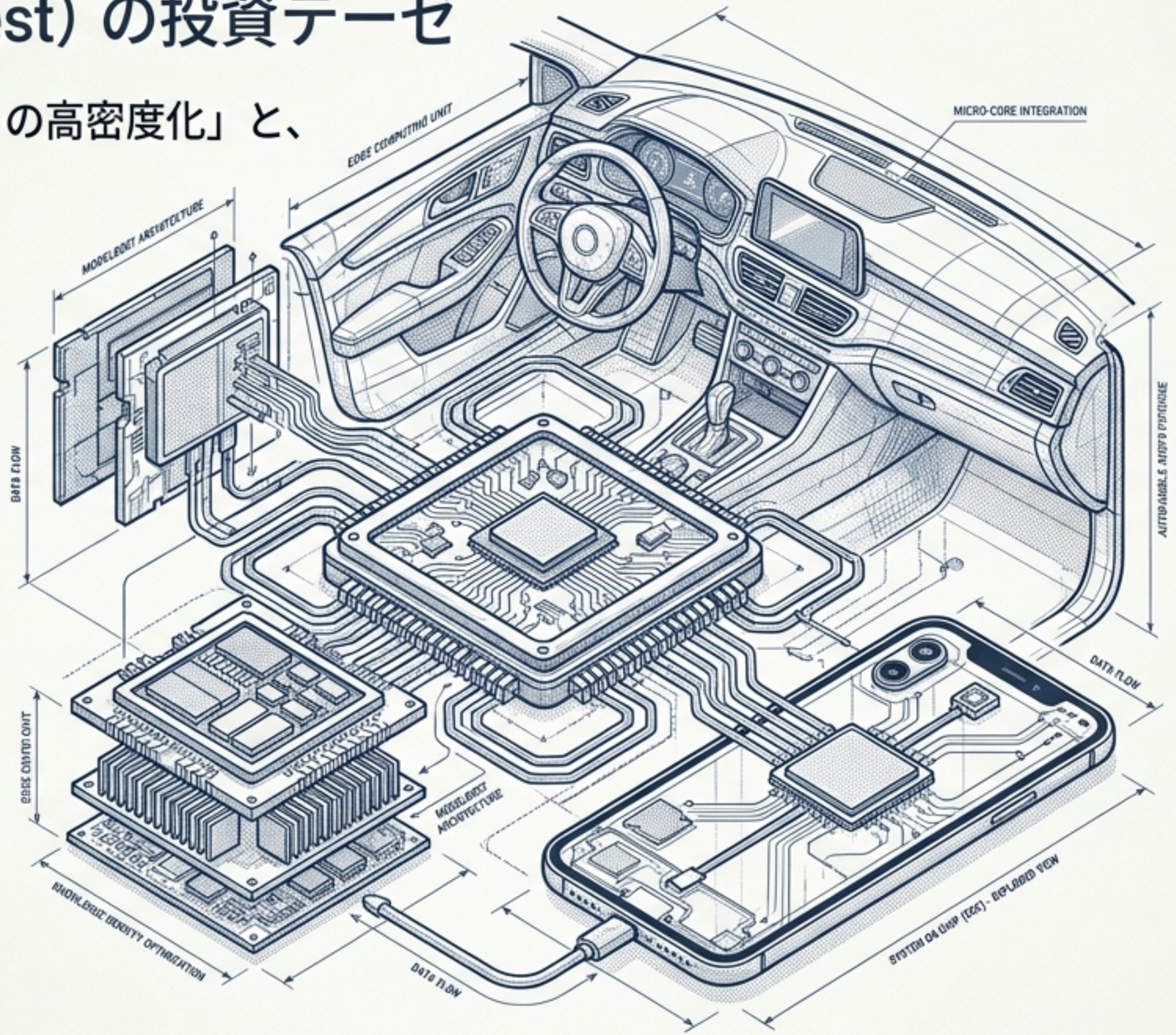
解体新書：面壁智能（ModelBest）の投資テーゼ

巨大LLM競争の裏で進行する「端側（エッジ）の高密度化」と、ハードウェア・エコシステムの掌握

なぜ10億元超を調達した清華大学発のユニコーンは、
パラメータ数ではなく「知識密度」を追求するのか。



The Verdict：総合評価

技術力はトップティア、収益可視性は発展途上

