

NVIDIA RTX Spark（128GB）による知財業務向けローカル LLM の選定と注意事項

— 2026 年 10 月時点の調査報告 —

2026 年 10 月 10 日

Claude Opus 5.5

凡例 本文の「（分析）」「（推測）」は筆者の分析・推測を示し、それ以外は出典に基づく事実である。表中の記号は、○＝確認済み、△＝一部が二次情報、？＝未確認、×＝不可を表す。上付きの番号は文末の参考文献に対応し、参考文献の「一次資料」は公式発表・モデルカード・論文・公的文書を指す。

1. 結論（要約）

主力候補は 2 つである。中国製モデルを許容する場合は Qwen3.5-122B-A10B（NVFP4/Q4 量子化）、許容しない場合は gpt-oss-120b（MXFP4）またはその日本語強化版の GPT-OSS Swallow 120B である。これに高速な補助モデル（Qwen3.6-35B-A3B または国産の LLM-jp-4.1）と日本語向け埋め込みモデル（ruri-v3）を組み合わせる。RTX Spark の性能を決めるのは計算力よりメモリ帯域（公称 300GB/s）であるため、総パラメータは大きく、1 トークンあたりに動くパラメータは小さい MoE（Mixture of Experts）モデルが最も適している（分析）。

- ・ **ハードウェア**：「NVIDIA RTX Spark」は実在する正式製品名で、2026 年 5 月 31 日に発表された（チップ名 N1X）^[1]。Windows 11 on Arm 向けの SoC で、Blackwell GPU（6,144 CUDA コア）、20 コアの Grace CPU、最大 128GB のユニファイドメモリ、最大 1PFLOP（FP4）を備える^[2,3]。DGX Spark（GB10、Linux/DGX OS、273GB/s）と中身はほぼ同じだが、OS と製品の位置付けが異なる^[4]。ネイティブ Linux の公式サポートは発表されていない^[5]。
- ・ **推奨構成**：主力は Qwen3.5-122B-A10B または gpt-oss-120b（いずれも約 60～75GB で、長いコンテキスト用の余裕も残る）。補助は Qwen3.6-35B-A3B、LLM-jp-4.1-32B-A3B、Qwen3 Swallow。RAG 用の埋め込みは ruri-v3-310m、補助に PLaMo-Embedding-1B。中国製を避ける場合は、gpt-oss 系、Nemotron 3 Super、Gemma 4、LLM-jp-4.1 で全体を構成できる（分析）。
- ・ **注意点**：ローカル運用でも安全とは限らない。テレメトリやクラウドへの自動ルーティング、改ざんされた派生モデル、エージェントの権限、ハルシネーション（特許番号・判例・条文の誤り）に注意が必要である。日本弁理士会のガイドラインは、生成 AI から得られた生成物を活用する際には「弁理士としてその正確性を確認する必要がある、最終的には弁理士が責任をもって提供するべきである」としている^[6]。

- ・ 使い分け：無効論や FTO の結論など最も難しい判断はクラウドのフロンティアモデルや専門家に任せ、ローカル LLM は下ごしらえ・一次整理・ドラフト作成に使うのが現実的である（分析）。

2. ハードウェアの確認

2.1 確認済みの事実

表 1 RTX Spark と DGX Spark の比較

項目	RTX Spark (N1X)	DGX Spark (GB10)
発表・発売	2026 年 5 月 31 日発表（GTC Taipei／Computex、NVIDIA と Microsoft の共同発表） ^[1,7]	2025 年 3 月発表、2025 年 10 月 15 日発売 ^[8]
GPU	Blackwell RTX、6,144 CUDA コア、第 5 世代 Tensor コア（FP4 対応） ^[7]	Blackwell、第 5 世代 Tensor コア（FP4／NVFP4） ^[4]
CPU	20 コア Grace（MediaTek と共同設計）。最上位 SKU は Cortex-X925×10+A725×10 ^[7,9]	20 コア Arm（X925×10+A725×10） ^[4]
メモリ	最大 128GB、256bit、ユニファイド（コヒーレント） ^[9]	128GB LPDDR5x、256bit ^[10]
メモリ帯域	300GB/s ^[9]	273GB/s ^[4]
AI 演算	最大 1PFLOP（FP4） ^[2]	1PFLOP（FP4、スパース時。密行列ではその約半分） ^[11]
CPU-GPU 接続	NVLink-C2C（600GB/s。メモリ帯域とは別） ^[12,13]	NVLink-C2C
OS	Windows 11 on Arm（Copilot+ PC 扱い） ^[1]	DGX OS（Ubuntu 24.04 ベース） ^[14]
形態・価格	ノート PC／小型デスクトップ。Surface Laptop Ultra は 2,599 ドルから（10 月 16 日発売）、Surface RTX Spark Dev Box（128GB）は 5,999 ドル（11 月出荷） ^[15-17]	小型デスクトップ。発売時 3,999 ドル、2026 年 2 月に 4,699 ドル、10 月 2 日に 6,950 ドルへ値上げ（発売時から約 74%増） ^[11,18,19] 。64GB 版が OEM 各社から 10 月 23 日に 4,999 ドルで発売予定 ^[19,20]

- ・ 構成は 2 種類ある（2026 年 9 月 3 日に NVIDIA が公表）。上位は 20 コア CPU+6,144 コア GPU で最大 128GB、下位は 18 コア CPU+5,120 コア GPU のノート PC 専用である^[3,21]。下位のメモリ上限は報道によって「24/32GB」と「最大 64GB」で食い違っており、128GB を選べるのは上位構成だけである^[5,21]。
- ・ 発売時期：ノート PC は 2026 年 10 月 16 日に出荷開始、小型デスクトップは 11 月に販売開始と NVIDIA が告知している^[12]。
- ・ 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実行可能」と説明している^[7]。ベンダーの主張であり、速度の条件は示されていない。

- ・ Linux：ネイティブ Linux の公式サポートは発表されていない（報道によれば、XDA の質問に NVIDIA はコメントしなかった）^[5]。一方、Microsoft は Surface RTX Spark Dev Box の WSL 2 を GPU パススルーと CUDA サポート付きで構成していると説明している^[22,23]。
- ・ 未確認・矛盾：一部報道の「600GB/s のメモリ帯域」は、NVLink-C2C の帯域の誤読と見られる^[13,24]。ノート PC が筐体の電力（45～80W）に関係なく 300GB/s で動くかは確認できていない^[5]。RTX Spark 実機での LLM 速度の実測は、本調査の時点では見つからなかった。

