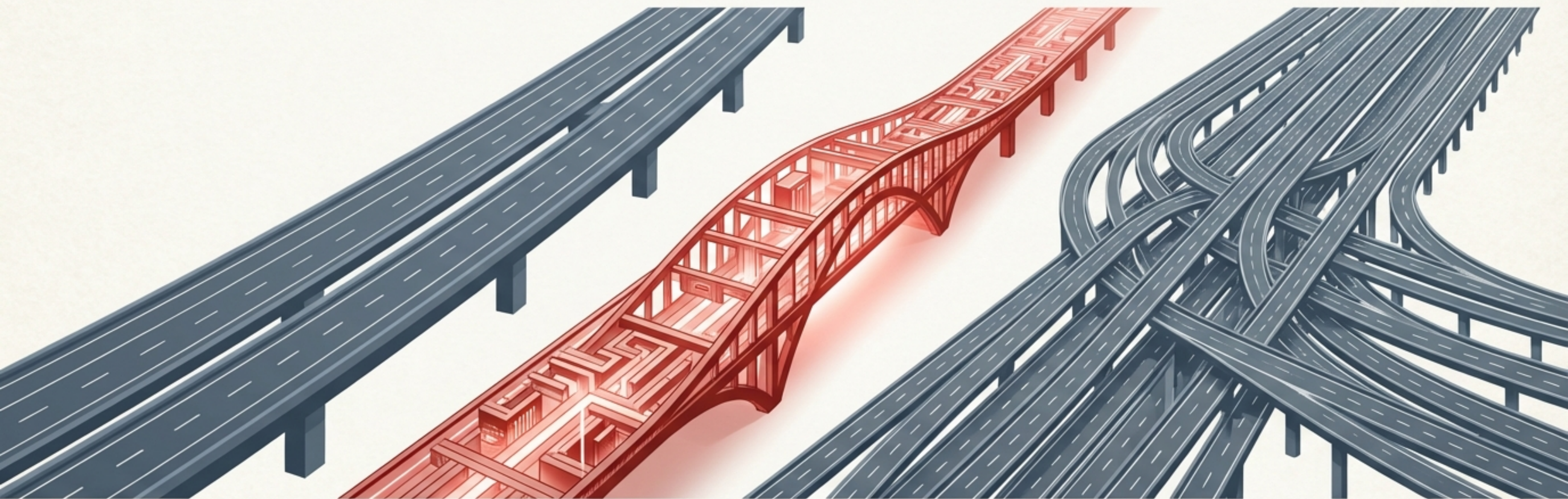


2025年ソブリンAI戦略：日本の「**第三の道**」が 拓く、**高信頼・高密度・高自律**の未来

技術的到達点とグローバル競争における戦略的ポジショニング



2025年12月

なぜ今「ソブリンAI」か？： 経済安全保障から「認知主権」の確保へ

2025年、読売新聞社とNTTの共同提言は、AI戦略の焦点を物理インフラから、国民の意思決定を支える基盤モデルそのものへとシフトさせた。海外製ブラックボックスへの依存は、国家の自律性に対する潜在的リスクである。

ソブリンAIの技術的要件（2025年時点）

フルスタックの統制：学習データ収集、アーキテクチャ設計、学習プロセス、推論インフラまで、国家の司法・技術的統制下にあること。

目的：地政学的変動や海外プラットフォーマーの方針変更に左右されず、AIを持続的に運用・改善できる能力（Capability）の確保。



世界のAI開発は二極化：米国の「スケール」と中国の「コスト」が覇権を争う

「汎用**超巨大モデル**
(AGI **via Scale**)」



「国家主導コスト破壊モデル
(Dominance via Cost)」



日本の選択肢は？

戦略：スケーリング則の追求 (Pursuit of Scaling Law)

代表：OpenAI GPT-5

哲学：圧倒的計算量による汎用性の実現

戦略：徹底的な推論コスト削減 (Relentless Inference Cost Reduction)

