

NVIDIA RTX Spark搭載PCが拓く Local LLMと知財業務の可能性

Executive Summary

結論から言えば、RTX Sparkは「クラウドLLMを完全に置き換えるPC」というより、これまでサーバーやクラウドに出す必要があった70B級モデルと機密データ処理を、知財担当者の机上・部門内に持ち込むための新しい境界装置になり得る。とりわけ知財業務では、未公開発明、出願前明細書、クレーム案、ライセンス契約、侵害鑑定資料など、「外部LLMに送りにくいデータ」が多いため、一般的なオフィス業務以上にLocal LLMとの相性がよい。

RTX Spark N1Xの上位構成は、**6,144 CUDAコアのBlackwell RTX GPU、20コアArm Grace CPU、最大128GBのコヒーレント統合メモリ、300GB/sのメモリ帯域、最大1 PFLOPのFP4 AI性能**をWindows 11 PCに統合する。CPUとGPUはNVLink-C2Cで接続され、5世代Tensor CoreとFP4を備える。デスクトップ版SoCのTDPは140W、ノート版は45～80Wである。〔注1〕 ¹

この構成の本質は、1 PFLOPという演算性能そのものより**128GBの統合メモリ**にある。従来の24～32GB級PC GPUでは困難だった70BモデルをQ4/FP4だけでなくQ8/FP8級まで載せられ、120B級も4bit量子化やMoEなら現実的な領域に入る。一方、メモリ帯域は300GB/sなので、70Bのdenseモデルを「高速な対話モデル」として動かすには帯域が制約となる。RTX Sparkは「大モデルが動く」ことと「大モデルが高速に動く」ことを分けて評価すべき製品である。NVIDIA自身も最大120B・最大1M token contextをうたうが、モデル、量子化、KV cache、Windowsのメモリ配分によって実用上の上限は変わる。〔注2〕 ²

2026年10月9日時点ではRTX Sparkそのものの第三者LLM実測値はまだ少ない。したがって本報告の性能値は、同じBlackwell系・128GB unified memoryを持つDGX SparkのNVIDIA公式実測を「proxy baseline」とし、DGX Sparkの273GB/sに対してRTX Sparkが300GB/sであることを踏まえて推定した。DGX Sparkでは、Llama 3.1 8B NVFP4が38.65 token/s、Qwen3 14B NVFP4が22.71 token/s、GPT-OSS-120B MXFP4が55.37 token/s、32K入力のLlama 3.3 70B NVFP4は1ノードでTPOT 269ms、すなわち約3.72 token/sを記録している。〔注3〕 ³

したがってRTX Spark 128GBデスクトップの現実的な期待値は、**7～8B Q4/FP4で約38～45 token/s、13～14Bで約22～30 token/s、70B dense Q4/FP4でおおむね4～6 token/s**である。ただし長いコンテキストでは70Bが4 token/s前後まで落ちる可能性が高い。7～14Bは個人コパイロットとして十分実用域、70Bは「品質優先・機密文書の深い分析」向け、120B denseはバッチ処理寄りという位置付けになる。

知財分野で最も有望なのは、LLM単独に検索・判断を任せるのではなく、**特許・契約・社内知財文書を対象にしたLocal RAG+引用検証+Human-in-the-loop**である。特許調査、外国特許の翻訳、OA/拒絶理由通知の要約、明細書とクレームの整合性検査、クレームチャート作成、契約条項レビュー、社内知財Q&Aは特に適合度が高い。逆に、最終的な新規性・進歩性判断、FTO/侵害結論、出願クレームの最終確定をLLMだけで行うのは推奨しない。

経営判断としては、**まず128GB RTX Sparkを1台導入し、機密文書の翻訳・要約、社内知財RAG、クレーム整合性チェックの三用途から始めるのが妥当**である。Surface RTX Spark Dev Box 128GB/2TBの米国価格は2026年10月時点でUS\$5,999.99であり、仮に予算換算レートを150円/US\$と置けば約90万円のハードウェア

となる。〔注4〕⁴ これはA100/H100/B200級サーバーを部門導入する場合と全く異なる調達単位であり、「一人一台～数人一台のLegal/IP AI」を現実にする点に市場的意思がある。

ただし、RTX Sparkが最良のLocal LLMハードウェアとは限らない。2026年のApple M5 Ultra Mac Studioは最大512GB unified memory・1.2TB/sで、純粋な大規模Local LLMの容量とメモリ帯域ではRTX Sparkを大幅に上回る。一方RTX SparkにはCUDA/TensorRT/NVIDIA AIスタックとWindows業務環境との親和性という大きな優位性がある。〔注5〕⁵ 知財部門では最高token/sではなく、「Windows業務環境＋CUDA＋128GB＋ローカル機密性」の組合せがRTX Sparkの競争力と考えるべきである。

技術基盤とLocal LLMへの意味

RTX Spark N1Xの主要仕様

RTX Sparkは、DGX Sparkを単純にWindows化した製品ではなく、Windows on Armを前提とするPC向けSoCプラットフォームである。NVIDIAの2026年9月更新Porting Guideによれば、最上位CPUはArm v9.2の20コアで、**10コアクラスタ×2、それぞれCortex-X925×5＋Cortex-A725×5**、すなわち性能コア10＋効率コア10という構成である。GPUはBlackwell統合GPUである。〔注6〕⁶

項目	RTX Spark N1X 上位構成	Local LLMへの意味
GPU	Blackwell RTX、6,144 CUDA cores	5世代Tensor CoreでFP4推論を高速化
AI演算性能	最大1 PFLOP FP4	Prefill、小中規模モデル、量子化モデルに有利
Tensor Core	第5世代	NVFP4等の低精度推論が中心
CPU	Arm v9.2 20 cores	10×Cortex-X925＋10×Cortex-A725
CPU-GPU	NVLink-C2C	CPU/GPUを単一SoCとして高効率接続
メモリ	最大128GB LPDDR5X unified	70B Q8/Q4、120B Q4/MoEが視野
メモリIF	256bit	—
帯域	300GB/s	大型dense LLMのdecodeでは主要ボトルネック
PCIe	PCIe Gen5対応	SSD・OEM拡張性。ただしlane/configurationはOEM依存
OS	Windows 11 on Arm	知財部門のOffice/DMS環境との統合に有利
デスクトップTDP	140W	一般オフィスで導入しやすい電力レンジ
ノートTDP	45～80W	携帯型Local AIが可能。ただし持続性能は冷却依存
公称LLM能力	最大120B、最大1M context	「搭載可能」であって全モデルが快適とは限らない