2.2 DGX Spark との関係（分析）

RTX Spark は GB10 と同系統のシリコンを Windows PC 向けに再構成した製品と見られ、報道でも DGX Spark の改良版と位置付けられている^[5,25]。LLM 推論の性能（帯域と容量）はほぼ同等と考えてよい。違いはソフトウェア環境にある。DGX Spark は vLLM、TensorRT-LLM、NIM といった Linux 向けの成熟したツール群がそのまま使える^[26]。RTX Spark は Windows on Arm で、ツールの対応はまだ整備途上である。サーバーとして使い込むなら DGX Spark、Windows の業務 PC として報告書作成と同居させるなら RTX Spark という使い分けが妥当である。

2.3 128GB で動かせる規模の目安（試算）

OS やアプリが使う分を差し引くと、モデルと KV キャッシュに使えるのはおおむね 100～115GB と見るのが安全である^[27,28]。Windows は Linux より多くのメモリを使う可能性がある（推測）。

表 2 量子化方式ごとの収容規模の目安^[29]

量子化	1 パラメータあたり	重みだけで収まる上限	実用上の目安（KV キャッシュ込み）
FP16／BF16	2 バイト	約 55B	密モデル～32B 程度
FP8／Q8_0	約 1～1.06 バイト	約 100B	～70B（MoE なら～100B）
Q4_K_M	約 0.6 バイト	約 170B	～122B 級の MoE（約 61～75GB）
NVFP4／MXFP4	約 0.53～0.56 バイト	約 190B	gpt-oss-120b（約 61～66GB）、Nemotron 3 Super（約 64～72GB）
Q2～Q3	約 0.3～0.45 バイト	約 235B 級の MoE（約 100GB）	品質低下が大きく、知財業務には非推奨

KV キャッシュの試算（BF16、各モデルの設定値に基づく概算、分析）：

- ・ 従来型の密モデル 70B（80 層、KV ヘッド 8、ヘッド次元 128）：約 320KB／トークン。128K トークンで約 40GB となり、長文処理では大きな負担になる。
- ・ Qwen3.5-122B-A10B（48 層のうち標準の Attention 層は 12 層、KV ヘッド 2、ヘッド次元 256）^[30,31]：約 24KB／トークン。262K トークンでも約 6～7GB で済む。ハイブリッド構造（Gated DeltaNet）の大きな利点である。

- ・ gpt-oss-120b（スライディングウィンドウ層と全文Attention層が交互）：128Kトークンで約5GB前後。

したがって、Qwen3.5／3.6 系や Nemotron 3 Super（Mamba-2 ハイブリッド）は、長い明細書を扱う際にメモリ面で有利である。

2.4 生成速度（DGX Spark の実測で代替）

生成速度の上限は、おおよそ「メモリ帯域÷1 トークンあたりに読み出す重みのバイト数」で決まる^[5]。RTX Spark の実測がないため、DGX Spark の実測値で代用する。

表 3 DGX Spark での生成速度（単ーストリーム）

モデル	実測値
gpt-oss-120b（MXFP4、llama.cpp）	生成 58.7 tok/s、プロンプト処理 2,444 tok/s ^[32]
gpt-oss-120b	約 55 tok/s（NVIDIA 公表値）。エンジンにより 14.5～50 tok/s ^[33]
Qwen3.5-35B-A3B	48.2 tok/s ^[33]
Qwen3.5-122B-A10B	20.1 tok/s ^[33]
Nemotron 3 Super 120B-A12B（NVFP4、vLLM）	22.7～23.7 tok/s ^[26]
Qwen3.5-27B（密）	10.4～11.1 tok/s ^[33]
Gemma 4 31B（密、NVFP4）	6.85 tok/s ^[34]
密モデル 70B（FP8）	約 2.7 tok/s ^[35]

- ・ RTX Spark の帯域が公称どおり 300GB/s であれば、上記より 1 割程度速くなる可能性がある（推測。実測なし）。
- ・ 日本語は 1 トークンあたりの文字数が英語より少なく、体感速度は tok/s の数値ほど出ない（分析）。
- ・ プロンプト処理（明細書の読み込み）は計算力が効くため速く、数万トークンの明細書でも十数秒～数十秒で読み込める（gpt-oss-120b の 2,444 tok/s からの試算）。

3. 候補 LLM の比較

3.1 比較表

表 4 128GB 環境での候補モデル（2026 年 10 月時点）

モデル（開発元・国）	規模・方式	128GB での量子化と可否	コンテキスト	日本語	ライセンス	知財業務での推奨用途
Qwen3.5-122B-A10B （Alibaba・中国） ^[36-38]	総 122B／アクティブ 10B。MoE＋GDN ハイブリ	NVFP4 （NVIDIA 公式）／Q4 で約 61～	262K（最大 1.01M）	良（201 言語）△	Apache 2.0 ○	汎用の主力：クレームの分節、対比表、要約、長文読解

モデル（開発元・国）	規模・方式	128GB での量子化と可否	コンテキスト	日本語	ライセンス	知財業務での推奨用途
	ッド、画像入力可	75GB。○余裕あり				
gpt-oss-120b (OpenAI・米国) [39,40]	116.8B／5.1B。 MoE、テキストのみ	MXFP4 がネイティブ、約 61～66GB。○	128K	中～良（公式の日本語評価なし）？	Apache 2.0+使用ポリシー ○	主力（非中国） ：推論、ツール呼び出し、速度重視
GPT-OSS Swallow 120B （東京科学大学・産総研・日本） [41,42]	120B 級／約 5B。gpt-oss に継続学習（2026 年 2 月公開）	gpt-oss-120b と同等 ○	元モデルに準拠？	強化済み（開発元評価）△	Apache 2.0 ○	主力（非中国かつ日本語重視）。科学分野は元モデルより低下との報告
Nemotron 3 Super 120B-A12B (NVIDIA・米国) [43,44]	120B／12B。LatentMoE（Mamba-2+ MoE+ Attention）、MTP	NVFP4（モデルカードに「最小 1× DGX Spark」と記載）○	最大 1M	公式対応言語に日本語あり ○	NVIDIA Nemotron Open Model License（商用可）○	長文読解、多数文献の横断、エージェント
Qwen3.6-35B-A3B (Alibaba・中国) [45,46]	35B／3B。MoE + GDN、マルチモーダル	FP8 で約 35GB、NVFP4 で約 25GB。○	262K（最大約 1M）	良 △	Apache 2.0 ○	高速な補助 ：一次スクリーニング、分類、抽出
Qwen3.8-27B (Alibaba・中国) [47,48]	27B（密）。ハイブリッド Attention、画像入力可	Q8 で約 28GB、BF16 で約 55GB。○ただし遅い	262K	良 △	Apache 2.0 △	品質重視のドラフト作成（夜間バッチなど）
Gemma 4 31B／26B-A4B （Google・米国） [49,50]	30.7B（密）／26B（MoE、A4B）。画像入力可	31B は Q8 で約 33GB ○。26B-A4B は高速	256K	良（140 以上の言語）△	Apache 2.0 ○	非中国の中型モデル。図面読み取りの補助
LLM-jp-4.1 （NII・日本） [51-53]	8B／32B-A3B／33B。密／MoE（thinking 版あり）	いずれも余裕で収まる ○	約 65K（LLM-jp-4 の値）△	日本語特化 （LLM-jp-4 33B は日本語 MT-Bench 8.00）△	Apache 2.0 ○	日本語文書の生成、国産重視の場面。ツール呼び出しに対応
Qwen3 Swallow （東京科学大学・産総研・日本） [41,54]	8B／30B-A3B／32B。Qwen3 に継続学習	余裕で収まる ○	元モデルに準拠	強化済み（開発元評価）△	Apache 2.0 ○	日本語の要約、報告書の下書き
PLaMo 3（PFN／NICT・日本） [55]	2B～31B。ベースモデルのみ	収まる	？	日本語特化	PLaMo Community License（条件要確認）△	指示調整がなく、そのままでは業務に使いにくい
Sarashina2.2（SB Intuitions・日本） [55]	0.5B～3B	収まる	？	日本語特化	MIT △	分類などの軽作業
Llama 4 Scout (Meta・米国) [56]	109B／17B。MoE	Q4 で収まる	10M（公称）	中	Llama License（月間アクティブ 7 億超は別途	優先度は低い