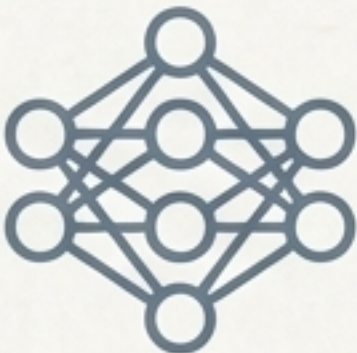
代表：DeepSeek V3.2

哲学：高性能AIを安価に提供し、応用層で圧倒

米国戦略：GPT-5が示す「力技」による汎用人工知能（AGI）への道

OpenAI GPT-5（2025年8月リリース）

アーキテクチャ



数兆パラメータ規模の
MoE (Mixture of
Experts) と推測

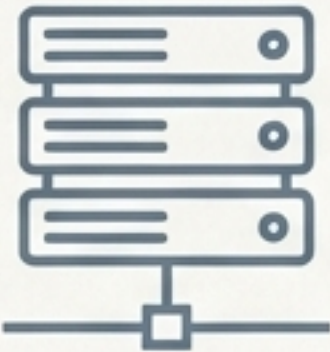
主要機能



コンテキスト長:
40万トークン

新技術: 適応型推論
(Adaptive Reasoning)
ー タスク難易度に応じて
思考時間を動的に調
整

インフラ

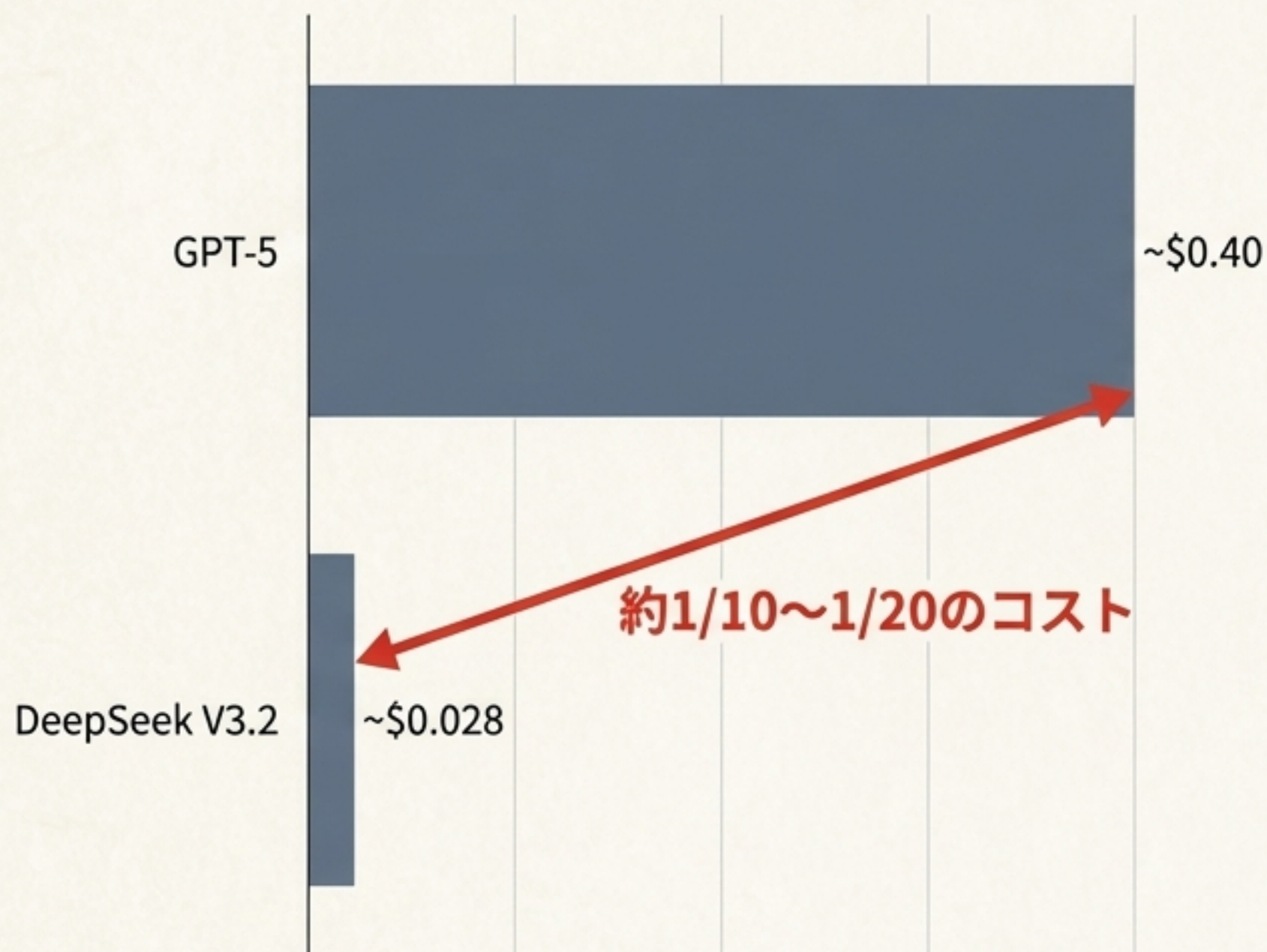


NVIDIA
H200/Blackwell世代
のGPUクラスタをフル
活用

戦略的要点:「米国の資本力とインフラがあって初めて成立する戦略。他国が同じ土俵で競争することは極めて困難。」

中国戦略：DeepSeek V3.2が仕掛ける「非対称な競争」とコストの破壊

推論コスト比較（入力100万トークンあたり）



モデル：DeepSeek V3.2

パラメータ：6,710億（アクティブ 370億）

推論コスト：GPT-5の**約1/10～1/20**
（入力100万トークンあたり ~\$0.028）

基盤技術：

