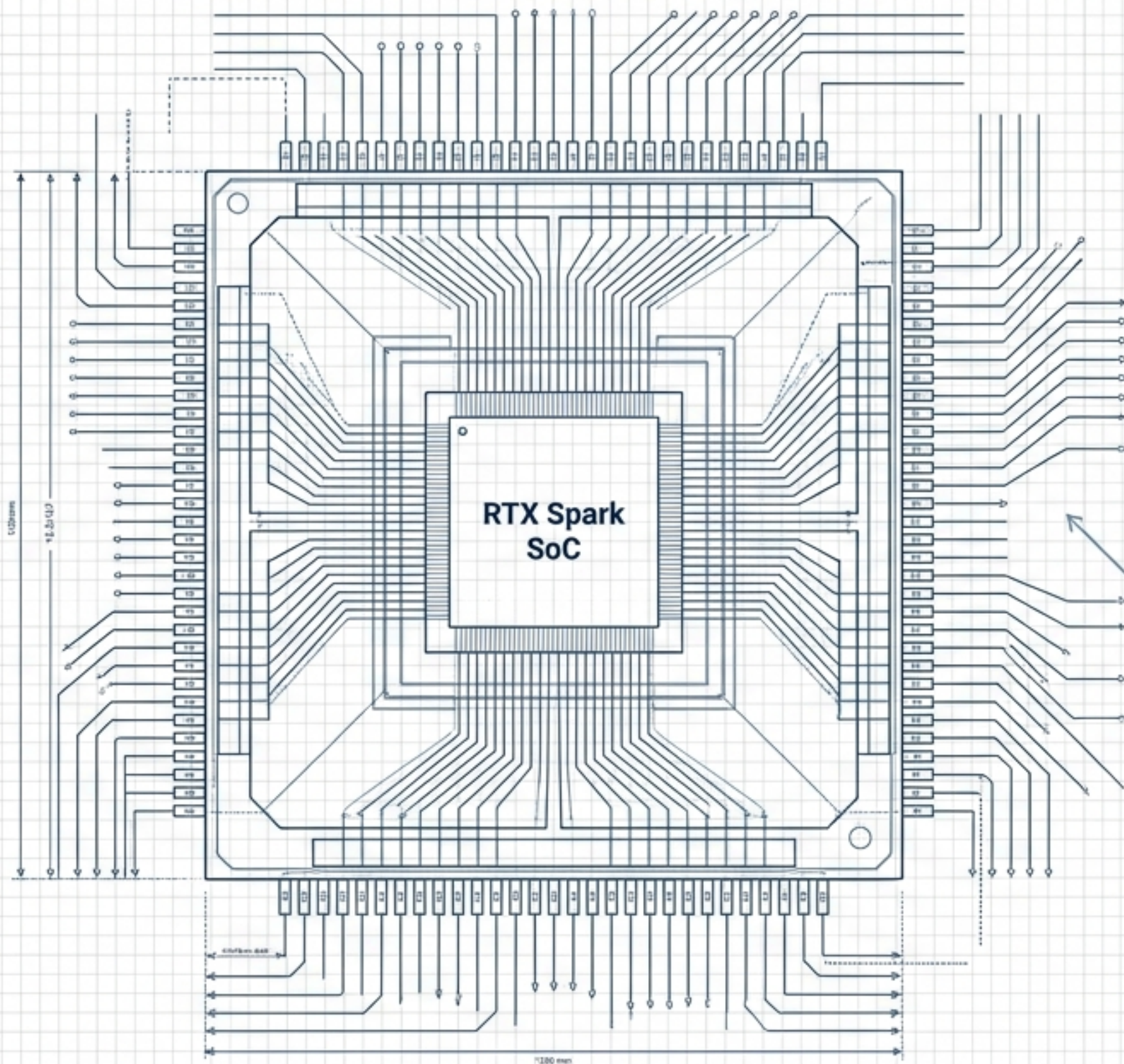




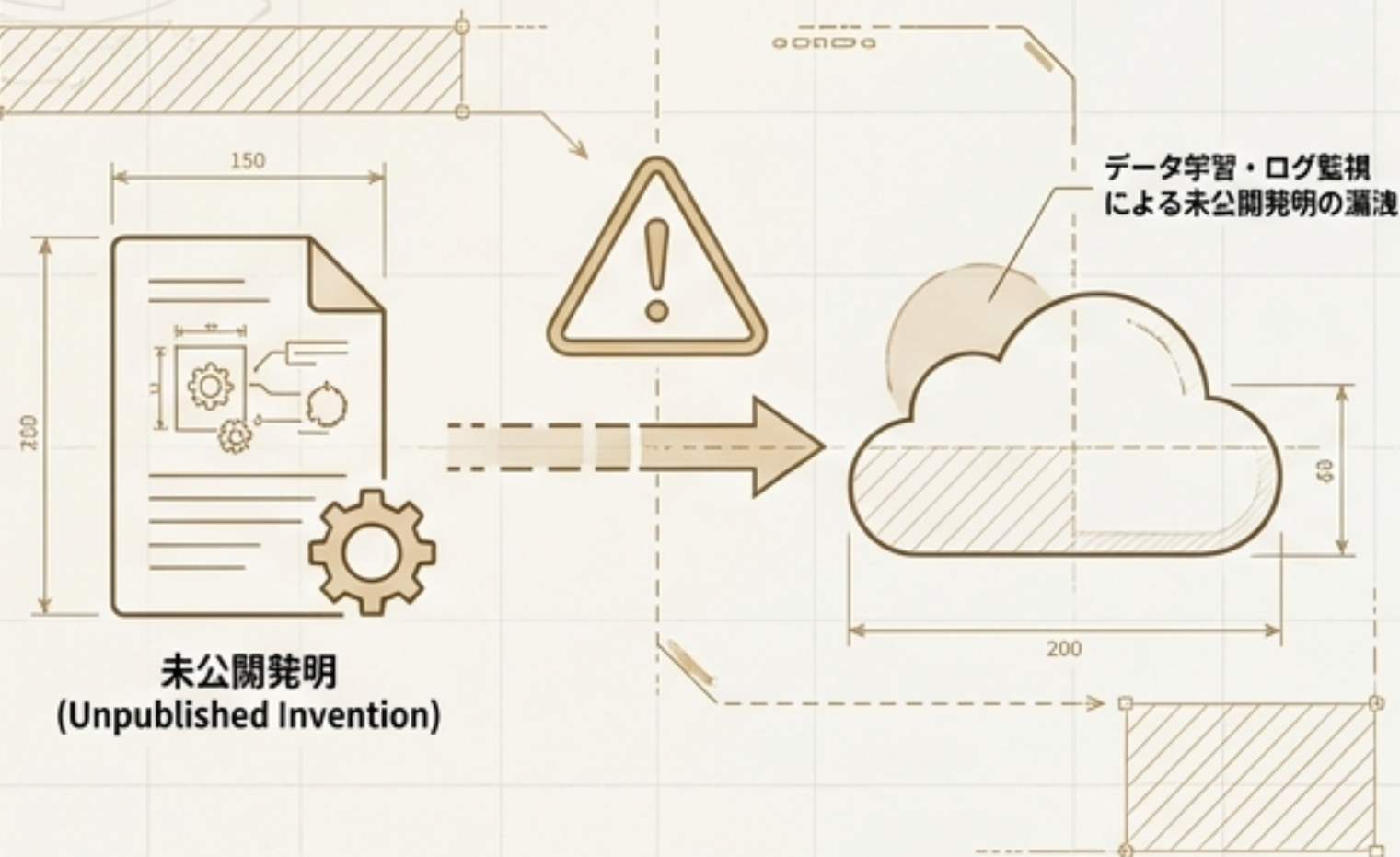
# 知財専用ローカルAI基盤 「RTX Spark」 構築・運用ブループリント

128GBユニファイドメモリ環境における  