モデル（開発元・国）	規模・方式	128GB での量子化と可否	コンテキスト	日本語	ライセンス	知財業務での推奨用途
					許諾）△	
Kimi K3、GLM-5.2、MiniMax M3、Qwen3.8-2.4T、DeepSeek V4 Flash（中国） ^[57]	数百 B～数 T。 超大型 MoE	×（128GB1 台には収まらない）	—	—	各社独自	対象外

- ・ Qwen3.8-2.4T 以外の超大型モデル（Kimi K3、GLM-5.2、MiniMax M3、DeepSeek V4 Flash）の規模は、本調査では一次資料で確認できていない。
- ・ tsuzumi（NTT）と ELYZA の最新版がオープンウェイトで公開されているかは、一次資料で確認できていない。PLaMo 3.0 Prime は法人向けの API／オンプレミス提供で、オープンウェイトではない^[58]。
- ・ Qwen3.8-27B の公開日（2026 年 8 月 14 日、Apache 2.0）は二次情報で、「当初は公開されていなかった」とする記事もある^[59,60]。ただし、LLM-jp の公式ブログ（9 月 30 日追記）が評価対象に含め^[61]、NVIDIA も DGX Spark 64GB 版で動かせるモデルの例に挙げていることから^[48]、公開済みと判断した。
- ・ Qwen3.8-Flash-Next（180B）は NVIDIA の開発者フォーラムで言及があるが、仕様とライセンスは確認できていない^[62]。

3.2 知財業務への適性（分析）

- ・ **長い明細書の読解**：262K～1M のコンテキストを持つ Qwen3.5 や Nemotron 3 Super が有利である。ただし公称の上限まで品質が保たれるわけではない。1 件ずつ（～6 万トークン）処理し、複数件は 1 件ずつまとめた結果を統合する運用を推奨する。
- ・ **クレーム解析・構成要件の分節**：Ishimaru らは、日本語と英語の特許クレーム各 100 件で並列構造の抽出を評価し、日本語タスクでは継続学習・ファインチューニングした 8B 級のオープンモデル（Llama-3-Swallow-8B など）が GPT-5 を F1 で約 0.02～0.05 上回ったと報告している^[63]。用途を絞れば、小型の国産系モデルでも十分戦えることを示す結果である。
- ・ **サポート要件（クレームと明細書の対応）**：日本語特許の公開ベンチマーク「JAPAS」（LREC 2026、2,000 件超に人手で注釈）がある^[64]。記載要件のチェックを支援するツールを評価する際の参考になる。
- ・ **日本語の法律文書**：Cao らの「JBE-QA」は、日本の司法試験短答式（2015～2024 年）から作成した 3,464 問で 26 モデルを評価し、推論機能を有効にしたプロプライエタリモデルが最も高い成

績を示したと報告している（F1 は一部の外れ値を除き 0.602～0.861）^[65]。ローカルモデルとの差を示す材料である。

- ・ 特許庁（JPO）や Japio による LLM ベンチマークは、本調査では見つからなかった。導入前に、自社の過去案件（守秘範囲内）で評価セットを作ることを強く推奨する。

3.3 用途別のおすすめ

表 5 用途別の推奨モデル

用途	第一候補	非中国製の代替
汎用の主力（分節、対比表、論点整理）	Qwen3.5-122B-A10B（NVFP4）	gpt-oss-120b／GPT-OSS Swallow 120B
高速な軽量モデル（スクリーニング、抽出、分類）	Qwen3.6-35B-A3B	gpt-oss-20b、Gemma 4 26B-A4B、LLM-jp-4.1-32B-A3B
日本語重視（報告書、判例評釈の下書き）	LLM-jp-4.1-33B／Qwen3 Swallow 32B	LLM-jp-4.1（国産、Apache 2.0）
長文読解・複数文献の横断	Nemotron 3 Super（1M）	同左（米国製）
埋め込み（RAG）	ruri-v3-310m（JMTEB 77.24、8,192 トークン） ^[66]	PLaMo-Embedding-1B（76.10） ^[66] 、Sarashina3 embedding ^[67]
多言語の先行技術（英・中・韓の文献）	Qwen3-Embedding-4B／8B（Apache 2.0、32K、100 以上の言語） ^[68]	ruri-v3 と BGE-M3（疎ベクトル）の併用
再ランキング	Qwen3-Reranker 系、bge-reranker-v2-m3 など（一次資料未確認）	ruri 系のリランカー（要確認）

JMTEB の Retrieval スコアは、Qwen3-Embedding-0.6B が 72.81^[69]、ruri-v3-310m が 81.89 で^[70]、日本語の検索では日本語特化モデルがはっきり優位である。ただし、自社データで実測すると順位が入れ替わる例も報告されており^[70]、自前での評価が欠かせない。

4. 構築環境

4.1 推論エンジンの対応状況

表 6 推論エンジン・ツールの対応状況

ツール	RTX Spark（Windows on Arm）	DGX Spark（Linux）
CUDA	○ CUDA 13.4 で Windows on Arm に対応。R616 系の開発者向けドライバが必要との報道あり ^[71,72]	○
llama.cpp	○ NVIDIA が Windows on Arm 版を公式に案内 ^[73]	○
Ollama	○ NVIDIA が公式に案内（llama.cpp ベース） ^[73]	○
LM Studio	△ NVIDIA が最適化の提供に一般的に言及。RTX Spark	○（Linux ARM64 版） ^[75]

