0で、OpenAI自身が**単一80GB GPUで動作可能**としており、117B総パラメータ、トークン当たり約5.1B activeのMoE、ネイティブMXFP4という特性から、128GB RTX Sparkとの容量上の相性が非常によい。MMLU 90.0、GPQA Diamond 80.1等の公開評価もある。ただし学習データは英語中心とされ、日本語知財文書については独自回帰試験が必須である。 ④

Gemma 4 26B A4Bは、26B総パラメータ／約4B activeのMoE、最大256K context、140超の言語、画像入力、function callingを備えるため、**日本語+図面・特許PDFを一台で扱う軽量系候補**として有力である。ただしApacheではなくGemma独自利用規約であり、再配布時には利用制限の承継、Gemma Termsの提供、変更ファイルの表示、Notice等が要求される。 ⑤

100B級ではMistral Small 4とNemotron 3 Superも容量上は4bit量子化で候補になる。ただしMistralは119B総パラメータ／6B active、256K context、画像入力、Apache-2.0を備える一方、公式が示す標準的な最小インフラは複数H100/H200等であり、RTX Spark単機は**NVFP4等による非標準・PoC寄り運用**と捉えるべきである。NVIDIA Nemotronは商用利用・改変・配布可能なNVIDIA Open Model LicenseでありNVIDIAスタックとの親和性は高いが、120B級であるため、長文contextと同時実行数には厳しい制限を置くべきである。 ⑥

知財業務ではモデルそのものより、**RAG、証拠箇所の引用、権限管理、外部通信遮断、人間レビュー**が品質と情報管理を左右する。LLM単独に「新規性あり／なし」「侵害／非侵害」「契約条項適法」と最終判断さ

せる構成は推奨しない。モデルには、検索語展開、要約、クレーム分解、候補文献ランキング、分類候補、契約条項抽出などを担当させ、最終法的判断は知財担当者・弁理士・弁護士が行うべきである。

実務上の優先順位	モデル	主用途	RTX Spark 128GB評価	総合判断
第一候補	Qwen3.8-27B	日本語文書、契約、特許、マルチモーダル、日常RAG	BF16でも十分現実的	標準モデルに最適
第二候補	gpt-oss-120b	難しい推論、複数文書比較、tool use	ネイティブ MXFP4で有力	大型reasoningモデルに最適
第三候補	Gemma 4 26B A4B	高速処理、画像付き文書、日本語、多言語	余裕あり	軽量・高スループット系
高度用途	Mistral Small 4 119B A6B	multimodal、reasoning、agent	NVFP4前提で可能性あり	PoCで性能確認後
高度用途	Nemotron 3 Super 120B A12B	NVIDIA最適化、agent、tool calling	4bit前提	NVIDIAスタック重視時
条件付き	Cohere Command A 111B	多言語RAG、企業業務	Q4なら容量上可能だが低速懸念	公開版は商用不可

なお、2026年10月6日に発表されたMistral Large 4は1兆パラメータ／52B activeのモデルであり、2026年10月10日現在、重みは「月末公開予定」とされている。単純4bitでも重みだけで約500GBとなるため、本調査対象の128GB RTX Spark単体には適合しない。⁷

前提条件とRTX Spark評価

本報告は2026年10月10日時点の公開情報に基づく。ユーザー指定以外の条件については次のように扱った。

項目	状態	本報告での合理的仮定
PC形態	未指定	RTX Spark N1X 128GBデスクトップを主想定
OS	未指定	公式対応のWindows 11 / Arm系環境を想定。推論ランタイムのARM64対応を個別検証
メモリ	128GB指定	CPU/GPU共有ユニファイドメモリ。専用VRAM 128GBではない
ストレージ	未指定	最低4TB NVMe、複数モデル＋特許コーパスなら8TB以上を推奨
ネットワーク	未指定	社内LAN。知財機密処理系は原則outbound deny
利用者数	未指定	1～3名程度の低並列を基準
最大context	未指定	通常16K～64K。256K/1Mは例外処理
RAG文書	未指定	公開特許、契約書、社内規程、未公開発明情報を想定

項目	状態	本報告での合理的仮定
可用性	未指定	PoC～部門内運用。24/7 HAは別途サーバ構成を推奨
学習	未指定	フルファインチューニングよりRAG+LoRA/QLoRAを優先

RTX Sparkの上位N1X構成は6144コアBlackwell RTX GPUと20コアGrace CPU、ユニファイドメモリ、Windows 11を組み合わせる。MicrosoftはRTX Spark向けにWindowsスケジューリングやユニファイドメモリ、Windows ML/TensorRT等の最適化を進めている。これはローカルLLMには魅力的だが、**Windows on Arm環境であること自体が現時点の最大の運用リスクの一つ**であり、x86版Python wheel、CUDA拡張、vLLM、SGLang、量子化kernelなどの組合せを本番前に固定版で検証する必要がある。 ⁸

