

NVIDIA RTX Spark (128GB) 知財特化型ローカルLLM運用戦略

未公開発明などの極めて高い秘匿情報を護るため、128GB統合メモリを持つNVIDIA RTX Sparkを活用したオンプレミスLLM運用が有効です。本図は、メモリ帯域制約を踏まえたモデル選定基準と、知財タスクの処理効率を高める階層型ルーティング設計の要点をまとめたものです。

モデル選定とライセンス要件



日本語知財の主力はQwen 2.5 (72B/32B/14B)

完全無償のApache 2.0で、高い日本語力とクレーム整合性検証力を備える。



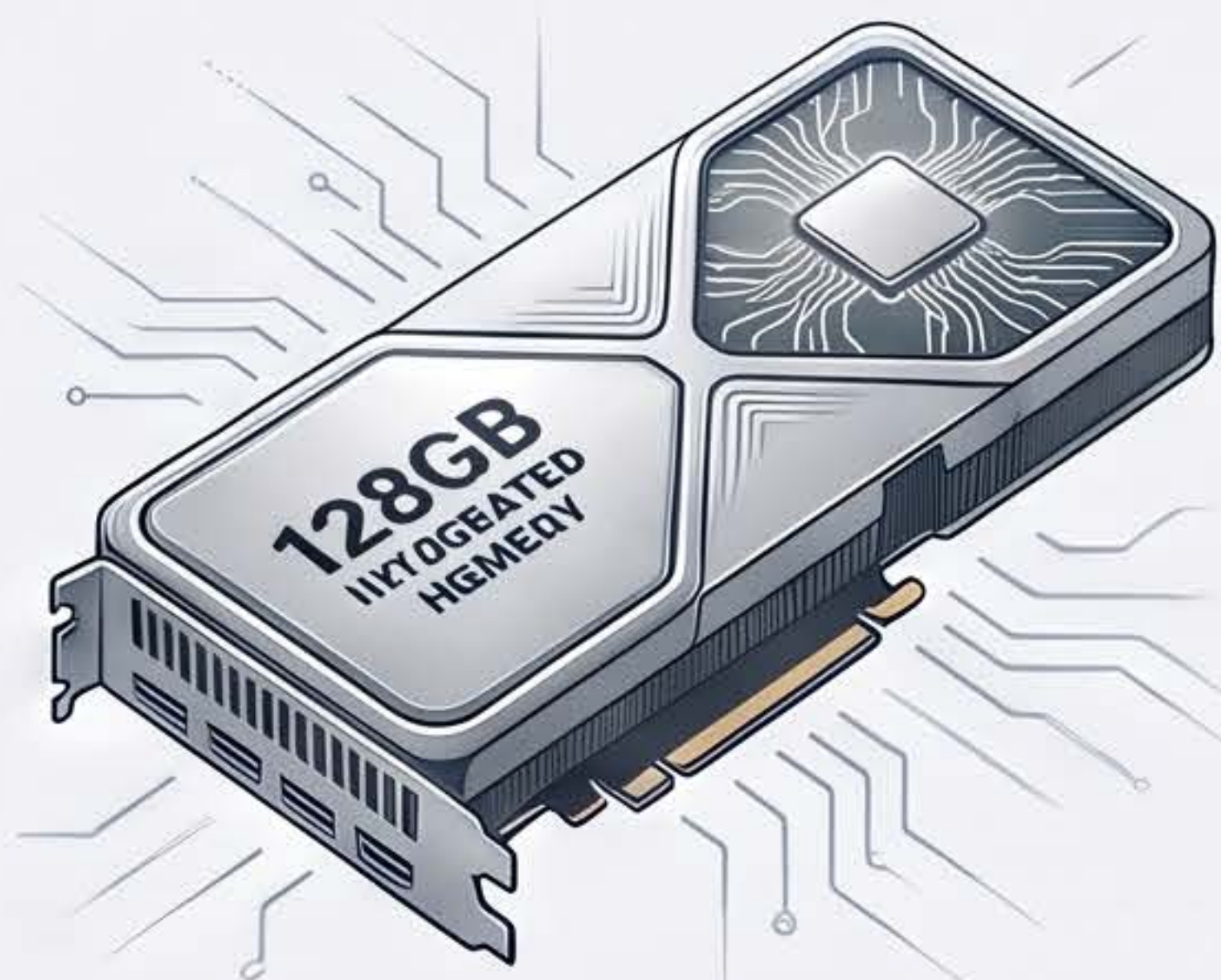
英語特許はLlama 3.3、推論反論はDeepSeek-R1

外国公報読解にはLlama 3.3、拒絶理由通知への論理構築にはDeepSeekが適合。



Command R+ の無断商用利用は法的に不可

CC-BY-NC-4.0ライセンスのため、企業利用には商用契約が必須となる。



階層型ルーティングとタスク適合度



Tier 1 (14B級) : 高速応答の日常コパイロット

38~45 token/s

対話型要約や簡易翻訳、社内問合せに対応。



Tier 2 (32B~70B級) : 高精度なコア知財業務



明細書・クレームの記載不備検証や引用文献との差異分析をローカル完結。



最終判断は必ず人間の専門家 (Human-in-the-loop) が実施

FTO判定や出願クレーム最終確定へのLLM単独利用は法的リスクから非推奨。