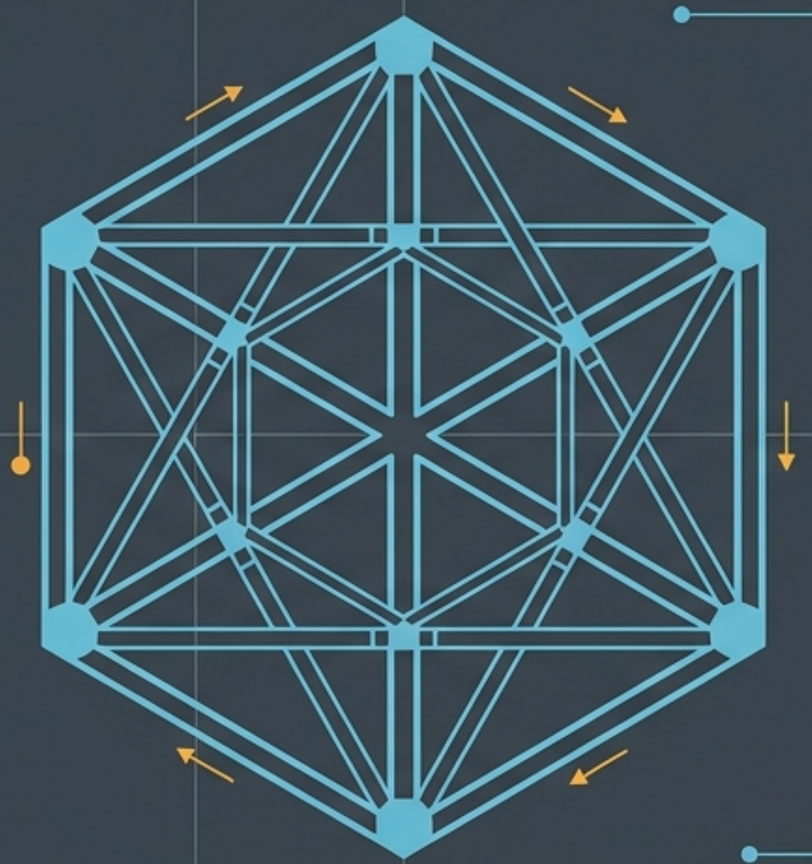


ELYZA公開推論モデルの技術境界と実務有用性

[DIAGNOSTIC_BRIEFING_v2026.10.03]

- 対象モデル: ELYZA-Thinking-1.0-llm-jp-4 (33B / 32B-A3B)
- 比較対象: OpenAI o1 / o3 / Qwen3.5
- 評価視点: 汎用推論性能および知的財産業務への適用性

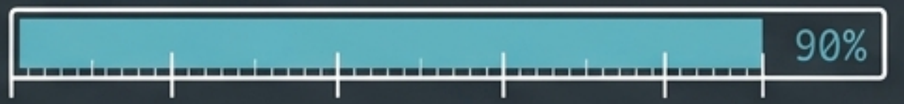
Gemini Notebook

結論：日本語特化の実務特化型としては極めて優秀だが、o1/o3を凌駕する汎用推論モデルではない。

The Verdict

- メディアで散見される「o1を超えた」「完全国産」という言説は、部分的な数値の切り取りや誤解を含む。
- 実際には、日本語の指示追従や文書処理に最適化された「実務特化モデル」であり、数学や科学などの純粋な難問推論ではOpenAIの最新フロンティアモデル（o3）に及ばない。