クラウドの巨大モデルを追わず、スマホ・車載・司法など「オフライン/閉域網」での実装力に特化。技術志向の研究組織から、B2B/B2Gのシステムベンダーへの転換期。

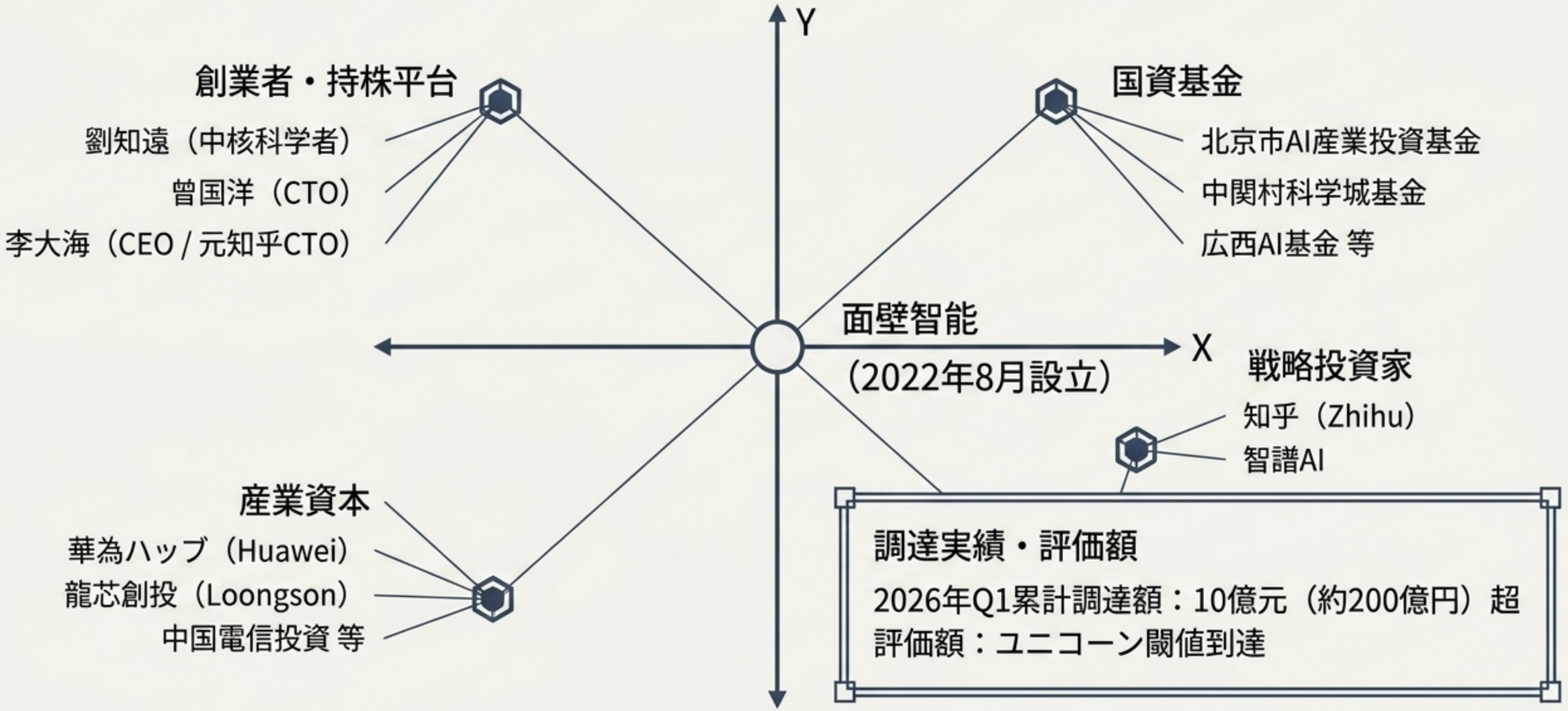
The Moat：競争優位性

- 極小パラメータでの高い「知識密度」（MiniCPMシリーズ）
- Apache-2.0によるオープンソース・エコシステムの掌握（2,400万DL超）
- 車載（吉利・長安マツダ）および司法網での先行量産・配備実績