国産LLMの選定と完全遮断ワークフロー設計

機密性レベル：最高水準 / 外部ネットワーク：遮断必須

## The Threat

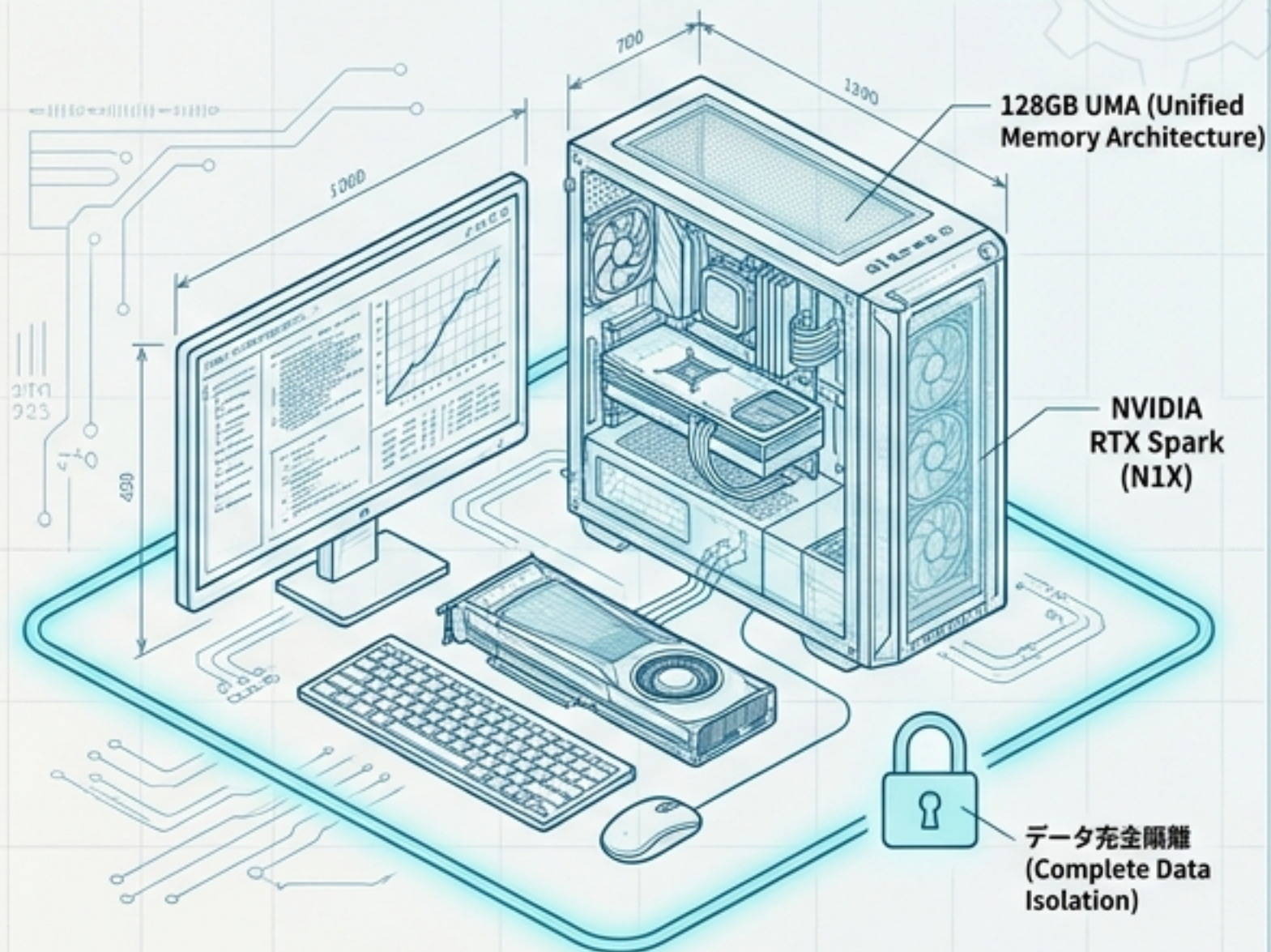
# クラウドAIのリスク：新規性喪失 (特許法第29条)