日本語業務処理



高い有用性。発明届の整理やデータ抽出に即座に活用可能。

汎用難問推論



限界あり。STEM分野や外部ツール連携は改善の余地あり。

RSI（再帰的自己改善）

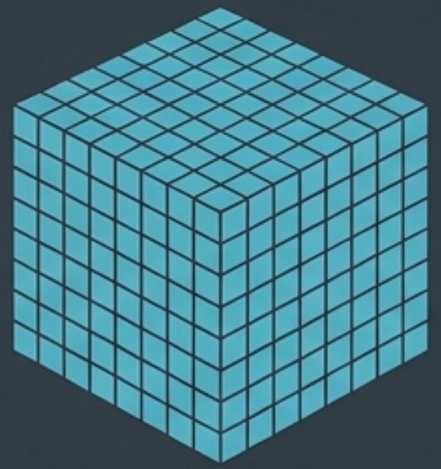


研究段階。設立された研究組織の目的であり、現行モデルに実装・達成されたものではない。

Gemini Notebook

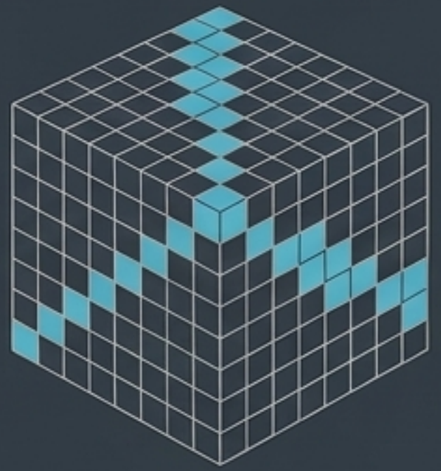
公開された2つの推論モデル：性能特化のDense型と計算量最適化のMoE型

Dense Engine



全パラメータ同時稼働：性能特化

MoE Engine



専門家経路のみ稼働：計算量最適化（約12%）

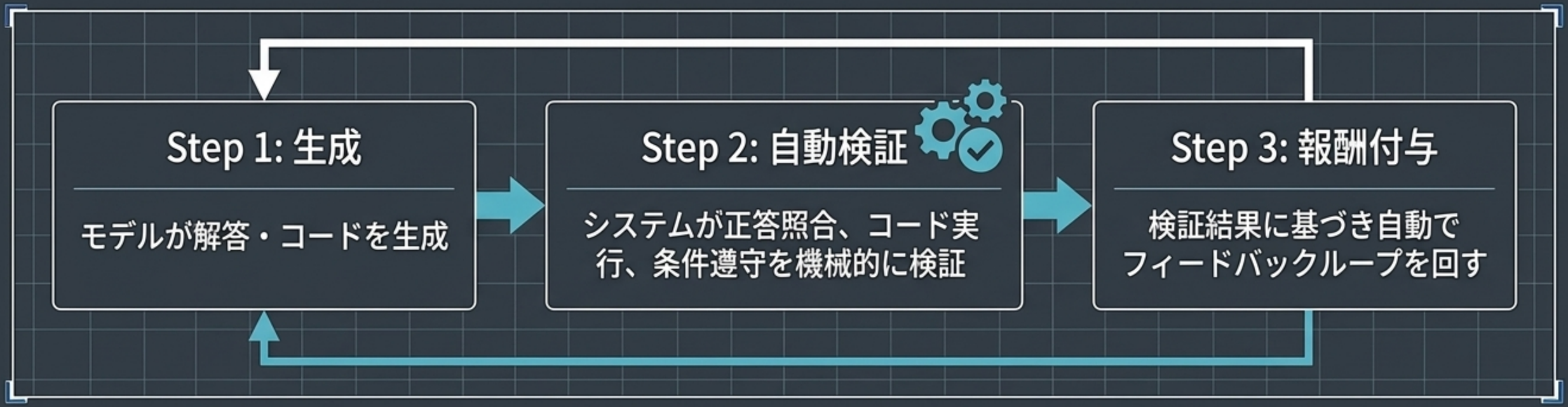
	ELYZA-Thinking-1.0-llm-jp-4-33b	ELYZA-Thinking-1.0-llm-jp-4-32b-a3b
アーキテクチャ	Dense（密結合）型	MoE（Mixture of Experts）型
総パラメータ数	約330億	約321億
稼働パラメータ数	約330億	約38億（総パラメータの約12%）
設計思想	性能重視のフラッグシップ	一部パラメータのみを稼働させ計算量を抑制
ライセンス	Apache 2.0（商用利用可）	Apache 2.0（商用利用可）