ツール	RTX Spark (Windows on Arm)	DGX Spark (Linux)
	専用の発表はない ^[74]	
vLLM	? WSL2 (Arm) 経由で動く可能性はあるが未確認。 NVIDIA は vLLM には Linux が必要としている ^[76]	○ NVFP4 への最適化が進んでいる ^[26,34]
TensorRT-LLM/ NIM	? 公式に案内されているのは「TensorRT for RTX」 ^[73]	○ NIM のサポート表には DGX Spark のみ記載 ^[77]
SGLang	?	○ (コミュニティでの実績あり)
WSL2	○ Surface RTX Spark Dev Box で GPU パススルーと CUDA が利用可能 ^[22,23] 。NVIDIA の WSL ガイドは x86 向けの記述のみ ^[78]	—

RTX Spark での実務的な選択肢は、**llama.cpp**（または **Ollama**）と **GGUF 形式 (Q4_K_M~Q6_K)** である。NVFP4 や vLLM の利点を十分に生かしたい場合や、複数人で同時に使う場合は、DGX Spark (Linux) が適している（分析）。

4.2 特許文献 RAG の要点（分析・推奨）

- ① **取り込み**：J-PlatPat や各国特許庁の XML（全文テキスト）を優先し、PDF の OCR は最後の手段とする。公報番号、出願日、優先日、IPC/FI/F ターム、出願人、請求項番号、段落番号【00xx】をメタデータとして必ず保持する。
- ② **チャンク分割**：請求項は 1 項ずつ、明細書は段落番号単位（数百～1,000 字程度、前後を重ねる）で分ける。固定長での機械的な分割は避ける。従属項の引用関係（「請求項 1 に記載の」）も保持する。
- ③ **図面**：符号の説明（「10…筐体」）をテキストとして索引に入れる。図そのものの読み取りは、Qwen3.5/3.6 や Gemma 4 の画像入力で補助的に行う程度にとどめる（精度は未検証）。
- ④ **検索方式**：密ベクトル（ruri-v3）に加えて BM25 などのキーワード検索を併用し（技術用語の完全一致が重要なため）、再ランキングをかける。
- ⑤ **出典の強制**：回答には必ず「公報番号＋段落番号」を出させ、原文の該当箇所と自動で突き合わせる仕組みを入れる。

5. 注意事項チェックリスト

5.1 機密保持・セキュリティ

- ☐ **ネットワークを遮断して運用する**（ファイアウォールで推論サーバーの外部通信を禁止し、モデル導入時だけ接続を許可する）。

- クラウドへの自動ルーティングを無効にする。NVIDIA OpenShell には、プライバシーポリシーに基づいてローカルモデルへ振り分ける機能や、クラウドモデルに送る個人情報を隠す機能がある^[7,79]。つまりクラウドへ送る経路が存在するということであり、ローカル専用に設定する。Windows のエージェント機能や診断データの送信設定も確認する。
- Ollama、LM Studio などの自動更新・利用統計の設定を確認する。
- モデルは公式の組織リポジトリから入手する（Qwen/、openai/、nvidia/、google/、llm-jp/など）。Hugging Face 上には安全調整を外した「abliterated」「Uncensored」などの派生モデルが多数あり、ダウンロード数も多いが^[30,80,81]、業務では使わない。
- 形式は safetensors/GGUF に限り、pickle 形式（.bin/.pt）は使わない。trust_remote_code は原則として無効にし、ハッシュ値を検証する。
- エージェント／ツール呼び出しには、ファイル削除、外部送信、メール送信の権限を与えない。サンドボックス（OpenShell のポリシー機能など）を使い、読み取り専用から始める。文献中に埋め込まれた指示文によるプロンプトインジェクションにも注意する。
- ディスクを暗号化（BitLocker）し、ベクトル DB とログも暗号化する。案件ごとにデータを分離する（コンフリクト対策）。

5.2 ライセンス

- Apache 2.0（Qwen3.5/3.6、Gemma 4、gpt-oss、LLM-jp-4.1、Swallow 系、ruri-v3）は商用利用可。著作権表示と変更点の明示が必要である^[41,49,82]。Qwen3.8-27B の Apache 2.0 は二次情報による。
- gpt-oss は、Apache 2.0 に加えて**使用ポリシー**がある^[83,84]。
- Nemotron は「NVIDIA Nemotron Open Model License」（商用可。条項は個別に確認）。Ollama 版の表示は「NVIDIA Software and Model Evaluation License」で、配布元により表示が食い違う。**公式の Hugging Face 版の規約に従う**^[43,85]。
- Llama 系は、月間アクティブユーザー7 億人超の制限など独自の条項がある^[56,84]。PLaMo Community License も、商用条件を個別に確認する^[55]。
- Qwen3.8-2.4T（Max 系）は独自ライセンス（報道では収益分配型）で、小型版と条件が異なる^[57,86]。
- 派生モデル（Swallow など）は元モデルの規約を引き継ぐ^[55]。

5.3 中国製モデルの論点

確認済みの事実：個人情報保護委員会は 2025 年 2 月 3 日、DeepSeek のクラウドサービスについて、

データが中国のサーバーに保存され中国の法令が適用されるとして注意を呼びかけた^[87,88]。続いてデジタル庁（2月6日）が政府機関などに注意喚起し、要機密情報は扱えないとする既存ルールの遵守を改めて求めた^[89,90]。

分析：ローカルで重みを動かすだけなら、モデル自体が外部にデータを送る仕組みはない（重みは数値の集まりで、通信機能を持たない）。上記の注意喚起は主にクラウドサービスを対象としたものである。残るリスクは次の3つである。①出力の偏り（中国の政治的・社会的話題での検閲、中国の出願人や技術に関する記述の偏りの可能性。知財業務への影響は限定的と見られるが、検証データはない）、②リモートコードやツールチェーン経由のサプライチェーンリスク、③クライアントの方針や経済安全保障上の受け止め方（防衛・重要技術分野のクライアントは、中国製モデルの使用自体を禁止している可能性がある）。

推奨：クライアントとの契約や社内規程で、使用するモデルの開発国を明示できるようにしておく。防衛・経済安全保障に関わる案件では、**米国製または国産のモデルだけで構成した環境**に切り替える。

5.4 ハルシネーションと実務での限界

- 特許番号・公報番号・判例（事件番号、判決日）・条文番号・引用文は必ず原典と照合する。モデルの知識から出させず、RAG で取り込んだ原文からのみ引用させる。
- 進歩性・新規性・侵害の有無・FTO の「結論」は LLM に出させない。構成要件の対比と論点の洗い出しにとどめる。
- 法改正や審査基準の改訂は学習データの締切日より後のことがあるため、最新の法令・基準は RAG で与える。弁理士会のガイドラインも、更新された情報に基づくファクトチェックを弁理士自身が行うよう求めている^[6]。
- 検証の手順（①出典の自動突き合わせ→②担当者レビュー→③責任者承認）を記録に残す。