The Risks：懸念事項

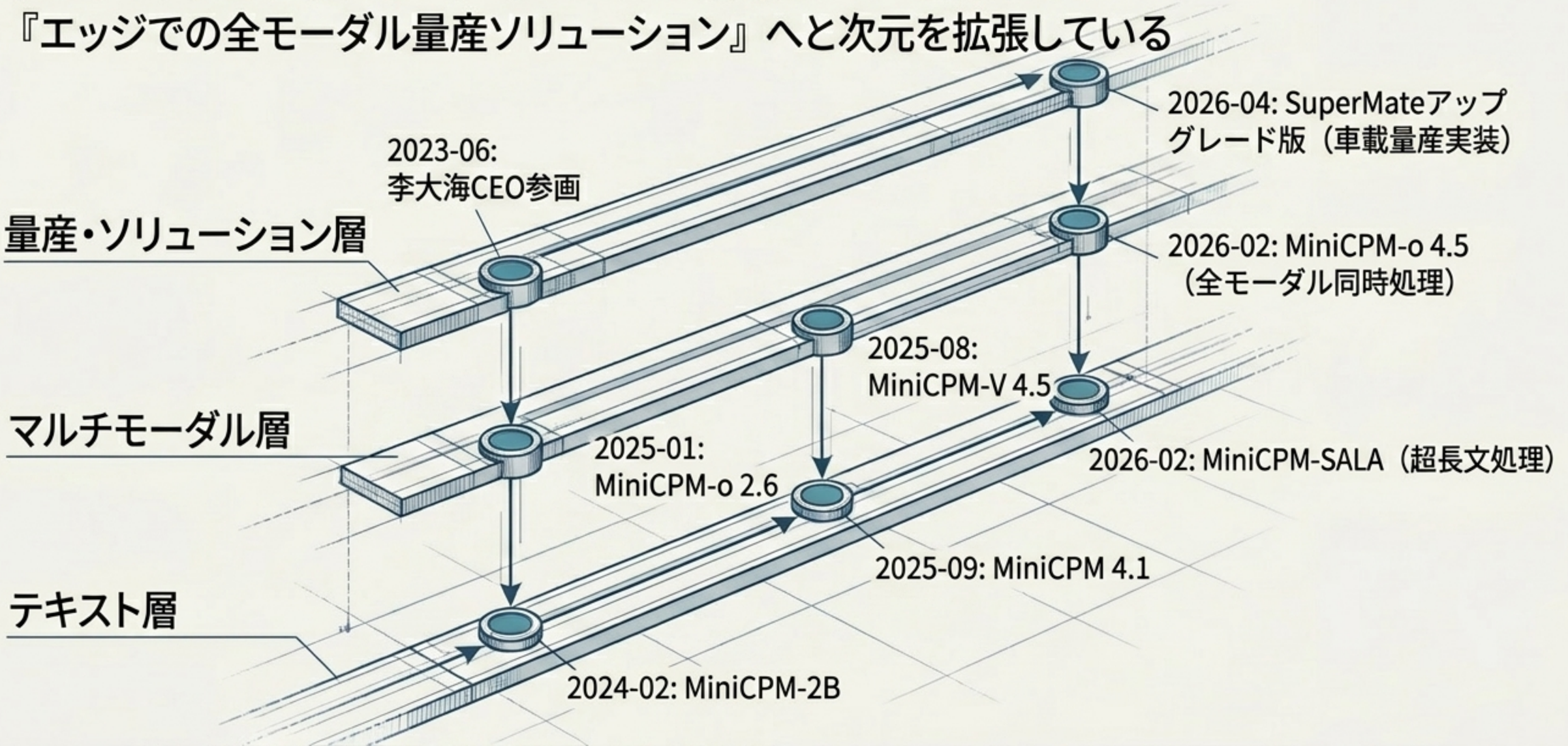
- ARR、粗利率など財務指標の不透明性
- OEM特有の長いセールス/量産サイクル
- 中国独自の備案（ファイリング）規制コストとハードウェア断片化の壁