※稼働パラメータの少なさと、モデルの保存・実行に必要なメモリ量が同比率で減少することは別である点に留意。

Gemini Notebook

開発パイプライン：人間介在を極小化する検証可能な報酬による強化学習 (RLVR)

重点学習領域 = 日本語化された数学・科学・プログラミング + 日本語の複雑な指示・固有知識



Mythbuster：「人間が介在する作業を1〜2人日程度に抑えた」という真意

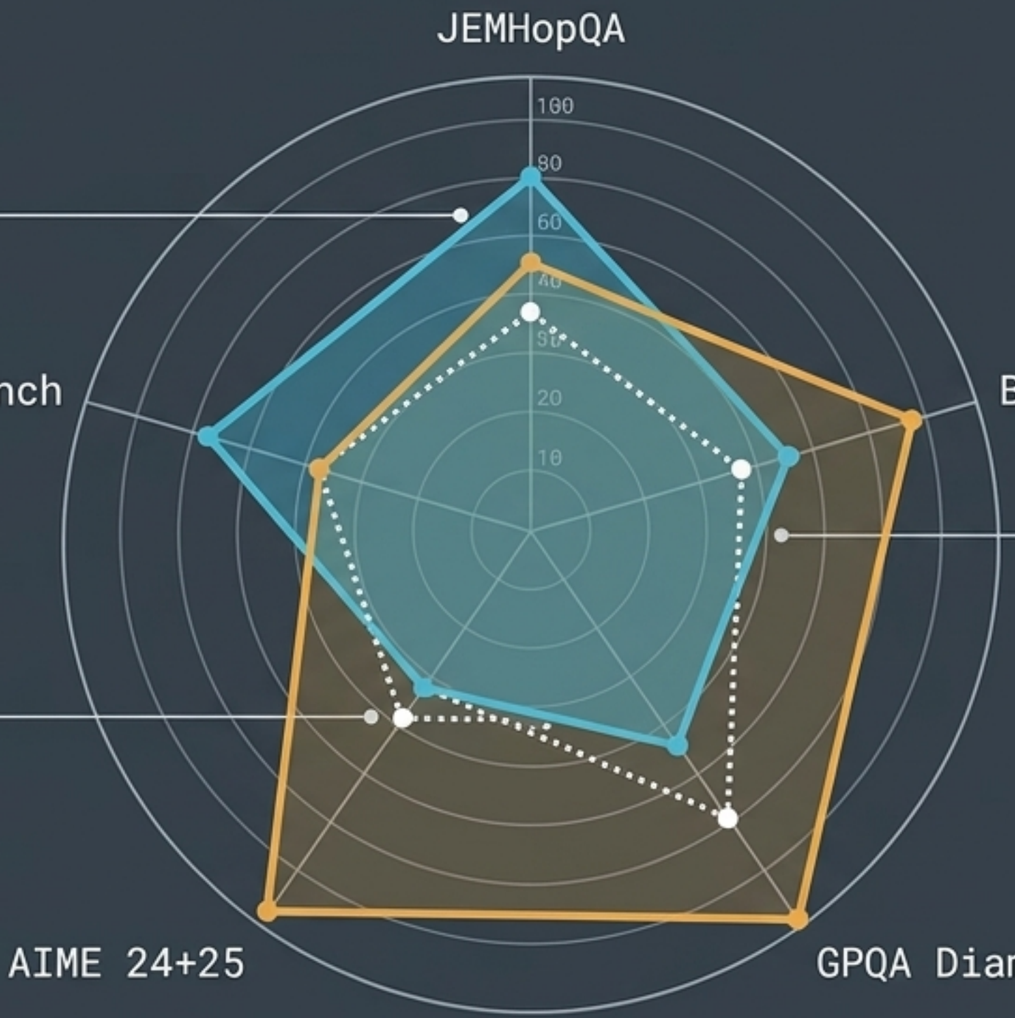
整備済みの学習基盤を活用し、中間学習・事後学習・評価における手作業を極小化したという意味。
基盤モデルの構築から学習設備を含めた「全開発が1〜2日で完了した」というわけではない。