5.5 精度・性能の限界

- Q4 未満の量子化は避ける。NVIDIA は NVFP4 について、FP8 に近い精度（劣化 1%未満）としているが、ベンダーの評価である^[91]。
- 長いコンテキストでは、公称の上限より手前で精度が落ちやすい。重要な対比は分割して処理する。
- 推論モード（thinking）は品質を上げる一方で、出力トークンが数倍に増え、待ち時間が延びる。

5.6 法務・倫理（断定は避ける）

- ・ 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）は、生成物の正確性の確認と最終的な責任を弁理士に求めている。民法 644 条の善管注意義務を根拠に、AI の生成結果を検

討・精査せずにそのままクライアントに提供することは、同義務に違反するおそれがあるとしている^[6]。

- ・ 同ガイドラインは、外部事業者が提供する生成 AI に秘密情報を入力することは守秘義務（弁理士法 30 条）違反のおそれがあるとし、クライアントから得た秘密情報を入力する場合は、学習の有無にかかわらずクライアントの合意を取るべきとしている。秘密保持規約のないプラットフォームへの入力が新規性喪失につながりうる点にも注意を促している^[6]。
- ・ ローカル運用は第三者への開示に当たらないため、このリスクを大きく下げる（分析）。ただし、弁理士法 30 条自体は AI への入力について直接定めておらず^[92]、バックアップ、共用 PC、ログなど運用次第でリスクは残る。クライアントが AI 利用自体を制限している場合もあるため、ローカルであっても事前の合意を推奨する（分析）。
- ・ 利用者が弁理士でない場合、他人の求めに応じて報酬を得て特許庁への書類を作成することは、弁理士法 75 条の業務制限に触れうる。AI で出願書類を作成する範囲については、弁理士と役割を分ける必要がある（個別の判断が必要）。
- ・ AI の発明者性については、知財高裁が令和 7 年 1 月 30 日判決（令和 6 年（行コ）第 10006 号、ダバス事件）で、特許法上の発明者は自然人に限られるとの判断を示した^[93,94]。最高裁第二小法廷は 2026 年 3 月 4 日付で上告を受理しない決定をし、判決は確定した^[95]。AI 生成物の著作権や、AI が関与した発明の発明者の扱いは、議論が続いている分野である。
- ・ クライアントへの説明として、AI を使う範囲と使うモデルを委任契約や報告書に明記することが望ましい（分析）。

5.7 クラウドのフロンティアモデルとの使い分け（分析）

- ・ 学術ベンチマーク（JBE-QA など）でも、法的推論の最上位はプロプライエタリモデルである^[65]。ローカルの 120B 級 MoE は、下ごしらえの作業（分節、抽出、翻訳、要約、対比表のたたき台）では実用水準だが、高度な論証の質では差があると思われるべきである。
- ・ 推奨：①機密情報（出願前・クライアント固有）はローカルで処理する。②公開済みの文献だけを使う一般的な調査や、高度な推論が必要な論点は、学習に使わない設定とデータ保存の条件を確認済みの法人契約クラウドで処理する。③クラウドに送るときは、案件を特定できる情報を除いて抽象化する。

6. 推奨構成（まとめ）

表 7 推奨構成

区分	構成 A：性能重視	構成 B：非中国製（経済安全保障に配慮）
主力	Qwen3.5-122B-A10B（NVFP4／Q4_K_M、約 70GB）	gpt-oss-120b（MXFP4、約 63GB）または GPT-OSS Swallow 120B
長文用	Nemotron 3 Super（主力と入れ替えて使用）	Nemotron 3 Super
補助（常駐）	Qwen3.6-35B-A3B（Q4、約 20～25GB）	LLM-jp-4.1-32B-A3B／Gemma 4 26B-A4B
日本語の文章生成	LLM-jp-4.1-33B／Qwen3 Swallow 32B	LLM-jp-4.1-33B
埋め込み・再ランキング	ruri-v3-310m＋Qwen3-Embedding-4B（多言語）＋リランカー	ruri-v3-310m＋PLaMo-Embedding-1B
実行環境	RTX Spark：llama.cpp／Ollama（Windows）。 DGX Spark：vLLM（NVFP4）	同左

メモリ配分の例：主力（約 70GB）、補助（約 20GB）、埋め込み（約 2GB）、KV キャッシュと作業領域（約 15GB）で合計 100GB 強となり、128GB に収まる。ただし、120B 級を 2 つ同時に常駐させることはできない（分析）。

導入の手順：

- ① 自社の過去案件（守秘範囲内）から評価セットを 30～50 問作る（分節、対比、要約、条文引用）。
- ② 構成 A と B を比較する。
- ③ 検証手順を規程にまとめる。
- ④ パイロット運用を経て本格導入する。

7. 確認できなかった事項

- ・ RTX Spark 実機での LLM 速度の実測値、ノート PC で帯域が常に 300GB/s で動くかどうか
- ・ LM Studio、vLLM（WSL2 経由）、TensorRT-LLM、NIM、SGLang の RTX Spark への対応
- ・ ネイティブ Linux の公式サポート
- ・ tsuzumi と ELYZA のオープンウェイト版の最新状況
- ・ Qwen3.8-Flash-Next の仕様、Sarashina3 embedding のライセンス
- ・ Kimi K3、GLM-5.2、MiniMax M3、DeepSeek V4 Flash の規模とライセンス
- ・ 特許庁（JPO）や Japio による LLM 評価、各モデルの日本語特許タスクでの直接比較データ
- ・ 参考文献のうち「※URL 未確認」と記したものは、記載元の資料から URL を転記したもので、アクセスを確認していない

参考文献

(URL の最終確認日：2026 年 10 月 10 日。本文中の上付き番号に対応)