- **MLA**（Multi-head Latent Attention）：
KVキャッシュのメモリ使用量を劇的に削減
- **DeepSeekMoE**：
高度に最適化されたMoEアーキテクチャ

戦略的要点：「高性能なAIを、湯水のように安価に使う環境を実現。日本の高コストなソブリンAIにとって無視できない脅威。」

日本の戦略的回答「第三の道」：産業特化型ハイブリッドモデル



高信頼 (High-Trust)

- ・著作権をクリアした「クリーンなデータ」
- ・技術的透明性とコンプライアンス遵守

高密度 (High-Density)

- ・パラメータ効率を極限まで高め、軽量でも高性能
- ・日本の文化・商習慣への深い適合性 (Cultural Alignment)

高自律 (High-Autonomy)

- ・国内の分散インフラ (IOWN等) を活用
- ・特定事業者に依存しない、強靱な運用体制

国内主要プレイヤーは、それぞれ独自のアプローチで「第三の道」を切り拓く



軽量・高密度

(Lightweight & High-Density)

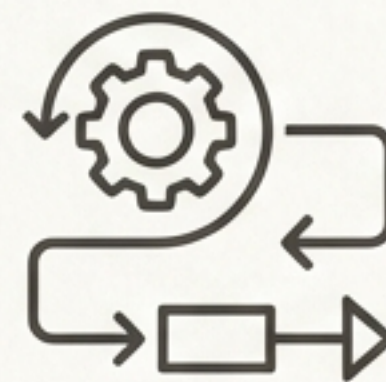
パラメータ効率、
IOWN連携、
分散知能



規模と蒸留

(Scale & Distillation)

国産最大規模、
Teacher-Studentモデル、
データ経済圏



自律エージェント

(Autonomous Agent)

