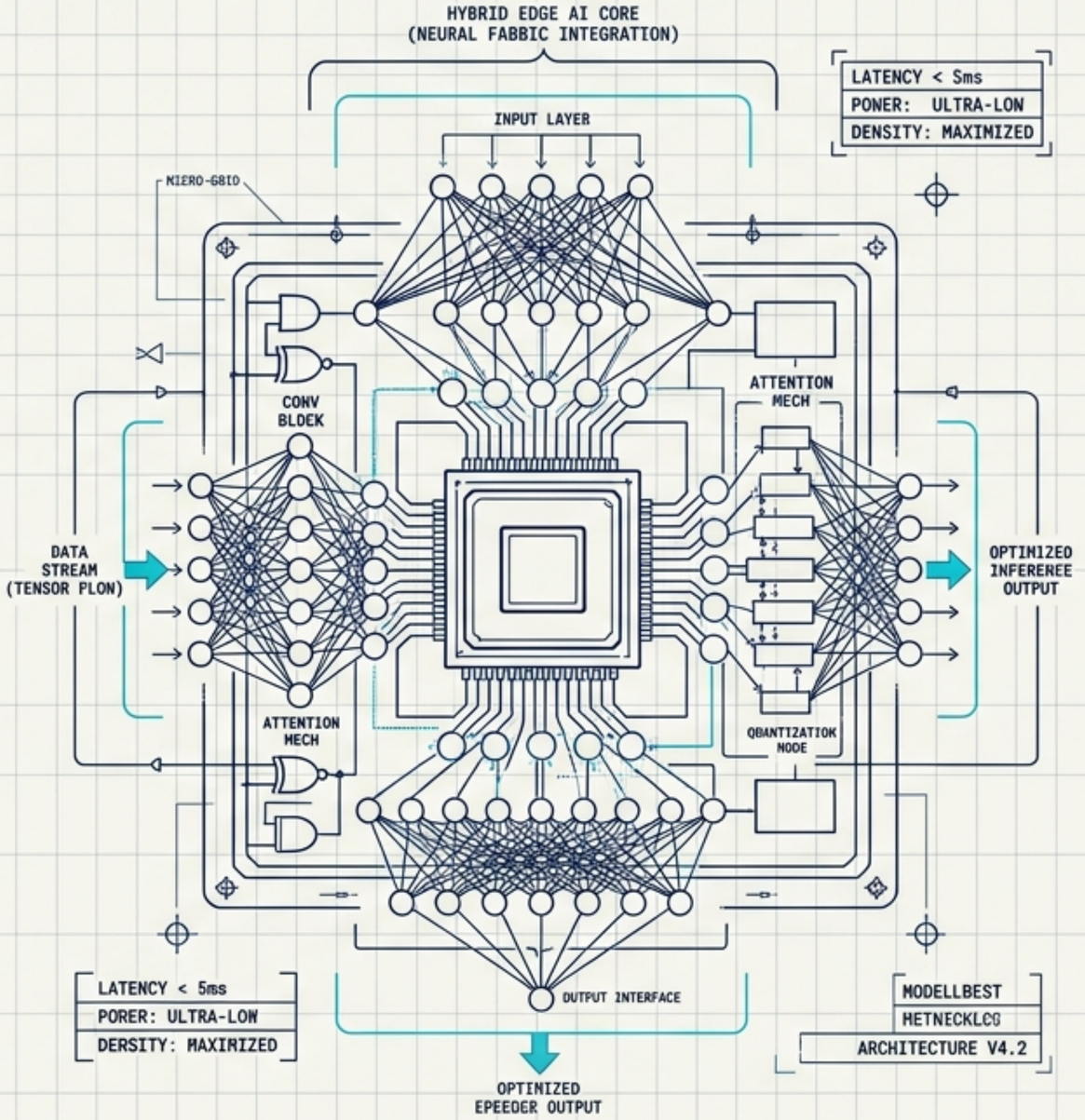
徹底解剖：面壁智能（ModelBest）の技術戦略

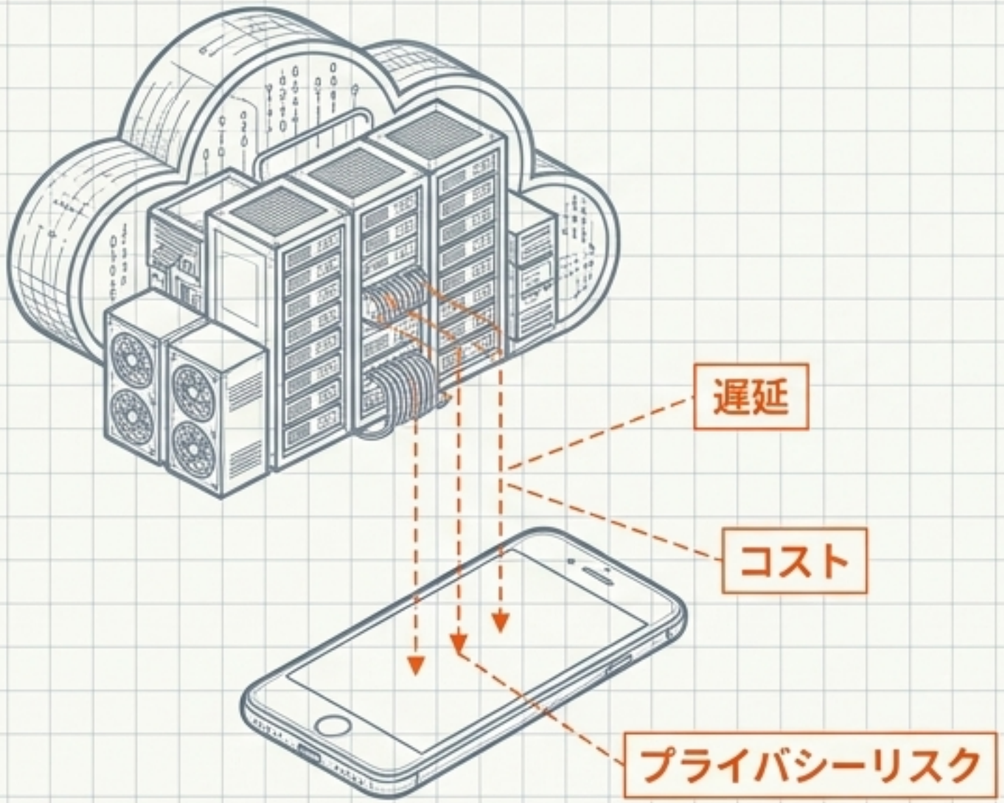
クラウド依存を打破する「エッジLLM」の台頭とパラダイムシフト

パラメータ競争から「知識密度」の極限へ。
設立数年でユニコーンへと変貌した清華大学発スタートアップの全貌。