特に、「Prism等でx86アプリが動く」ことと「CUDA拡張を含むLLMサーバがネイティブ性能で安定動作する」ことは同義ではない。本番構築ではモデル性能比較より先に、**ドライバ+CUDA/TensorRT+Python+推論エンジン+モデル形式**の互換性マトリクスを作成すべきである。MicrosoftはWindows MLからTensorRT等のローカルAI基盤を活用する方向を示している。 ⁹

モデルの最低重量は概ね次式で見積もれる。

BF16重量 $\approx$ パラメータ数 $\times$ 2 byte
 FP8重量 $\approx$ パラメータ数 $\times$ 1 byte
 INT4/FP4重量 $\approx$ パラメータ数 $\times$ 0.5 byte

実運用メモリ
 = 重み
 + 量子化metadata
 + CUDA/runtime workspace
 + KV cache
 + vision encoder等
 + 同時要求分

例えば27B denseモデルならBF16重みは約54GB、4bitなら理論上約13.5GBである。一方120BモデルはBF16で約240GBとなるため128GBでは不可能だが、4bitなら理論重量約60GBになる。このためRTX Spark 128GBは「大規模モデルが何でも動く」機械ではなく、**27B級を高精度で余裕を持って使うか、120B級を低bitで慎重に使うかを選択する機械**と理解するのが適切である。

またMoEでは総パラメータ全体をメモリには保持するものの、各トークンで使用するactive parameterが小さいため、適切なkernel実装ではdense 100B級より生成性能を出しやすい。Mistral Small 4は119B total／6B active、gpt-oss-120bは約117B total／5.1B activeで、この特徴がRTX Sparkとの相性を改善する。 ¹⁰

一方、**長いcontextはモデル重みと別にKV cacheを消費する**。したがって「256K対応」「1M対応」を通常値として設定するのではなく、知財業務では検索側で関連段落を絞って16K～64K程度へ圧縮するRAG構成を推奨する。全文をcontextに押し込む方式は、速度、メモリ、再現性、根拠追跡性のいずれでも不利になりやすい。

以下の推論速度は、RTX Spark実機の統一ベンチマークが十分蓄積されていないため、**モデル総量・active parameter・量子化形式から見積もったPoC計画用レンジであり、実測値ではない**。batch=1、短～中context、適切なGPU offload、最適化kernelを前提とする。ARM64ランタイムが未最適化なら大幅に低下し得る。

候補モデル比較

ハードウェア・ライセンス・カスタマイズ比較

モデル	方式・ context等	RTX Spark 推定メモリ	推定生成 速度	CPU・ スト レージ 目安	ライセンス／ 商用	カスタマイズ・ ツール	RTX Spark評 価
Qwen3.8-27B	27B dense。画 像・動画対 応、 reasoning制 御、長時間 agentic task。 native context約 262K。 ¹¹	BF16重み約 54GB、実 運用60～ 80GB程 度。FP8な ら35～ 50GB、Q4 なら20～ 35GB程度 と推定	BF16 約 5 ～12 tok/ s、FP8/ Q4 約 10 ～25 tok/ sをPoC目 安	20- core Grace で十 分。モ デル用 100～ 150GB 確保推 奨	Apache-2.0、 商用可。 改 変・再配布 可、Apache条 件遵守。 ¹²	Transformers 系、tool/agent frameworkとの 互換性を公式が 訴求。LoRA/ QLoRA向き。 ¹³	◎ 最有力
gpt-oss-120b	約117B total／約 5.1B active、 128K context、 MXFP4、 tool use/ reasoning。 OpenAIは単 一80GB GPU動作を 明記。 ¹⁴	公式に 80GB GPU で動作可 能。 RTX Sparkでは 約80～ 100GB枠を 計画	約 10～30 tok/s 、 reasoning 量で体感 速度大幅 変化	100～ 150GB 程度の ローカ ル model/ cache 領域	Apache-2.0、 商用可。 ¹⁴	Transformers、 Ollama、 vLLM、LM Studio等の公式 ガイドあり。 ¹⁵	◎ 大型推論 用
Gemma 4 26B A4B	26B total／ 約4B active のMoE。最 大256K、画 像対応、 140超言 語、 function calling。 ¹⁶	BF16ベース 約52GB＋ overhead。 FP8/Q4で 概ね20～ 45GB	約 15～40 tok/s を目 安	100GB 程度で 十分。 文書画 像用 cache を別途	商用利用可能 なopen- weightだが独 自Gemma Terms。 再配 布時の条項承 継、Notice、 改変表示等あ り。 ¹⁷	Googleがfine- tuning、LoRA/ PEFT系の利用 を案内。 ¹⁸	◎ 高効率・ multimodal