長文脈対応、
業務プロセス自動化、
産業特化

NTT「tsuzumi 2」：IOWNと連携する軽量分散知能

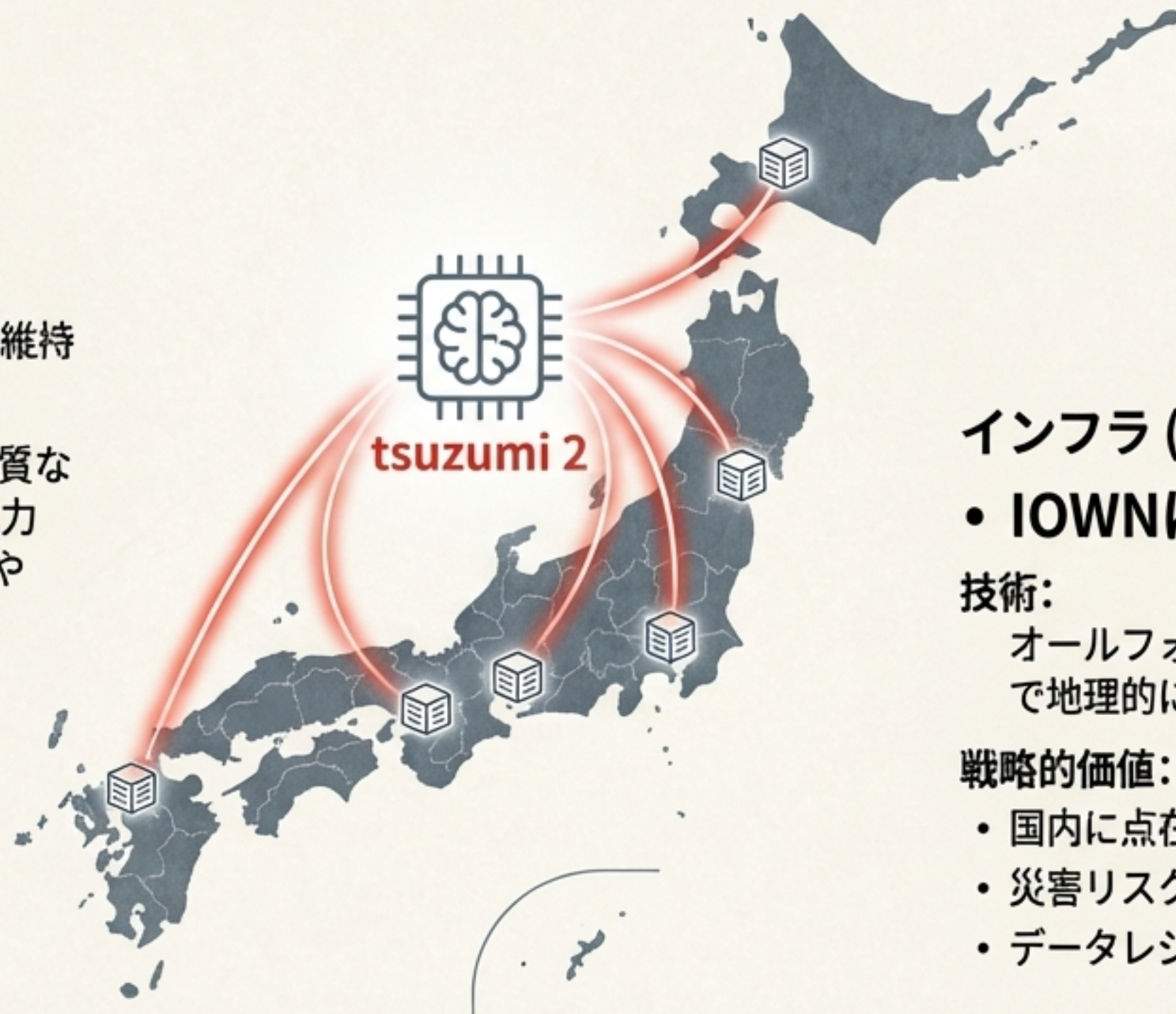
世界的な「巨大化競争」とは一線を画す、「軽量・高密度」な設計思想。

モデル (Model)