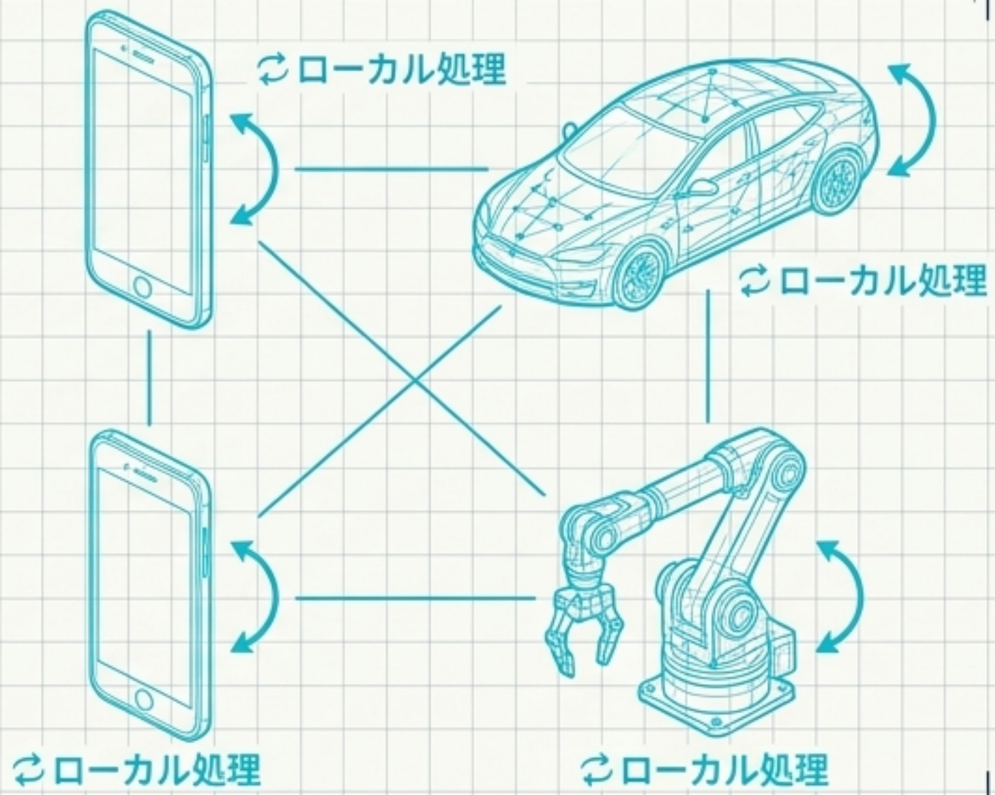
巨大クラウドLLMの構造的限界とエッジAIの到来

クラウドLLMの限界



パラメータ規模の指数関数的拡大に依存する従来型LLMは、ネットワーク遅延、膨大な運用コスト、機密データ流出リスクという物理的・経済的限界に直面。

エッジAI（SLM）の到来