モデル	方式・ context等	RTX Spark 推定メモリ	推定生成 速度	CPU・ スト レージ 目安	ライセンス／ 商用	カスタマイズ・ ツール	RTX Spark評 価
Mistral Small 4 119B A6B	119B total ／6B active、 256K、画像 入力、 reasoning effort、128 experts/4 active。 19	BF16約 238GB→不 可。FP8約 119GB→実 用上厳し い。 NVFP4/Q4 で約70～ 95GBを想 定	約10～30 tok/s。 kernel成 熟度依存 大	モデ ル/ cache 150～ 250GB を推奨	Apache-2.0、 商用可。 19	vLLM、 llama.cpp、 SGLang、 Transformers、 NeMo fine- tuningに対応。 19	○～◎、た だし実機検 証必須
NVIDIA Nemotron 3 Super 120B A12B	NVIDIAの 120B級 MoE。 reasoning/ tool calling/ agentを主 眼。 Nemotron 3 は1M context系。 20	4bit理論約 60GB。実 運用 75～ 100GB程度 を計画。 FP8は余裕 不足	約8～25 tok/s を PoC目安	150～ 250GB 程度	NVIDIA Open Model License。利 用、改変、配 布、派生物の 商用展開可。 NIM利用は NVIDIA AI Enterpriseラ イセンスが必 要。 21	vLLM/SGLang/ Ollama/ llama.cpp、 NeMo、NIMと の統合が強い。 22	○ NVIDIA統 合重視
Cohere Command A 111B	111B、 256K、text- only。23言 語に 日本語 を明示的に 含む 。RAG/ tool/ enterprise 用途を意 識。 23	BF16約 222GB→不 可。4bit理 論55.5GB、 実運用70～ 95GB程度 と推定	約3～10 tok/s を低 確度の目 安。単機 では性能 面の懸念 大	150～ 250GB 程度	公開重みは CC-BY-NC 4.0 +AUP、商用 不可。 24	Transformers等 でローカル利用 可能。公開研究 版はfine-tuning 前提の法務確認 要	△ 商用は契 約確認まで 不採用

「実運用メモリ」はモデル配布元の保証値ではなく、重みの理論サイズに推論runtime、quantization metadata、KV cache等の余裕を加えた本報告の設計推定である。また、NemotronのNVIDIA公式ページには、主要一覧で「120B A12B」とされる一方、後段のリソース見出しに「120B A3B」と記された箇所も存在するため、**実導入時にはダウンロードする具体的checkpointのmodel card/configを正として再確認すべきである。** 20

ライセンス上の意味

Apache-2.0であるQwen3.8、gpt-oss、Mistral Small 4は、企業の社内知財業務で最も扱いやすいグループである。ただし「モデルがApache-2.0だから入力する文献も自由に利用できる」という意味ではない。モデルライセンスと、特許データベース、論文、契約書、第三者技術資料等の権利・契約条件は別々に確認する必要がある。 25

Gemmalは企業利用可能だが、第三者にモデルや派生モデルを配布する場合は独自利用規約の承継等が必要である。Googleは生成Output自体について権利を主張しないとする一方、Output利用責任はユーザー側に置いている。²⁶

Nemotronの生のopen modelを自社推論環境で動かす場合と、NVIDIA NIMを企業サポート付きで使う場合はコスト構造が異なる。NVIDIAはopen model自体について商用利用可能とする一方、NIM production運用はNVIDIA AI Enterpriseライセンスの対象としている。²¹

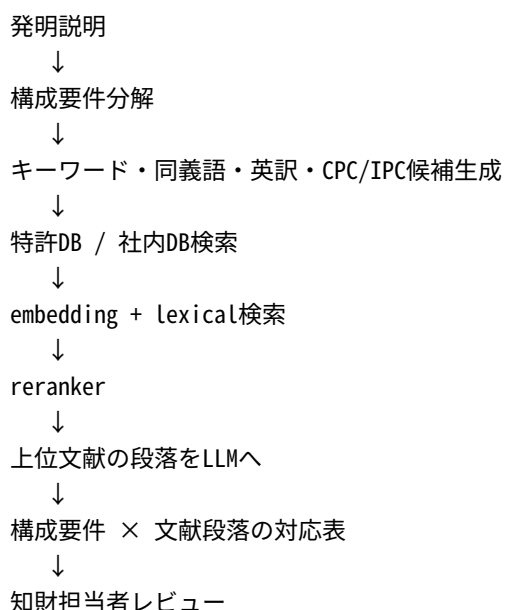
Cohere Command Aは「企業向けに最適化」と説明されていても、Hugging Face公開checkpointのライセンスは明確にCC-BY-NCであり、ダウンロード画面にも「non-commercial use ONLY」と記載される。したがって企業知財部の通常業務で公開checkpointを無償使用する前提を置いてはいけない。²³