資本・組織の陣営：国家と産業の強力なバックアップ



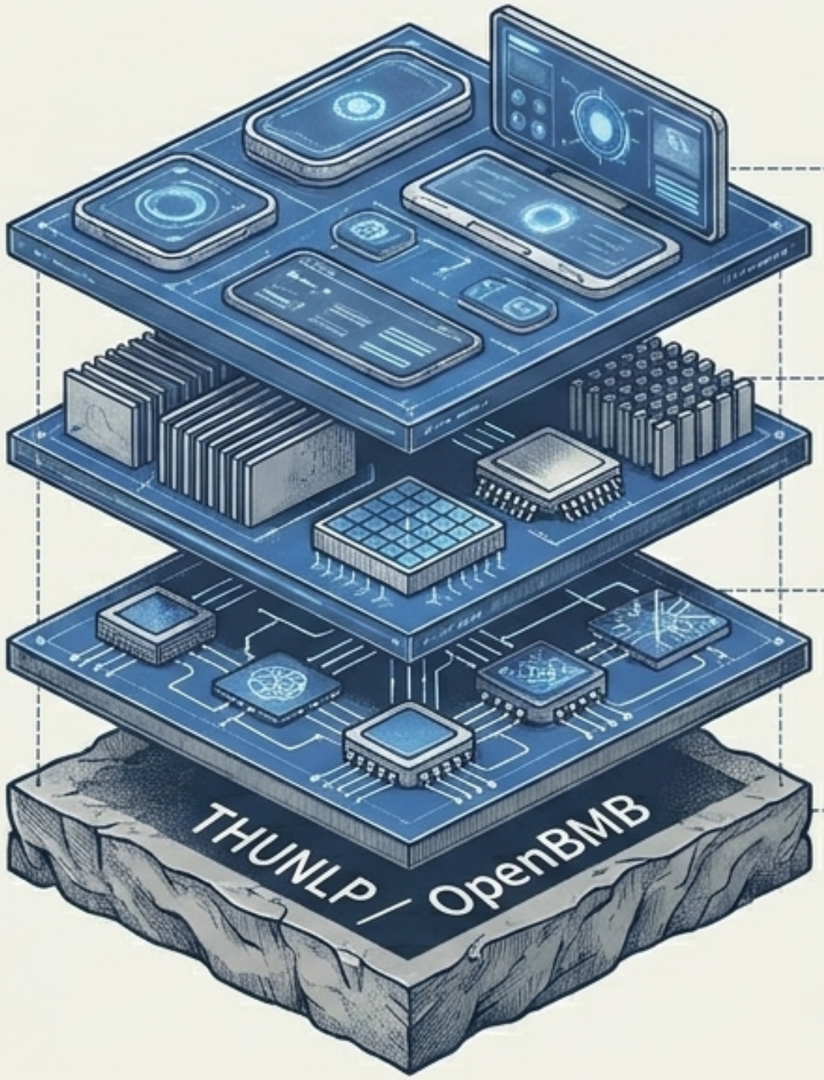
Acceleration Track：技術と実装の多次元的拡張

単なるモデルのアップデートではなく、2年足らずで『言語』から
『エッジでの全モーダル量産ソリューション』へと次元を拡張している



The Edge Architecture Blueprint

研究のオープンソース化による認知獲得と、推論・業界ソリューションでの高単価収益化の二段構え



Layer 4 : 業界別ソリューション

SuperMate（スマートコックピット）、Pinea（端側アシスタント）、
Lantay（文書エージェント）、法院専網（法律AI）

Layer 3 : 推論・軽量化基盤（Core Moat）

BitCPM（量子化）、InfLLM-V2, CPM.cu, ArkInfer

Layer 2 : 基礎モデル & モーダル拡張

MiniCPM4 (Text), MiniCPM-V (Vision), MiniCPM-o (Omni-modal)

