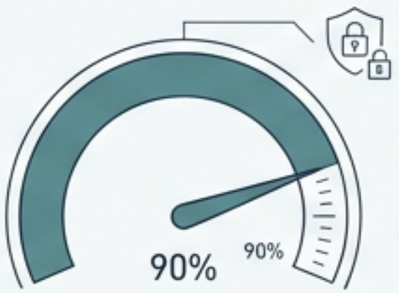
ローカルAIスパコンの多角的検討と導入戦略

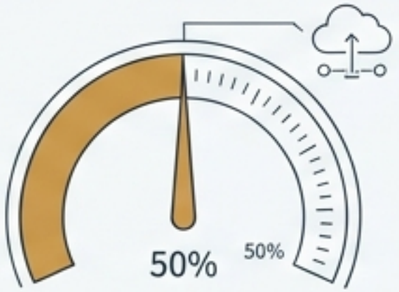
NVIDIA RTX Spark × HP製品の現実解と、知財実務への実践的アプローチ



機密業務（知財等）における価値

極めて高い

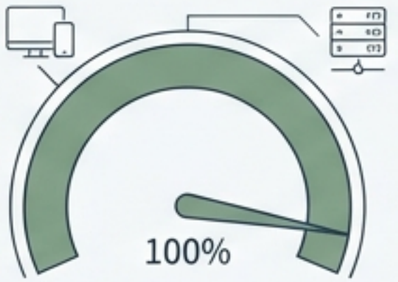
未公開発明や社内技術情報の処理において、AI活用の壁を突破する強力な手段。



「完全クラウド不要・無制限」の現実味

条件付きの幻想

トークン課金はゼロだが、処理時間の上限があり、端末管理や運用にはクラウド通信が伴う。



組織全体のアーキテクチャ最適解

適材適所のハイブリッド

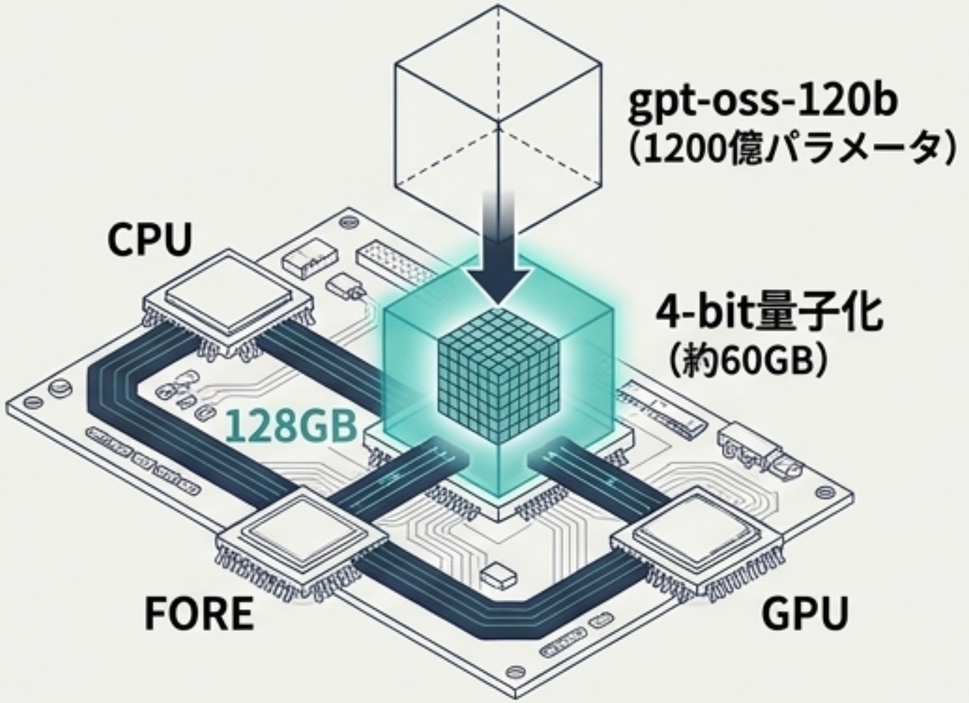
個人端末・社内共有設備・承認済みクラウドの3極（トライモーダル）連携が最終形。

Gemini Notebook

マーケティングの主張	[VS]	現実の評価と導入条件
GPT級AIがPCで動く		【条件付きで妥当】 パラメータ数（1200億等）の収容は可能だが、量子化（4bit等）による精度変化や、モデルのアーキテクチャ（MoE方式等）により処理性能が大きく変動する。