クラウドAIの利用は、学習データへの取り込みやログ監視による「未公開発明の漏洩」リスクを孕む。

## The Solution

# RTX Spark (128GB) ローカル完結環境



NVIDIA RTX Spark (N1X) は、128GBのユニファイドメモリ (UMA) を搭載し、クラウドに依存せず100Bクラスの巨大モデルをセキュアなローカル環境で単独稼働させる次世代エッジ基盤である。

THOLETYPE:

PROJECT:  
SECURE AI WORKSTATION

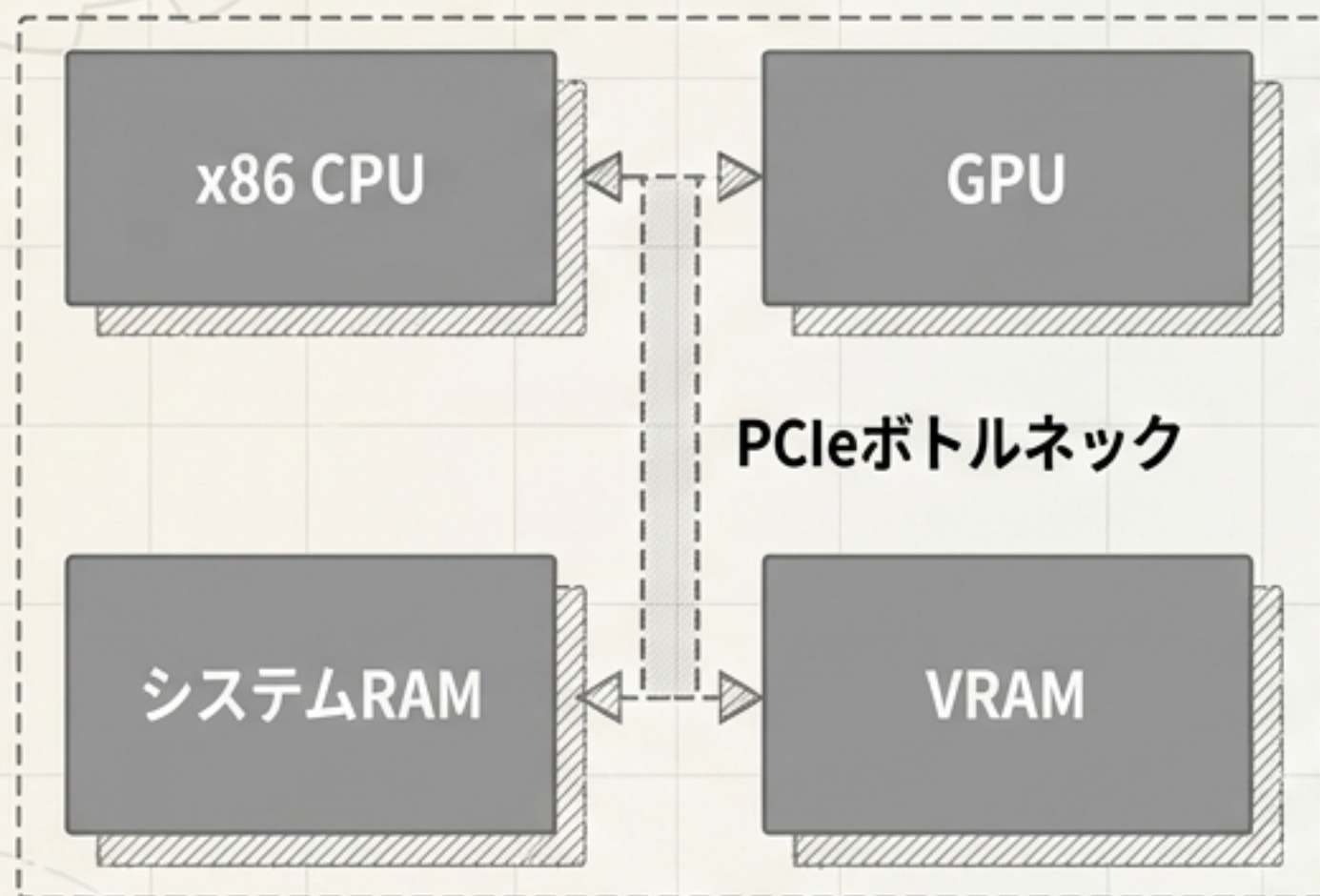
DRAWING NO:  
AI-SEC-601-REV A

DATE:  
OCT 26, 2024

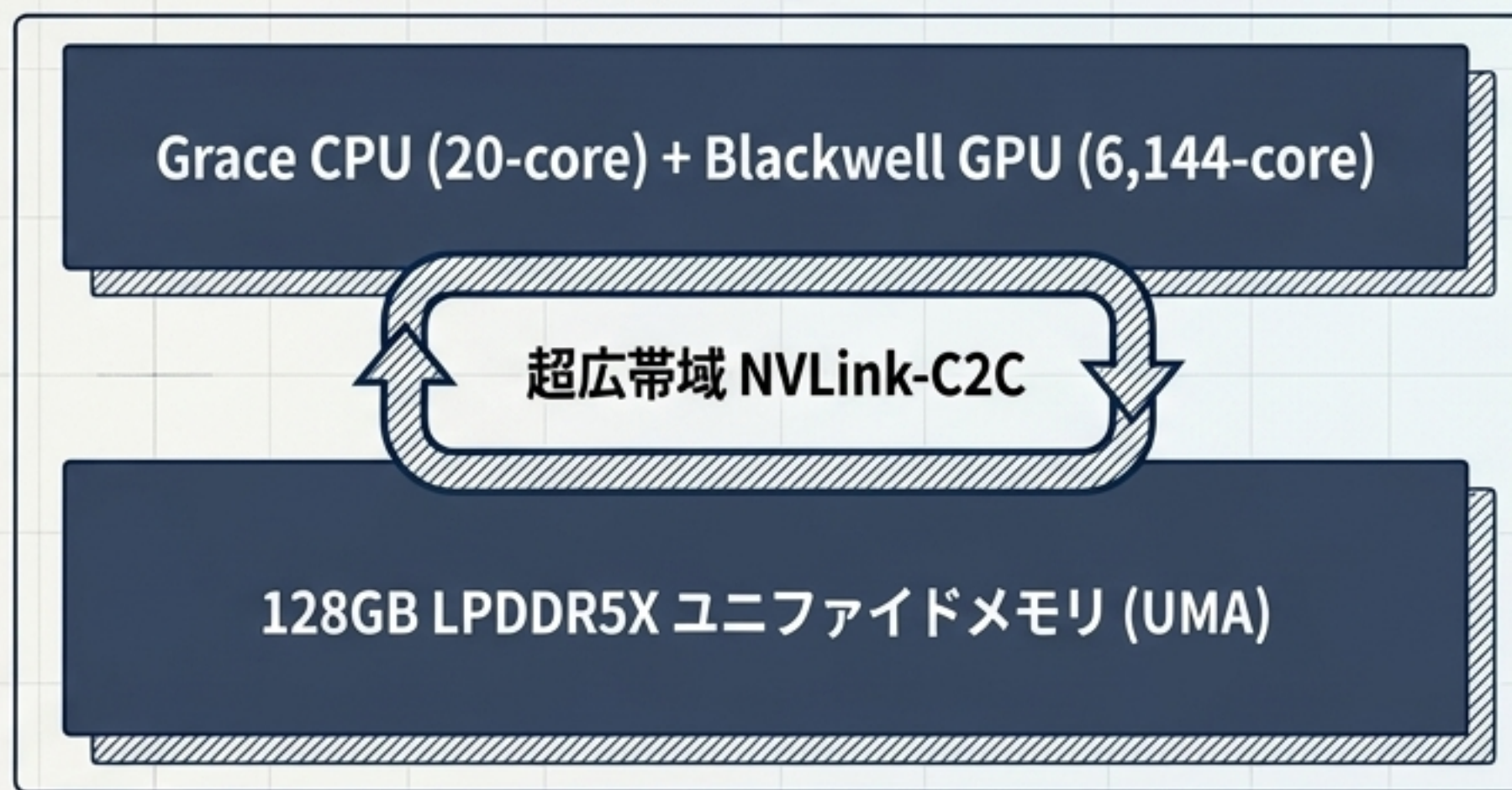