デバイス上で独立稼働する小規模・高性能な「エッジAI（SLM）」。個人データ流出リスクを根本的に排除し、通信遅延のないリアルタイム応答を実現する次世代の主戦場。

面壁智能（ModelBest）：清華大学発のユニコーンと資本戦略

清華大学NLPラボ発
（2022年8月設立）

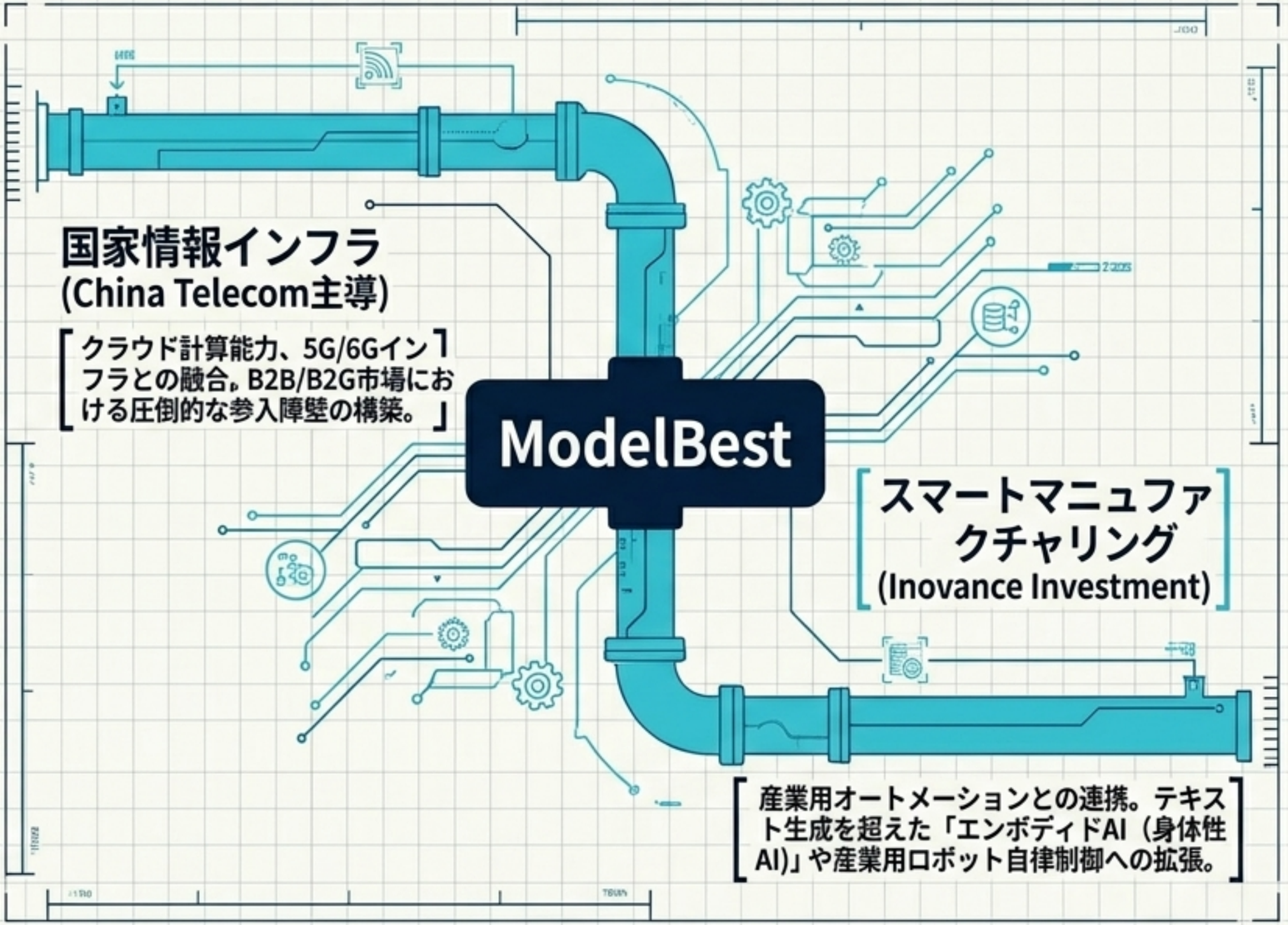
THU-NLP Lab Origin (Aug 2022 Est.)