NVIDIA公式仕様による。PCIe Gen5についてNVIDIA自身がOEMごとのconfiguration確認を求めており、**PCIe lane数や拡張カード構成をNVIDIA共通仕様として仮定すべきではない**。またNVLink-C2Cの存在は公式に確認できるが、本稿で参照したRTX Spark文書ではC2Cリンクの実効/理論帯域値までは公開されていない。〔注7〕⁷

なお、1 PFLOPは**FP4 AI性能**である。調達比較でFP16、BF16、FP32性能へ単純換算することは不適切で、NVIDIAのRTX Spark公開仕様にも本稿執筆時点でそれらのピーク値は示されていない。LLMについてはむしろFP4/FP8性能、メモリ容量、メモリ帯域、KV cache、runtime実装を見る必要がある。〔注8〕 ⁸

128GB unified memoryが最大の変化

128GBは通常の「RAM 128GB+VRAM 24GB」という構成とは意味が異なる。RTX SparkではCPUとiGPUが同じ物理メモリを共有し、ハードウェアとしてコヒーレンシが維持される。したがって巨大なモデルをPCIe越しにCPU RAMへoffloadする従来型PCより、モデル配置を単純化できる。〔注9〕 ⁹

ただしWindows側からは、物理メモリが、

GPU dedicated carveout → CPU/GPU shared memory → CPU-only memory

の三領域に分けて見える。NVIDIA文書では、carveout後のshared memoryについて「残量から16GBを引いた値を基準とし、50~80%にclamp」と説明している。つまり、「**128GB搭載=任意のCUDAプロセスが128GB全部をVRAMとして自由に確保できる**」わけではない。120Bモデルや1M contextを採用する案件では実機でのメモリ設定確認が必須である。〔注10〕 ⁹

知財LLMでは「演算能力より帯域」が重要になる場面が多い

LLM推論は大きくprefillとdecodeに分かれる。長い明細書や大量の先行技術を最初に読むprefillは演算能力の恩恵を強く受ける。一方、一語ずつ文章を生成するbatch=1のdecodeでは、巨大なweightを繰り返し読むためメモリ帯域が支配的になりやすい。

これは知財用途で重要である。例えば、

70Bモデルが128GBに「載る」ことと、70Bモデルが30~50 token/sで「快適に走る」ことは別問題である。

RTX Sparkの300GB/sは統合メモリとしては高いが、後述するM5 Ultraの1.2TB/sやデータセンターBlackwell B200の約8TB/sとは桁が違う。NVIDIA自身の比較でもDGX Sparkは273GB/s、B200は約8TB/sである。〔注11〕 ¹⁰

したがってRTX Sparkに適した設計は、「最大モデルを常時使う」のではなく、

8~14B高速モデルを日常作業に使い、70Bを高難度案件にroutingし、さらに難しい問題だけクラウド frontier modelへ匿名化して送る

という階層型である。

冷却とArm移行で見落としやすい点

デスクトップSoC TDPは140W、ノートは45~80Wなので、ラックGPUのような専用空調や液冷を前提とする製品ではない。とはいえ長時間のLLM推論では持続的な高負荷となるため、OEMのactive cooling能力と吸排気スペースが実効性能を左右す。ASUS等もRTX Spark機を長時間AIワークロード向けに設計している。〔注12〕 ¹¹

またCPUはx86ではなくArmである。NVIDIAはWindows on Arm向けの専用Porting Guideを公開しており、x86エミュレーションだけで既存LLMスタックがそのまま最高性能になるとは考えない方がよい。本番採用

ではPython wheel、CUDA extension、vector DB client、PDF parser、OCR、暗号/DLP agent、EDRなどをARM64で一つずつ確認すべきである。〔注13〕 6

性能評価と最適化

公開実測値と本報告の推定方法

RTX Sparkは2026年秋の新製品で、2026年10月9日時点では成熟した独立ベンチマークが十分ではない。そのため以下では、NVIDIA公式の**DGX Spark実測値を下限/基準にしたengineering estimate**を用いる。

DGX Sparkは1 PFLOP FP4、128GB unified memory、273GB/sである。RTX Spark上位SKUは300GB/sなので、decodeが純粋にmemory-boundなら理論上の帯域差は

$$300/273 = 1.099$$

すなわち最大約10%の改善余地がある。ただしWindows/Arm runtime、driver、kernel、量子化形式による差があるので、**+10%を保証性能として扱ってはいけない**。DGX Spark公式実測では2,048-token input、128-token output、batch 1で、Llama 3.1 8B NVFP4が38.65 token/s、Qwen3 14Bが22.71 token/sである。〔注14〕 12

モデル規模別の実用性

モデル規模・例	推奨精度	Weight概算	RTX Spark実行可否	単一ユーザーdecode見積り	2K前後入力の体感	快適な対話同時利用目安
7～8B：Llama 3.1 8B、Mistral 7B、Vicuna/Alpaca 7B	Q4/NVFP4	約4～5GB	容易	38～45 tok/s	TTFT 約0.3～1秒	4～6人
13～14B：Llama 2 13B、Qwen 14B系	Q4/NVFP4	約7～9GB	容易	22～30 tok/s	約0.5～1.5秒	2～4人
30～35B	Q4/FP8	約18～35GB	容易	約10～20 tok/s*	1～数秒	1～3人
70～72B：Llama 2/3 70B	Q4/NVFP4	約35～45GB	実用可能	約4～6 tok/s	数秒、長文で増加	1人、低rateなら2人
70B	Q8/FP8	約70～85GB	可能	Q4より低下傾向	長文ではKV余裕に注意	原則1人
70B	FP16/BF16	約140GB+overhead	不可/非現実的	—	—	—
120B dense	Q4/FP4	約60～80GB	容量上は可能	約2～3 tok/s級*	バッチ処理寄り	1人
120B MoE	MXFP4等	モデル依存	非常に有望	DGX proxyで55.37 tok/sの例	active paramsに大きく依存	複数可