# なぜ128GBが知財業務を変革するのか？

## Unified Memory Engine

### Legacy dGPU



### RTX Spark



### 【圧倒的な優位性】

CPUとGPUが物理メモリ空間をシームレスに共有。複数枚のグラフィックスカードを要する100B規模の特許深層解析を、単一の端末上で実現。

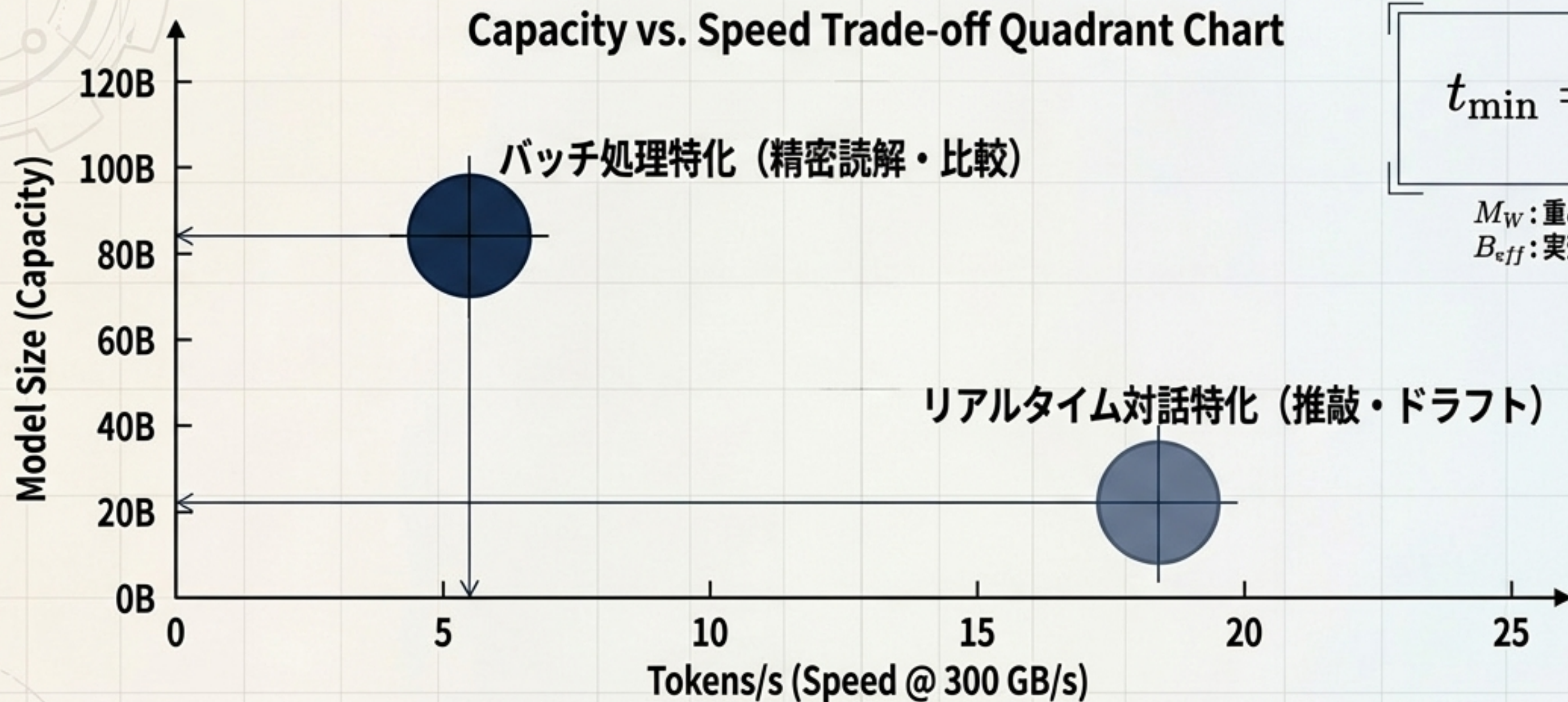
|            |                                   |                       |
|------------|-----------------------------------|-----------------------|
| TH0ELPPPE: | PROJECT:<br>SECURE AI WORKSTATION |                       |
|            | DM00RS AG<br>AI-SEC-601-REV A     | DATE:<br>OCT 25, 2024 |

