



主権AIの具現化

- 2026年10月3日（独統一の日）公開
- 法的・物理的な完全統制



エンタープライズ最適化

- 78.1B総容量 / 3.46B稼働の極小MoE
- ハルシネーションを抑止する新アーキテクチャ



地政学的再編と波及

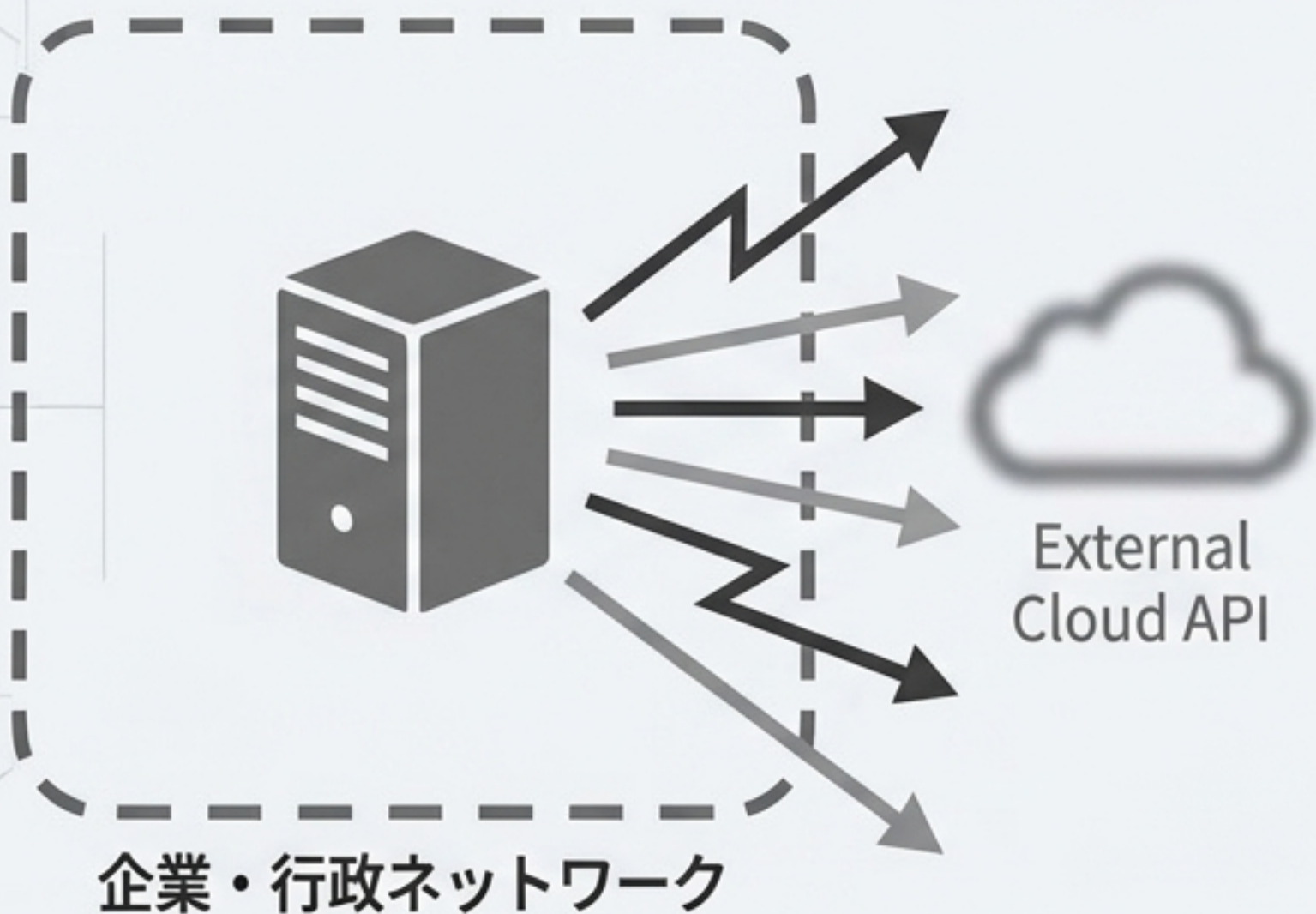
- Cohere社との事業統合
- 富士通「Takane」への技術還流

「政治的スローガン」であった主権AIが、厳格な「工学的要件」として実装された分岐点。