- [1] Wikipedia「Nvidia RTX Spark」. https://en.wikipedia.org/wiki/Nvidia_RTX_Spark
- [2] NVIDIA「NVIDIA RTX Spark Unveiled (GeForce News, Computex 2026)」. <https://www.nvidia.com/en-us/geforce/news/gfecnt/20266/computex-2026-nvidia-geforce-rtx-announcements/> (一次資料)
- [3] Tom's Hardware「Nvidia's RTX Spark N1X launches in October for laptops and desktops — 18 or 20 CPU cores, paired with 5,120 or 6,144 CUDA cores, up to 128GB of unified memory」. <https://www.tomshardware.com/laptops/nvidias-rtx-spark-n1x-launches-in-october-for-laptops-and-desktops-18-or-20-cpu-cores-paired-with-5-120-or-6-144-cuda-cores-up-to-128gb-of-unified-memory>
- [4] GPUWerk「DGX Spark specs: GB10, 128 GB memory, 273 GB/s explained」. <https://gpuwerk.com/dgx-spark/specs/>
- [5] ComputingForGeeks「NVIDIA RTX Spark Laptops: N1X Specs, CUDA and Linux」. <https://computingforgeeks.com/nvidia-rtx-spark-linux/>
- [6] 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> (一次資料)
- [7] NVIDIA Newsroom「NVIDIA and Microsoft Reinvent Windows PCs for the Age of Personal AI」. <https://nvidianews.nvidia.com/news/nvidia-microsoft-windows-pcs-agents-rtx-spark> (一次資料)
- [8] AI Tool Discovery「NVIDIA DGX Spark: Specs, Price, and Who Should Buy It」. <https://www.aitooldiscovery.com/ai-infra/nvidia-dgx-spark-explained>
- [9] NVIDIA「System Overview — NVIDIA Windows on Arm Porting Guide」. <https://docs.nvidia.com/rtx-spark/rtx-spark-porting-guide/latest/overview.html> (一次資料)
- [10] GPUSmith「NVIDIA DGX Spark Review: Price, Specs & Performance (2026)」. <https://gpusmith.com/articles/en/nvidia-dgx-spark-review-specs-performance>
- [11] IntuitionLabs「NVIDIA DGX Spark Review: \$4,699 Price & Benchmarks」. <https://intuitionlabs.ai/articles/nvidia-dgx-spark-review>
- [12] NVIDIA Blog「NVIDIA, Microsoft Kick Off a New Beginning for Windows PCs with RTX Spark and AI Agents」. <https://blogs.nvidia.com/blog/local-ai-rtx-spark-microsoft-windows-event/> (一次資料)
- [13] LeCompute「RTX Spark vs DGX Spark: specs, memory, price (2026)」. <https://lecompute.fr/en/silicon/rtx-spark-vs-dgx-spark/>
- [14] NVIDIA「NVIDIA DGX Spark Porting Guide」, 2026 年 5 月 24 日. <https://docs.nvidia.com/dgx/dgx-spark-porting-guide/dgx-spark-porting-guide.pdf> (一次資料)
- [15] Microsoft (Eleven Forum 転載)「You can now Pre-order Surface Laptop Ultra and Surface RTX Spark Dev Box」, 2026 年 10 月 7 日. <https://www.elevenforum.com/t/you-can-now-pre-order-surface-laptop-ultra-and-surface-rtx-spark-dev-box.50005/> Microsoft Devices Blog の告知文の転載
- [16] TweakTown「Microsoft opens Surface Laptop Ultra preorders at \$2,599 with RTX Spark N1X inside」. <https://www.tweaktown.com/news/113971/microsoft-opens-surface-laptop-ultra-preorders-at-dollars2599-with-rtx-spark-n1x-inside/index.html>
- [17] Notebookcheck「Microsofts Mac Studio Konkurrent mit 1.000 Lüftungsöffnungen startet in den Verkauf」, 2026 年 10 月. <https://www.notebookcheck.com/Microsofts-Mac-Studio-Konkurrent-mit-1-000-Lueftungsoeffnungen-startet-in-den-Verkauf.1418423.0.html>

- [18] AI Weekly 「Nvidia Launches 64GB DGX Spark at \$4,999 as 128GB Hits \$6,950 (Tom's Hardware 記事の要約)」 . <https://aiweekly.co/alerts/nvidia-launches-64gb-dgx-spark-at-4999-as-128gb-hits-6950>
- [19] The Register 「Nvidia debuts \$4,999 DGX Spark with half the RAM and storage, amid memory crunch」 , 2026 年 10 月. <https://www.theregister.com/a/5300622>
- [20] TweakTown 「NVIDIA's new DGX Spark has half the RAM and costs more than the original 128GB model did at launch」 , 2026 年 10 月. <https://www.tweaktown.com/news/113885/nvidias-new-dgx-spark-has-half-the-ram-and-costs-more-than-the-original-128gb-model-did-at-launch/index.html>
- [21] Windows Central 「NVIDIA confirms RTX Spark configurations and availability: First devices expected to begin shipping as soon as next month, with two N1X configs on offer」 .
<https://www.windowscentral.com/hardware/nvidia/nvidia-confirms-rtx-spark-configurations-and-availability-first-devices-expected-to-begin-shipping-as-soon-as-next-month-with-two-n1x-configs-on-offer>
- [22] TechPowerUp 「Microsoft Introduces New Surface RTX Spark Dev Box for Developers」 .
<https://www.techpowerup.com/349644/microsoft-introduces-new-surface-rtx-spark-dev-box-for-developers>
- [23] Microsoft Windows Blog 「Building the next generation of devices for developers: Surface RTX Spark Dev Box」 , 2026 年 6 月 2 日. <https://blogs.windows.com/devices/2026/06/02/building-the-next-generation-of-devices-for-developers-surface-rtx-spark-dev-box/> (一次資料)
- [24] NVIDIA Developer Forums 「Spark CPU - NVIDIA RTX Spark Reinvents Windows PCs for the Age of Personal AI」 . <https://forums.developer.nvidia.com/t/spark-cpu-nvidia-rtx-spark-reinvents-windows-pcs-for-the-age-of-personal-ai/371908>
- [25] Notebookcheck 「Nvidia N1X officially confirmed to arrive as the RTX Spark」 .
<https://www.notebookcheck.net/Nvidia-N1X-officially-confirmed-to-arrive-as-the-RTX-Spark.1312010.0.html>
- [26] vLLM Blog 「vLLM on the DGX Spark: Architecture, Configuration, and Local Evaluation」 , 2026 年 6 月 1 日.
<https://vllm.ai/blog/2026-06-01-vllm-dgx-spark> (一次資料)
- [27] Pinggy 「Best Local LLM Tools and Models in 2026: What Fits in 128GB」 .
https://pinggy.io/blog/top_5_local_llm_tools_and_models/
- [28] Conselara 「DGX Spark Model Comparison: What Fits and What Runs (SM121, 128 GB)」 .
<https://conselara.dev/notes/dgx-spark-model-comparison/>
- [29] LLMRequirements.com 「Best AI models that run on NVIDIA DGX Spark (128 GB)」 .
<https://llmrequirements.com/hardware/dgx-spark-128>
- [30] Hugging Face 「HauhauCS/Qwen3.5-122B-A10B-Uncensored-HauhauCS-Aggressive」 .
<https://huggingface.co/HauhauCS/Qwen3.5-122B-A10B-Uncensored-HauhauCS-Aggressive> 安全調整を外した派生モデルの例。設定値の参照にのみ使用
- [31] APXML 「Qwen3.5-122B-A10B: Specifications and GPU VRAM Requirements」 .
<https://apxml.com/models/qwen35-122b-a10b>
- [32] Eurokommerz 「DGX Spark Benchmarks: Tokens per Second and TDP」 . <https://eurokommerz.at/blog/nvidia-dgx-spark-real-benchmarks/>
- [33] Tokenstead 「DGX Spark Benchmarks: Which Local Model Wins in 2026」 . <https://tokenstead.ai/guides/dgx-spark-benchmarks-2026>
- [34] ai-muninn 「[vLLM] Nemotron 3 Nano on DGX Spark: 74.75 tok/s NVFP4 — 11.5% Past the Public Baseline」 . <https://ai-muninn.com/en/blog/dgx-spark-nemotron-3-nano-w4a16-74-toks>
- [35] LeCompute 「DGX Spark vs RTX 5090: capacity or speed, the choice that decides your local LLM」 .