Gemini Notebook

ベンチマークから見える「形状」：日本語タスクへの突出とSTEM領域の課題

JEMHopQA:
ELYZA（73.85）がQwen3.5（62.87）を圧倒。

JFBench:
ELYZA（39.02）がNII版（32.79）から大幅向上。



BFCL v4:
ELYZA（38.73）はQwen3.5（69.37）に大きく遅れをとる。

ELYZA 33Bは総合スコア(61.69)でNII版(60.64)を上回るが、一律に全能力が高いわけではなく、明確な得意・不得意のプロファイルを持つ。

Gemini Notebook

進化の分岐：LLM-jp-4基盤からの異なるチューニングの方向性

[基盤]
国立情報学研究所（NII）
LLM-jp-4（2026年9月）

[NII LLM-jp-4.1]

方向性: STEM分野、ツール呼び出し機能(BFCL)、回答の簡潔さの改善に注力。

[ELYZA-Thinking-1.0-llm-jp-4]

方向性: 日本語の複雑な指示追従、多段的な質問応答、日本固有の知識の強化に注力。

「どちらが上位モデルか」という単純な比較は無意味である。両者は同じ基盤から出発し、全く異なるユースケースに向けて調整された「姉妹モデル」として捉えるのが技術的に適切である。

Gemini Notebook

報道と実態の乖離：公開直後の過熱に対する技術的ファクトチェック

[ナラティブ] ✗	[実態] ✓
完全な国産AIである	国内主体が開発したことは事実だが、学習データ、教師モデル、技術要素のすべてが国内由来であると広義に解釈することはできない。(ITmediaも10/2に表現を修正済)
RSI（再帰的自己改善）を達成した	ELYZA RSI Researchは「RSIの前段階の研究」組織として設立された。現行の公開モデルがAIによるAIの自律的改善を達成したという発表はない。
ローカル環境で即座に無料で実運用可能	モデルの「配布」が無料(Apache 2.0)なだけである。Hugging FaceにGGUF版等の量子化モデルはあるが、実行インフラのコストや量子化後の実用性能は別途検証が必要。