* は公開RTX Spark直接測定ではなく、本報告の帯域・類似モデルからのengineering estimate。

根拠となるNVIDIA実測では、Llama 3.3 70B NVFP4、32K input/1K output、DGX Spark 1ノードでTPOT 269ms、つまり約3.72 token/sである。RTX Sparkが帯域に比例すると仮定した上限寄り推定は約4.1 token/sとなる。したがって70Bについて上表の4~6 token/sは「短~中contextでの期待レンジ」であり、32K以上では4 token/s前後を計画値とする方が安全である。〔注15〕 13

一方、GPT-OSS-120B MXFP4はDGX Spark 1台・llama.cppで55.37 token/sを記録している。これは120Bという名称だけを見てdense 120Bと同じ速度だと考えてはいけな好例であり、**MoEでは総parameter数より「1 tokenあたり何parameterをactivateするか」が重要**である。〔注16〕 14

長い特許文書ではTTFTが別問題になる

知財では出力速度以上に「何十ページの文書を読ませるか」が問題になる。

DGX Sparkの128K-token input+1K output公式試験では、

モデル	128K prompt処理	Generation	E2E
Nemotron 3 Super 120B NVFP4 / TensorRT-LLM	2,855 tok/s	18 tok/s	99秒
Qwen3.5 35B A3B FP8 / vLLM	3,080 tok/s	35.75 tok/s	73秒
Qwen3 Coder Next 80B FP8 / vLLM	2,390 tok/s	28.95 tok/s	89秒

となっている。〔注17〕 15

つまり、「1M context対応だから会社の全特許をプロンプトに入れる」という設計は得策ではない。**特許調査ではRAG、claim-aware chunking、family/IPC/CPC/assignee/date metadata filtering、rerankerによって入力を数千~数万tokenまで絞る方が高速かつ検証しやすい。**

同時ユーザーについて

NVIDIAのDGX Spark/vLLM試験ではQwen3 Coder Next、32K input+1K outputで同時タスク数を1→2→4と増やすとaggregate generation throughputが38→47→53 token/s、median TTFTが9→12→15秒となった。4並列で処理量は増えるが、各ユーザーの応答開始は遅くなる。〔注18〕 16

従って知財部門では、RTX Spark一台を「20~50人向けLLMサーバー」と考えるより、

- 7~14B中心なら数人で共有
- 70B中心なら一人~少人数
- 大人数には**複数nodeへのrequest-level load balancing**
- 本格的な20人以上の同時利用にはGPU server/cloud

という設計が現実的である。

なおDGX SparkはConnectX-7/RoCEを使い最大4ノードまでTensor Parallelを拡張できるが、**これはRTX Spark PCの共通仕様ではない**。RTX Sparkに同じConnectX-7 multi-node性能があるとは仮定せず、当初は各PCにモデルを複製してrequest単位で振り分ける方が安全である。DGX Sparkでは4ノードで最大700B級までの実験例がNVIDIAから示されている。〔注19〕 13

量子化・蒸留・runtime最適化の効果

最適化	メモリ効果	速度効果	知財品質への注意	RTX Sparkでの位置付け
FP16/BF16	2 byte/parameter前後	遅い/容量大	基準品質	7～14B向け。70Bには不向き
FP8/Q8	約1 byte/parameter	中	品質低下は一般に小さいが要評価	70B品質優先に有力
NVFP4/Q4	約0.5 byte/parameter+ metadata	大	数字・引用・細かな法律論を要検証	RTX Sparkの本命
KV cache量子化	長context容量を削減	状況依存	長文精度への影響を試験	特許全文処理に有用
Distillation	モデル自体を縮小	極大	教師モデルの能力を完全には維持しない	日常14B+高難度70B構成に有効
LoRA/PEFT	base model共用	adapterは小さい	学習データ品質が重要	部門別adapterに向く
Speculative decoding	容量影響小	条件次第で向上	出力品質を保ちながら高速化可能	TensorRT/vLLMで検討

NVIDIAはDGX SparkでNVFP4についてFP8に近い精度を保ちながらfootprintを削減すると説明しており、TensorRT-LLMはKV cache reuse、in-flight batching、speculative decoding、LoRAなどを備える。〔注20〕
 17 ただし「一般ベンチマークで誤差1%未満」と「日本語特許クレームで法的に重要な語を誤らない」は同義ではないので、**Q4採用の可否は知財専用evaluation setで決めるべき**である。

ソフトウェア選択は次のように整理できる。

Stack	推奨用途	評価
TensorRT-LLM	NVIDIA GPUで最大性能、multi-user serving	RTX Sparkの第一候補。量子化・batching・KV最適化が強い
llama.cpp / GGUF	PoC、幅広いopen model、簡易ローカル利用	導入が容易。NVIDIA公式benchmarkにも採用
vLLM	API server、continuous batching、multi-user	部門共用サーバー化に有力。ただしWindows/Arm対応を実機確認
NeMo	SFT/LoRA/PEFT、データ整備、評価	domain adaptation側に強い。TensorRT-LLM/vLLMへdeployment可能
ONNX Runtime	embedding、classifier、reranker等	汎用性が高い。主LLM decoderより周辺NLPモデルで有用
Neural Magic 系	vLLM中心のopen serving	Neural MagicはRed Hatに買収済みで、現在はvLLM/Red Hat AIの文脈で評価するのが適切