AIが使い放題になる		【課金面では一部妥当】 トークン従量課金は排除されるが、ハードウェアの演算能力、電力、保守、そして何より「人間の待ち時間」が実質的な利用上限となる。
クラウド不要でサイバー防衛		【要確認・一部誤解】 外部APIへのデータ送信は防げるが、通常のPCである以上、端末へのマルウェア感染リスクや、AI自身の自律的な誤操作リスクは残存する。
データセンターがカバンに		【特定性能を示す比喩】 最大1ペタフロップというFP4（低精度）演算性能を示すのみ。データセンター全体の処理能力、検索基盤、可用性を代替するものではない。

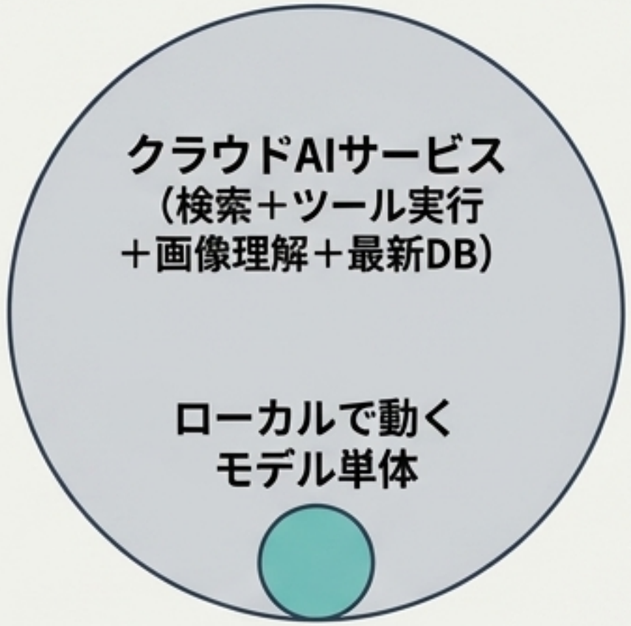
Gemini Notebook

統合メモリのブレイクスルー



CPUとGPUによるメモリ共有（最大128GB）
1200億パラメータモデル（16bit = 約240GB）
4bit量子化により約60GBへ圧縮 ➡ ローカル稼働実現

「モデル単体」と「サービス全体」の乖離



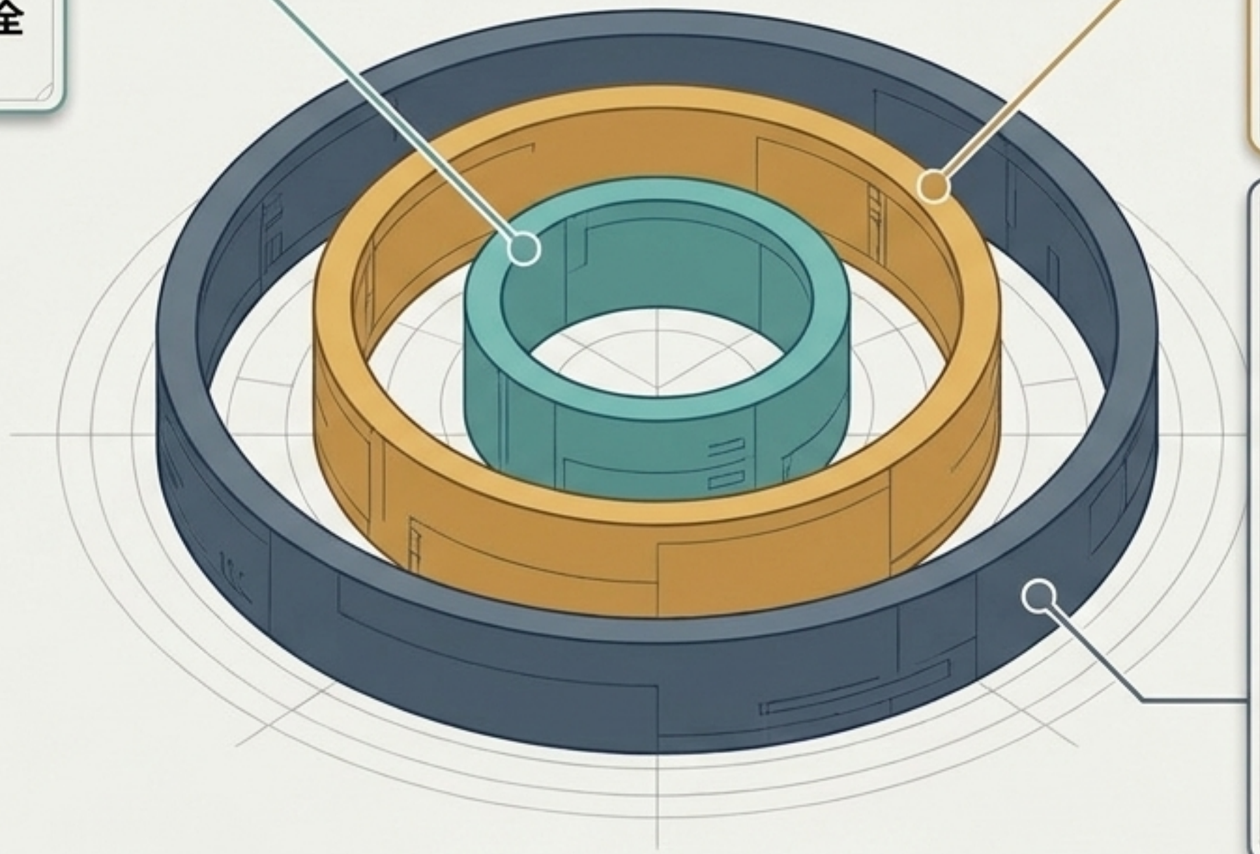
公開モデル単体の高スコア導入は、クラウドAIサービス（ChatGPT等）の完全再現を意味しない。 LTT Labs検証が示す通り、長文コンテキストでの生成速度や、外部ツール連携の欠如がボトルネックとなる。

