

NVIDIA RTX Spark搭載PC(128GB)における知財業務特化型ローカルLLM構築要件と国産モデル選定レポート

Gemini 3.8 Flash

ハードウェア特性と知財LLM推論基盤の適合性分析

NVIDIA RTX Spark(開発コード名:N1X)は、20コアのArmベースGrace CPUと、6,144基のCUDAコアを備えるBlackwell世代のRTX GPUを集積したSoCであり、最大128GBのLPDDR5Xユニファイドメモリ(Unified Memory Architecture: UMA)を採用したWindows on Arm対応の次世代エッジコンピューティング基盤である¹。本プラットフォームが知財業務のローカルAI実行基盤として注目される最大の要因は、CPUとGPUが単一の物理メモリ空間をNVLink-C2C相互接続を介してシームレスに共有する点にあり、ディスクリートGPU(dGPU)では複数枚のグラフィックスカードを要する70B~100Bパラメータ規模の大規模言語モデルを単一の端末上で展開できる点にある¹。

ローカル推論基盤の設計において最も厳密に評価すべき技術的ボトルネックは、メモリ帯域幅である。RTX Sparkのメモリ帯域幅は約273~300 GB/sに設計されており⁵、dGPUのGeForce RTX 5090(1,792 GB/s)やデータセンター向けGPU(3 TB/s超)と比較すると帯域幅の制限が存在する⁵。自己回帰的な言語モデルのトークン生成フェーズは典型的なメモリバウンド(Memory-bound)処理であり、出力スループットの上限はパラメータの読み出し速度に厳格に律速される。モデル全体の重みデータサイズを M_{weights} (GB)、実効メモリ帯域幅を B_{mem} (GB/s)とした場合、1トークン生成に要する最小時間 T_{token} は理論上以下の関係式で定まる。

$$T_{\text{token}} \geq \frac{M_{\text{weights}}}{B_{\text{mem}}}$$

4ビット量子化(INT4/Q4_K_M)を施した70Bモデル(重み容量約40~43 GB)を読み出す場合、300 GB/sの帯域幅下における理論上の生成速度は約4.5~6.5 tokens/s

にとどまり、1,000トークンの出力に数分を要する計算となる[cite: 5]。一方で、20B前後のモデル(重み容量約12~15GB)であれば毎秒

15~22 tokens/sの実用的な対話速度が確保される⁹。したがって、リアルタイムな対話性を要する推敲タスクと、バッチ的に実行される精密な文献解析タスクとで要求仕様を明確に分離する必要がある。

また、知財業務においては特許出願明細書や審査請求書などの長大なテキストを扱うため、KVキャッシュ(Key-Value Cache)が消費するメモリ領域の計算が不可欠である。Grouped Query

Attention(GQA)を採用したモデルにおいて、トランスフォーマー層数を N_{layer} 、Key-Valueヘッド数を $N_{\text{kv_heads}}$ 、ヘッド次元数を D_{head} 、コンテキスト長を L 、トークンあたりの精度バイト数を

P_{bytes} と置いた場合、KVキャッシュ容量 M_{kv} は次式で算出される。

$$M_{\text{kv}} = 2 \times N_{\text{layer}} \times N_{\text{kv_heads}} \times D_{\text{head}} \times L \times P_{\text{bytes}}$$

96B～100B規模のモデルにおいてコンテキスト長32,768トークン（FP16精度、2バイト）を確保する場合、KVキャッシュだけで約11.5 GBを占有する。RTX Sparkの128GB構成は、OSおよびシステム予約領域（約8～12 GB）を除いた後も約115 GBの実効領域を保持できるため、100B級モデルの4ビット量子化（約55～60 GB）を展開した上で、32k超のKVキャッシュ領域を余裕をもって確保できる卓越したキャパシティを提供する⁴。