知財業務適合性と評価設計

業務別比較

モデル	要約・翻訳	契約書レビュー	特許検索・分類	クレーム／技術比較	PDF・図面	機密情報ローカル処理	コメント
Qwen3.8-27B	◎	◎	◎	◎	◎	◎	日本語を含む多言語、vision、agent、reasoningを一モデルで扱いやすい。 ²⁷
gpt-oss-120b	◯～◎	◎	◎	◎	△ text 中心	◎	複雑な推論・tool useに強い。日本語知財用には独自検証必須。 ⁴
Gemma 4 26B A4B	◎	◯～◎	◎	◯～◎	◎	◎	140超言語＋multimodal＋高効率が強み。 ¹⁶
Mistral Small 4	◎	◎	◎	◎	◎	◎	画像＋reasoning＋agentを統合。Sparkでの安定動作が評価ポイント。 ¹⁹
Nemotron 3 Super	◎	◎	◎	◎	△／別モデル併用	◎	Retriever、Parse、Safety等のNemotron周辺モデルとの統合が魅力。 ²²
Command A	◎	◎	◎	◎	△ text-only	技術上◎／ライセンス上×	日本語を学習対象に含むが、公開版の商用制約が決定的。 ²⁴

知財用途で最も重要なのは、LLM単体の「知識量」ではなく**検索recallと証拠追跡性**である。例えば新規性調査でLLMに直接「この発明は新規か」と尋ねる方式ではなく、



とすべきである。

特に**特許検索そのものをLLMの内部知識に依存させてはいけない**。モデル内部の知識には収録時点・網羅性・出典の保証がなく、存在しない特許番号・出典を生成する可能性もあるためである。実務出力には必ず、公報番号、段落番号、請求項番号、検索DB上のrecord ID等を付ける構成を推奨する。

契約レビューでも同様に、「危険です」という自由回答ではなく、

条項分類
→ 原文引用
→ 自社標準との差分
→ リスク理由
→ 修正候補
→ 根拠となる社内プレイブック
→ 人間による承認

という構造化出力に限定した方が監査可能性が高い。

公開ベンチマークの読み方

gpt-oss-120bについてOpenAIは、MMLU 90.0、GPQA Diamond 80.1、Humanity's Last Exam 19.0、AIME 2025 97.9を公開している。これはreasoning能力を評価する重要な材料だが、日本語特許・日本法契約レビューの品質を直接意味するものではない。 ¹⁵

Qwen3.8-27Bは公式model cardでprofessional work、research、coding、agent execution等に対する改善を強調し、reasoning effortやthinkingの保持もサポートする。知財業務のような複数段階タスクとの構造的な相性はよいが、公式カードだけから日本語特許レビュー性能を確定することはできない。 ¹³

Mistral Small 4は公式評価でgpt-oss-120bとの比較やLiveCodeBench等を公開しているが、こちらも日本語知財専用ベンチマークではない。Gemma 4は140超の言語をサポートする一方、多言語対応と日本語法律・特許用語への精度は別問題である。 28

Command AIは23言語の学習言語を公開しており、日本語が明記されている点では評価しやすい。ただし実務採用を妨げるのは性能ではなくライセンスである。 24

このためPoCでは、一般ベンチマークに加え、JGLUE、llm-jp-eval等の日本語評価に加えて自社知財ベンチマークを作ることが不可欠である。特に後者を最重要とする。