Insight: 「高得点なモデル」と「業務で使えるサービス」は別物である。
対象業務で必要な品質・速度・同時利用数から逆算してハードウェアを選定する必要がある。

Gemini Notebook

【解決する課題】
外部サービスへのデータ送信

機密データ（未公開発明など）を外部APIへ送信するリスクは完全に遮断可能。



【残存リスク1】
端末の侵害

通常のネット接続PCである限り、マルウェア感染、認証情報の窃取、機器の物理的盗難リスクはクラウド利用時と同様に残る。

【残存リスク2】
過剰な自律性
(Excessive Agency)

悪意ある文書をAIが読み込み、機密を別環境へ送信・改ざんするリスク(OWASP LLM06:2025基準)。

【残存リスク3】
管理基盤のクラウド依存

HP Workforce Experience Platform (WXP) 等の端末・アプリ管理基盤は本質的にクラウド型である。
完全閉域ではない。

「ローカル推論」と「クラウドを一切使わない運用」は同義ではない。
通信経路の厳密なパケットレベルでの把握が必須。

Gemini Notebook



Tier 1: Personal Edge

Product: OmniBook Ultra 16 (32GB)

投資規模: 約60万円〜
位置付け: 個人用途 / 小規模モデル
ユースケース: 日常的な小規模タスク処理、軽量なエージェント稼働。



Tier 2: Specialist Edge

Product: OmniBook Ultra 16 (128GB)