# ハードウェア推論基盤の比較評価

|          | RTX Spark                 | RTX 5090        | RTX 6000 Ada       |
|----------|---------------------------|-----------------|--------------------|
| 構成       | Grace CPU + Blackwell GPU | x86 + Blackwell | x86 + Ada Lovelace |
| メモリ形態    | ユニファイド UMA                | 独立 GDDR7        | 独立 GDDR6 ECC       |
| 有効VRAM容量 | 最大128 GB（共有）              | 32 GB           | 48 GB              |
| メモリ帯域幅   | 約273～300 GB/s             | 1,792 GB/s      | 960 GB/s           |
| 最大稼働モデル  | 100B～120B級（INT4）          | 32B級（INT8）      | 70B級（INT4）         |

帯域幅（生成速度）ではdGPUに劣るが、  
特許文献の深層解析に不可欠な「100Bモデルの搭載能力」  
においてRTX Sparkが唯一の選択肢となる。

# メモリバウンド特性とタスク別モデル最適化



$$t_{\min} = \frac{M_W}{B_{\text{eff}}}$$

$M_W$ : 重みデータサイズ  
 $B_{\text{eff}}$ : 実効メモリ帯域幅

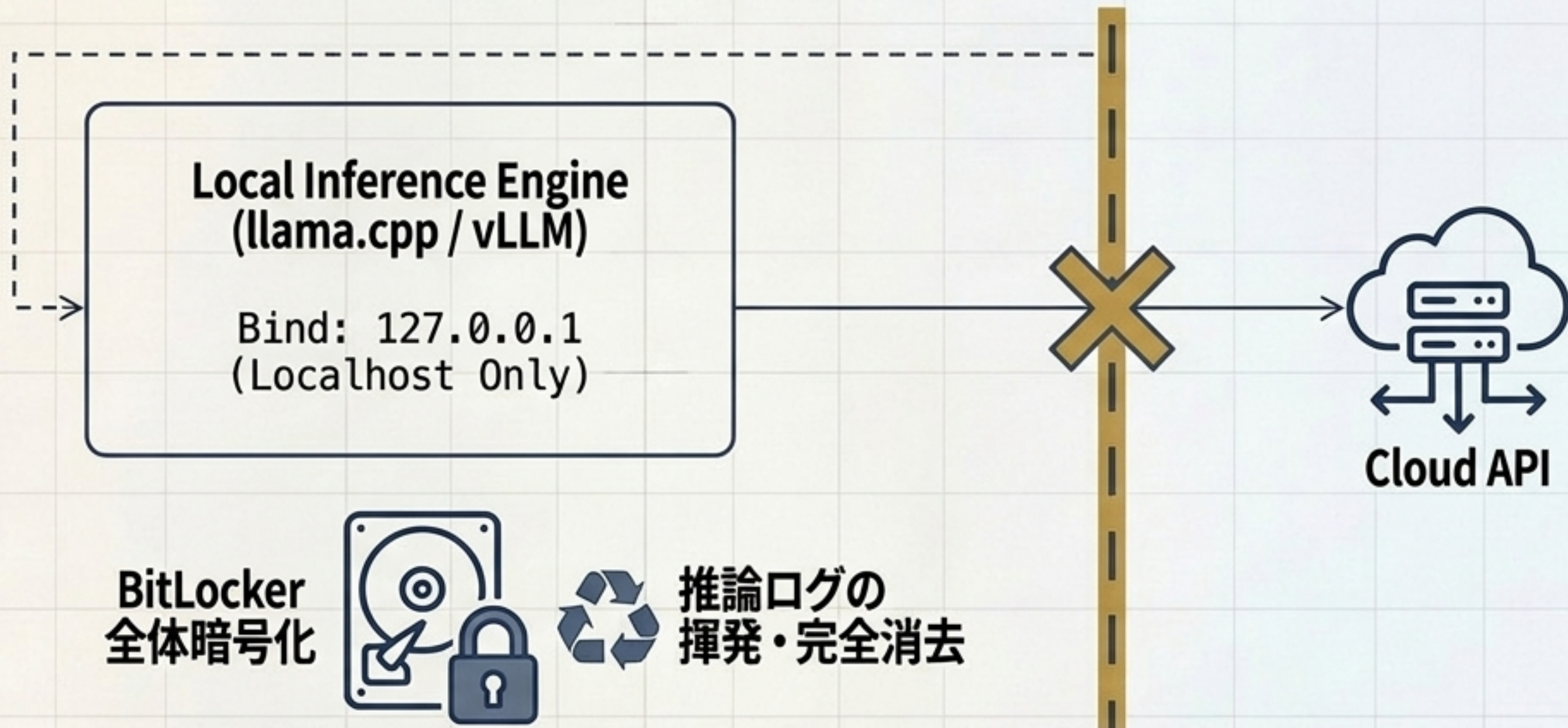
RTX Sparkの出力スループットは読み出し速度に律速される。  
タスクの要求速度（リアルタイム対話 vs バッチ解析）に  
応じたモデル規模の使い分けが必須。