NeMoはPEFTでbase weightを固定しadapterのみを学習でき、公式文書ではtrainable parameterをしばしば1%未満に抑えられるとしている。またNeMoモデルをTensorRT-LLMへexportできる。〔注21〕¹⁸ Neural MagicはRed Hatが2025年に買収を完了し、同社のvLLMへの技術貢献をRed Hat AIへ統合している。〔注22〕¹⁹

知財業務ユースケース

Local LLMが特に効くのは「機密性×大量文書×反復処理」

知財部門には、一般部門と比べてLocal LLMに向く条件が揃っている。価値の中心は「LLMが弁理士や弁護士を置き換える」ことではなく、**調査・読解・比較・草案・引用確認という大量の中間作業を、機密データを外部へ送らず反復できることにある。**

EPOを含む知財機関ではAIによるprior-art探索や文書検索が既に実務に取り込まれつつあり、WIPOも特許庁におけるsemantic searchや機械翻訳などのAI利用を整理している。一方、WIPOの特許ドラフティング関連資料でも生成AIにはなお制約があることが指摘されており、AIを最終判断者にするより「検索・草案・検証補助」に置く方が適切である。〔注23〕²⁰

知財業務	推奨Local AIシナリオ	精度・速度への期待	機密性・コスト効果	必須ガードレール
特許調査	keyword+embedding+reranker+LLMで検索式展開	意味的類似・多言語検索を補強	大量queryを定額化	LLM検索だけでrecallを保証しない
先行技術調査	発明を構成要件へ分解し、各elementごとに検索	調査者の読み込み時間を短縮	未公開発明をlocalで保持	原文・公開番号を必ずcitation
明細書作成	invention disclosure→実施形態候補→記載骨格	初稿・variation作成が高速	出願前情報を外部送信しない	発明者/弁理士レビュー必須
クレーム草案	independent claim→dependent alternatives生成	言い換え・fallback position探索	多数案を低限界費用で生成	support/antecedent basis 自動検査
クレーム整合性	claim ↔ specification ↔ drawings mapping	特に高適合	機械的レビューを自動化	ページ/段落 citationを保存
侵害/FTO分析	claim element→製品資料のevidence抽出	claim chart作成を高速化	訴訟・鑑定情報をlocal保持	結論は法律家が承認
契約レビュー	IP ownership、license、indemnity、AI/data条項抽出	定型条項比較に強い	契約内容を外部送信しない	approved clause libraryとのRAG
要約・翻訳	foreign patent/OA/契約を日英相互変換	最も導入しやすい	API token 費を抑制	数字、claim dependencyのQA

知財業務	推奨Local AIシナリオ	精度・速度への期待	機密性・コスト効果	必須ガードレール
Knowledge Management	過去出願、OA、意見書、契約、鑑定書RAG	組織知を会話検索	社内資料を閉域運用	元文書ACLを検索時にも継承

特許調査では「LLMに検索させる」のではなく「検索系をLLMに操作させる」

最も危険なのは、LLMに「この発明の先行技術を教えて」とだけ聞き、モデルの内部知識を検索結果とみなす方法である。特許調査は**recallが重要**であり、hallucinated publication numberは致命的である。

推奨pipelineは、

発明説明 → claim-like element分解 → query expansion → keyword/IPC/CPC/semantic search → top-N retrieval → cross-encoder rerank → LLM比較 → 原文citation

である。

LLMは検索エンジンそのものではなく、検索戦略の作成、検索結果の比較、claim chart化に使う。これによりLLMが知らない最近の公開特許も扱え、回答の根拠を監査できる。

明細書とクレーム作成での最重要用途は「生成」より「検査」

生成AIという名称からdraft generationに目が行きがちだが、知財でROIが高く、リスクが比較的低いのは、

claim support checker

antecedent-basis checker

用語統一 checker

図面符号整合 checker

独立/従属claim dependency checker

明細書に記載されたvariantの漏れ検出

である。

これらは「正解の根拠が文書内にある」ため、RAGとrule engineを組み合わせると生成型の自由回答より品質管理しやすい。AI利用の成果を単なる文章生成量ではなく、**reviewerが見つける不整合件数、編集率、citation precision、処理時間**で測定できる点も重要である。

PoCで測るべき知財専用KPI

一般的なMMLUやchat benchmarkだけでは導入可否を決めない方がよい。推奨するacceptance testは次である。

業務	KPI
先行技術	Recall@20/50、既存調査結果に対するhit率
RAG	Citation precision、citation completeness
要約	unsupported statement率、重要claim omission率
クレーム	support mapping正答率、dependency error率
翻訳	claim terminology consistency、数字/符号誤り
契約	mandatory clause検出recall
ドラフティング	弁理士による編集率、完成までの時間
セキュリティ	外部egress件数=0、ACL violation=0

「LLM回答のaccuracy 95%」のような一つの数字ではなく、業務ごとに失敗モードを定義することが導入成功の鍵になる。

法務・データ保護・ガバナンス

日本ではLocal処理の価値が大きい、「localだから適法」ではない

個人情報保護委員会は生成AI利用について、個人情報をpromptに入れる際は利用目的の範囲内か確認すること、本人同意なしに個人データを投入する場合にproviderが応答生成以外、例えば機械学習に利用しないこと等を確認するよう注意喚起している。〔注24〕 ²¹

この観点からLocal LLMには非常に明確な利点がある。

prompt、RAG文書、embedding、responseを原則として端末/社内ネットワークから出さないことで、第三者providerへのprompt送信、provider側logging、海外data centerへの転送、provider-side trainingという経路そのものを減らせる。

ただしLocal LLMでも、

- prompt log
- vector database
- embedding
- output
- fine-tuning dataset
- backup
- crash telemetry
- cloud fallback

に個人情報や機密情報が残る可能性がある。従って個人情報保護法上の安全管理、アクセス制御、保存期間、削除、漏えい対応等が不要になるわけではない。PPCサイトでは既に令和8年改正個人情報保護法も独立項目として扱われており、2026年以降の施行規則・ガイドライン更新も継続確認すべきである。〔注25〕