項目	NVIDIA RTX Spark (128GB構成)	一般的ハイエンドPC (RTX 5090単体)	サーバー向け構成 (RTX 6000 Ada単体)
プロセッサ構成	Grace CPU (20コア) + Blackwell RTX GPU ¹	x86 CPU + Blackwell RTX GPU (21,760コア) ⁵	x86 CPU + Ada Lovelace GPU (18,176コア)
メモリ形態	ユニファイド LPDDR5X (UMA) ¹	システムRAM + 独 立GDDR7 ⁵	システムRAM + 独 立GDDR6 ECC
有効VRAM容量	最大128 GB (共有) ¹	32 GB ⁵	48 GB
メモリ帯域幅	約273～300 GB/s ⁵	1,792 GB/s ⁵	960 GB/s
最大動作モデル規模	100B～120B級 (INT4) ¹	32B級 (INT8) / 70B (オフロード必須) ⁵	70B級 (INT4～ INT8)
知財タスク適性	長大特許文献の一 括読解・深層解析	高速な短文対話・リ アルタイム推敲	安定した中規模並列 処理

検討対象とすべき主要国産LLMの選定

知財業務における特許請求の範囲（クレーム）の分析、構成要件分解、先行技術調査の比較検討、拒絶理由通知への対応案作成といった実務への適合性を踏まえ、RTX Spark（128GB）上で稼働可能な主要国産オープンモデルを精査した。

モデル名	開発組織	パラメータ数	推奨量子化形	推定必要メモリ	最大コンテキ	ライセンス形態	知財業務にお
------	------	--------	--------	---------	--------	---------	--------

			式	(重み +KV)	スト長		ける主 な適合 領域
Stockm ark-2-1 00B-Ins truct	ストック マーク	96B (公 称 100B)	INT4 (Q4_K_ M)	約68～ 75 GB (32k時)	32,768 ¹ ₁	MIT License ₁₁	特許文 献精 読、構 成要件 分解、 先行技 術比較 ₁₃
Llama- 3.1-Sw allow-7 0B-Inst ruct	東京科 学大 / 産総研	70B	INT4 (Q4_K_ M)	約50～ 65 GB (32k時)	128,000 ₁₄	Llama 3.1 Commu nity License ₁₅	意見書・ 補正書 の論理 構成推 敲、拒 絶理由 通知の 反論検 討
CALM3 -22B-C hat	サイ バー エージェ ント	22B	INT8 (Q8_0) / FP16	約28～ 48 GB (16k時) ₁₀	8,192～ 16,384	Apache License 2.0 ¹⁶	高速な クレーム 推敲対 話、発 明提案 書の要 約、知 財契約 レビュー
PLaMo 2 8B	Preferre d Network s	8B	FP16 / INT8	約16～ 22 GB (32k時)	32,768 ¹ ₉	PLaMo Commu nity License ₂₀	英文特 許・和文 特許間 の高精 度な翻 訳支援 ₁₉

Sarashina2.2-Vision-3B	SB Intuitions	3.8B (VLM)	FP16 / BF16	約8～10 GB ²¹	画像入力対応 ²¹	MIT License ²¹	特許図面、回路図、動作フローチャートの画像解析 ²¹
Fugaku-LLM-13B-instruct	富士通 / 理研 他	13B	INT8 / FP16	約16～28 GB	2,048～4,096 ²⁴	Fugaku-LLM Terms of Use ²⁵	定義文抽出、短文での構成要素抽出(長文処理には非推奨)

ピックアップモデルの詳細評価と知財業務適合性

Stockmark-2-100B-Instruct

ストックマークが経済産業省およびNEDOのGENIACプロジェクトを通じてスクラッチ開発したStockmark-2-100B-Instructは、知財実務において最も直接的な適合性を有する基盤モデルである¹¹。本モデルの最大の特徴は、2兆トークンに及ぶ事前学習データの中に約375億トークンの日本語特許データを含んでいる点にある¹¹。一般的なLLMが苦手とする特許文献独特の文体、すなわち「～を備え、前記～は…することを特徴とする」といった厳密な係り受け構造や従属クレームの多重引用構造に対して、極めて低いパープレキシティを示す。