・究極のパラメータ効率

サイズ：単一GPUで推論可能なサイズを維持
(前世代は0.6B/7B)

性能：パラメータ数を抑えつつ、高品質な学習データにより日本語処理能力はGPT-3.5級を凌駕。特に敬語や専門用語に強み。



インフラ (Infrastructure)

・IOWNによる分散学習

技術：

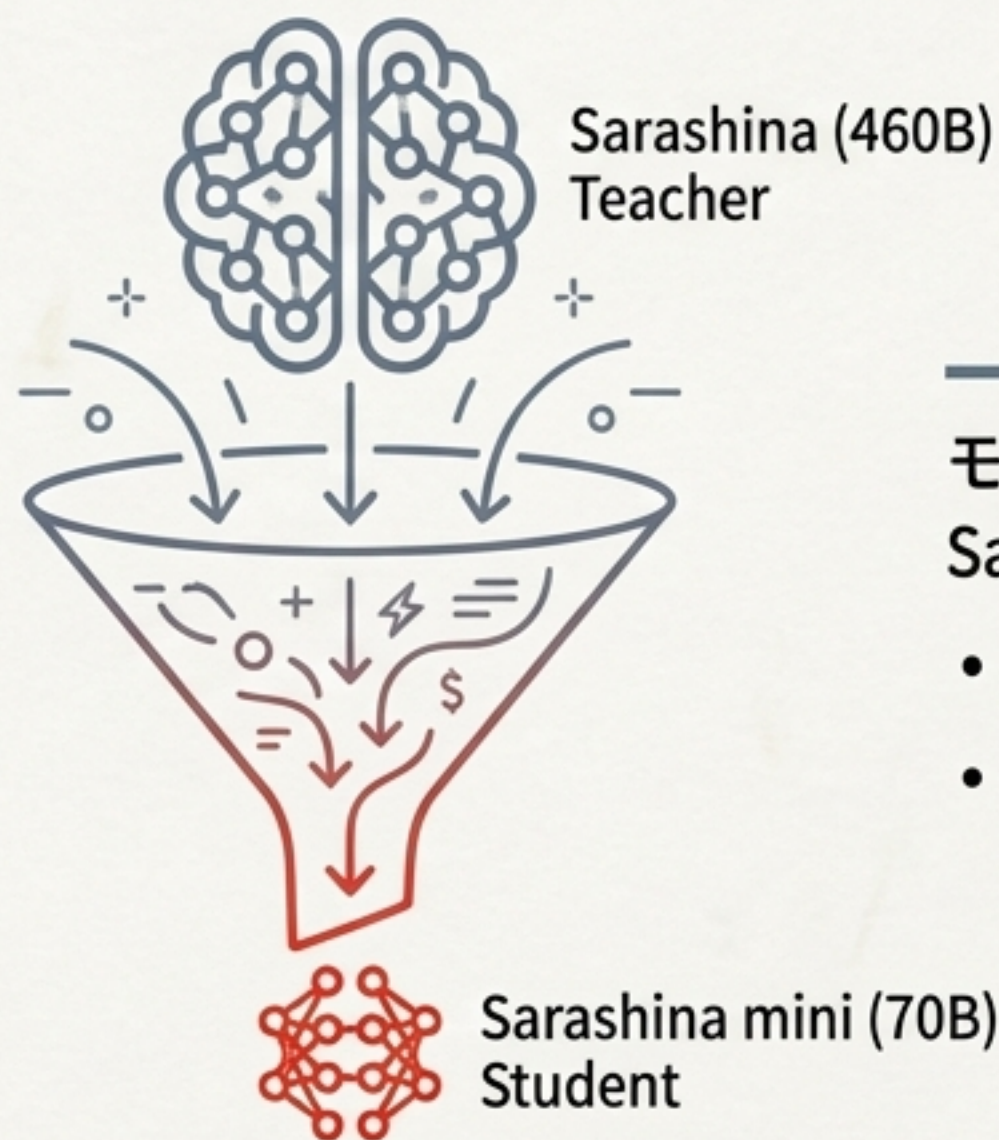
オールフォトニクス・ネットワーク(APN)で地理的に分散したDCを低遅延で直結。

