

NVIDIA RTX Spark (128GB) で実現する完全ローカル知財LLM活用ガイド

128GBユニファイドメモリ環境における知財業務向け国産ローカルLLMの選定基準と複合運用アーキテクチャ



NVIDIA RTX Spark (128GB)

128GB UMAにより100B級LLMと
超長文処理を単一端末で実現

約115GBの領域を保持

100B級モデルと32k超のKVキャッシュを確保可能

知財タスクの要求水準（精度 vs 速度）に応じたモデルの最適選定



Stockmark-2-100B

96B / INT4

特許文献の精読、構成要件分解、
クレームチャート作成



CALM3-22B-Chat

22B / INT8

高速対話推敲、発明提案書の
要約、契約書レビュー



Sarashina2.2-Vision-3B

3.8B / FP16

特許図面・回路図・動作
フローチャートの画像解析

複合モデル連携パイプラインと安全運用



図面抽出
(3.8B)



構造化・対比
(100B)



対話推敲
(22B)

総容量約87GBの複数モデル常駐による協調パイプライン構築

図面抽出→構造化・対比→対話推敲をローカル内で一気通貫



ネットワーク隔離による出願前機密保護と根拠明示の義務付け



外部通信を遮断して
新規性喪失リスクを防ぐ



出力には段落番号
【00XX】を必須化