

RTX Spark 128GBにおける知財業務向け国産Local LLM選定ガイド

推奨国産LLMモデルの階層設計

第一候補（通常業務の主力）：
Qwen3-Swallow-32B (AWQ 4-bit)

性能・速度・メモリ残量のバランスが最も
良く、RAGや長文読解に最適。

第二候補（難案件・高精度）：
Stockmark-2-100B-Instruct (4-bit)

約2兆トークンで学習された100B級
フルスクラッチ国産モデル。

第三候補（高速・ソブリン）：
CALM3-22B-Chat

完全国内開発の22Bモデル。海外基盤に
依存しない安定運用・フォールバック用。

知財ローカルLLMの運用・実装原則

128GB統合メモリの実用安全圏は100～112GB以下

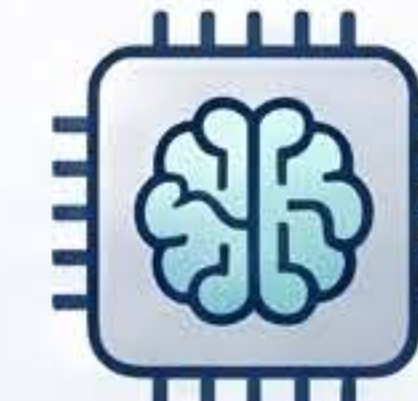


OSやKVキャッシュ等と競合するため、22B～32Bの4-bit量子化がスイートスポット。

J-PlatPat等と連携するRAG＋機械検証ワークフロー



J-PlatPat



RAG



LLMの内部知識に依存せず、一次データ取得と独立コードによる引用照合を実施。

モデル出自（ライセンス）と法務ガバナンスの徹底



Apache-2.0



Apache-2.0や独自規約、輸出管理・著作権の観点からモデル管理（SBOM化）を行う。