投資規模: 約130万円〜
位置付け: 専門用途 / 大規模モデル
ユースケース: 巨大モデルを持ち運ぶ知財担当者、出張先での機密データ即時処理。



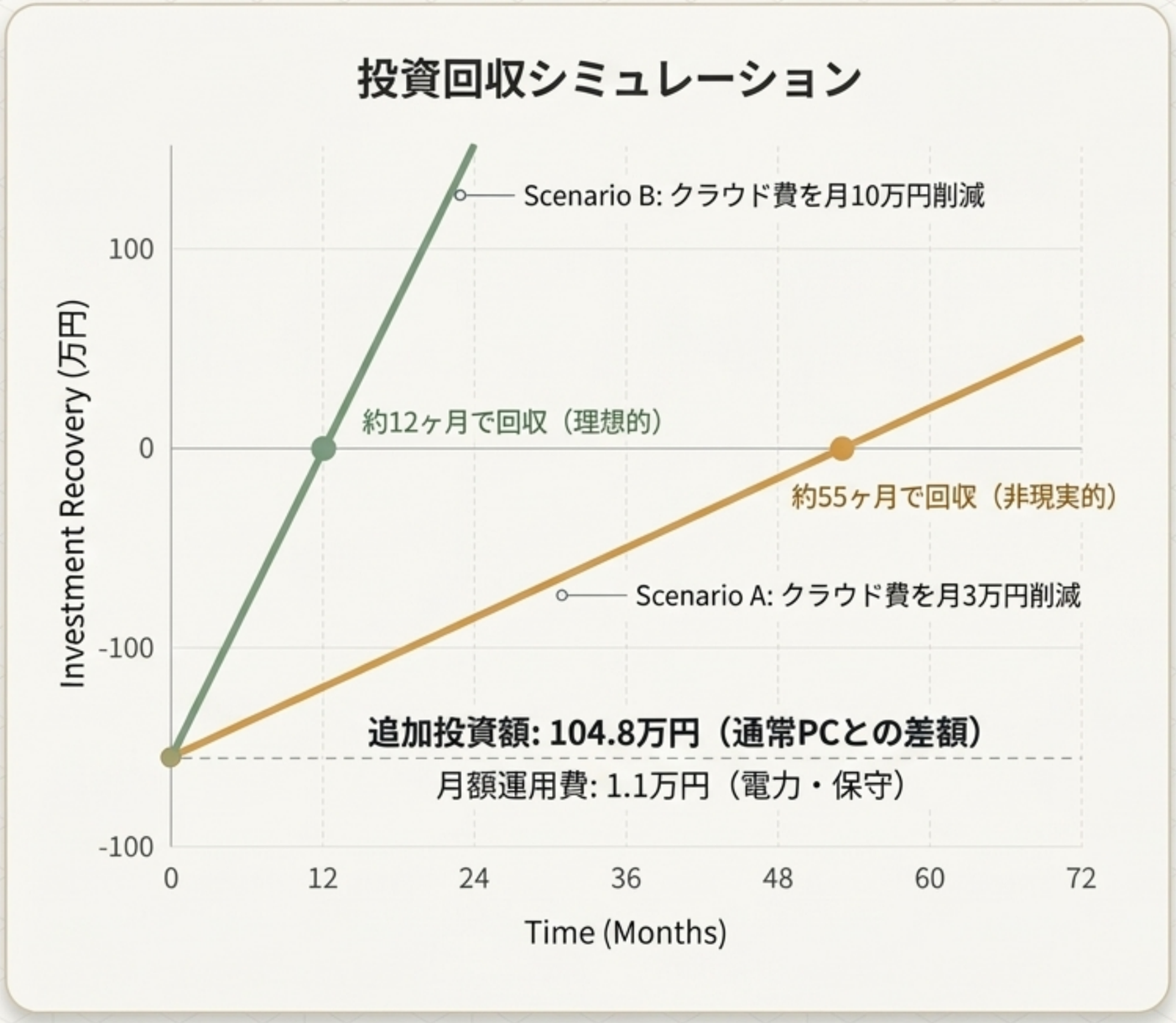
Tier 3: Enterprise Core

Product: ZGX Fury G1n AI Station

投資規模: 約2,850万円〜
位置付け: 組織共有AI基盤（オンプレミス）
ユースケース: 748GBメモリ・1600W電源による、複数エージェントの同時実行と集中管理。