²²

日本のAIガバナンスは2026年に一段進んでいる

日本では「人工知能関連技術の研究開発及び活用の推進に関する法律」が2025年に公布され、2025年9月に全面施行された。2026年にはAI政策・指針の整備がさらに進められている。〔注26〕 23

経産省・総務省のAI事業者ガイドライン第1.2版（2026年3月）は、privacy、security、transparency、accountability等を含むリスクベースの管理を求める方向にあり、内閣府のAI関連指針ページも、AIと知財、個人情報、政府生成AI調達、ISO/IEC 42001、AISIの生成AI品質管理・red teaming等を横断的に参照している。〔注27〕 24

したがって知財部門では、Local LLMを単なるPCソフトとして導入するのではなく、少なくとも、

model card、model/version registry、利用目的、承認データ、評価結果、変更履歴、人間の最終承認責任者

を管理する「小型AI management system」として運用する方がよい。

著作権と二次利用

文化庁は2024年に「AIと著作権に関する考え方」を取りまとめ、AI学習段階と生成・利用段階を区別して整理している。文化庁のガイダンスでは、AI学習のための著作物利用について著作権法30条の4等が問題となり得る一方、条件を満たさない場合には許諾が必要となること、海賊版サイトと認識しながら学習data sourceにすること等を避けること、data provenanceを確保することがリスク低減策として挙げられている。〔注28〕 25

また生成物を利用する場面では、既存著作物との類似性と依拠性が侵害判断の重要な要素となる。AI生成物自体に著作物性があるかという問題と、既存作品を侵害しているかという問題は別である。〔注29〕 26

知財部門で特に注意すべきなのは、外部公開された特許公報だけではない。契約書ひな形、有料databaseの全文、法律出版社の解説、他事務所作成の鑑定書などを勝手にfine-tuning corpusへ入れると、契約・著作権上の問題が生じ得る。

従ってtraining/RAG dataは、

出所、license、利用目的、保存期限、公開/社外秘区分

をmodel registryと同様に記録するべきである。

出力の説明責任

LLMに「なぜそう判断したか」を自然言語で説明させるだけでは説明責任を満たさない。LLMの理由説明自体がhallucinationするからである。

知財用途では、回答ごとに少なくとも、

model/version → prompt template → retrieval documents → source paragraph/page → generated answer → reviewer → final disposition

を追跡できるようにする方が実務的である。

ただしraw prompt logを無制限に保存すると、逆に発明・契約・個人情報の巨大な機密databaseを作ることになる。通常は**監査metadataを保存し、raw contentは短期retentionまたは案件DMSへ紐付ける方が安全**である。

海外法制

地域	2026年時点の重要点	知財Local LLMへの含意
日本	AI法、AI事業者ガイドライン、個人情報保護法、著作権法等を組合せて管理	Local処理は有利だが、data governanceは必要
EU	AI Actは2026年8月から多くの規定の適用・執行段階。GDPRはAIにもtechnology-neutralに適用	EU拠点・EU個人dataを扱う場合はAI ActとGDPR双方を確認
米国	NIST AI RMF/GenAI Profileはvoluntary risk-management framework	多国籍企業の共通governance baselineとして有用
中国	生成AIサービス暫定管理弁法に加え、public-facing serviceでは备案/登記制度が運用	中国で外部提供する場合、純粋な社内local利用とは別途scope確認が必要

EU AI Actは2024年8月に発効し、2026年8月2日から執行権限が本格化している。2026年にはAI Omnibusによる時期・手続の変更もあるため、古い「2024年時点のタイムライン」をそのまま使わない方がよい。〔注30〕²⁷ GDPRは処理技術にかかわらずpersonal dataに適用される。〔注31〕²⁸

米国ではNIST AI RMFとGenerative AI Profileが、Govern、Map、Measure、Manageを中心とする実務的な共通枠組みを提供している。〔注32〕²⁹ 中国では2023年の生成AI暫定管理弁法に加え、2026年にも备案・登記制度が現実に運用されている。〔注33〕³⁰

コスト・運用・競合・リスク

導入コスト

MicrosoftのSurface RTX Spark Dev Boxは、6,144-core GPU+20-core CPU、128GB unified memory、2TB Gen5 SSD、Windows 11 Proで**US\$5,999.99**とされている。〔注34〕⁴

日本価格が確定していない/販売経路により異なる機種もあるため、以下は**予算化用のシナリオ**であり実売価格ではない。

費目	PoC一台の予算モデル	コメント
RTX Spark 128GB PC	約90万円	US\$5,999.99×150円/\$という計画仮定
SSD/NAS/backup	10～30万円	patent corpusとmodel cache
UPS/周辺機器	5～15万円	desktop運用
Security/EDR/DLP	既存契約～追加費	ARM64対応確認
OSS inference	0円～	llama.cpp/vLLM等。ただしmodel licenseは別
RAG/vector DB	OSS～enterprise契約	部門規模次第
構築・評価	hardware以上になり得る	data整備、ACL、evaluationが主費用

費目	PoC一台の予算モデル	コメント
保守	OEM/enterprise support次第	3～5年運用を想定

つまり本番導入で支配的なのは、最終的にはGPU代より**文書connector、ACL、評価data、security、運用**になりやすい。

電力・冷却

140WはSoC TDPであり、PC全体のコンセント消費電力そのものではない。それでも最低限の規模感を把握するには使える。NVIDIA公式のdesktop TDP 140Wを基準に、電力単価を仮に35円/kWhとすると、

利用形態	SoC相当電力量	電気代試算
8時間×250営業日	280kWh/年	約9,800円/年
24時間×365日フル負荷	1,226kWh/年	約42,900円/年
上記24/7＋冷却を20%上乗せする計画仮定	—	約51,500円/年

TDP根拠はNVIDIA公式仕様、電力単価35円/kWhと冷却20%は本報告の予算仮定である。〔注35〕 31

従ってデータセンターGPUと違い、**電気代はRTX Spark導入の主要障壁にはなりにくい**。むしろ冷却不良によるthrottling、長時間運転時のfan noise、office UPS、故障時の代替機が運用上の論点となる。

