NVIDIA「RTX Spark」の衝撃：
ローカルAI PCが変える知財実務の未来

RTX Sparkの正体と驚異のスペック

ローカルで1200億パラメータ・100万トークンを実行

20コアのArm CPUと6144 CUDAコアのBlackwell GPUを搭載し、巨大なモデルをWindows PCで直接動かすことが可能です。

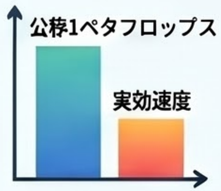
DGX Sparkのシリコンを Windows向けに転用

開発者向けのDGX Spark（GB10）と寸分同一のチップを、消費者・ビジネス向けのWindows on Arm環境へ最適化しています。

1ペタフロップスのAI計算性能

TSMC 3nmプロセスを採用し、最大128GBのユニファイドメモリを備え、Microsoftと共同開発したセキュリティ基盤を搭載しています。

【重要】スペックと実効性能の乖離



生成速度は15〜31トークン/秒

公称1ペタフロップスでも、メモリ帯域（273GB/s）がボトルネックとなり、120B級モデルの実行速度には限界があります。

有効文脈長は公称の50〜70%

100万トークンの文脈が可能とされていますが、実際には「lost in the middle」現象により、中間情報の活用率は低下します。

RAG（検索拡張生成）との併用が現実的

長大な明細書や契約書の分析には有用ですが、額面通りの性能を過信せず、既存の技術と組み合わせるべきです。

知財実務への3つの本質的影響

機密保持リスクの構造的低減

未公開の発明や営業秘密を外部送信せずローカル処理できるため、漏洩や学習利用のリスクを根本から解消します。

知財部門の「二層組織モデル」への移行

定型業務を担う「運用AI層」と、熟練企業や最終判断を使う「戦略人間層」への分類が加速します。

出願の「民主化」と新たなガバナンス

中小・個人事務所でも高性能AIが利用可能になる一方、管理外の「野良AIエージェント」の統制が新たな課題となります。

導入への3段階ロードマップ

第1段階：事実認識の修正と期待値管理

第1段階：事実認識の修正と期待値管理
公称値と実効性能の差を正しく理解し、社内での誤解（漏洩）を防ぐための周知を行います。

第2段階：高機密業務でのPoC（実証実験）

第2段階：高機密業務でのPoC（実証実験）
出願戦略のドラフトや自社特許分析など、機密性の高い領域からローカルLLMの試用を開始します。

第3段階：二層体制の構築とリスキリング

第3段階：二層体制の構築とリスキリング
AI出力を戦略的に評価できる人材を育成し、ローカル環境での監査ログや権限管理を徹底します。

主要なローカルAI実行環境の比較

	NVIDIA RTX Spark	AMD Ryzen AI Max+	Apple M4 Ultra
特徴	CUDAエコシステム対応	高い帯域(256GB/s)	圧倒的帯域(800GB/s+)
強み	業界標準の互換性	消費者向けモデル	巨大なメモリ容量
弱点	消費電力/熱（推定）	CUDA非対応	CUDA非対応/高価格