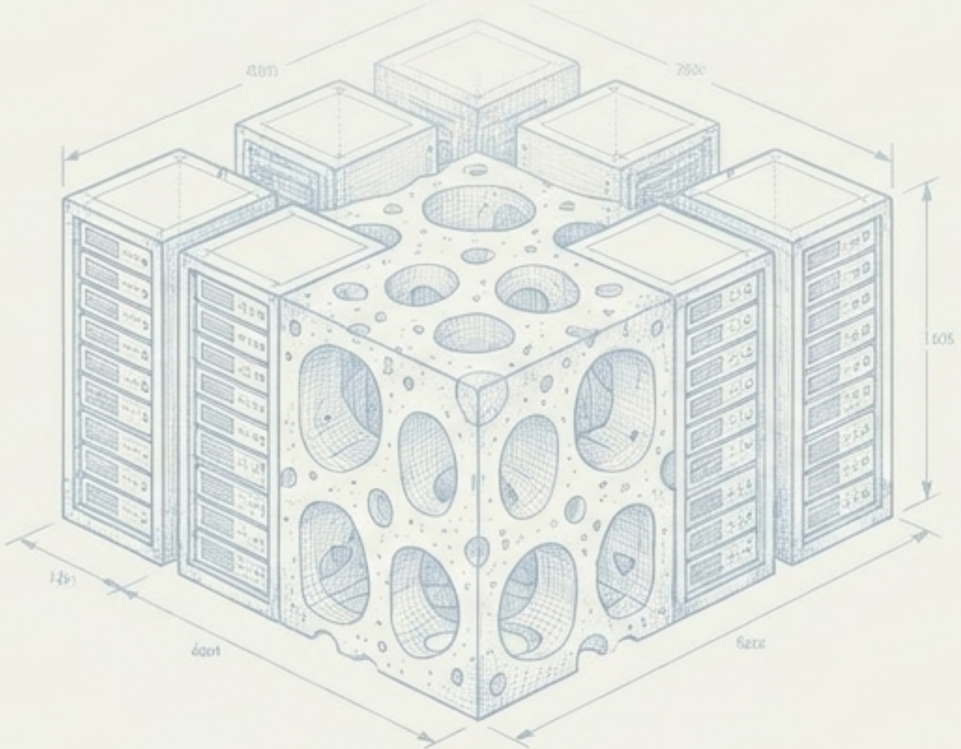
Layer 1 : 研究源流

研究源流による[illegible]LO)が獲得、[illegible]モデルを実現し、
重力し、年上市を投資する高単価収益化を備え

「密度法則」の視覚メタファー：大きさ（Scale）ではなく密度（Density）

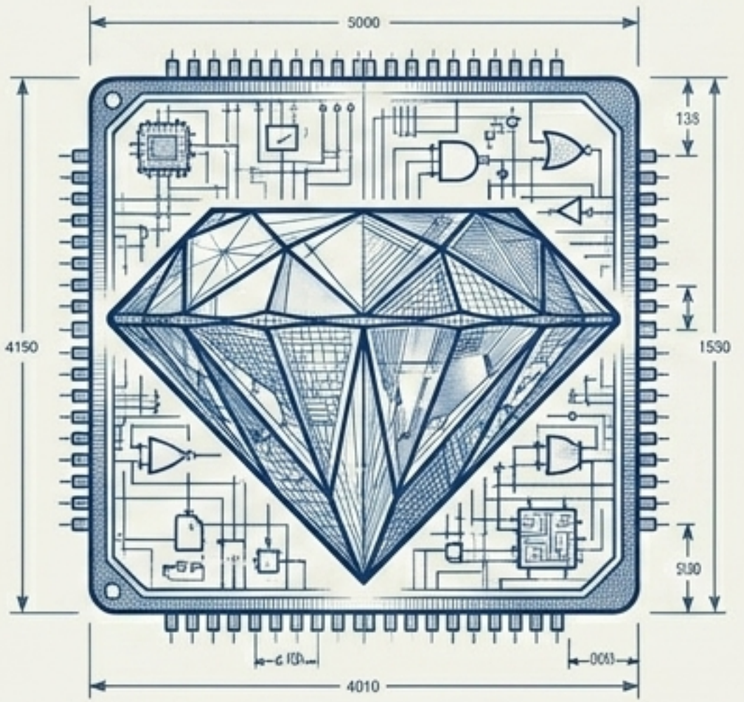
限られた算力でモデルの知識密度とタスク性能を最大化する

クラウド巨大LLM



- 巨大だが隙間の多いスポンジ
- パラメータは多いが非効率
- 高コスト・高遅延

MiniCPM（マイクロコア）



- 極限まで圧縮された高密度のダイヤモンド
- Trainable sparse attention + 量子化による極限の知識密度
- MiniCPM4.1-8B: 64K事前学習 + YaRNによる128K拡張
- MiniCPM-SALA 9B: Hybrid attention (InfLLM-V2 25% + Lightning 75%) により、コンシューマGPUで1Mトークン処理