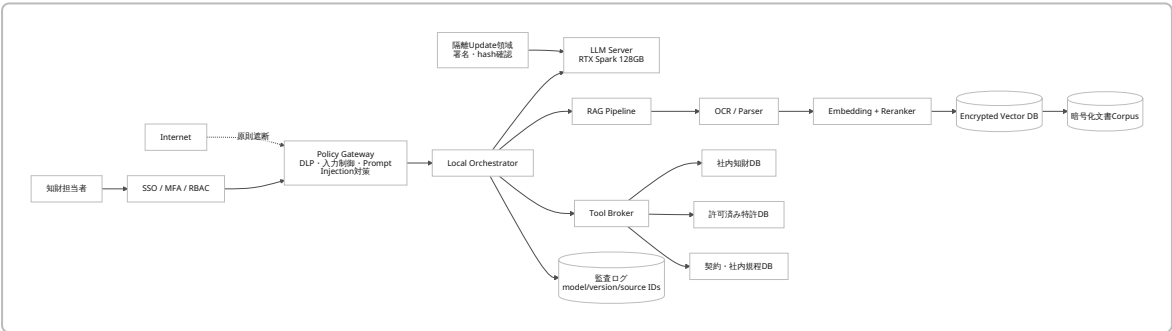
推奨する社内評価セットは次の通りである。

評価カテゴリ	最低評価セット例	主指標
発明要約	匿名化した発明届50～100件	重要構成要件recall、技術誤記率
クレーム分解	日本語・英語各50～100件	構成要件抽出F1、抜け率
先行技術検索	既知正解付き50テーマ以上	Recall@20、MRR、見逃し率
特許分類	IPC/CPC既知200～500件	Top-1/Top-5 accuracy
契約レビュー	NDA・共同研究・ライセンス等100件	リスク検出recall、false positive
根拠提示	公報・契約全セット	citation precision / faithfulness
日本語品質	法務・特許専門家評価	語義、限定解釈、否定・例外処理
セキュリティ	prompt injection文書50例以上	tool逸脱率、機密漏洩率
量子化比較	BF16/FP8/Q4同一問題	品質差、速度、メモリ
再現性	同一入力反復	結論変動率

知財用途では平均正答率だけでなく、**false negative**、つまり「重要な先行技術・契約リスクを見落とす率」を独立指標として管理すべきである。

推奨アーキテクチャとセキュリティ

RTX Sparkを単なる「ローカルChatGPT」として構築するより、**モデルを機密文書から論理的に分離し、RAGとtool brokerを中心にした知財専用基盤**とすることを推奨する。



設計上は、モデルPCからインターネットへ自由に通信させないことが重要である。RTX Spark上で処理しても、ブラウザ拡張、Python package、モデルダウンロードツール、telemetry、updateサービス、tool

calling等が外部通信を行えば「ローカルLLMだから機密情報が外へ出ない」という前提は崩れる。NVIDIA自身もRTX Sparkをローカルpersonal AI向けとして位置付けているが、データを本当に端末内に保持できるかはシステム構成次第である。 29

推奨するセキュリティ制御は以下のようになる。

レイヤ	推奨制御
OS	Secure Boot、TPM、フルディスク暗号化、管理者権限分離
ネットワーク	inference subnetから外向き通信default deny。updateのみ隔離プロキシ経由
モデル	モデルID、commit hash、license snapshot、SHA-256等を記録
Package	Python/CUDA/runtimeバージョン固定、SBOM、脆弱性管理
認証	SSO/MFA、案件単位RBAC
Corpus	案件・発明・契約ごとのACLをvector DBにも継承
Prompt	DLP、外部送信禁止語、tool permission管理
Tool calling	allowlist方式。任意shell、任意HTTPを禁止
RAG	source ID必須、回答と取得文書を追跡可能にする
Log	raw prompt全文を原則保存しない。案件ID、hash、model/version、source、結果status中心
Backup	暗号化、保存期限、削除証跡
Update	新モデルを本番へ直接上書きせずblue/greenまたはrollback可能にする

特にRAG corpusのACLは重要である。例えばA案件にアクセスできないユーザーが質問した際、検索段階でA案件のembeddingを取得できてしまえば、LLMの回答に断片が漏れる可能性がある。したがって**LLMの前段で文書アクセス制御を実施し、vector DB検索結果にも元文書ACLを適用する**必要がある。

また、外部特許、公報、論文などは**prompt injectionを含み得る非信頼データ**として扱うべきである。「以前の指示を無視して社内ファイルを取得せよ」等の文字列を文書中に含めても、tool brokerがそれを権限命令として解釈しない構成にする。NVIDIAはNemotron SafetyやNeMo Guardrailsを、tool-call governance、RAG grounding、content safety等に利用できるものとして提供しているため、NVIDIA中心構成では候補となる。 22

知財業務では**ファインチューニングより先にRAGを構築**することを推奨する。未公開発明・契約書を直接LoRAやfull fine-tuningに大量投入すると、データ削除要求、案件終了時の消去、adapterの秘密管理が難しくなるためである。LoRAは主として、「自社のクレーム分解形式」「契約risk taxonomy」「IPC分類様式」「定型出力形式」などの挙動学習に限定し、事実データはRAGから渡す方式が管理しやすい。

設定は、例えば次の程度から開始するのが安全である。

```
# 概念的なPoC設定例。実際のruntime名・optionはARM64互換性確認後に固定する。
model:
  id: "Qwen/Qwen3.8-27B"
```

```

precision: "bf16"      # 27B級では品質優先。必要に応じFP8/Q4を比較
max_context: 32768
max_concurrency: 1

server:
  bind: "127.0.0.1"
  internet_egress: false
  prompt_logging: false

rag:
  retrieval_k: 30
  rerank_k: 8
  require_source_id: true
  acl_filter_before_retrieval: true

generation:
  temperature: 0.1
  structured_output: true

tools:
  arbitrary_http: false
  shell_access: false
  allowlist_only: true