累計資金調達額
10億元（約230億円）超

Cumulative Funding > 1B CNY
(Approx. 23B JPY)

累計2400万DL突破 /
評価額ユニコーン到達

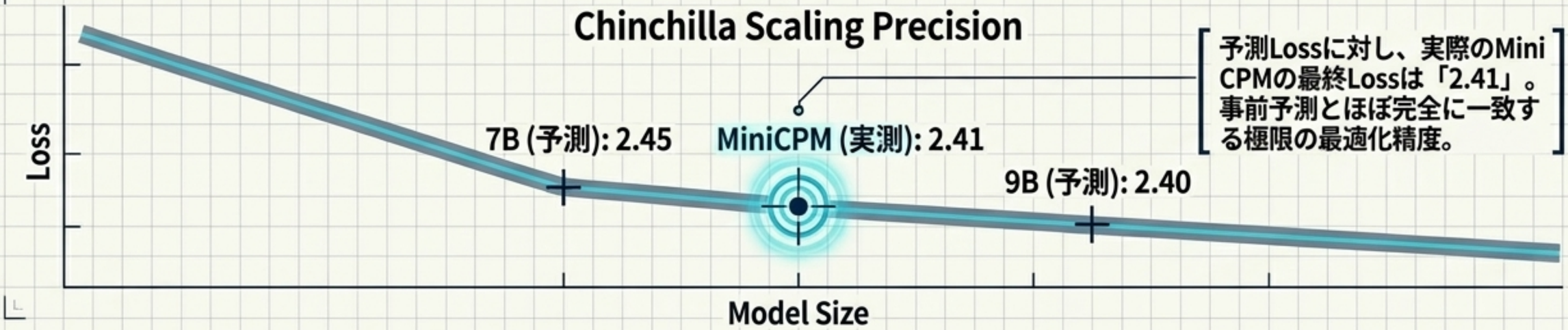
24M+ Downloads /
Unicorn Valuation Achieved



核心設計思想：「知識密度法則（Density Principle）」

無闇な巨大化を避け、限られたパラメータ空間に知識と推論能力を高密度に圧縮する。

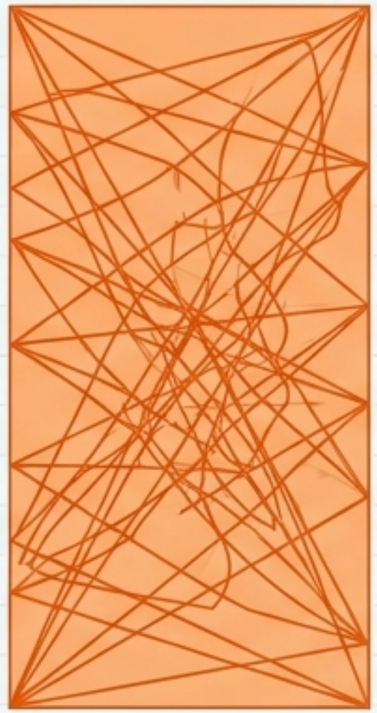
モデル風洞実験：
航空宇宙工学に着想を得たアプローチ。0.04Bの小規模テストで最適学習率（0.01）を特定し、2.1Bまで一貫して適用。計算資源の浪費をゼロに。



予測Lossに対し、実際のMiniCPMの最終Lossは「2.41」。事前予測とほぼ完全に一致する極限の最適化精度。

メモリ限界を突破するハイブリッド構造「MiniCPM-SALA」

OOM / メモリ不足
(Quadratic Explosion)