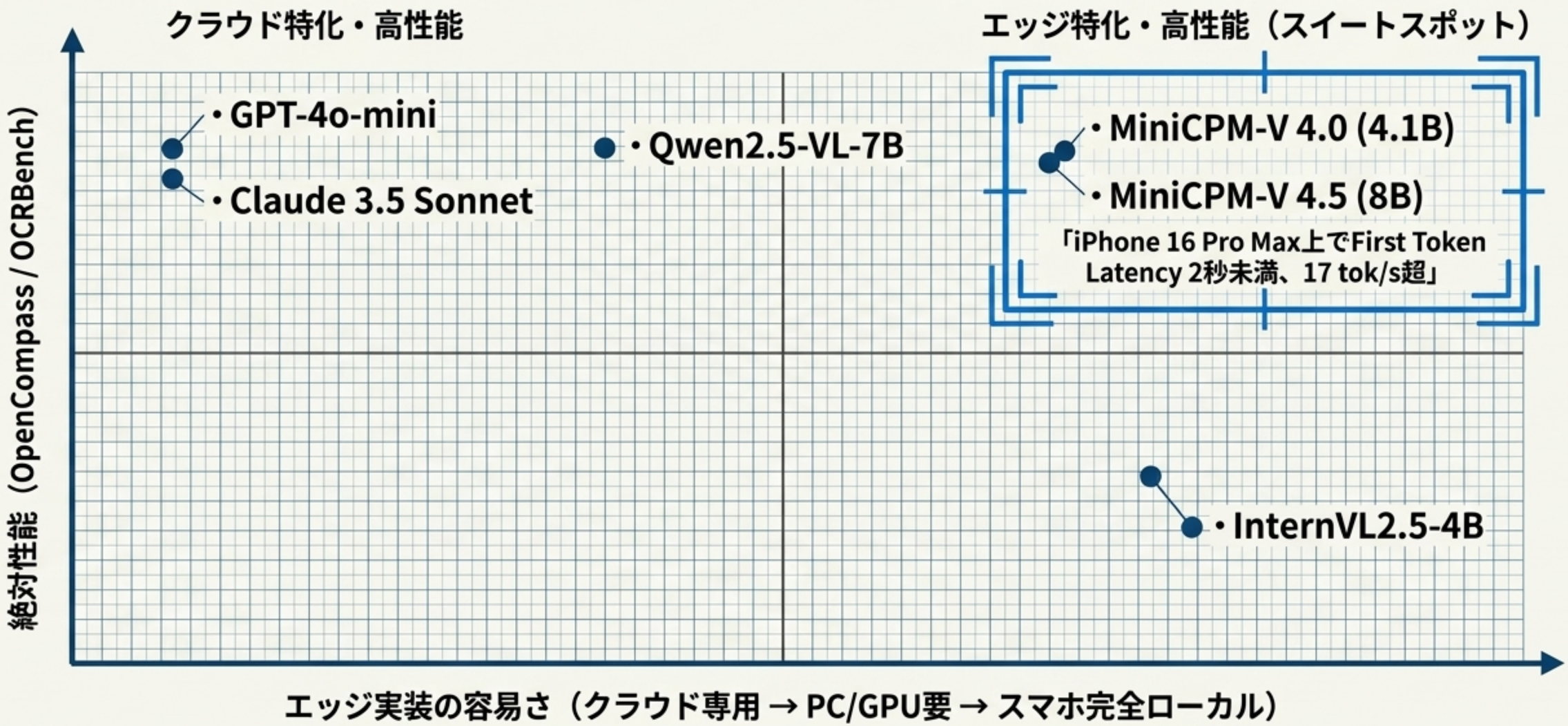
軽量テキストモデルの「下克上」：パラメータサイズと性能の反比例

量子化（Int4）しても性能劣化がほぼゼロ。パラメーターサイズが半分以下でありながら、
Llama 3.2やQwen3を圧倒する基礎推論力

モデル名	サイズ	オンデバイス適性	MMLU	CMMLU
MiniCPM4-0.5B	0.5B	非常に高い	55.55	65.22
MiniCPM4-0.5B Int4	0.5B	極めて高い	55.46	63.91
Qwen3-0.6B	0.6B	高い	42.95	42.05
Llama 3.2 1B	1B	高い	46.89	23.73
Gemma 3 1B	1B	高い	41.64	25.09