特許請求の範囲から発明の構成要件を過不足なく分節化し、引用文献の記載事項と1対1で対応付けられるクレームチャート作成業務において、他の汎用オープンモデルを明確に上回る解釈精度を発揮する。GQAの採用によりKVキャッシュの肥大化が抑制されており、32kトークンの長大なコンテキストを4ビット量子化重みとともにRTX Sparkの128GBメモリ上へ余裕をもって常駐させることが可能である¹¹。メモリ帯域幅の制限によりトークン生成速度は毎秒3～5トークン程度に制限されるが、バックグラウンドでの先行技術文献の自動読解や、無効審判資料の対比作業といったバッチ指向の重タスクにおいて極めて高い実用性を提供する。

Llama-3.1-Swallow-70B-Instruct

東京科学大学(旧東京工業大学)と産業技術総合研究所のSwallowプロジェクトが開発した本モデルは、MetaのLlama 3.1 70Bをベースに大規模な日本語コーパスによる継続事前学習を実施したモデルである¹⁵。ベースモデルが元来備えている優れた論理的推論力、コード理解力、数学的处理能力を保持したまま、自然な日本語表現および日本固有の文脈理解が注入されている¹⁵。

知財業務における主要なユースケースは、特許庁からの拒絶理由通知書に対する意見書および補正書のドラフティングである。新規性・進歩性の否定論理に対し、本願発明と主引用発明・副引用発

明の間の課題の差異、設計事項の範囲、動機付けの有無、阻害要因の存在といった複雑な法的論理を精緻に組み立てるタスクにおいて高い追従性を示す。ベースモデル由来の128kコンテキスト処理能力を活かし、明細書全体(数万字規模)を一括入力した上で、明細書の「背景技術」「課題を解決するための手段」「発明の効果」の間に論理的な断絶やサポート要件(特許法第36条第6項第1号)違反が存在しないかを網羅的に検証する用途に極めて適している¹⁴。

CALM3-22B-Chat

サイバーエージェントが自社開発したCALM3-22B-Chatは、22Bパラメータという中規模クラスでありながら、高度な日本語指示チューニングとDPO(Direct Preference Optimization)が施された実用モデルである¹⁶。本モデルの特筆すべき優位性は、RTX Spark環境における圧倒的な実行効率にある。重みサイズがFP16でも約45 GB、INT8量子化であれば約24 GBに収まるため、300 GB/sのメモリ帯域幅であっても毎秒15～22トークンという快適な生成速度を維持できる⁹。

弁理士や社内知財担当者が端末に向かい、発明者からのヒアリングシートを対話形式で推敲する作業や、クレームの構成要素をより広範な上位概念へと拡張するアイデア出し、または他社との秘密保持契約書(NDA)や共同研究開発契約書の条項レビューなど、即時応答が不可欠なインタラクティブ業務において中核的な役割を担う。ライセンスがApache License 2.0で配布されており、商用利用における法的リスクが極小化されている点も企業知財部にとって大きな利点である¹⁶。

Sarashina2.2-Vision-3B

SB Intuitionsが開発したSarashina2.2-Vision-3Bは、3.8Bパラメータという極小サイズながら、日本語の図表理解や文書OCRベンチマークにおいて同クラス最高峰のスコアを達成しているマルチモーダル言語モデル(VLM)である²¹。特許業務はテキストのみで完結するものではなく、明細書に添付された構造図、回路図、ブロック図、シーケンス図といった「図面」の読解が権利解釈に決定的な影響を与える。

本モデルをRTX Spark上に常駐させることで、特許図面内の引き出し線と符号(符号10、符号12等)の対応関係を自動抽出する前処理や、制御ロジックのフローチャート画像から分岐条件を自然言語のテキスト記述へ変換する処理が可能となる²¹。抽出されたテキスト情報をStockmark-2-100BやSwallow-70Bへ引き渡す多段階パイプラインを組むことにより、テキストと図面が高度に融合した知財解析環境がローカル完結で実現する²¹。