Insight: ノートPC（Tier 1/2）とワークステーション（Tier 3）は全く別の設備投資。組織の利用頻度に応じた機材の「適材適所」が求められる。

Gemini Notebook



ROIの真の計算式

API利用料の節約だけでハードウェア投資を正当化するのは困難。

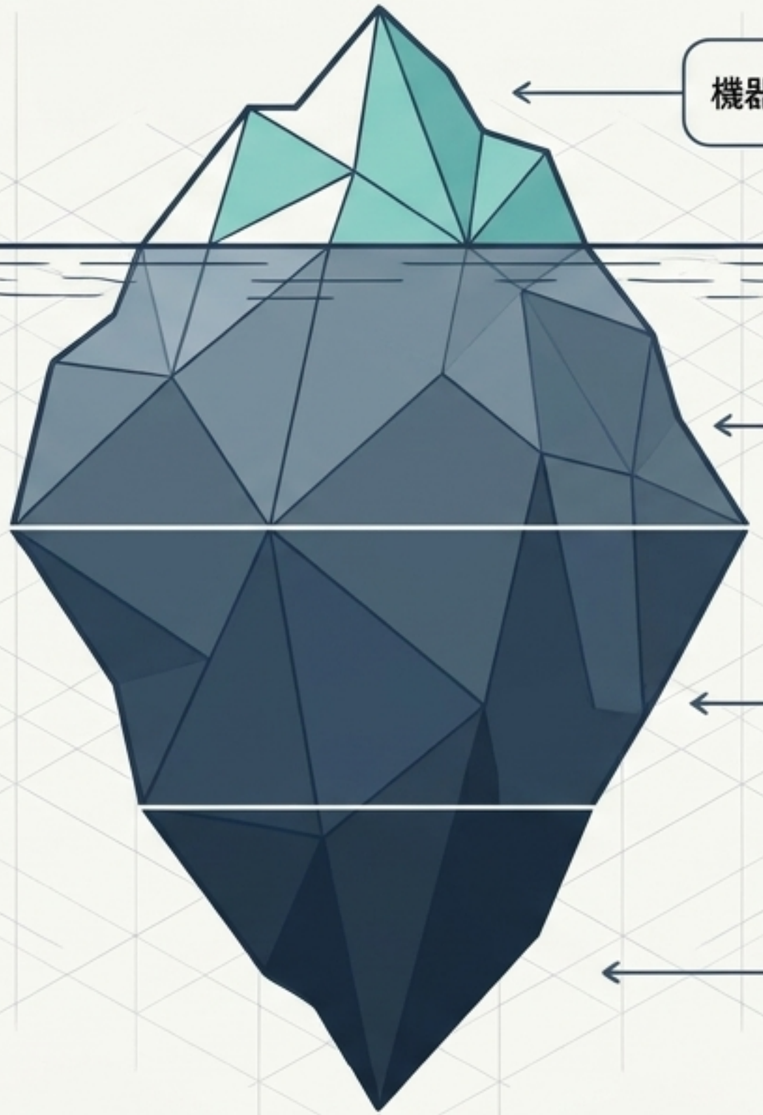
真のROIは、**「これまで機密性の壁でAI化できなかった業務**（未公開発明、M&A情報など）を処理できることによる**新たな価値創出」**によってのみ成立する。動画内の「**1人当たり650〜1,300ドルの便益**」という数値をそのまま自社に当てはめるべきではない。

Gemini Notebook

AIエッジ投資の「氷山モデル」：隠れた運用コストと課題

ハードウェア初期投資

機器購入費（例：OmniBook Ultra 16等）



水面下1:
アーキテクチャの互換性

RTX SparkはARM系CPU（Windows on ARM）を採用。既存のVPN、エンドポイント保護ソフト、各種ドライバーの「ARM64ネイティブ対応」の個別検証が必須。