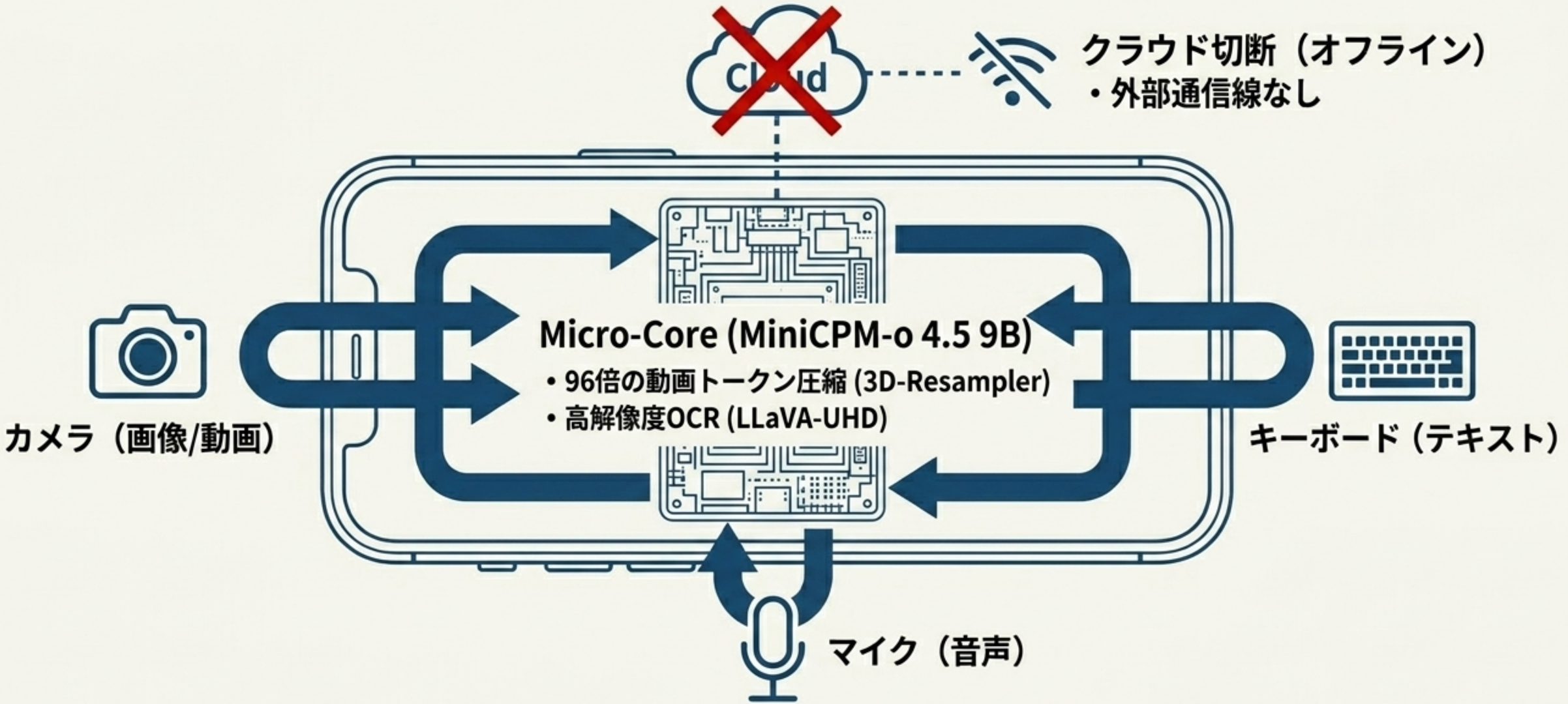
マルチモーダルモデルの実装現実性マップ

クラウド上の巨大モデルを追わず、「完全ローカルでGPT-4クラスのマルチモーダルを動かす」独自の戦場



オンデバイス「完全同時全モーダル」処理の解剖

通信遅延ゼロ、完全なプライバシー保護。閉域・オフライン環境で面壁智能が選ばれる理由

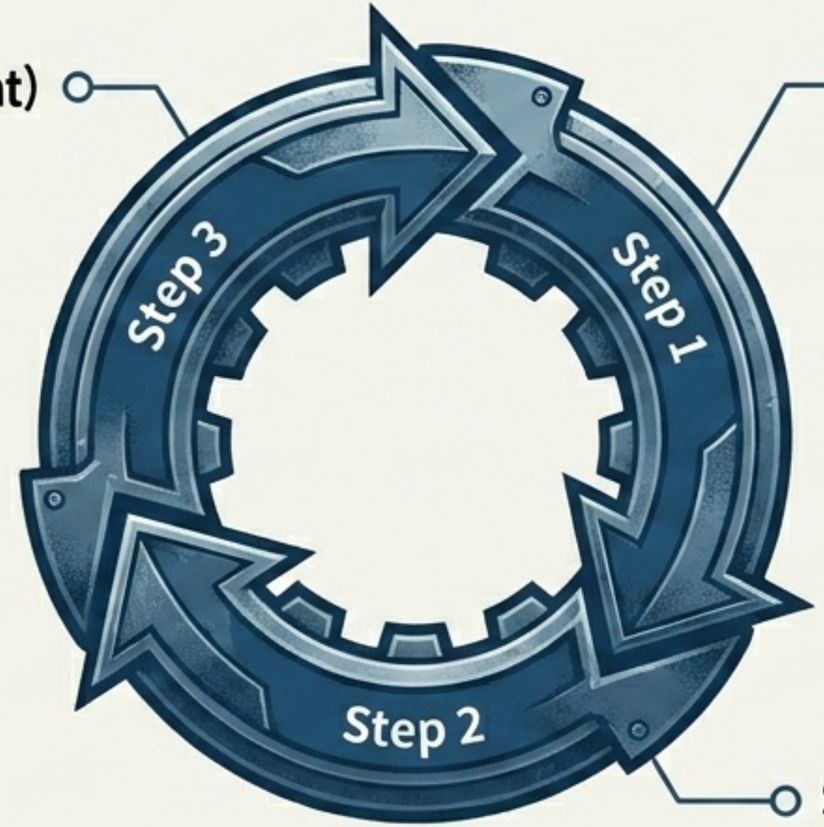


The Edge-AI Flywheel：端側AIの独自エコシステム

OSSコミュニティを巨大な『ハードウェア実証実験室』として使い、その成果を高利益のエンタープライズ量産案件へ直結させる

Step 3: 高単価なB2B/B2G量産導入（Moat）

ハード・ソフト協調設計の知見を武器に、長安マツダ、吉利汽車、法院（裁判所）専網などの「弱ネット・セキュリティリティ必須」な閉域案件を独占。



Step 1: 圧倒的なオープンソースの浸透

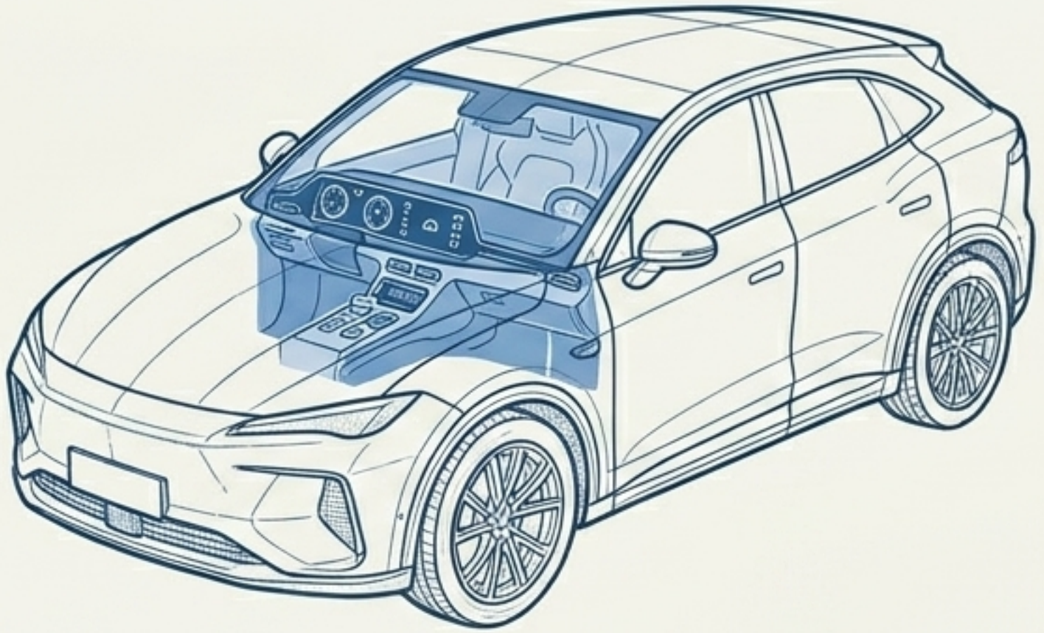
Apache-2.0による公開。Hugging Face/GitHubで累計2,400万回以上のダウンロード。

Step 2: 末端ハードウェアの最適化データ獲得

多様なエッジチップ（Apple, Snapdragon, Intel, Jetson）での推論データ蓄積と研ぎ澄まされるモデル（CPM.cu / ArkInfer）。

商業トラクションの実態：システムベンダーへの脱皮

認証と参入障壁が極めて高い『車載・行政・司法』のB2B/B2Gレイヤーへの深い浸透



【車載量産実装】

- プロダクト: 端側スマートコックピット「SuperMate」
- 実績: 長安マツダ（EZ-60）、吉利汽車（吉利銀河 M9）へすでに量産実装済み（PoCではなく実車搭載）。
- ハードウェア協業: 華為（AI Phone）、聯想（AI PC）、長城汽車、メディアテック等