# 知財実務向け国産LLM：特性と適合領域マップ

| モデル名                      | パラメータ<br>(量子化)    | VRAM要件    | 最大<br>コンテキスト | 知財業務の適合領域                | ライセンス      |
|---------------------------|-------------------|-----------|--------------|--------------------------|------------|
| Stockmark-2-100B-Instruct | 96B (INT4)        | 約68-75 GB | 32k          | 特許文献精読、構成要件分解、<br>先行技術比較 | MIT        |
| Swallow-70B-Instruct      | 70B (INT4)        | 約50-65 GB | 128k         | 意見書・補正書の論理構築             | Llama 3.1  |
| CALM3-22B-Chat            | 22B<br>(INT8)     | 約28-48 GB | 16k          | 高速な推敲対話、<br>発明提案書要約      | Apache 2.0 |
| PLaMo 2 8B                | 8B<br>(FP16/INT8) | 約16-22 GB | 32k          | 日英特許の高精度翻訳支援             | PLaMo      |
| Sarashina2.2-Vision-3B    | 3.8B<br>(FP16)    | 約8-10 GB  | -            | 特許図面、フローチャートの<br>画像解析    | MIT        |
| Fugaku-LLM-13B            | 13B<br>(INT8)     | 約16-28 GB | 4k           | 定義文・構成要素の短文抽出            | Fugaku     |

# 出願前機密保持と新規性喪失リスクの完全遮断

## Network Quarantine Schematic



特許法第29条に基づく新規性喪失を完全に防ぐため、推論エンドポイントをローカルホストに限定し、ストレージの暗号化とセッションキャッシュの完全破棄を徹底する。

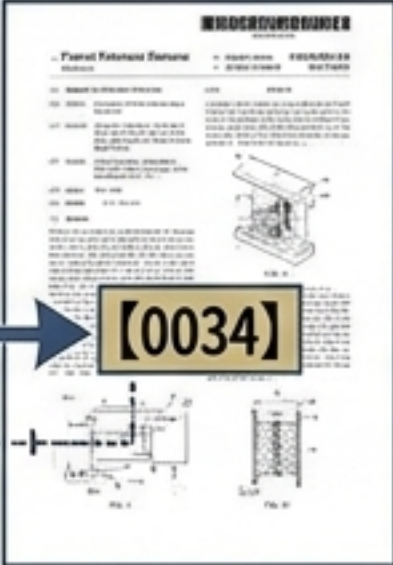
## 一般的な自由記述生成

Lorem ipsum dolor sit amet consectetur  
adipiscing elit. Suspendisse potenti.  
Sed ut perspiciatis unde omnis  
iste natus est error qui consequatur  
autem vel illum qui dolorem eum fugiat  
quia voluptas tempora et vero  
quod autem sunt. Lorem ipsum  
dolor sit amet consectetur adipiscing  
elit. Sed ut perspiciatis unde omnis  
iste natus est error qui consequatur  
autem vel illum qui dolorem eum fugiat  
quia voluptas tempora et vero quod  
autem sunt.