On-prem、cloud、hybrid

モデル	長所	弱点	知財向け評価
RTX Spark完全 local	機密性、固定費、offline	concurrencyと最大品質に限界	非常に高い
社内GPU server	高throughput、一元管理	CAPEX、運用人材、冷却	大規模知財部に有力
Cloud GPU	elastic、最新GPU	時間課金、data governance	burst用途
Managed LLM API	frontier quality、運用不要	data送信、token課金、provider dependency	非機密/匿名化案件
Hybrid	local機密＋cloud高性能	policy routingが必要	最も推奨

クラウドとのbreak-evenは単にGPU時間で比較するのではなく、

$$\text{月間Local費用} = \text{ハード償却} + \text{電力} + \text{運用保守} + \text{人件費}$$

と、

月間Cloud費用 = 入力token + 出力token + embedding/rerank + storage + network + enterprisesecurity
を比較すべきである。

仮に90万円のPCを36か月償却すればhardware部分は**月2.5万円**なので、高頻度の内部翻訳・要約・RAGではlocalの経済性が成立しやすい。一方で利用が週数回ならcloudの方が合理的である。

競合・代替技術

Platform	Local AI memory / bandwidth	強み	弱み	知財用途評価
NVIDIA RTX Spark	128GB / 300GB/s	CUDA、TensorRT、Windows、FP4	70B dense decode帯域は中程度	Windows企業では最有力
AMD Ryzen AI Max+ 395	最大128GB LPDDR5X、256bit/8000MT/s	低cost PC選択肢、large unified memory	CUDA系 softwareとの互換性	高
Apple M5 Max	最大128GB / 614GB/s	RTX Sparkより高帯域、MLX	macOS、CUDAなし	高
Apple M5 Ultra	最大512GB / 1.2TB/s	巨大モデル容量・帯域で突出	価格、macOS、CUDAなし	Local LLM性能重視なら最有力候補
Intel Arc Pro B60	24GB/card / 456GB/s	multi-GPU、Windows/Linux	一枚の容量が小さい	7～14B向け
Graphcore Bow IPU	datacenter/IPU Pod型	独自AI architecture	PC代替でない、ecosystem限定	特殊案件
Cerebras CS-4/Cloud	wafer-scale/rack/cloud	非常に高いinference speed	desktop localではない	cloud/rack代替
Cloud B200等	HBM、大容量・TB/s級	throughput・scale	機密、OPEX、data residency	burst/frontier用

AMD Ryzen AI Max+ 395は最大128GB、256-bit LPDDR5X-8000、Radeon 8060S 40 graphics coresで、AMDは大規模ローカルLLM用途を明示している。〔注36〕 ³²

Appleは2026年のM5 Ultra Mac Studioで最大512GB unified memory、1.2TB/s、36-core CPU、最大80-core GPUを実現し、Apple自身が「巨大なLLMを完全にdevice上で動かせる」と位置付ける。M5 Maxでも614GB/sである。〔注37〕 ⁵ **大きなdense LLMのtoken generationだけを重視するなら、RTX SparkよりAppleの方が構造的に有利な可能性が高い。**

Intel Arc Pro B60は24GB GDDR6、456GB/sでmulti-GPU inferenceを想定している。一枚の帯域はRTX Sparkより大きい容量が24GBなので、70Bを一枚で保持するRTX Sparkとは設計思想が異なる。〔注38〕 ³³

Graphcoreの現行Bow IPUはPod/server型であり、机上のWindows Local LLM PCを直接置き換える製品ではない。〔注39〕 ³⁴ CerebrasのCS-4もrack-scale infrastructureで、同社はcloud inferenceも展開しているため、超低latencyが必要な中央サービスの代替候補であって、RTX Sparkの「個人/部門local PC」とは競争軸が異なる。〔注40〕 ³⁵

リスクと緩和策

リスク	知財で起こる例	緩和策
Hallucination	存在しない特許番号、判例、引用	RAG citation必須、retrieved source以外を根拠にさせない
Recall不足	重要先行技術を見逃す	keyword+classification+semanticのhybrid search、人間の調査継続
モデル劣化/更新	version更新後にclaim解析品質低下	model pinning、golden datasetによる regression test
Prompt injection	外部文書内の命令をagentが実行	untrusted data扱い、tool allowlist、文書とinstructionの分離
RAG poisoning	悪意あるPDFをknowledge baseへ投入	ingestion quarantine、署名/hash、source provenance
機密log漏えい	未公開発明がprompt historyに残る	raw-log最小化、暗号化、retention、RBAC
Supply chain	改変model/adaptor/library	model hash、SBOM、署名repository、malware scanning
Vendor lock-in	TensorRT依存で他GPUへ移行困難	OpenAI-compatible内部API、GGUF/ONNX/vLLM abstraction
ARM compatibility	DMS connectorやsecurity agentが動かない	PoCでARM64 support matrixを作成
Capacity不足	部門利用でqueueが長くなる	model replica、load balancing、cloud burst
著作権	training/outputが第三者権利を侵害	dataset provenance、類似性検査、人間による公開前review