戦略的価値：

- ・国内に点在する計算資源を有効活用
- ・災害リスクを分散
- ・データレジデンシーを完全に国内に保持

ソフトバンク「Sarashina」：規模の経済と戦略的「蒸留」技術

グローバル級の「規模」を追求しつつ、その知能を効率的なモデルに凝縮する。



モデル名 (Model Name): Sarashina

- 規模 (Scale): 4,600億 (460B) パラメータ - 国産最大級
- 役割 (Role): 複雑な推論と広範な知識を保持する「教師」

モデル名 (Model Name):

Sarashina mini (2025年秋予定)

- 規模 (Scale): 70B パラメータ
- 役割 (Role): 教師モデルから知識を「蒸留」し、高性能と低運用コストを両立

Data Strategy (データ戦略):

ソフトバンクグループの持つ良質な日本語データを活用し、日本の商習慣や文化的ニュアンスを巨大パラメータ空間に高解像度でマッピング。

NEC「cotomi」：業務プロセスを遂行する「自律エージェント」への進化

単なる言語モデルから、社内システムと連携し業務を自律的に実行するAIへ。

1 長文脈対応 (Long Context) :

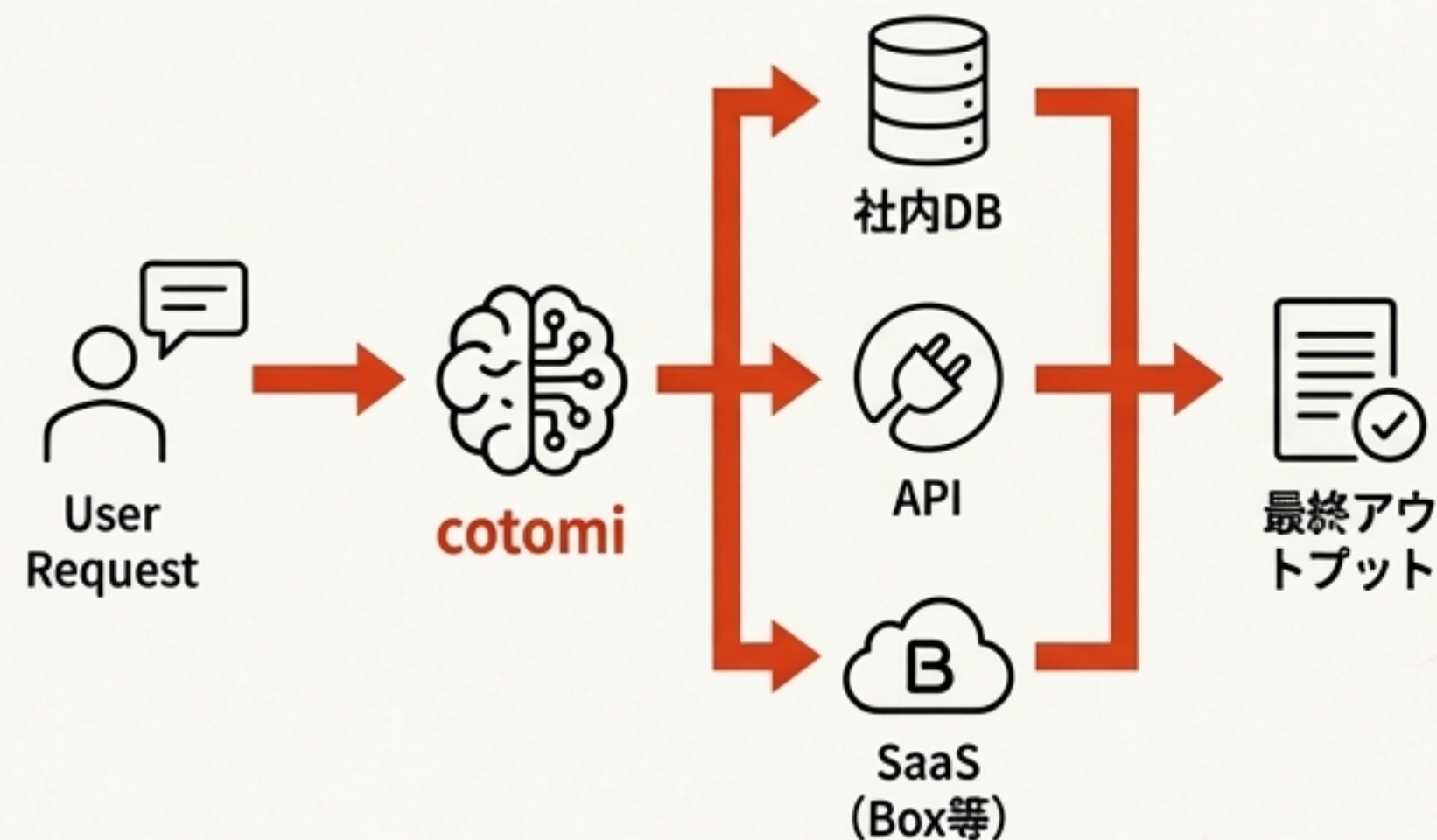
- コンテキスト長: 128kトークン (日本語約20万文字)
- ユースケース: 膨大なマニュアルや契約書を丸ごと読み込み、複雑な照会応答が可能に。

