ELYZA-Thinking-1.0 (LLM-jp-4版) 戦略的評価レポート

国産基盤モデルの現在地と、エンタープライズAI導入における戦略的ポジショニング

2026年10月 / 分類：戦略分析レポート

Gemini Notebook

実用的な日本語推論とデータ主権を両立する特化型アプローチ



完全な国内開発パイプライン

国立情報学研究所（NII）の12兆トークン事前学習モデル「LLM-jp-4」をベースに開発。真の意味ででの「国内基盤」を確立。



固有コンテキストでの圧倒的優位

日本語マルチホップQAや厳格な書式指定（JFBench）において、海外の最先端オープンモデルを凌駕するスコアを記録。



Sovereign AIとしての真価

OpenAI o3のような極限的な数理推論ではなく、機密データを保護できる「オンプレミス環境でのApache 2.0運用」に最大の戦略的価値が存在。

Gemini Notebook

NIIの基盤をベースとした極めて効率的な事後学習パイプライン

RSI（再帰的自己改善）の要素技術：

LLM学習基盤の自動化により、事後学習から評価までを人間介在「1〜2人日」の極限効率で完遂。

フェーズ1: NII開発ベース
「LLM-jp-4」
（12兆トークンによるスクラッチ事前学習）

フェーズ2: ELYZA独自事後学習

① 中間学習 (Mid-training)

② 教師あり微調整 (SFT)

③ 報酬に基づく強化学習 (RLVR)

フェーズ3: 完成モデル
ELYZA-Thinking-1.0
（思考プロセスと最終回答を分離するharmony形式を採用）

Gemini Notebook

運用環境に応じて選択可能な2つのアーキテクチャ

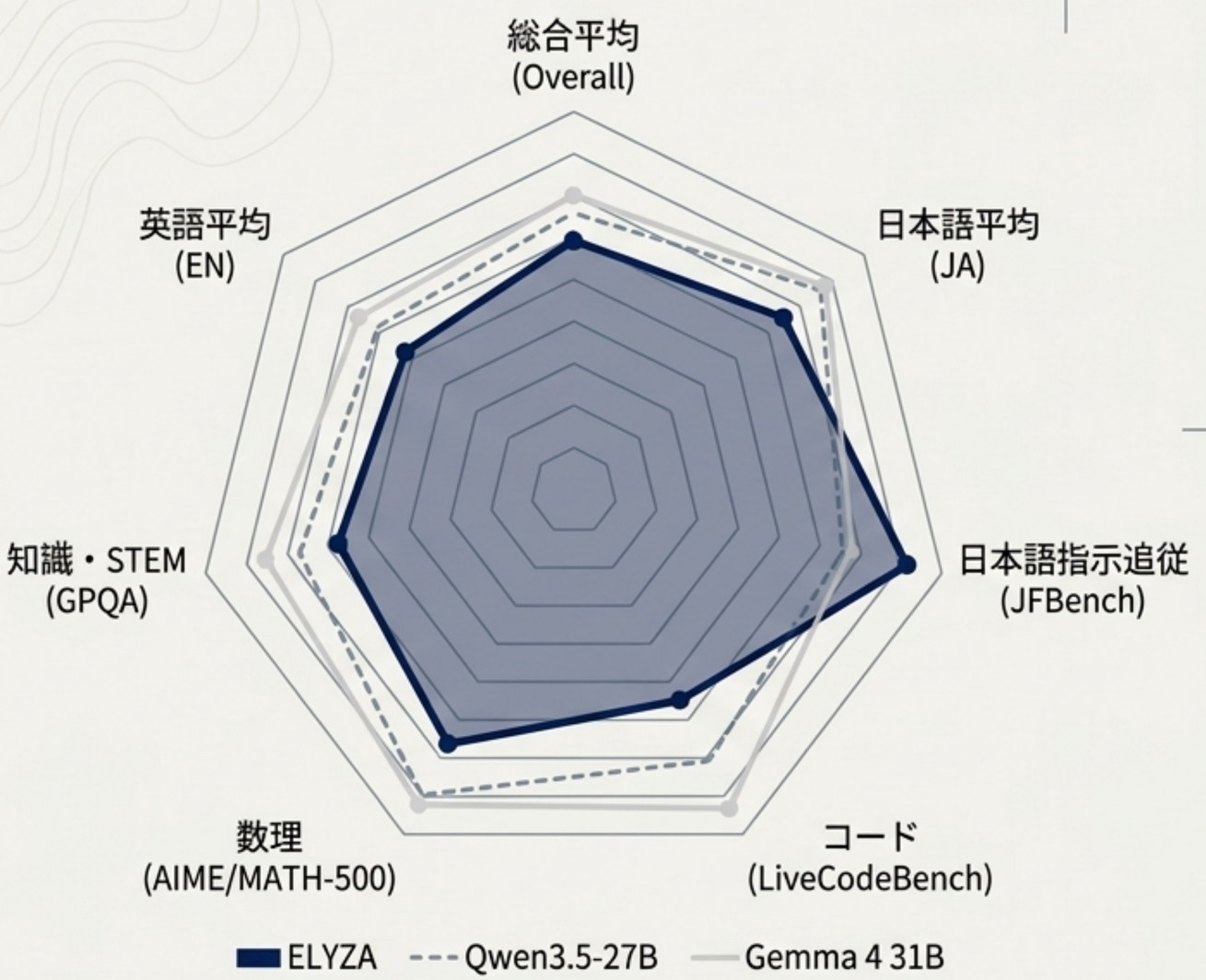
ELYZA-Thinking-1.0-llm-jp-4-33b	ELYZA-Thinking-1.0-llm-jp-4-32b-a3b
ベースモデル: llm-jp-4-33b-base (NII)	**ベースモデル:** llm-jp-4-32b-a3b-base (NII)
アーキテクチャ: Dense（全パラメータ常時稼働）	**アーキテクチャ:** MoE（128エキスパート中8つを活性化）
パラメータ数: 約332億（33B）	**パラメータ数:** 約321億（活性化時 約38億）