ブラックボックス化された依存

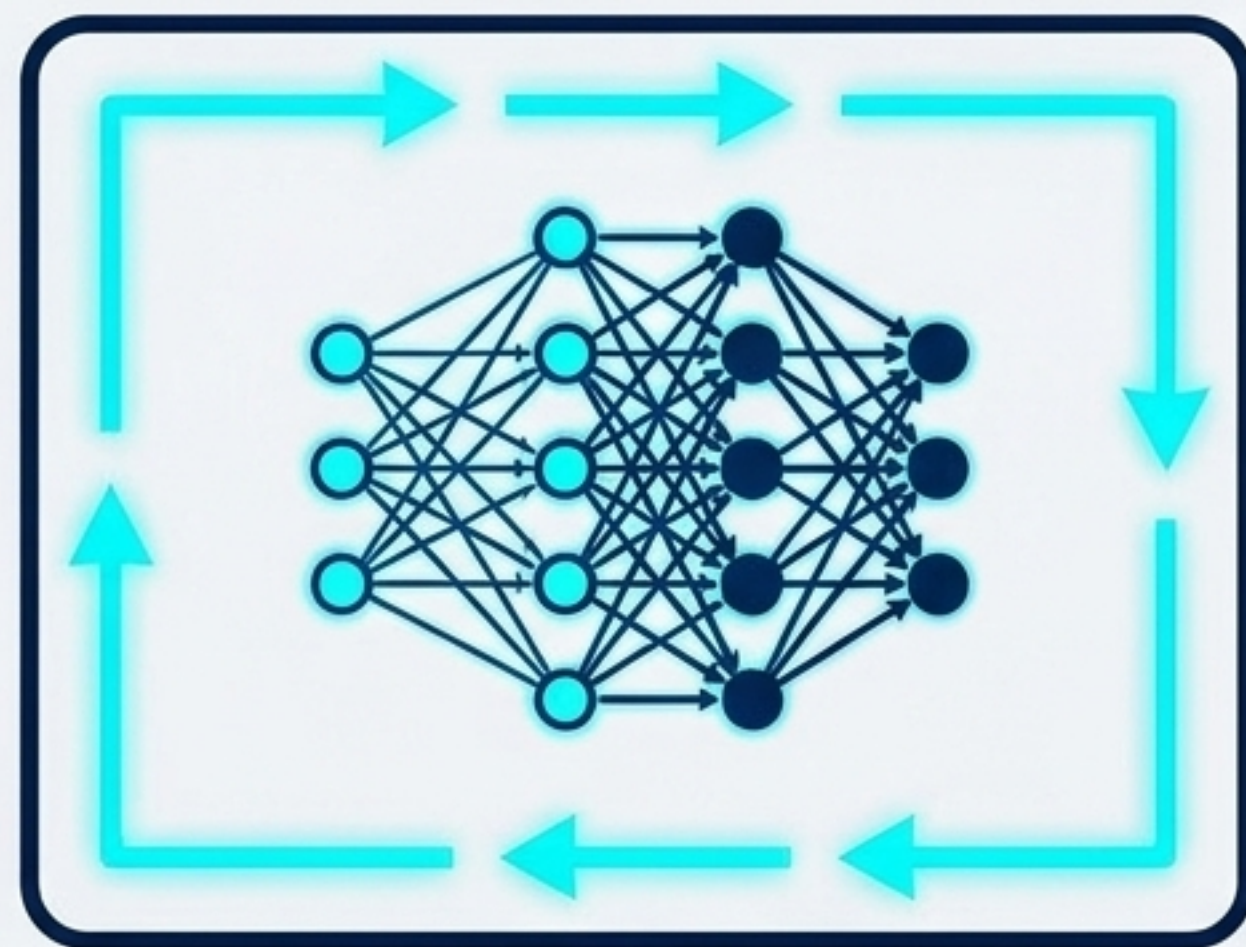
ターゲット領域:
行政、防衛、重工業、医療

ブラックボックス化された依存



巨大テック企業のクラウドAPIへの依存。厳格な規
領域においては「制御なき性能」は負債となる。

完全なエアギャップ統制 (Kolibri)



データ抽出から推論まで、すべてのプロセスを
ファイアウォール内部（オンプレミス）で完結。

主権AI (Kolibri)