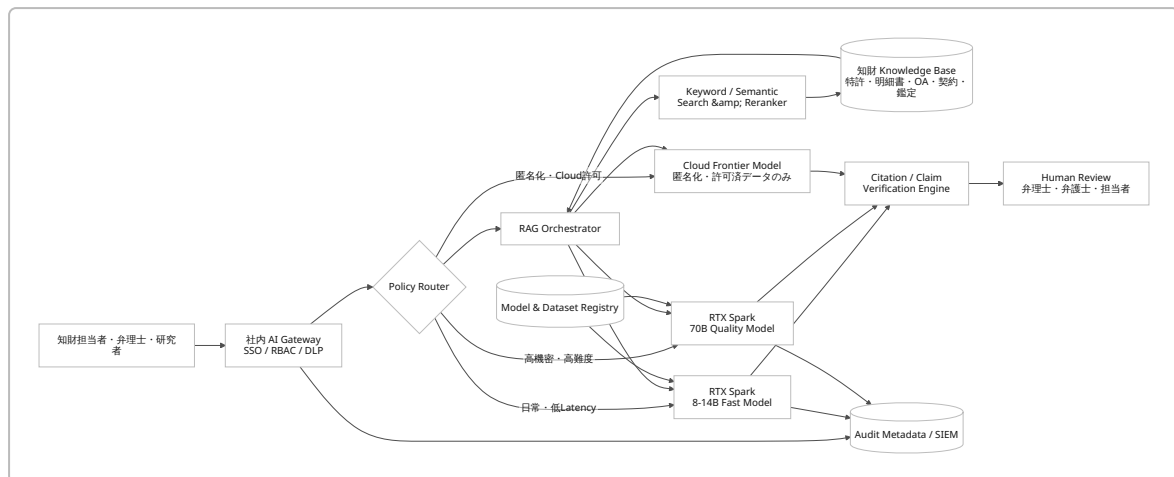
特にagent化する場合、LLMの「文章が間違う」リスクから「ファイル削除、メール送信、DMS更新、出願操作を間違う」リスクへ質的に変わる。NVIDIA/MicrosoftもRTX Sparkでagent向けidentity、containment、policy controlやOpenShellによるlocal/cloud routingを前面に出している。〔注41〕 36

したがって知財では、**read-only agent → draft agent → tool-using agent → autonomous agent**の順に権限を広げ、出願・契約承認など不可逆操作には人間承認を残すべきである。

推奨アーキテクチャと導入戦略

推奨する知財Local AIアーキテクチャ

最も推奨するのは、「ユーザーが各LLMを直接選ぶ」方式ではなく、社内gatewayが機密性・task・必要品質によってmodelを選ぶ構成である。



このarchitectureの要点は四つある。

第一に、機密区分がmodel selectionより先に来る。「70Bで足りなければクラウド」という単純fallbackではなく、cloud許可dataかをPolicy Routerが判定する。

第二に、fast modelとquality modelを分ける。7～14Bは検索query作成、翻訳、分類、要約、metadata抽出に使い、70Bは複雑なclaim comparison、契約推論、難しいdraftingに限定する。

第三に、RAG結果には必ずsource identifierを付ける。LLM回答から元特許のpublication number、paragraph、claim番号、契約条項へ戻れることを本番要件とする。

第四に、modelとdataを同じくらい厳格にversion管理する。AIの出力はモデルだけでなく、retrieval corpus、embedding model、chunking、reranker、prompt templateで変わるからである。

PoCから本番までのロードマップ

Phase	推定期間	対象	Exit criteria
PoC	6～8週間	RTX Spark 1台、2～3モデル、翻訳/RAG/claim check	品質・速度・ARM compatibility確認
Pilot	8～12週間	5～20名、DMS/RAG/SSO連携	golden test、security test、利用率・時間短縮を確認
Production準備	8～12週間	logging、DR、model registry、DLP	法務・security・知財責任者承認
本番拡張	継続	複数RTX Spark+cloud routing	SLA、capacity、model regressionを継続監視

従って、単純なdesktop demoなら2か月程度、**知財機密dataを扱う本番systemとしては概ね6～9か月**を計画値とするのが妥当である。これは製品vendorの公称導入期間ではなく、本報告で想定する企業向け工程見積りである。

PoCでは「とにかく最大120Bを動かす」より、

8B/14B Q4、70B Q4、cloud frontier model

の三者を同じ知財evaluation setで比較する方が有益である。14Bで95%の案件が解けるなら、70Bを常用しない方がthroughputも電力も良い。

導入判断チェックリスト

判断質問	YesならRTX Spark適性
未公開発明・契約・鑑定など外部送信しにくいdataが多いか	非常に高い
Windowsを中心とした既存知財workflowを維持したいか	高い
CUDA/TensorRT資産・人材を活用したいか	高い
70Bモデルを一台で保持したいか	高い
70Bを30 token/s以上で常時使いたい	低い。別GPU/serverを検討
120B以上のdense modelを高速に使いたい	低い
一台を20～50人が同時利用したい	低い
offline/air-gapped環境が必要か	非常に高い
Mac/MLXへの移行が可能で、最大Local LLM性能が最優先か	M5 Ultraも必ず比較
利用量が非常に少ないか	Cloudの方が経済的な可能性
出力根拠をcitation付きで監査する必要があるか	Local RAG構成と好相性

短期戦略

今後0～3か月では、**128GB desktopを一台だけ導入する**。モデルは「小型高速」「70B品質優先」の二系統にし、最初の業務を①外国特許/OA翻訳・要約、②社内特許RAG、③claim/specification consistency checkに限定する。

この三つは、機密性の価値が高く、正解を原文と照合しやすく、AIが誤っても最終提出前に人間が発見できるため、risk-adjusted ROIが高い。

この段階ではfine-tuningを急がず、まず**RAG+prompt+reranker**でどこまで精度が出るかを見る。fine-tuningしてしまうと知識更新、説明可能性、dataset copyright/provenanceという新しい運用負担が生じる。

中期戦略

3～12か月では、知財部門共通の**Local AI Gateway**を設ける。3～5台程度まで増える場合、70Bを全nodeに分散して一つのmodelとして動かすより、まず各nodeへモデルを複製し、requestをload-balanceする。

ここで、

- confidential → local only
- internal but sanitizable → local first/cloud fallback
- public → best available model

というdata-classification routingを実装する。

必要ならNeMo/LoRAで、社内approved exampleに基づく「claim drafting style」「office-action response style」「contract issue spotting」adapterを作る。NVIDIAのNeMo/NIM stackはLoRAをbase modelと分離し、dynamic multi-LoRA inferenceもサポートしているため、知財部、法務部、研究部ごとにbase modelを共用する設計が可能である。〔注42〕 37

長期戦略

12～24か月以降は、RTX Sparkを「PC」としてではなく**企業内のPersonal/Departmental AI tier**として位置付けるべきである。