共通のエンタープライズ仕様

コンテキスト長: 65,536 トークン（約64K） | ライセンス: Apache 2.0（完全な商用利用・オンプレミス展開が可能）

Gemini Notebook

グローバル最先端モデルとの性能トポロジー比較



明確なトレードオフ構造

総合平均スコア（61.69）ではQwen3.5-27B（68.18）やGemma 4（70.54）に遅れをとる。

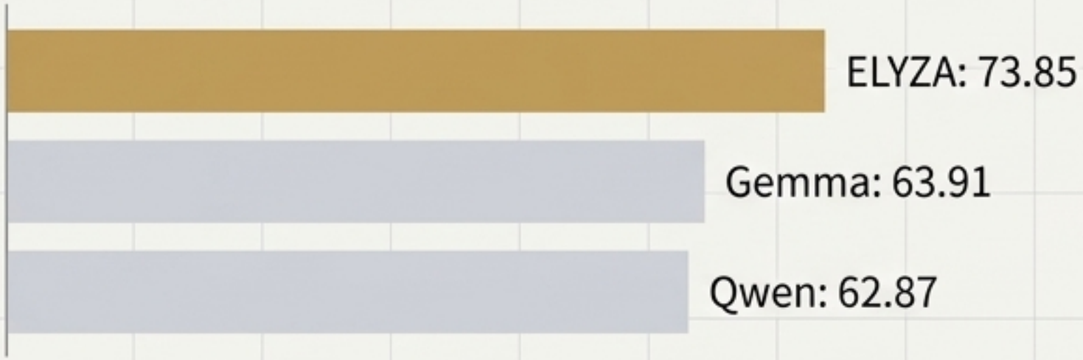
非対称な性能曲線