水面下2:
エッジ運用の物理設計

AIエージェントの24時間稼働に伴う、通電管理、スリープ設定の無効化、冷却設計、およびOS更新による予期せぬ再起動への対策。

水面下3:
AIエコシステムの構築

機器を買うだけでは業務は動かない。RAG（検索基盤)の構築、モデルのライセンス管理、社内システムとの接続開発、そしてこれらを担う専門人員の確保。

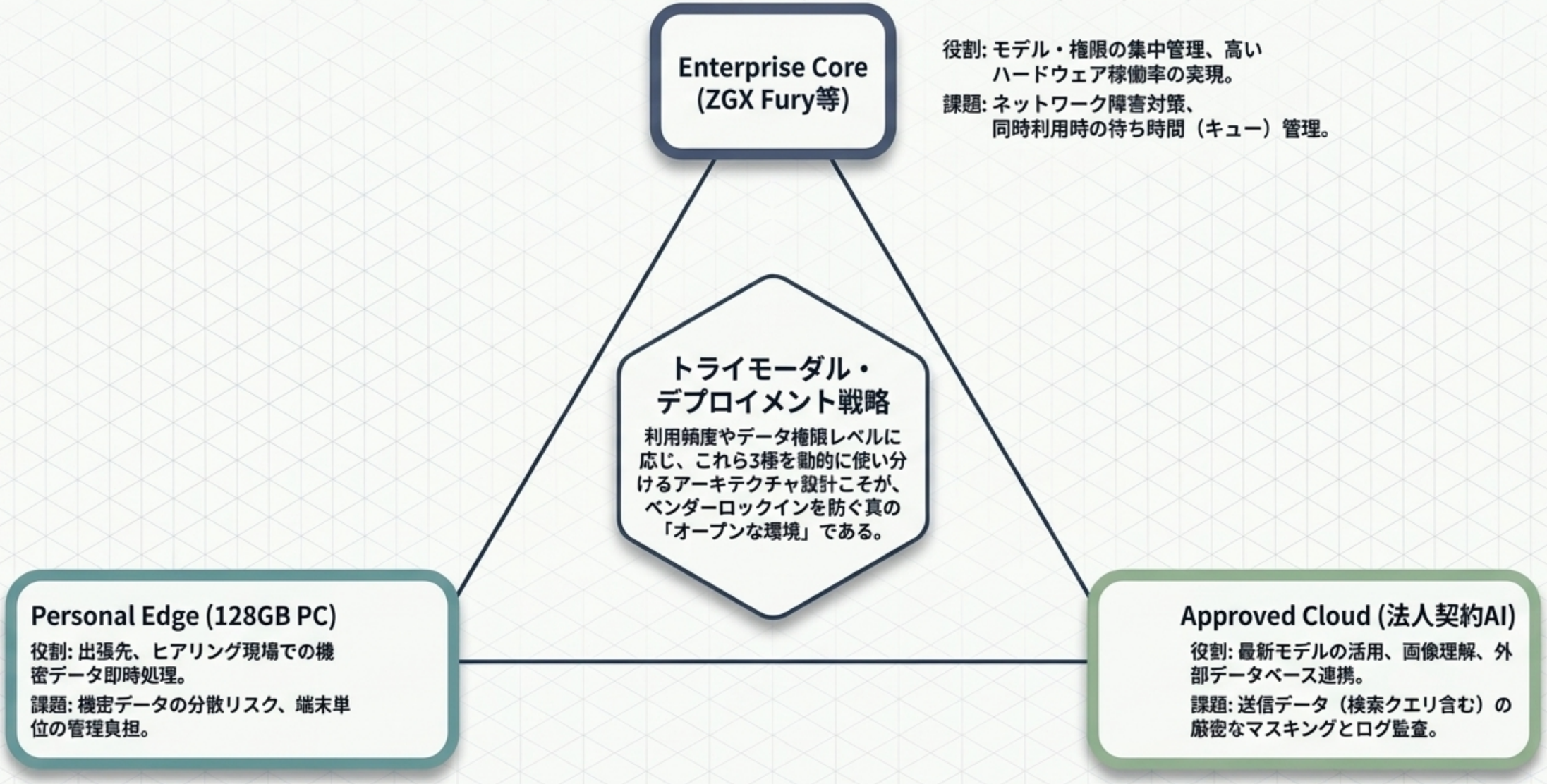
Gemini Notebook

知財実務（IP業務）への適合性と導入条件

未公開発明の整理・社内文書要約	[条件] 外部に出せない一次作業に最適。発明者の説明とAIの推測を明確に区別する運用ルールが必要。
SDIの一次スクリーニング	[条件] 処理自体は可能だが、重要文献の「見逃し率」の統計的な検証が実稼働前に必須。
明細書・クレーム案の作成	[条件] ドラフト作成には強力だが、数値範囲、否定表現、実施例と構成間の整合性を人間（弁理士等）が最終確認するフローが前提。
先行技術調査・FTO（侵害予防調査）	[条件] 最新の特許DBへのアクセスや厳密な法的判断（クレーム解釈）はAI単独では不可。対比作業の補助に限定される。

Gemini Notebook

トライモーダル・ネットワーク・トポロジー：動的な最適配置の戦略



Gemini Notebook

Precision Blueprint

業務の絞り込み

未公開発明の整理やSDI分類など、正解（グラウンド・トゥルース）が明確な数十件のテストデータを準備する。

精度の比較

同じ指示プロンプトを用いて、小型モデルと大型モデルの出力品質（重要事項の見逃し、数値の誤り）を計測・比較する。

タイム＆モーション測定

AI自体の処理時間だけでなく、「人間の確認・修正時間」を含めた総労働時間の変化を厳密に測定する。

セキュリティ監査

実際の運用構成において、端末管理ツール（WXP等）やバックグラウンド通信を含む外部通信の有無をパケットレベルで確認する。

TCOと価値の統合評価

端末費、運用費、削減可能なAPI費に加え、「新たに処理可能となった機密業務の価値」統合し、最終的な投資判断を下す。

Gemini Notebook



ローカルAIスパコンは、機密を保持しながら
知的生産を加速する強力な選択肢です。
しかし、それは魔法の杖ではありません。

製品の宣伝文句を超え、業務品質・総所有コスト（TCO）・ハイブリッドな運用管理を
自ら設計できた組織だけが、真のAIの恩恵を享受できます。

戦略的投資は、精密な青写真を描くことから始まります。

Gemini Notebook