知財実務におけるローカル運用の技術的・法的注意事項

構成要件解釈におけるハルシネーションと根拠明示の義務付け

LLMの生成テキストに不可避免的に生じるハルシネーションは、厳密な文言解釈が求められる知財実務において致命的な法的リスクを招く。特に特許請求の範囲の文言解釈においては、一般的な日常語と特許法上の技術用語とで概念の射程が異なるケースが多々存在する。モデルが学習データ内の一般的な文脈に引きずられ、出願人が明細書中で明示的に定義した「辞書的定義」を無視して一般的な解釈を優先してしまう現象や、クレームの構成要件を無意識のうちに限定解釈・広義解釈にすり替えてしまうリスクが潜在する。

この課題を回避するためには、モデルに対するゼロショットでの自由記述生成を厳格に制限し、出力を構成要件ごとの構造化フォーマット(JSONや表形式)に限定するプロンプトエンジニアリングが必須となる。具体的には、モデルが提示するすべての解釈や比較判断に対して、根拠となった明細書中の「段落番号【00XX】」および「請求項番号」を明示的に出力させ、人間の知財担当者がワンクリックで原文と対照確認できるグラウンディング体制を設計することが実務上の絶対条件となる。

文脈埋没現象の克服と長文コンテキスト管理

RTX Sparkの128GBユニファイドメモリは、32,768トークンから128,000トークンにおよぶ長大なコンテキストの展開を可能にするが¹、モデルのアーキテクチャに起因する「Lost in the Middle (文脈中央部の情報埋没)」現象への警戒を怠ってはならない。特許明細書は全体の構造が「技術分野」「背景技術」「課題」「手段」「実施例」「特許請求の範囲」と体系化されているが、入力コンテキストの中央付近に配置された実施例の具体的数値や比較例の記載を、モデルのアテンション機構が見落とす傾向が知られている。

したがって、明細書全体を無加工のままプロンプトに一括投入する運用は推奨されない。特許文献のXML構造や段落構成をパースし、クレーム分析時には「特許請求の範囲」と「課題を解決するための手段」を優先的にプロンプトの先頭および末尾に配置するレイアウト最適化や、実施形態ごとに分割して局所的な要約・構成抽出を行った後に統合するチャンキング戦略を導入することが、解析精度の向上に不可欠である。

出願前機密保持と新規性喪失リスクの完全遮断

企業知財部門が外部のクラウドAIサービスを忌避し、RTX Sparkを用いたローカルLLM構築を選択する根本的な動機は、未公開発明に関する最高レベルの機密保持である。特許法第29条第1項が規定する新規性要件において、特許出願前に発明が公然知られた状態になった場合、その発明は特許を受けることができない。仮に機密情報がクラウド事業者のログや学習データに吸い上げられた場合、新規性喪失の法的リスクに直面する可能性がある。

ローカル環境を構築するにあたっては、推論エンジン (llama.cpp、vLLM、Ollama等) がバインドするエンドポイントをローカルホスト (127.0.0.1) に厳格に限定し、Windows Defender Firewall等のパケットフィルタリングによって当該プロセスの外部アウトバウンド通信を完全に遮断するネットワーク隔離を施さなければならない。加えて、Windowsのページファイルや一時キャッシュへの平文書き込みを通じた情報漏洩を防ぐため、BitLockerによるストレージ全体の暗号化を施すとともに、推論ログをセッション終了時に揮発・完全消去する運用ポリシーの徹底が求められる。

オープンモデルの商用ライセンス条項と企業コンプライアンス

オープンソースまたはオープンウェイトとして配布されている国産モデルであっても、その利用規約はモデルごとに大きく異なるため、知財部門自身による契約条項の厳密な精査が求められる。Stockmark-2-100BやSarashina2シリーズに適用されているMIT License、およびCALM3-22Bに適用されているApache License 2.0は、商用利用や内部業務利用が無償で許諾されており、改変や社内ツールへの組み込みに対する法的制約が最も少ない¹¹。