完全な域内完結

データ・インフラの独立

- 欧州およびドイツ法の完全適用
- 768基のNVIDIA B200でスクラッチ学習
- 外部モデルからの蒸留なし

完全な運用管理権

コントロールの移譲

- Apache 2.0 ライセンス
- 利用規約・MAUの制限なし
- 完全オフラインのエアギャップ運用保証

監査可能性

規制適合と透明性

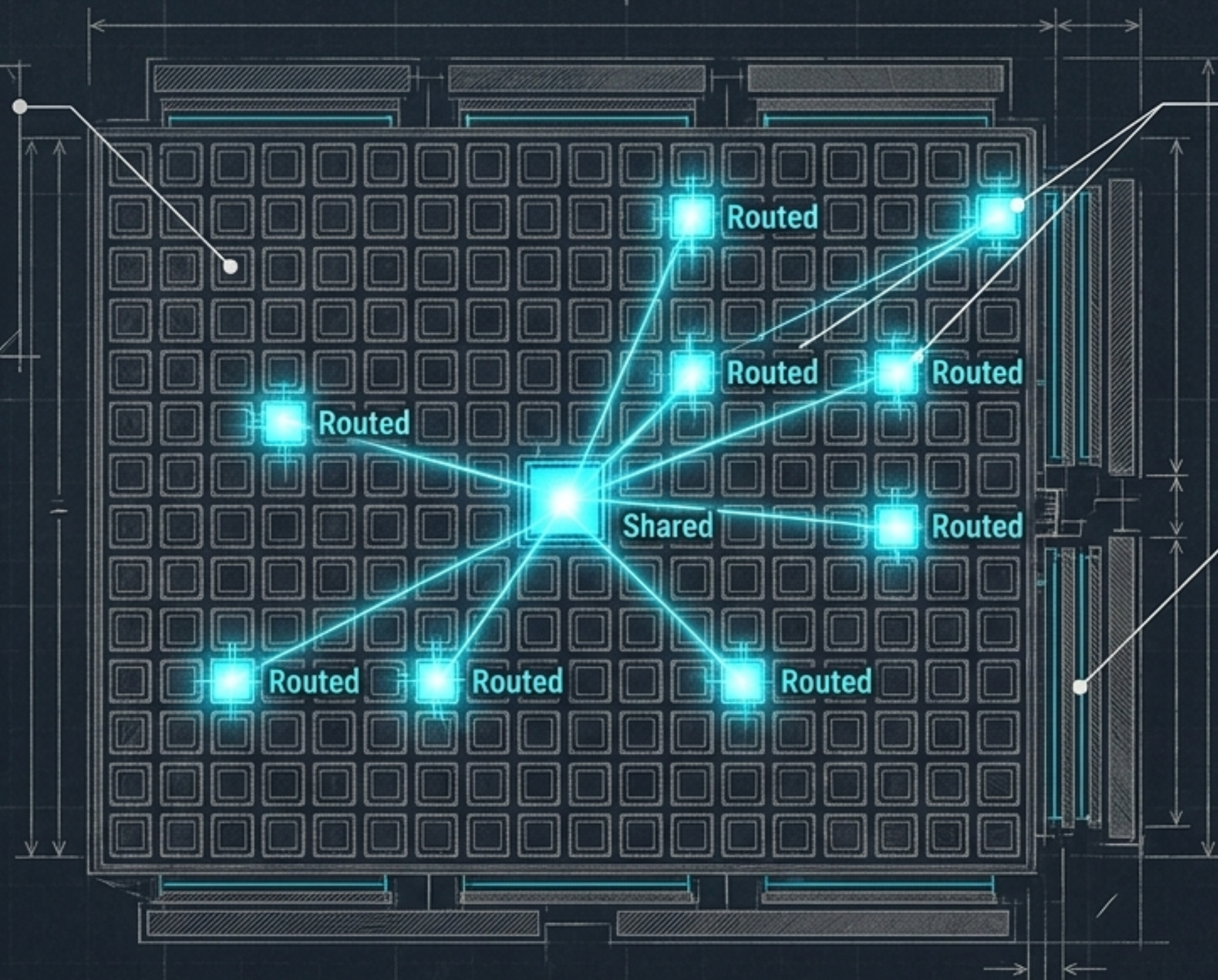
- 欧州AI法 (GPAI) 行動規範準拠
- 20兆トークンの学習データ内訳を完全開示

主権とは法的制約ではなく、システム設計そのものである。

総パラメータ数

78.1B

フロンティアモデルクラスの
推論能力と表現力



アクティブパラメータ数

3.46B (全体のわずか4.4%)