<https://lecompute.fr/en/silicon/dgx-spark-vs-rtx-5090/>

- [36] Hugging Face 「nvidia/Qwen3.5-122B-A10B-NVFP4」 . <https://huggingface.co/nvidia/Qwen3.5-122B-A10B-NVFP4> (一次資料)
- [37] Spheron 「Deploy Qwen 3.5 on GPU Cloud: GDN Hybrid Architecture, 262K Context, and vLLM Setup (2026)」 . <https://www.spheron.network/blog/deploy-qwen-3-5-gpu-cloud/>
- [38] NVIDIA API Documentation 「qwen / qwen3-5-122b-a10b」 . <https://docs.api.nvidia.com/nim/reference/qwen-qwen3-5-122b-a10b> (一次資料)
- [39] OpenAI 「gpt-oss-120b & gpt-oss-20b Model Card」 , 2025 年 8 月 5 日 . https://cdn.openai.com/pdf/419b6906-9da6-406c-a19d-1bb078ac7637/oai_gpt-oss_model_card.pdf (一次資料)
- [40] Hugging Face 「openai/gpt-oss-120b」 . <https://huggingface.co/openai/gpt-oss-120b> (一次資料)
- [41] GIGAZINE 「日本語能力を強化した AI 「GPT-OSS Swallow」 と 「Qwen3 Swallow」 を東京科学大の研究チームが公開」 , 2026 年 2 月 24 日 . <https://gigazine.net/news/20260224-gptoss-qwen3-swallow-japanese-ai/>
- [42] Swallow LLM (東京科学大学・産総研) 「GPT-OSS Swallow」 . <https://swallow-llm.github.io/gptoss-swallow.ja> (一次資料) ※URL 未確認 (文献[41]に記載された URL を転記)
- [43] Hugging Face 「nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-BF16」 . <https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-BF16> (一次資料)
- [44] Hugging Face 「nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4」 . <https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4> (一次資料)
- [45] Qwen 「Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All」 . <https://qwen.ai/blog?id=qwen3.6-35b-a3b> (一次資料)
- [46] Hugging Face 「Qwen/Qwen3.6-35B-A3B」 . <https://huggingface.co/Qwen/Qwen3.6-35B-A3B> (一次資料)
- [47] Yotta Labs 「Qwen 3.8 27B: Specs, Hardware Requirements, and How to Run It (2026)」 . <https://www.yottalabs.ai/post/qwen-3-8-27b-specs-hardware-requirements-how-to-run-2026>
- [48] Guru3D 「NVIDIA launches 64GB DGX Spark desktop AI system」 , 2026 年 10 月 . <https://www.guru3d.com/story/vidia-launches-64gb-dgx-spark-desktop-ai-system-for/>
- [49] Google 「Gemma 4: Byte for byte, the most capable open models」 . <https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/> (一次資料)
- [50] OpenRouter 「Gemma 4 31B - API Pricing & Benchmarks」 . <https://openrouter.ai/google/gemma-4-31b-it:free>
- [51] 国立情報学研究所 大規模言語モデル研究開発センター 「大規模言語モデル 「LLM-jp-4.1」 シリーズを公開」 . <https://llmc.nii.ac.jp/topics/llm-jp-4-1/> (一次資料)
- [52] AI Revolution 「LLM-jp-4 33B とは? 国立情報学研究所の国産オープン LLM ・ Apache 2.0 ・ gpt-oss-20b 超えの日本語性能を解説 【2026 年 8 月速報】」 . <https://ai-revolution.co.jp/media/llm-jp-4-33b/>
- [53] PC Watch 「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>
- [54] npaka (note) 「Timeline of Local LLM Releases」 . <https://note.com/npaka/n/n1d99253ae2cf>
- [55] Coplus One 「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/>
- [56] Digital Applied 「Google Gemma 4: Apache 2.0 Open-Source Complete Guide」 . <https://www.digitalapplied.com/blog/google-gemma-4-apache-2-open-source-complete-guide>
- [57] Yotta Labs 「Qwen 3.8-Max: Specs, Pricing, Benchmark Status, and How to Access It (2026)」 .

<https://www.yottalabs.ai/post/qwen-3-8-max-release-date-specs-how-to-access-2026>