Full-Attention

MiniCPM

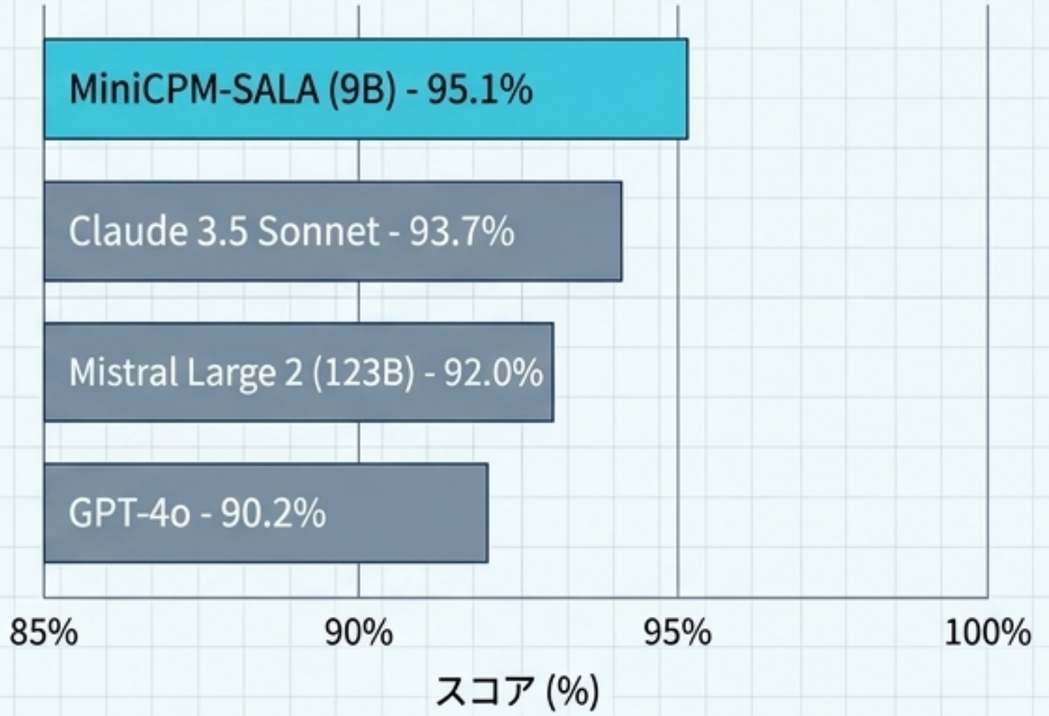
25%
Sparse Attention
(局所的詳細情報)

75%
Linear Attention
(広範なコンテキスト)