**特許法上の辞書的定義を無視した  
広義解釈や、学習データに起因する  
ハルシネーションの法的リスク。**

## 構造化出力 (JSON形式) による根拠明示

```
{
  "構成要件": "弾性部材を備え",
  "解釈": "パネまたはゴム材料に限定される",
  "根拠段落": "[段落番号【0034】]"
}
```



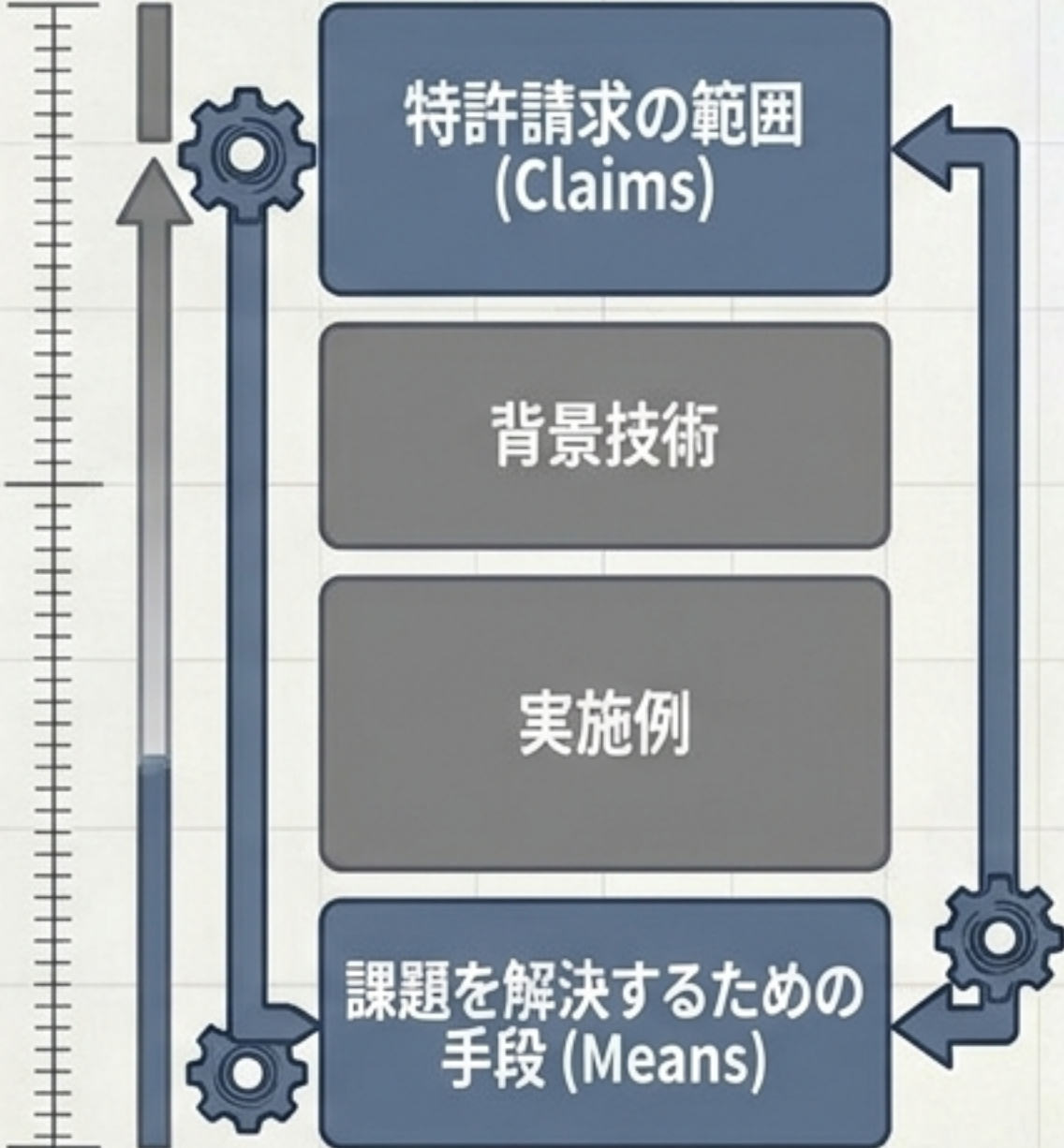
**自由記述を厳格に制限し、全ての解釈に対して明細書中の段落番号と請求項番号を紐付ける出力を強制する。**

# 文脈埋没現象 (Lost in the Middle) の克服

最適化: XMLパース&再配置



非推奨: 一括投入  
(文脈埋没)



最適化:  
XMLパース&再配置

長大な明細書を一括投入せず、  
XML構造をパースして  
レイアウトを最適化。  
最も重要な「特許請求の範囲」を  
プロンプトの先頭と末尾に配置  
し、アテンション機構の  
見落としを防止する。

# オープンモデルの商用利用トラップとコンプライアンス



## MIT / Apache 2.0

(Stockmark, CALM3, Sarashina)

無償かつ商用・内部業務利用に法的な制約なし。安全に社内ツールへの組み込みが可能。



## Llama 3.1 Community

(Swallow)