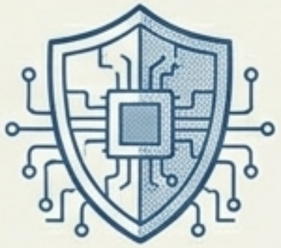
【垂直産業の閉域網展開】

- プロダクト: 法院办案流程大模型
- 実績: 最高法・深圳中院など、完全な「法院専網（閉域網）」への私有配備（全国初）。セキュリティとコンプライアンスの厳格な要件をクリア。

投資リスクとオポチュニティの診断

技術的防御力は極めて高いが、収益性と中国独自の規制環境が変数となる

オポチュニティ・防御力（Opportunities）



- 米国の半導体規制に対するヘッジ
  巨大GPUクラスターへの依存度が低く、スマホ・PC・車載の分散算力で戦える強靭なアーキテクチャ。
- 独自の巨大オフライン市場
  中国の「データ域内化」「弱ネット環境」要請に完璧に合致。

リスク・不確実性（Risks）

- 収益性のブラックボックス
  売上高、ARR、粗利率、継続率が未開示。OEM特有の長い量産サイクルによるキャッシュフローのラグ。
- 中国の規制・コンプライアンスコスト
  《深度合成管理規定》に伴う備案（ファイリング）負担。ハードウェア断片化による開発・適応負荷の増大。

The Verdict: 今後18ヶ月の再投資判断トリガー

今後のバリュエーションの跳躍は、モデル精度ではなく『ハード・ソフト協調の実装速度と有償案件のスケール』にかかっている

現在確認済みのシグナル (Clear Signals)		今後注視すべきKPIトリガー (Crucial KPIs to Watch)	
☑	[確認済] **コミュニティ需要：累計2,400万DL超**（先行需要は強力）		[注視] **車載・端末の実搭載/出荷台数**（ 実売上化の規模を測る）
☑	[確認済] **量産採用実績：**EZ-60、吉利河M9への搭載（技術の商業化証明）		[注視] **法院・企業内案件の有償継続率**（SaaS的ARR/粗利の代理指標）
☑	[確認済] **推論効率の優位性：**Jetson / iPhone等での速度改善持続		[注視] **ハード断片化への対応力**（SoC多様化による開発コスト増大の有無）