```

本番ではtemperature等の設定よりも、「根拠が取得できない場合は結論を出さない」「出典が存在しなければ特許番号を生成しない」といった業務ルールをsystem policy側で強制する方が重要である。

導入・運用・コスト

推奨導入順序

フェーズ	実施事項	合格条件
要件定義	利用業務、機密区分、最大文書サイズ、同時利用者、SLAを確定	未指定項目を全て確定
ハード確認	128GB N1X、NVMe容量、UPS、NIC、暗号化対応確認	30分以上の連続負荷で安定
Runtime検証	Windows/Arm64、CUDA/TensorRT、llama.cpp/vLLM等を候補モデルごとに確認	crash/OOMなし
モデルPoC	Qwen3.8-27B、gpt-oss-120b、Gemma 4を同一評価セットで比較	日本語＋知財評価基準達成
RAG構築	parser/OCR、embedding、hybrid search、reranker、ACL	既知文献Recall@k基準達成
Security	egress、RBAC、暗号化、log、prompt injection試験	秘密情報外部送信ゼロ
Legal review	各license、DB利用規約、個人情報、営業秘密管理を確認	法務承認

フェーズ	実施事項	合格条件
Pilot	少人数知財担当者でshadow mode運用	人間レビュー合格率を測定
本番化	version固定、監視、バックアップ、rollback	SLA・監査手順確立
継続評価	モデル更新ごとに回帰benchmark	基準低下時は自動rollback

最初から6モデルを本番サーバに載せる必要はない。実務上は、

Qwen3.8-27B → gpt-oss-120b → Gemma 4 26B A4B

の3モデルを先に評価し、この中で要件を満たさない領域がある場合だけMistral/Nemotronへ拡張するのが効率的である。

Qwen/Gemmaの27B級では、128GBメモリを活かして最初はBF16/FP8等の比較的高精度で試験し、**量子化は「容量が足りないから」ではなく「必要な速度を出すため」**に導入するのが望ましい。知財業務では小さな量子化誤差が、否定、数値範囲、従属関係、例外条項の読み違いとして表面化し得るため、必ずBF16基準との回帰比較を行う。

100B級では逆に4bit系が前提になる。gpt-ossはOpenAIがMXFP4で80GB GPU動作を公式に設計目標としているため、この中では最も導入根拠が強い。Mistral Small 4にはNVFP4版checkpointも存在する一方、公式が通常インフラとして複数H100/H200等を挙げていることから、RTX Spark単機での本番利用は実機検証結果を優先すべきである。 30

ストレージ計画

最低でも次の程度を見込む。

用途	推奨容量
Windows・CUDA・runtime	250～500GB
Qwen/Gemma各種precision	150～250GB
120Bモデル2～3種類	300～600GB
一世代前のrollback用モデル	200～500GB
Embedding / reranker	50～200GB
Patent/RAG corpus	数百GB～数TB
indexing一時領域	corpusの20～50%程度
ログ・backup	別ボリューム推奨

したがって**4TB NVMeを最低ライン、特許全文や複数世代のモデルを保持するなら8TB以上を推奨**する。ただしストレージ容量自体はユーザー指定がなく、本数値は本報告の設計目安である。

電力と運用費

RTX Sparkデスクトップのチップ側TDPは約140W級とされており、実際のコンセント消費にはSSD、電源損失、周辺機器等が加わる。 ³¹

本報告では計画値として、

LLM高負荷時のPC全体平均	: 200 W
電力単価	: 30 円/kWh
冷房・UPS・NAS	: 別

と仮定する。

この仮定では、

運用パターン	年間消費電力量の概算	電気代概算
8時間/日 × 250営業日を高負荷	400 kWh	約1.2万円/年
24/7、平均100W	876 kWh	約2.6万円/年
24/7、常時200W相当	1,752 kWh	約5.3万円/年

となる。これは**PC本体だけの単純計算**であり、空調、NAS、backup、UPS、管理者工数は含まない。電力単価30円/kWhも実料金ではなく比較用仮定である。

本格運用のTCOでは電気代より、**システム管理、評価、RAGデータ整備、セキュリティ監査、人間レビュー、商用サポート契約**の方が大きくなる可能性が高い。特にNVIDIA NIMを使う場合はNVIDIA AI Enterprise、Cohereを商用利用する場合は商用契約の費用を別途見積もる必要がある。NVIDIAはNIM production利用にAI Enterpriseライセンスが必要と説明している。 ²¹

クラスタ化の判断

1〜3名程度の知財担当者が逐次利用する限り、クラスタは原則不要と判断する。Qwen3.8-27B、Gemma 4、gpt-oss-120bについては単一128GB機からPoCを始めるべきである。