月間アクティブユーザー7億人超の事業者に対する特有の許諾申請義務。生成AIモデルの競合改善への利用制限あり。



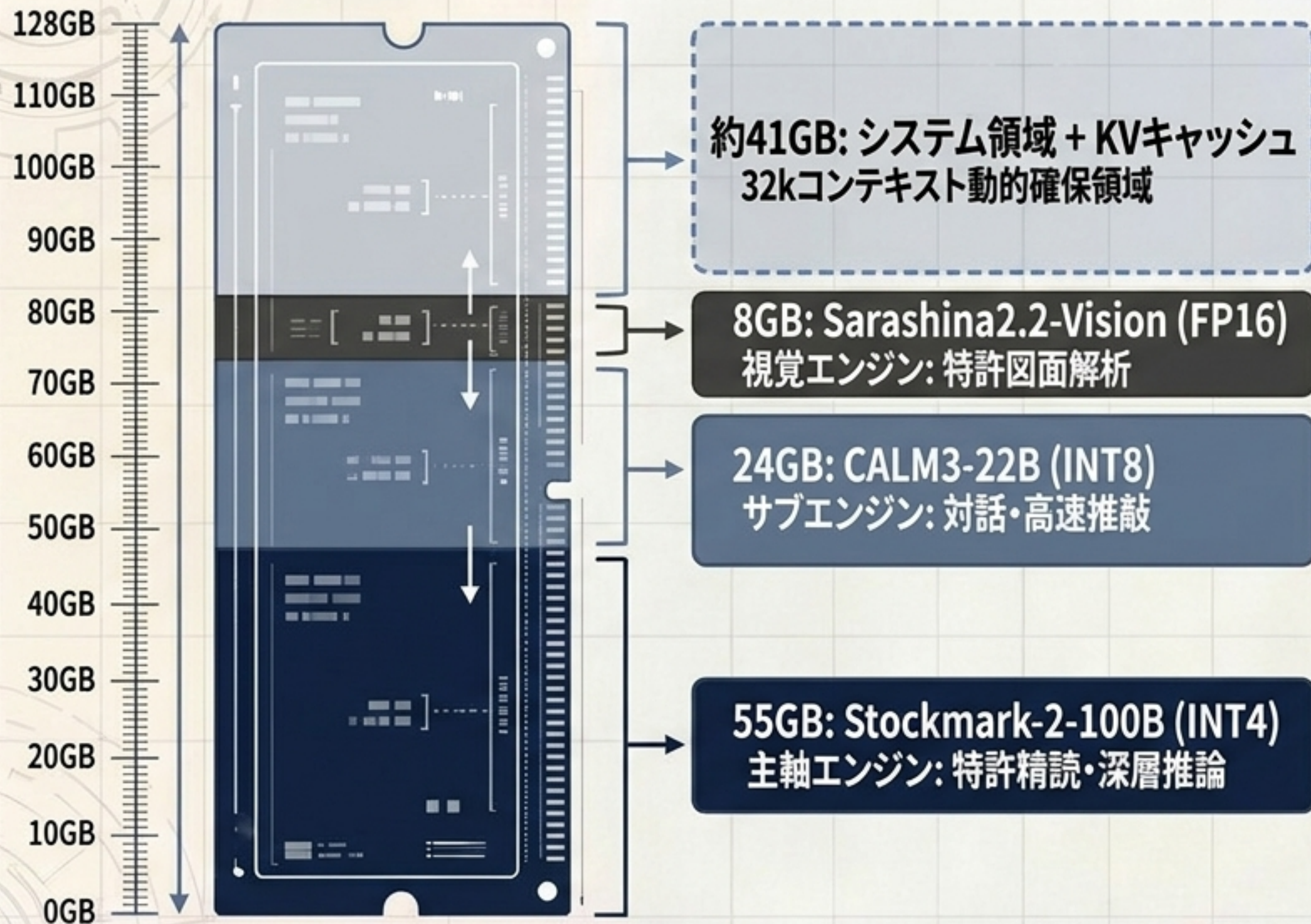
## PLaMo Community

(PLaMo 2 8B)

直近事業年度の売上高「10億円」を超える企業の業務利用は無償対象外。別途有償契約の締結が必須。

オープンウェイト＝無条件の商用利用可能ではない。知財部門自身による厳密な規約照合が必須。

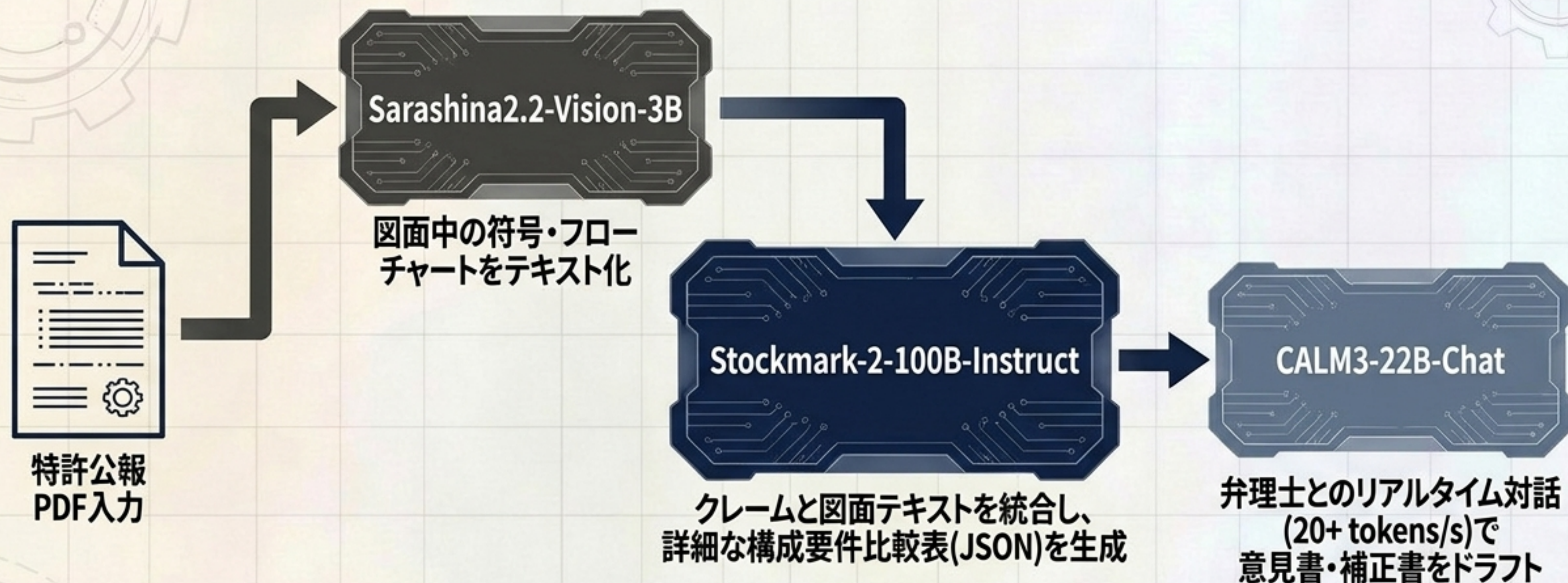
# 複合モデル協調アーキテクチャの実現（128GBの割り当て）



128GBの真価は単一モデルの稼働ではなく、「AIチーム」の同時常駐にある。  
重み総容量約87GBを展開しても、長文解析用のKVキャッシュを余裕で確保可能。

|             |                                   |              |
|-------------|-----------------------------------|--------------|
| T04GL177PL0 | PROJECT:<br>SECURE AI WORKSTATION |              |
|             | DMP CY5AC                         | 010-         |
|             | AI-SEC-633-REV A                  | OCT 25, 2634 |

# ローカルPC内で完結する多段階知財解析パイプライン



出願前機密の完全保護、多角的な図面処理、そして深層推論と対話応答性の両立。  
RTX Sparkは単なるハードウェアではなく、卓上の完結した知財エージェンシーとして機能する。

|           |                                 |              |
|-----------|---------------------------------|--------------|
| T04GLTFL6 | WO #01<br>SECURE AI WORKSTATION |              |
|           | DMETCS.AE                       | orla         |
|           | ローカル知財パイプライン                    | OCT 25, 2634 |