- [58] 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>
- [59] Codersera 「Qwen 3.5 to 3.8: Alibaba Open-Weights Guide (2026)」 . <https://codersera.com/blog/qwen-3-5-complete-guide-2026/>
- [60] OrcaRouter 「Qwen3.8-27B Open Weights: Max Shipped, 27B Delayed」 . <https://www.orcarouter.ai/blog/qwen-3-8-27b-open-weights-leak>
- [61] LLM-jp 「LLM-jp-4.1 モデルの公開」 . <https://llm-jp.nii.ac.jp/blog/llm-jp-4-1/> (一次資料)
- [62] NVIDIA Developer Forums 「Qwen3.8-27B coming next week - full 3.8 will go open-weights!」 . <https://forums.developer.nvidia.com/t/qwen3-8-27b-coming-next-week-full-3-8-will-go-open-weights/379613>
- [63] Ishimaru, T., Utsuro, T., Nagata, M. 「Coordinate Structure Extraction for Patent Claims Using Multilingual LLMs」 , Proceedings of LREC 2026, pp. 4931–4941, 2026 年 5 月. <https://aclanthology.org/2026.lrec-1.387/> (一次資料)
- [64] ACL Anthology 「JAPAS: A Benchmark and Neural Approach for Japanese Patent Support Relation Extraction」 , Proceedings of LREC 2026. <https://aclanthology.org/2026.lrec-1.330/> (一次資料)
- [65] Cao et al. 「JBE-QA: Japanese Bar Exam QA Dataset for Assessing Legal Domain Knowledge」 , arXiv:2511.22869. <https://huggingface.co/papers/2511.22869> (一次資料)
- [66] 土門大貴 (note) 「日本語埋め込みモデルは OpenAI API を超えたか。Ruri v3 と PLaMo-embedding の実力」 . <https://note.com/daikidomon/n/nc821db32813e>
- [67] SB Intuitions TECH BLOG 「Sarashina3 embedding: 日本語に強い最新のテキスト埋め込みモデル」 , 2026 年 7 月 2 日. <https://www.sbintuitions.co.jp/blog/entry/2026/07/02/101054> (一次資料)
- [68] Zenn (libercraft) 「埋め込みモデル 2026 総点検 : Qwen3-Embedding + Matryoshka で日本語 RAG のコストと精度を両取りする」 , 2026 年 7 月 9 日. <https://zenn.dev/libercraft/articles/20260709-embedding-models-2026-matryoshka>
- [69] secon.dev 「Qwen3 Embedding 文章ベクトルの日本語性能を JMTEB で測る」 , 2025 年 6 月 11 日. <https://secon.dev/entry/2025/06/11/100000-qwen3-embedding-jmteb/>
- [70] Zenn (shogo_h) 「日本語の埋め込みモデルって結局どれが良いの? JMTEB のスコアと実測で比べてみた」 . https://zenn.dev/shogo_h/articles/japanese-embedding-benchmark-2026
- [71] VideoCardz 「CUDA 13.4 released, ready for RTX Spark」 . <https://videocardz.com/newz/cuda-13-4-released-ready-for-rtx-spark>
- [72] Tech Insider 「NVIDIA Extends CUDA 13.4 to Windows on Arm Laptops [2026]」 . <https://tech-insider.org/nvidia-cuda-13-4-windows-on-arm-2026/>
- [73] NVIDIA Developer 「NVIDIA RTX Spark: Port Your Windows App to Arm64」 . <https://developer.nvidia.com/topics/ai/local-ai/port-apps> (一次資料)
- [74] NVIDIA Blog 「Sparks Fly: NVIDIA Accelerates Local AI at IFA 2026」 . <https://blogs.nvidia.com/blog/local-ai-ifa-next-gen-agents-nv-pair-rtx-spark/> (一次資料)
- [75] LM Studio Blog 「NVIDIA DGX Spark」 . <https://lmstudio.ai/blog/dgx-spark> (一次資料)
- [76] NVIDIA 「How to Get Started With Large Language Models on NVIDIA RTX PCs」 . <https://build.nvidia.com/rtx/llms> (一次資料)
- [77] NVIDIA 「Supported Models for NVIDIA NIM for LLMs」 . <https://docs.nvidia.com/nim/large-language-models/1.15.0/supported-models.html> (一次資料)

- [78] NVIDIA 「CUDA on WSL User Guide」 . <https://docs.nvidia.com/cuda/wsl-user-guide/index.html> (一次資料)
- [79] NVIDIA Blog 「NVIDIA Levels Up Local AI Agents Across RTX PCs and DGX Spark」 . <https://blogs.nvidia.com/blog/rtx-ai-garage-computex-spark-local-agents> (一次資料)
- [80] Hugging Face 「Quantized Models for Qwen/Qwen3.6-35B-A3B」 . https://huggingface.co/models?other=base_model%3Aquantized%3AQwen%2FQwen3.6-35B-A3B
- [81] Hugging Face 「wangzhang/Qwen3.6-35B-A3B-abliterated」 . <https://huggingface.co/wangzhang/Qwen3.6-35B-A3B-abliterated> 安全調整を外した派生モデルの例
- [82] Hugging Face 「nvidia/Qwen3.6-35B-A3B-NVFP4」 . <https://huggingface.co/nvidia/Qwen3.6-35B-A3B-NVFP4> (一次資料)
- [83] centron 「OpenAI gpt-oss Explained: Architecture, MXFP4 Quantization & 120B/20B Models」 . <https://www.centron.de/en/tutorials/openai-gpt-oss-explained-architecture-mx4p4-quantization-120b-20>
- [84] Ernest Chiang 「OpenAI GPT-OSS - Open weight models」 . <https://www.ernestchiang.com/en/notes/ai/openai-gpt-oss/>
- [85] Ollama 「nemotron-3-super:120b-a12b」 . <https://ollama.com/library/nemotron-3-super:120b-a12b>
- [86] ExplainX 「Qwen3.8-Max Open Weights Are Live (August 2026)」 . <https://www.explainx.ai/blog/qwen3-8-max-open-weights-live-hugging-face-august-2026>
- [87] ITmedia NEWS 「個人情報保護委員会、中国 AI 「DeepSeek」 利用に注意喚起 林官房長官も「留意を」」 , 2025 年 2 月 4 日 . <https://www.itmedia.co.jp/news/articles/2502/04/news140.html>
- [88] 日本経済新聞 「個人情報保護委員会、DeepSeek 利用に留意を「中国の法令が適用」」 . <https://www.nikkei.com/article/DGXZQOUA038GB0T00C25A2000000/>
- [89] 日本経済新聞 「DeepSeek の利用に注意 政府が各省庁に喚起」 . <https://www.nikkei.com/article/DGXZQOUA069NC0W5A200C2000000/>
- [90] Seculligence (note) 「デジタル庁、中国製 AI 「DeepSeek」 の利用に警鐘 データの中国サーバー保存などにリスク」 . <https://note.com/seculligence/n/nd87a34317f37>
- [91] NVIDIA Developer Blog 「How NVIDIA DGX Spark's Performance Enables Intensive AI Tasks」 . <https://developer.nvidia.com/blog/how-nvidia-dgx-sparks-performance-enables-intensive-ai-tasks> (一次資料)
- [92] Coplus One 「土業の AI 活用 | 守秘義務の条文と事務所の環境・ルール」 . <https://coplus-one.jp/shigyo-ai/>
- [93] 特許庁 「特許制度に関する検討課題について (産業構造審議会知的財産分科会 第 52 回特許制度小委員会 資料 1)」 , 令和 7 年 3 月 5 日 . https://www.jpo.go.jp/resources/shingikai/sangyo-kouzou/shousai/tokkyo_shoi/document/52-shiryuu/01.pdf (一次資料)
- [94] 知的財産高等裁判所 「令和 6 年(行コ)第 10006 号 出願却下処分取消請求控訴事件 判決 (令和 7 年 1 月 30 日)」 . https://www.ip.courts.go.jp/app/files/hanrei_jp/757/093757_hanrei.pdf (一次資料) ※URL 未確認 (特許庁の調査報告書の脚注に記載された URL を転記)
- [95] ビジネス+IT 「【AI は発明者になれるか裁判】 最高裁「発明者は人間に限られる」との判断」 , 2026 年 3 月 6 日 . <https://www.sbbt.jp/article/cont1/182034>