英語・数学・コード領域で生じるビハインドと、特定の日本語タスクにおける局所的な優位性が明確に分かれる**「推論特化型のスペシャリスト」**。

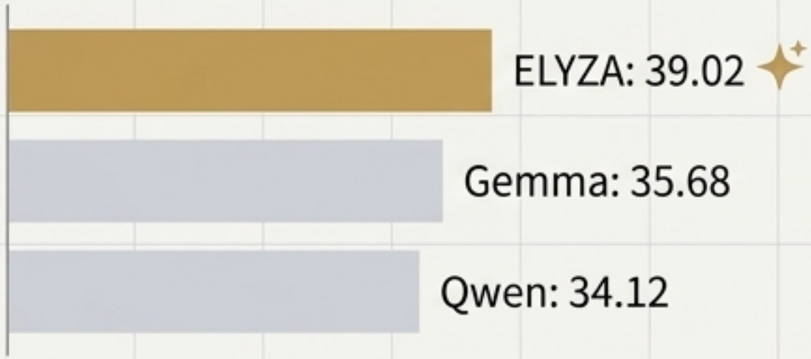
Gemini Notebook

ELYZA最大の防壁：日本語コンテキストと厳密な書式制御

日本語複数文書QA (JEMHopQA)



日本語指示追従 (JFBench)



優位性の源泉

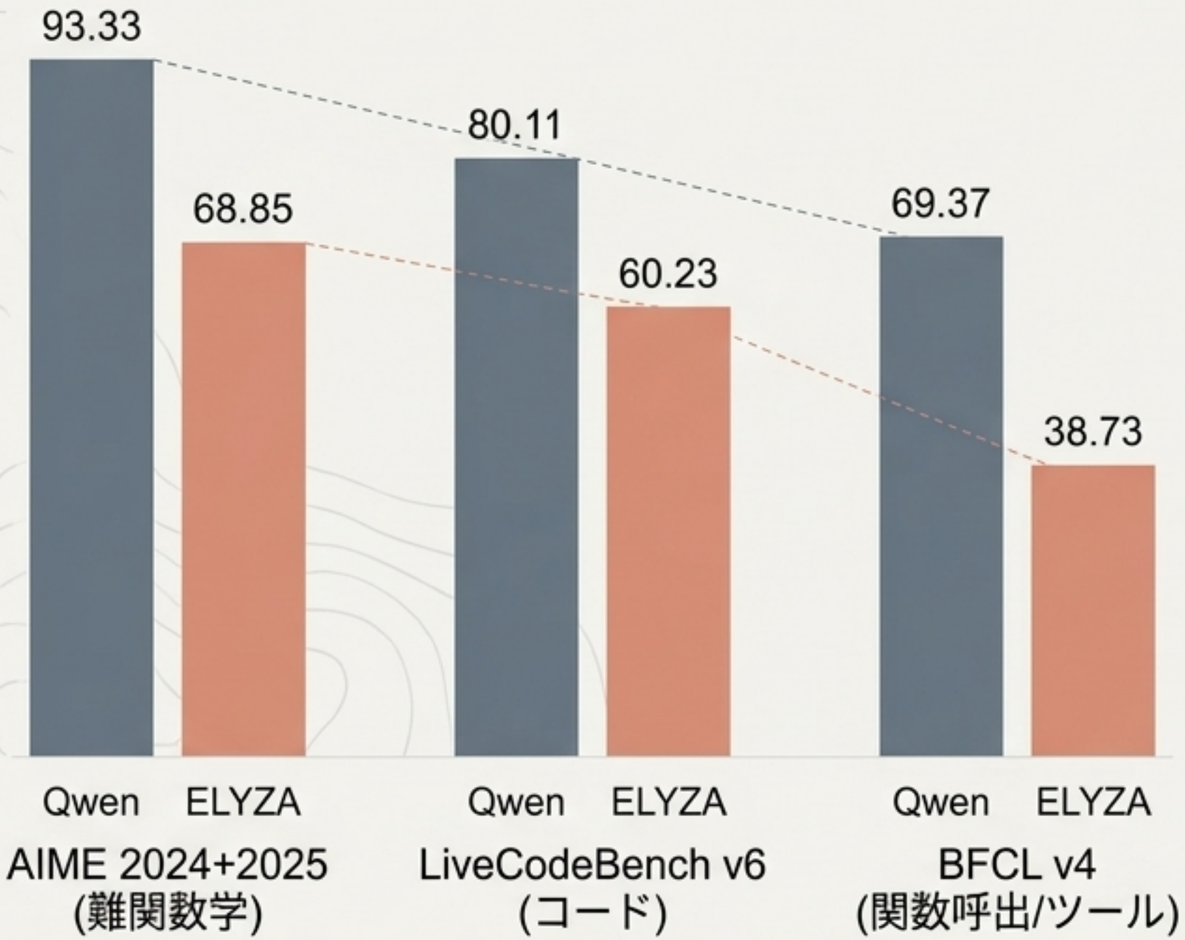
最新の日本語Wikipedia / Wikidataからの合成知識データの的確な注入。