一方で、Llama-3.1-Swallowに適用されるMeta Llama 3.1 Community Licenseでは、月間アクティブユーザー数が7億人を超える大規模事業者に対する利用許諾申請義務や、生成AIモデルの競合改善への利用制限といった特有の条項が含まれている¹⁵。さらに、Preferred NetworksのPLaMo 2系モデルに適用されるPLaMoコミュニティライセンス契約では、直近事業年度の売上高が10億円を超える企業が業務利用する場合、無償ライセンスの対象外となり別途有償契約の締結が義務付けられている点に格別の注意を要する²⁰。社内利用の規模や自社の資本構成と各モデルの利用規約を照合し、ライセンスコンプライアンスを完全に維持した運用体制を構築しなければならない。

128GBユニファイドメモリ環境を活かした複合モデル連携システム設計

知財業務における多様な要求仕様を単一のLLMで網羅しようとすることは、計算効率および業務品

質の双方において合理性を欠く。RTX Sparkの128GBユニファイドメモリが提供する真の価値は、特化領域の異なる複数のモデルをメモリ上に同時常駐させ、パイプライン的に連携動作させる「複合モデル協調アーキテクチャ」の実現にある¹。

具体的なシステム構成として、文献読解と精密な構成要件抽出を担う主軸エンジンには、特許データで事前学習されたStockmark-2-100B-InstructをINT4量子化（約55 GB）で展開する¹¹。これと並行して、明細書添付図面やブロック図の解析を担当するSarashina2.2-Vision-3BをFP16精度（約8 GB）でメモリ上に常駐させる²¹。さらに、ユーザーとの対話的なドラフティングや高速なテキスト要約・言い換え処理を担うサブエンジンとして、CALM3-22B-ChatをINT8量子化（約24 GB）で配備する¹⁰。これら3種類のモデルを同時にメモリへ展開した場合の重み総容量は約87 GBであり、OSおよび推論エンジンの基本フットプリントを考慮しても約25～30 GBの空きメモリ領域が残される。この余剰メモリは、Stockmark-2-100Bが32kコンテキストを処理する際のKVキャッシュ、およびCALM3-22Bの動的コンテキスト領域として配分するのに十分な容量である¹¹。

実際の知財ワークフローにおいては、まず入力された特許公報のPDFから図面データがSarashina2.2-Vision-3Bに送られ、図面中の符号一覧およびフローチャートのテキスト記述が生成される²¹。続いて、抽出された図面テキストと明細書の特許請求の範囲がStockmark-2-100Bに入力され、構成要件ごとの詳細なクレームチャートおよび先行技術との対比表が作成される¹¹。最後に、その構造化データを参照しながら、担当弁理士がCALM3-22B-Chatと対話することで、毎秒20トークンを超える高レスポンス環境下で意見書や補正書の文案を迅速に練り上げることが可能となる⁹。このようなマルチモデル協調パイプラインをローカルPC内部で完結させることにより、出願前機密の完全な保護、特許図面を含む多角的な情報処理、そして深層推論と対話応答性の両立という、高度知財業務に求められるすべての技術的・法的要件を達成することが可能となる。

引用文献

1. NVIDIA RTX Spark - Summary of Windows AI PCs with up to 128GB, <https://note.com/npaka/n/n82358351b2c5?hl=en>
2. Nvidia RTX Spark - Wikipedia, https://en.wikipedia.org/wiki/Nvidia_RTX_Spark
3. DGX Spark / RTX Spark / DGX Station の比較 | npaka - note, <https://note.com/npaka/n/nc8bef1e0a3ed>
4. Microsoft Surface Laptop Ultra with Nvidia RTX Spark chip now up for pre-order at \$2,599, <https://timesofindia.indiatimes.com/technology/laptops-pc/microsoft-surface-laptop-ultra-with-nvidia-rtx-spark-chip-now-up-for-pre-order-at-2599/articleshow/134785284.cms>
5. DGX SparkとRTX 5090の違い | ゲームは動く？AIの生成速度は5090, <https://gaming-st.com/gamingpc/dgx-spark-vs-rtx-5090-gaming-local-ai/>
6. Alibaba Prefired for NVIDIA's RTX Spark | by Ekansh Jain - Medium, <https://medium.com/@ekansh.jain2011/alibaba-prefired-for-nvidias-rtx-spark-aa8bd183ed81>
7. Nvidia RTX Spark - Hacker News, <https://news.ycombinator.com/item?id=48352939>
8. Reading the RTX Spark specs. 128GB of unified memory alone, <https://note.com/thegadget/n/n297abe57d793?hl=en>