MiniCPM

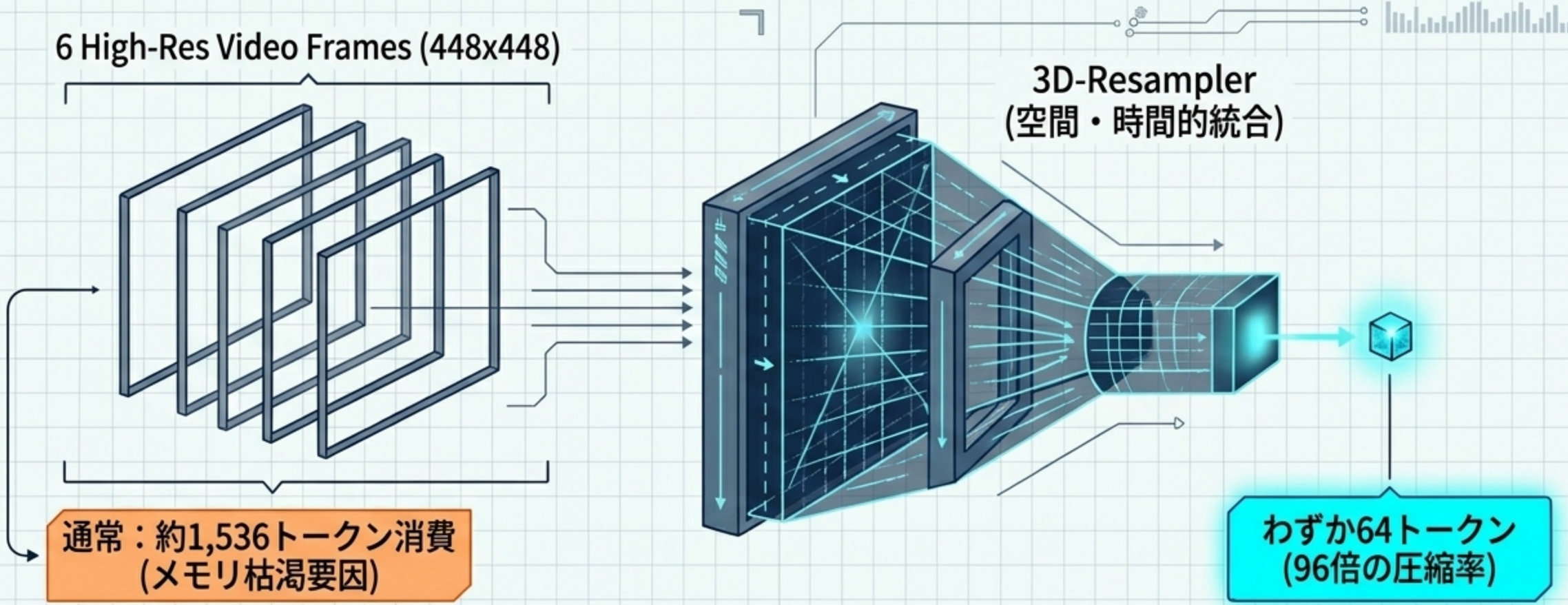
単一GPUで「100万（1M）トークン」処理を実現

HumanEvalベンチマークにおけるMiniCPM-SALAの性能比較



コーディング/推論タスクにおいて、パラメータ数が数十倍の世界最先端プロプライエタリモデルを凌駕。

エッジでのリアルタイム視覚処理：「3D-Resampler」による劇的圧縮



10FPS リアルタイム処理

計算量を増やさずに高フレームレートを処理

エッジキラーアプリ

車内外の監視カメラやロボットの視覚フィードバックをローカル単独で完結

エッジLLM 競合ベンチマーク比較（vs Meta, Microsoft）

指標	MiniCPM-SALA (9B)	Llama-3.1-8B	Phi-3.5-mini
最大コンテキスト長	1,000,000 (1M)	131,072 (128K)	128,000 (128K)
HumanEval (コード)	95.1%	72.6%	62.8%
APIコスト (1Mあたり)	$0 (ローカル運用主体)	$0.03	$0.10

Strategic Insights:

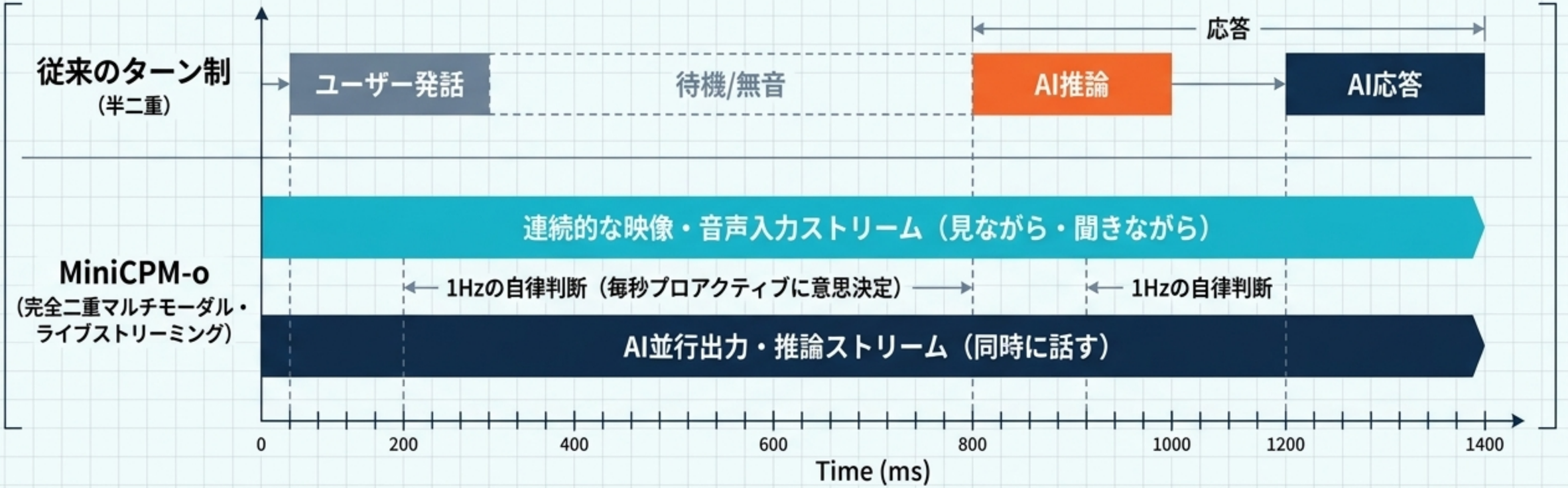
限界費用ゼロの経済性

オンデバイス推論によりAPIコストが完全ゼロ。持続的推論が必要なロボティクスにおいて圧倒的優位。

メモリ帯域幅の最適化

エッジ特有のメモリ帯域ボトルネックを、SALAおよび3D-Resamplerのアルゴリズム設計で完全に回避。

業界初 エッジ向け「全二重・全モーダル通信」（MiniCPM-o）



見ながら・聞きながら・同時に話す

入力と出力をミリ秒単位で同期し、互いをブロックせずに同時並行処理。

自発的プロアクティブ対話

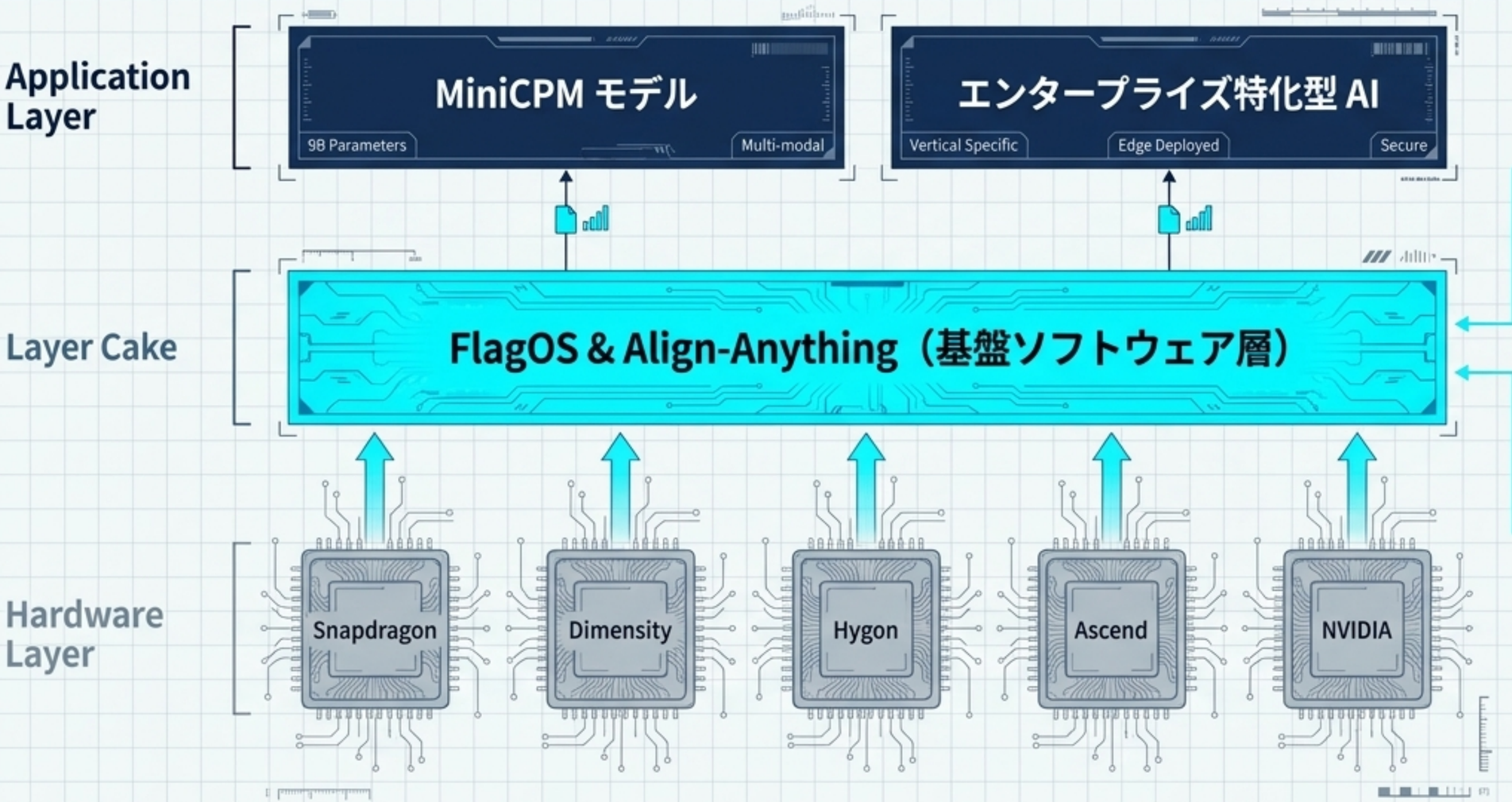
居眠り運転の検知・自発的警告や、調理中の能動的アドバイスをエッジで実現。

究極のプライバシー保護

llama.cpp-omniを活用しiPhoneでネイティブ動作。生体データや映像を外部に送信しない。

ハードウェア断片化を無効化する「基盤ソフトウェア抽象化」

米国輸出規制によるAIチップの乱立を克服する Develop once, migrate across chips 戦略。

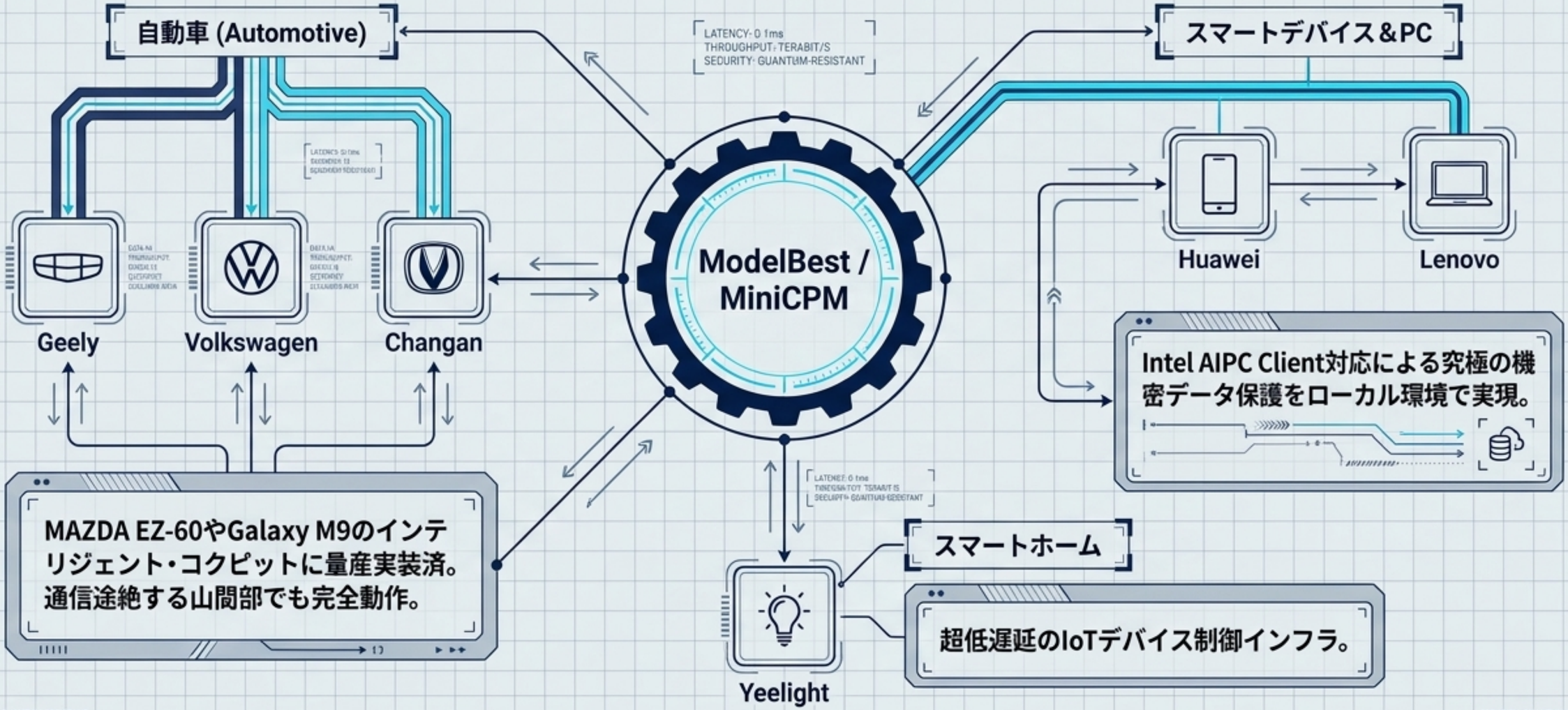


モバイルSoC最適化:
- SnapdragonとDimensityの熱制約（サーマルスロットリング）を深く理解したローカル最適化。

FlagOS:
- NVIDIAだけでなく、海光(Hygon)や昇騰(Ascend)など6種類の国産AIチップに標準対応し、企業をハードウェア制約から解放。

実体経済への急速な社会実装ネットワーク

アルゴリズムの優位性を机上の空論で終わらせず、強固な産業エコシステムを構築。



Apache 2.0ライセンスによる「デファクトスタンダード」戦略

A プロプライエタリAPI

⚠ 突然の価格改定

⚠ 規約変更による機能停止

依存企業はシステム移行を
余儀なくされる致命的リスク。

B MiniCPM (Apache 2.0)

✓ 改変後の公開義務なし

✓ 法的・経済的リスクゼロ

企業自身のサーバー/
デバイスで完全ホスト

「独自のクローズドな
商用製品」として開発可能

累計2400万DLによる
市場制圧の原動力

結論：次世代IoT空間を牽引する絶対的キープレイヤー

極限の知識密度 1
(Algorithm)

物理的限界を克服し、巨大モデルを凌駕。

2 技術の民主化
(Open Source)

Apache 2.0とFlagOSによる依存からの解放。

3 垂直統合エコシステム
(Infrastructure)

自動車・ロボティクスへの圧倒的スピードでの量産実装。

面壁智能は単なるLLM開発企業ではない。巨大モデルのスケーリング競争から、リアルタイムの「エッジ・エンボディドAI」へと不可逆的にシフトする次世代AIパラダイムにおいて、新たな標準インフラを再定義するアーキテクトである。