小型モデル並みのデコードス
ループットと低コスト

アーキテクチャ仕様

全50層 / FP8精度による
動的量子化

処理能力の絶対値を競う「力技」からの脱却。
78Bの推論力と3Bの実行コストを両立する極小エキスパートMoE。

形態素構造に適合する独自トークナイザ「UniBPE」

従来のBPE
(トークン消費大)

Bun des da ten schutz ge setz

UniBPE
(意味単位を保持)

Bundes

datenschutz

gesetz

驚異的な圧縮率:
4.90 bytes/token
(語彙数128,000)

周期的ハイブリッドアテンション

SWA: 512
(スライディングウィンドウ)

SWA: 512
(スライディングウィンドウ)

SWA: 512
(スライディングウィンドウ)

SWA: 512
(スライディングウィンドウ)

Full Attention
(全系列俯瞰)

← 4:1 ループ (全50層) →

時間 →

まとめ

独自の形態素圧縮とハイブリッドアテンションにより、
VRAM肥大化を防ぎながら262,144トークンのネイティブコンテキスト長を達成。

Merlin-Arthur プロトコル (ハルシネーション抑止機構)

入力: ユーザー質問 + 文脈

Merlin (証明者):
正解を導く完全な
文脈を提示

Morgana (対抗者):
根拠となる情報を
意図的に除去

Arthur (モデル本体):
両者の状態を比較・評価

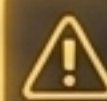
事実想起評価
(AA-Omniscience)

- 棄権成功率が旧モデルの15%から44%へ劇的向上
- RAG (検索拡張生成) における圧倒的な接地性

回答を生成

客観的証拠が存在する

証拠が存在しない



明示的な回答棄権
(Abstain)

性能プロファイル (vs Qwen3.6 35B)

タスク	Kolibri	Qwen3.6 35B
高度科学推論 (GPQA 英)	84.3	83.4
コード生成 (LiveCodeBench)	85.9	N/A
多段論理推論 (ドイツ語・英語)	トップクラス	劣後
世界知識を問う想起 (MMLU等)	限定的	トップクラス

汎用的な「トリビアボット」ではなく、
高度な推論とデータ抽出に特化した
実務特化型エンジン。

実務運用の物理的制約



Schwarz Group

(巨大流通・ITコングロマリット)

€500M
(約800億円)
の統合出資

最大10万基の
AIアクセラレータ
(STACKIT)

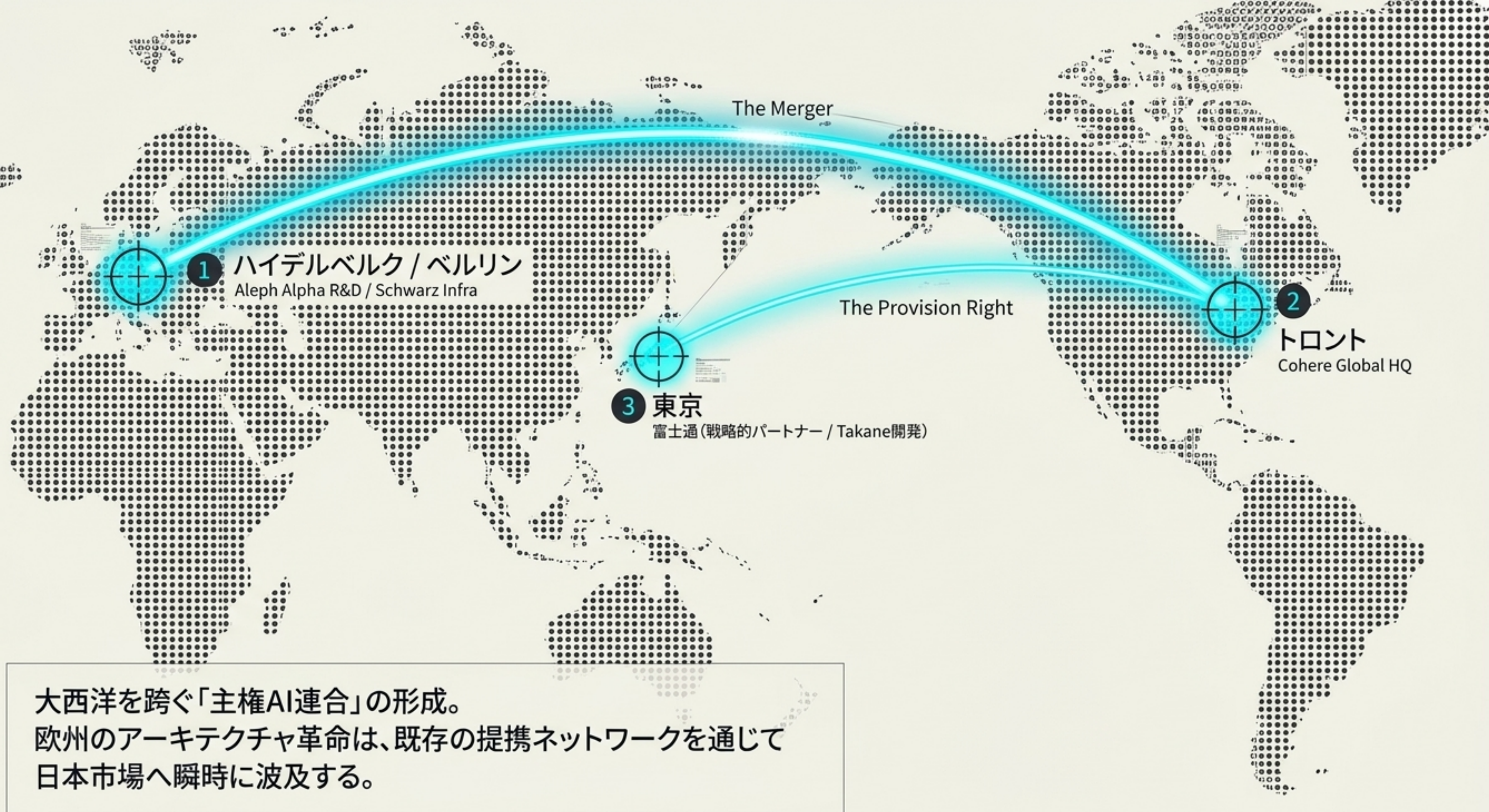
Aleph Alpha
(ハイデルベルク: 中核研究拠点)