言語的制約の支配力

「振り仮名の付与」や「漢数字の指定」など、グローバルモデルが取りこぼしやすい日本固有の厳密な出力フォーマット要件において、極めて高い追従性を発揮する。

Gemini Notebook

抽象的論理推論と自律的エージェント操作における課題



記号処理や多段階アルゴリズムの探索を要するタスクにおいて、グローバル水準との決定的な格差が存在する。

言語化された知識処理には長けるが、外部ツールを用いた自律的なエージェント操作や、極限的 数理推論においては海外の最先端モデルに劣後する構造。

Gemini Notebook

フロンティア推論モデル（OpenAI o1 / o3）との実力差

ベンチマーク	ELYZA-33b	OpenAI o1	OpenAI o3
AIME（数学オリンピック）	68.85%	83.3%	96.7%
GPQA Diamond（博士レベル理系）	69.95%	78.0%	87.7%
MATH-500（高校数学）	96.00%	96.4%	97.0%+

指標の飽和（天井効果）への注視

MATH-500ではo1に肉薄するスコアを示すが、これは近年の推論モデルにおける「指標の飽和」によるもの。極限的な推論深度を測るAIMEやGPQAのスコア乖離が、両者の決定的な推論格差（大学院・専門家レベルでの壁）を物語っている。

Gemini Notebook

設計思想の分岐：なぜ推論能力に決定的な差が生じるのか？



ELYZA-Thinking-1.0

コンセプト: パラメータ効率とローカル制御

- **設計アーキテクチャ:** 33Bのコンパクト設計。限られた容量を日本語コンテキストと指示適合に全振り。
- **運用ターゲット:** 企業内の単一ノード・手頃なGPU環境でのオンプレミス運用。



OpenAI o1 / o3

コンセプト: 圧倒的スケールと動的計算

- **設計アーキテクチャ:** 数千億パラメータの基盤。強化学習を通じた推論時計算量（Test-time Compute）の動的スケーリング。
- **運用ターゲット:** 膨大な計算資源を背景にした、クラウド上での極限的な記号論理・数学的証明の追求。

Gemini Notebook

AIコミュニティにおける評価と、市場の誤解の是正

市場の混同（誤解）

「OpenAI o1-miniに匹敵する推論力を持つモデルである」

それは2025年5月に公開されたQwen2.5-32Bベースの別モデルに対する評価。今回のLLM-jp-4版はアーキテクチャ・学習手法が完全に異なる別系統であり、過去の評価を適用するのは技術的に誤り。