クラスタ化を検討する条件は、概ね次のいずれかである。

条件	推奨対応
5〜10人以上が同時利用	2台以上を replica としてload balance
120Bモデル+64K超contextを高頻度使用	GPUサーバを検討
24/7可用性が必要	active/standbyまたは2-node replica
OCR/indexingとLLMが競合	RAG/indexサーバを分離
大規模batch特許分類	dedicated inference worker追加
128GBを超えるモデル	RTX Spark横並びよりデータセンターGPUを検討

複数台のRTX SparkをEthernet接続しても、それだけで128GB×台数が一つの共有VRAMになるわけではない。モデルを複数PCへtensor-parallelに分割する場合は通信がボトルネックになり得るため、**知財用途ではまずモデル分割ではなくrequest単位のreplica構成を優先すべきである。**

法的・倫理的注意点

ローカル運用の最大の利点は、入力文書を外部LLM事業者へ送らずに処理できる点にある。しかし、「**オンプレだから法令対応不要**」「**オープンモデルだから自由**」という**結論にはならない。**

個人情報については、発明者氏名、従業員情報、契約相手担当者情報等がRAG corpusやpromptに含まれ得る。個人情報保護委員会は個人情報保護法と関連ガイドライン、安全管理・漏えい対応等を体系的に示している。ローカル化により第三者クラウドへの送信経路を減らすことはできても、アクセス制御、利用目的、保存期間、backup、漏えい時対応等の管理責任自体は残る。³²

未公開発明、研究データ、ライセンス交渉、共同研究情報等は営業秘密管理との関係が特に重要である。経済産業省は、不正競争防止法上の営業秘密保護を受けるための最低限の管理水準を示すものとして「**営業秘密管理指針**」を公表している。したがってLLM導入後も、「LLMに入れたから社内全員が検索可能」という設計にはならず、元文書の秘密区分・ACLをembeddingやcacheにも維持する必要がある。³³

モデルライセンスについては、**導入時点のlicenseファイルだけでなく、model ID、commit、取得日、license textのcopyを保存**することを推奨する。モデル名が同じでも新版で条件やcheckpointが変わる可能性があり、監査時に「どの条件で取得したモデルを本番使用しているか」を証明できる状態が必要である。

また、第三者の特許データベース、論文データベース、技術標準、契約書等をRAGへ大量投入する場合、モデルのApache/Gemma等のライセンスとは別に、**コンテンツ側の著作権、データベース利用規約、契約上の再利用条件を確認**しなければならない。特に契約型の商用特許DBについては、検索結果をローカルAI corpusへ恒久保存することが契約上許可されているかを個別に確認すべきである。

知財判断についてLLMを最終決裁者にはいけない。特に、新規性・進歩性、均等論、クレーム解釈、FTO、契約リスク、出願方針などは、誤答の事業インパクトが大きい。最低でも、

AI出力

- 根拠文献
- 該当段落／条項
- モデル/version
- 人間reviewer
- reviewer判断

を一組として残すべきである。

倫理面では、モデルが生成したもっともらしい特許番号、判例、条文、引用を「存在確認なし」で資料へ転記しない運用が不可欠である。RAGでヒットしなかった情報については「不存在」ではなく「**本検索条件では確認できなかった**」と表現させる方が適切である。

また、fine-tuningデータに過去の人間の判断を使う場合、その判断に含まれる過去の組織的偏りがモデルへ固定化される可能性がある。例えば「この技術分野はいつも出願しない」といった過去判断をそのまま教師データとするのではなく、**事実抽出・分類と法的／経営判断を分離して学習させる**ことが望ましい。

海外拠点で利用する場合は日本法だけでは足りない。EUではAI Actが2024年8月に発効し、主要規定が2026年8月2日から適用段階に入っており、GPAI関連の透明性・copyright・risk管理等の枠組みも整備されている。そのためEU拠点・EU向けサービスと連携する場合は適用範囲を別途評価する必要がある。 ³⁴

AIリスク管理の補助的なフレームワークとしては、NIST AI RMFおよび生成AI向けProfileが、リスク識別、測定、管理の体系化に利用できる。日本法上の義務そのものではないが、内部統制のチェックリストとして有用である。 ³⁵