2 エージェント機能 (Agent Functionality) :

- 標準規格: MCP (Model Context Protocol) に迅速対応
- 能力: 社内DB、API、SaaS (Box等) と接続し、自律的にツールを選択・実行。

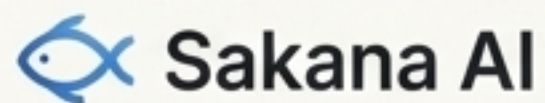
3 高速推論 (Fast Inference) :

- 性能: GPT-4比で87%~93%の高速化 (レイテンシ削減) を実現。



GENIAC世代の挑戦者たち：多様なアプローチがエコシステムを深化させる

経済産業省の支援プロジェクト「GENIAC」から、大企業とは異なる革新的な手法を持つスタートアップが台頭。



技術

**進化的モデルマージ
(Evolutionary Model Merge)**

コンセプト

既存モデルを遺伝的アルゴリズムで掛け合わせ、低コストで高性能な新モデルを創出。「柔よく剛を制す」アプローチ。



技術

**ドメイン特化型モデル
(Domain-Specific Model)**

コンセプト

製造業や材料開発に特化(100Bクラス)。特許データ等を事前学習ささせ、汎用モデルでは不可能な専門的回答を実現。



技術

マルチモーダル・End-to-End学習

コンセプト

完全自動運転を目指し、視覚情報と言語情報を統合したモデルを開発。

自律性を支える物理層：国内回帰する計算資源と「クリーンデータ」戦略

計算資源 (Compute)



ボトルネック解消：GENIACによるGPUクラスタへのアクセス提供。

恒久インフラ：

- ソフトバンク：北海道苫小牧市に国内最大級300MWのAIデータセンターを建設（2026年度稼働）。
- NTT：IOWN構想に基づき、全国のDCを統合運用。

データ主権 (Data Sovereignty)