真の実務的価値（事実）

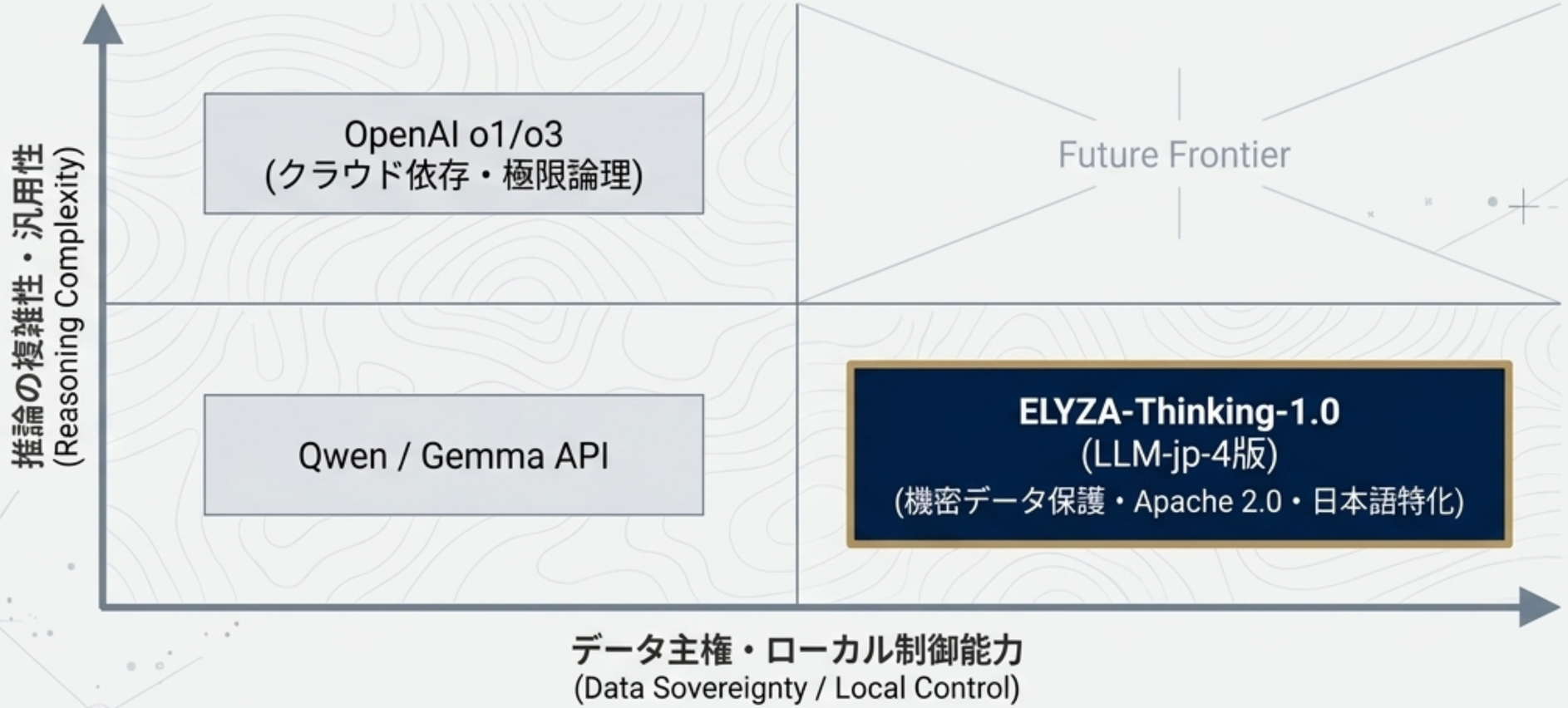
コスト効率の劇的改善とデータ主権の確保

推論コストを67%削減しつつエージェントの精度を維持した実績。完全な商用利用を担保するApache 2.0でのオンプレミス運用が可能。

※課題: 推論データ生成過程における海外教師モデルへの依存脱却が今後の自律的開発サイクルの鍵となる。

Gemini Notebook

エンタープライズAI戦略におけるポジショニング



Strategic Takeaway

ELYZAは「最高峰の論理推論」を競う領域にはない。しかし、金融・医療など「機密データを国外に出せない」環境下において、実用的な推論AIを配置できる唯一無二のポジションを獲得している。

Gemini Notebook

最終評価（Final Verdict）

ELYZA-Thinking-1.0の真の価値は、フロンティアモデルと「純粋な論理力」で競うことではない。

それは、企業が自社の管理下で極めて安全かつ低コストに、「高度な日本語対話・指示追従能力」をオンプレミス実装できる道を切り拓いたことにある。

データガバナンスとソブリンAIの確立を目指す国内エンタープライズにとって、実務導入可能な最高峰のオープン・スペシャリストである。

Gemini Notebook