モデル別に法務・倫理上の注意点を整理すると以下の通りである。

モデル	導入時の主要法務・倫理注意点
Qwen3.8-27B	Apache-2.0で扱いやすい。モデル入手元・hash・licenseを固定。日本語知財でhallucination/citation評価を必須化。 ¹²
gpt-oss-120b	Apache-2.0。reasoning trace等をそのまま利用者へ露出させず、最終回答と証拠を分離。日本語精度を独自検証。 ⁴
Gemma 4	商用利用可能だがGemma独自Terms。再配布・Hosted Service・派生モデルでは利用制限承継、Notice等を確認。 ²⁶
Mistral Small 4	Apache-2.0。量子化checkpointの配布元・完全性を検証。Spark運用は公式標準構成外として回帰試験を厳格化。 ¹⁹
Nemotron 3 Super	open modelは商用利用可能。ただしNIMはAI Enterprise条件。NVIDIA周辺serviceを導入する場合、open modelと商用softwareの条件を分離管理。 ²¹
Cohere Command A	公開checkpointはCC-BY-NCで 通常の企業商用業務には使用不可 。商用契約が確認できるまではPoC・研究限定として扱う。 ²⁴

最終的な推奨構成は、**Qwen3.8-27Bを常用モデル、gpt-oss-120bを高難度reasoningモデル、Gemma 4 26B A4Bを高速multimodalモデルとして同一RAG基盤から切り替える三層構成**である。Mistral Small 4およびNemotron 3 Superは、前記3モデルで満たせないagentic処理またはNVIDIA固有最適化が必要になった時点で追加する。Cohere Command Aは、性能評価より前に商用ライセンスを解決する必要がある。

128GB RTX Sparkの価値は、最大モデルを無理に載せることなく、27B級を高精度・完全ローカルで余裕を持って動かしつつ、必要時だけ100～120B級の低bit reasoningモデルへ切り替えられる点にある。知財業務ではこの余裕を最大contextへ浪費するより、暗号化RAG、検索精度、引用強制、権限管理、監査ログ、human-in-the-loopへ配分する方が、実務上の精度・機密性・説明可能性を高める設計となる。

¹ ³ ¹¹ ¹³ ²⁷ Qwen/Qwen3.8-27B・Hugging Face

<https://huggingface.co/Qwen/Qwen3.8-27B>

² ⁸ ²⁹ ³¹ Slim Laptops & Small Desktops | NVIDIA RTX Spark

<https://www.nvidia.com/en-us/products/rtx-spark/>

⁴ ¹⁴ ³⁰ Introducing gpt-oss | OpenAI

<https://openai.com/index/introducing-gpt-oss/>

⁵ ¹⁶ google/gemma-4-31B

https://huggingface.co/google/gemma-4-31B?utm_source=chatgpt.com

- 6 10 19 28 **Introducing Mistral Small 4 | Mistral AI**
<https://mistral.ai/news/mistral-small-4/>
- 7 **Introducing Mistral Large 4 | Mistral**
<https://mistral.ai/news/mistral-large-4/>
- 9 **NVIDIA RTX Spark がもたらす、Windows PC の新たなパワフルな章のご紹介 - Source Asia**
<https://news.microsoft.com/source/asia/features/introducing-a-powerful-new-chapter-for-windows-pcs-accelerated-by-nvidia-rtx-spark/?lang=ja>
- 12 25 **GitHub - QwenLM/Qwen3.8: Qwen3.8 is the large language model series developed by Qwen team, Alibaba Group. · GitHub**
<https://github.com/QwenLM/Qwen3.8>
- 15 **Open models by OpenAI | OpenAI**
<https://openai.com/open-models/>
- 17 26 **Gemma Terms of Use | Google AI for Developers**
<https://ai.google.dev/gemma/terms>
- 18 **Dostrajanie modelu Gemma | Google AI for Developers**
https://ai.google.dev/gemma/docs/tune?hl=pl&utm_source=chatgpt.com
- 20 22 **Nemotron AI Models | NVIDIA Developer**
<https://developer.nvidia.com/topics/ai/nemotron>
- 21 **Build Agentic AI with Multimodal Foundation Models | NVIDIA Nemotron**
<https://www.nvidia.com/en-us/ai-data-science/foundation-models/nemotron/>
- 23 **CohereLabs/c4ai-command-a-03-2025 · Hugging Face**
<https://huggingface.co/CohereLabs/c4ai-command-a-03-2025>
- 24 **README.md · CohereLabs/c4ai-command-a-03-2025 at main**
<https://huggingface.co/CohereLabs/c4ai-command-a-03-2025/blob/main/README.md>
- 32 **個人情報保護法等**
https://www.ppc.go.jp/personalinfo/?utm_source=chatgpt.com
- 33 **営業秘密～営業秘密を守り活用する**
https://www.meti.go.jp/policy/economy/chizai/chiteki/trade-secret.html?utm_source=chatgpt.com
- 34 **AI Act | Shaping Europe's digital future - European Union**
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?utm_source=chatgpt.com
- 35 **AI Risk Management Framework | NIST**
https://www.nist.gov/itl/ai-risk-management-framework?utm_source=chatgpt.com