9. N番煎じでCyberagent社のCALM3-22B-Chat(量子化版)をDatabricks,
<https://qiita.com/isanakamishiro2/items/5dea69a3cb33f8ac51e0>
10. Aratako/calm3-22b-RP-v2-GGUF - Hugging Face,
<https://huggingface.co/Aratako/calm3-22b-RP-v2-GGUF>
11. Stockmark-2-100B-Instruct - NVIDIA API Documentation,
<https://docs.api.nvidia.com/nim/reference/stockmark-stockmark-2-100b-instruct>
12. stockmark/Stockmark-2-100B-Instruct - Hugging Face,
<https://huggingface.co/stockmark/Stockmark-2-100B-Instruct>
13. stockmark/stockmark-100b - Hugging Face,
<https://huggingface.co/stockmark/stockmark-100b>
14. LLaMaモデル徹底比較！その進化と選び方,
<https://omomuki-tech.com/archives/2184>
15. 【Llama 3.1 Swallow】東工大の日本語に強いLLM！ | EdgeHUB,
<https://highreso.jp/edgehub/machinelearning/llama31swallow.html>
16. サイバーエージェントの日本語LLM「CyberAgentLM3」の使い方！,
<https://highreso.jp/edgehub/machinelearning/cyberagentlm3.html>
17. 日本語LLMの新バージョン「CyberAgentLM3」を公開 | gihyo.jp,
<https://gihyo.jp/article/2024/07/calm3>
18. schroneko/calm3-22b-chat - Ollama,
<https://ollama.com/schroneko/calm3-22b-chat>
19. PFNの国産大規模言語モデルPLaMo 2.0 Prime、2025年日経優秀,
<https://prtimes.jp/main/html/rd/p/000000020.000156310.html>
20. PLaMoコミュニティライセンス契約には,
<https://plamo.preferredai.jp/info/plamo-community-license-ja>
21. Sarashina2.2-Vision-3Bで社内資料を完全ローカル分析！日本語特化,
<https://gai.workstyle-evolution.co.jp/2025/12/02/sarashina2-2-vision-3b-japanese-vlm-local-deployment-enterprise-guide/>
22. Sarashina2.2-Vision-3B: コンパクトかつ性能が高いVLMの公開,
<https://www.sbintuitions.co.jp/blog/entry/2025/11/25/100000>
23. Sarashina2-Vision: 日本語特化の大規模視覚言語モデルの公開,
<https://www.sbintuitions.co.jp/blog/entry/2025/03/17/111659>
24. nebel/fugaku-llm - Ollama, <https://ollama.com/nebel/fugaku-llm>
25. ローカルLLM日本語対応モデルおすすめ8選,
<https://aoi-localllm.com/local-llm-japanese/>
26. 【Use Case vol.1】パナソニックHD×ストックマーク 国内最大規模,
https://www.meti.go.jp/policy/mono_info_service/geniac/geniac_magazine/usecase_01.html
27. パナソニックHDとストックマーク、国内最大規模(1000億,
<https://news.panasonic.com/jp/press/jn240702-3>
28. LLMのコード能力を向上させる取り組み - Zenn,
https://zenn.dev/yokota_lab/articles/65dbdcd56ab8bd
29. cyberagent/calm3-22b-chat-selfimprove-experimental - Hugging Face,
<https://huggingface.co/cyberagent/calm3-22b-chat-selfimprove-experimental>
30. sbintuitions/sarashina2-70b - Hugging Face,
<https://huggingface.co/sbintuitions/sarashina2-70b>

31. PLaMo 2の8Bモデルをオープンなライセンスで公開: 条件付きで商用,
<https://www.preferred.jp/blog/tech/plamo-community-license>