Cohere
(トロント: グローバル統合本社)

2026年9月 対等事業統合 (大西洋主権AI連合)

ビジネス戦略の転換

- Kolibriのオープンウェイト公開は、Schwarzの巨大ソブリンククラウド (プライベートホスティング) へエンタープライズ顧客を直接誘導するための戦略的布石。単独スタートアップのインフラ投資限界を突破。



Kolibri Innovations

- 極小エキスパート MoE
- UniBPE 形態素圧縮
- Merlin-Arthur 抑止機構

Cohere Platform

- グローバルな統合開発体制
- エンタープライズ基盤技術

富士通「Takane (高嶺)」

- 日本語特化型LLM
- グローバル独占提供権

日本エンタープライズ市場へのインパクト

欧州最高峰のハルシネーション抑制とコンテキスト圧縮技術が、
日本の重要インフラ企業向けソリューションに直接還流。
米系ハイパースケーラーのAPIに対抗し、高度な説明責任を果たす強力な武器となる。

従来の国内LLM開発



密結合 (Dense) モデル

数百億～千億パラメータの全稼働による、
恒常的なGPU枯渇と莫大な消費電力のボトルネック。

Kolibriが示す設計指針



極小エキスパートMoE + UniBPE

トークンあたり3.46Bのみを駆動。
限られた国内の計算インフラで
最大のコストパフォーマンスを引き出す現実解。

経済産業省・NEDOの「GENIAC」等に対する技術的参照点。
膠着語である日本語の語彙爆発を防ぎ、推論効率を劇的に改善する青写真。

日本の政策目標 (Policy Goal)

Kolibriの解決策 (Engineering Solution)

「開かれたAI主権」と有事の自律的代替

完全なオープンウェイト形式による
オンプレミスでのエアギャップ (閉域網) 運用保証

デジタル庁「源内」等の厳格な公共調達基準

EU AI法 (GPAI) に準拠した、20兆トークンの
学習データ追跡文書の完全開示と透明性

セキュリティと相互運用性の両立

学習パイプラインの完全自動化 (Model Factory)
による監査と更新のサイロ化防止

Kolibriは、防衛、金融、重要インフラにおける日本独自の
AIガバナンス策定に向けた「世界標準のモデルケース」として機能する。

パラダイムシフト

閉鎖的な米国APIモデルへ
単一依存する時代は終焉した。

日本への戦略的要請

大西洋同盟の技術的果実を活用し、
自国のデータと計算基盤の主権を
確保する次世代AIエコシステム
を早急に構築せよ。

アーキテクチャの勝利

主権は法規制の宣言ではなく、
MoEやUniBPEといった高度な
工学設計によって初めて実装される。

デーブ テチ・ナント