最終的な三層構造は、

Personal Edge: RTX Spark / Apple Silicon等

Department AI: GPU server / multi-node inference

Enterprise/Cloud Frontier: B200級cloud・managed frontier models

になる可能性が高い。

ここでRTX Sparkの戦略的価値は、最大LLM性能ではなく、「**機密データのそばに128GB級CUDA AIを配置できること**」にある。Apple M5 Ultraのように容量・帯域で明確に上回る競合が既に存在する以上、NVIDIAの長期優位性はハードウェア単体ではなく、CUDA、TensorRT-LLM、NeMo、vLLM周辺、Windows agent security、cloud Blackwellへのportabilityを含むecosystemに依存する。NVIDIA自身もRTX Sparkについて、localで構築し必要時にはcloudへoffloadできる同一CUDA stackを重要な特徴としている。〔注43〕 38

知財分野についての最終評価は「積極導入、ただし70B万能論は採らない」である。RTX Sparkが最も変えるのは、モデルの賢さそのものではない。これまで「機密だから生成AIに入れられない」「70Bはクラウドでしか動かない」という二つの制約を同時に緩める点である。

そのため最も合理的な投資方針は、

14B前後の高速Local LLMを日常の知財コパイロットにする → 70Bを高難度・高機密案件に使う → RAGで事実根拠を固定する → cloud frontier modelは匿名化可能な例外経路にする → 最終的な法的・知財判断は人間が保持する

という構成である。

この形であれば、RTX Sparkは単なる「高性能AI PC」ではなく、**企業の未公開知財を外部へ出さずに検索・読解・翻訳・比較・ドラフティングできる“知財AIワークベンチ”**を、**100万円前後のPCクラスに落とし込む転換点**になる可能性がある。一方、組織全体を対象とする大量同時推論やfrontier-class dense modelの高速処理には、Appleの大容量unified-memory機、GPU server、クラウドBlackwell/Cerebras等を併用するハイブリッド戦略が、中長期的には最も堅牢である。

1 2 7 11 31 38 Slim Laptops & Small Desktops | NVIDIA RTX Spark

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

3 12 14 17 How NVIDIA DGX Spark's Performance Enables Intensive AI Tasks | NVIDIA Technical Blog

<https://developer.nvidia.com/blog/how-nvidia-dgx-sparks-performance-enables-intensive-ai-tasks/>

- 4 Surface RTX Spark Dev Box: AI Development PC - Microsoft
<https://www.microsoft.com/en-us/store/configure/surface-rtx-spark-dev-box/8wg5vqp0x4gr>
- 5 Apple introduces new Mac Studio with M5 Max and M5 Ultra - Apple
<https://www.apple.com/newsroom/2026/08/apple-introduces-new-mac-studio-with-m5-max-and-m5-ultra/>
- 6 System Overview — NVIDIA Windows on Arm Porting Guide
<https://docs.nvidia.com/rtx-spark/rtx-spark-porting-guide/0.1.0/overview.html>
- 8 36 NVIDIA and Microsoft Reinvent Windows PCs for the Age of Personal AI | NVIDIA Newsroom
<https://nvidianews.nvidia.com/news/nvidia-microsoft-windows-pcs-agents-rtx-spark>
- 9 Unified Memory Architecture — NVIDIA Windows on Arm Porting Guide
<https://docs.nvidia.com/rtx-spark/rtx-spark-porting-guide/0.1.0/uma/index.html>
- 10 13 15 16 Scaling Autonomous AI Agents and Workloads with NVIDIA DGX Spark | NVIDIA Technical Blog
<https://developer.nvidia.com/blog/scaling-autonomous-ai-agents-and-workloads-with-nvidia-dgx-spark/>
- 18 PEFT in NeMo 2.0 — NVIDIA NeMo Framework User Guide
https://docs.nvidia.com/nemo-framework/user-guide/25.04/sft_peft/peft_nemo2.html
- 19 Red Hat Announces Definitive Agreement to Acquire Neural ...
<https://www.redhat.com/en/about/press-releases/red-hat-acquire-neural-magic>
- 20 AI use at the EPO | epo.org
<https://www.epo.org/en/news-events/in-focus/digital-innovations/ai-use-at-the-EPO>
- 21 22 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/
- 23 人工知能関連技術の研究開発及び活用の推進に関する法律 ...
https://www8.cao.go.jp/cstp/ai/ai_act/ai_act.html
- 24 [PDF] AI 事業者ガイドライン - 経済産業省
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20260331_1.pdf
- 25 AIと著作権について | 文化庁
<https://www.bunka.go.jp/seisaku/chosakuken/aiandcopyright.html>
- 26 https://www.bunka.go.jp/seisaku/chosakuken/pdf/94097701_02.pdf
https://www.bunka.go.jp/seisaku/chosakuken/pdf/94097701_02.pdf
- 27 AI Act | Shaping Europe's digital future - European Union
<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- 28 Data protection explained - European Commission
https://commission.europa.eu/law/law-topic/data-protection/data-protection-explained_en
- 29 AI Risk Management Framework | NIST
<https://www.nist.gov/itl/ai-risk-management-framework>
- 30 生成式人工智能服务管理暂行办法 - 中国网信网
https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
- 32 AMD Ryzen™ AI Max+ 395
<https://www.amd.com/en/products/processors/laptop/ryzen/ai-300-series/amd-ryzen-ai-max-plus-395.html>
- 33 Intel® Arc™ Pro B60 Graphics - Product Specifications
<https://www.intel.com/content/www/us/en/products/sku/243916/intel-arc-pro-b60-graphics/specifications.html>

34 **Lenovo servers now available for Graphcore IPU systems**

<https://www.graphcore.ai/posts/lenovo-servers-now-available-for-graphcore-ipu-systems>

35 **Cerebras**

<https://www.cerebras.ai/>

37 **Parameter-Efficient Fine-Tuning - NVIDIA Documentation**

<https://docs.nvidia.com/nim/large-language-models/1.8.0/peft.html>