✓ クリーンAI



⚠ 無断利用



課題：Perplexity等、海外AIによる報道機関記事の無断利用が訴訟に発展。

日本の回答：「クリーンAI」

- 国立情報学研究所(NII)などが著作権処理済みの日本語コーパスを整備・提供。
- NTT/SBは自社保有データを活用し、スクレイピングへの依存を低減。

競争優位性：コンプライアンスを重視する日本企業にとって、性能差を補う採用理由となる。

国産AIの現在地：日本語能力では世界と渡り合う一方、 明確な課題も残る

+ STRENGTHS | 強み

日本語能力：

- ベンチマーク（JMMLU, Rakuda）で**同パラメータ帯の海外モデルを凌駕**し、GPT-4**クラス**に肉薄。

文化的適合性（Cultural Alignment）：

- 敬語の使い分けや稟議書など、**日本固有の文脈・文書作成**で自然な出力を生成。
- 日本の法律・行政知識に強く、**ハルシネーション発生率が低い**。

— CHALLENGES | 課題

論理・数理能力の壁（Logic/Math Wall）：

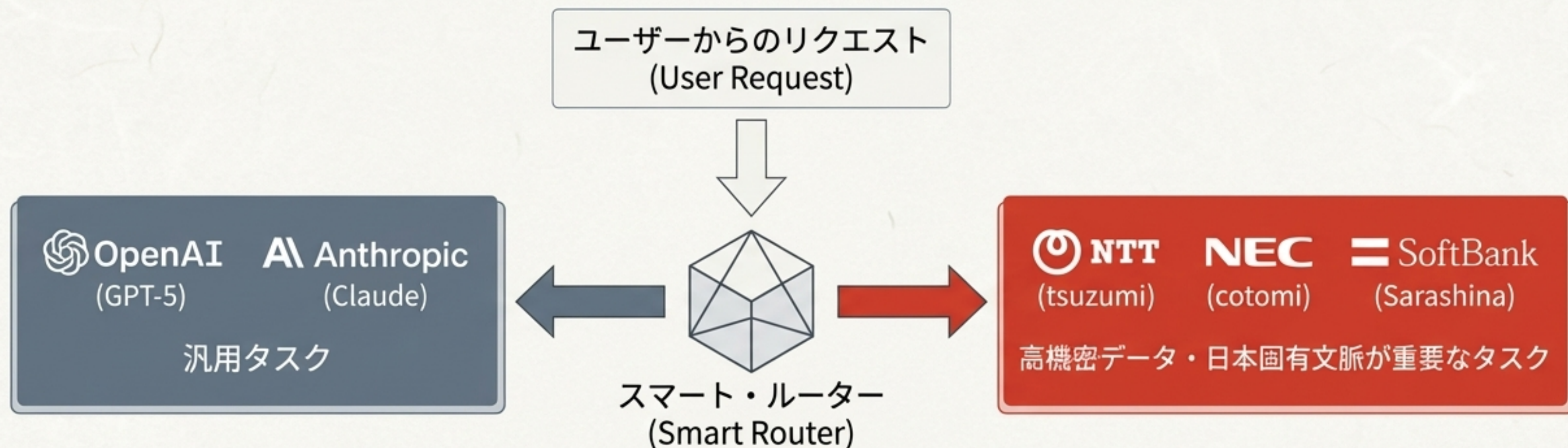
- 数学やコーディング能力では、学習データの絶対量で勝る**海外トップティアモデル**が依然優勢。

推論コストの壁（Inference Cost Wall）：

- DeepSeek等が実現した**桁違いの低コスト**に対し、**トークン単価での価格競争力**は課題。

戦略的展望：「鎖国」ではなく「連合型主権」による自律性の確立へ

「完全な純国産」はイノベーションを阻害する。目指すべきは、適材適所で国内外のモデルを使い分けるハイブリッドアーキテクチャである。



フェデレーテッド・ソブリンティ

汎用知能は海外の最先端モデルを活用しつつ、認知主権の根幹に関わる領域は、高信頼な国産モデルに処理を委譲する。

日本は「キャッチアップ」から、独自の価値を創造するフェーズへ



米国のような「圧倒的汎用性」でもなく、
中国のような「圧倒的低コスト」でもない。

日本が確立しつつある価値提案は、**「高信頼・高密度・高自律」なAI基盤**である。

この独自のAI基盤は、日本のデジタル社会における「認知主権」を技術的に担保し、持続可能な発展の土台となる。
2026年は、これらの技術が社会に実装され、その真価が問われる実行の年となるだろう。