Gemini Notebook

推論能力のベンチマーク比較における統計的陥穽（GPQA / AIME）





Why point-based comparison is invalid

プロンプトと設定の不一致

推論に使用するトークン量、試行回数、プロンプトの条件が統一されておらず、単純な引き算を能力差として確定できない。

AIMEの出題年度のズレ

ELYZAは2024+2025年版、o1は2024年版、o3は2025年版を使用しており、直接の順位表として成立しない。

評価軸の混同

日本語の指示追従性が高いことと、未知の数学・科学の難問を解く「推論力」は独立した能力である。

Gemini Notebook

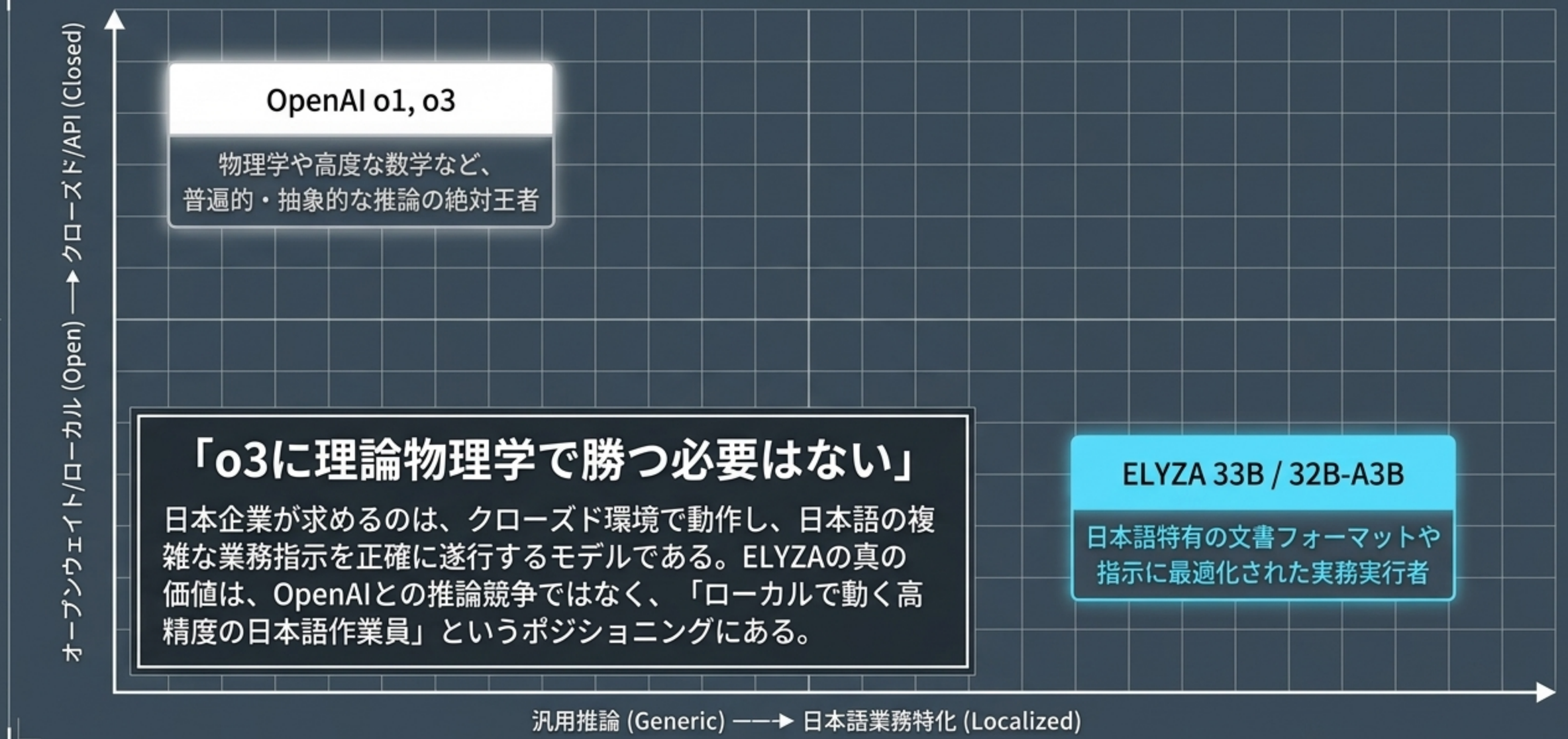
フロンティアモデルとの推論性能比較マトリクス（暫定評価）

[比較対象]	[ELYZAに対する推定優位性]	[判断の確からしさ(定性)]	[OpenAIモデルのコア優位点]
o1-preview (2024.9)	数学等でELYZAを上回る可能性はあるが、総合的な超過は未確認	中程度以下	早期の推論パラダイム導入
正式版 o1 (2024.12)	ELYZAの一部の能力は接近しているが、汎用的な難問推論でELYZAが上回ったとする根拠は不足	中程度	安定した高度な論理展開
o3 (2025.8)	ELYZAが汎用的な難問推論でo3を上回っている可能性は極めて低い	中程度～比較的高い	汎用的な難問推論における圧倒的性能 (ツールなしAIME 88.9)

結論：独立した評価者による同条件の実機比較が存在しない現在、ELYZAがo1/o3の推論能力を上回ったと断定する証拠はない。

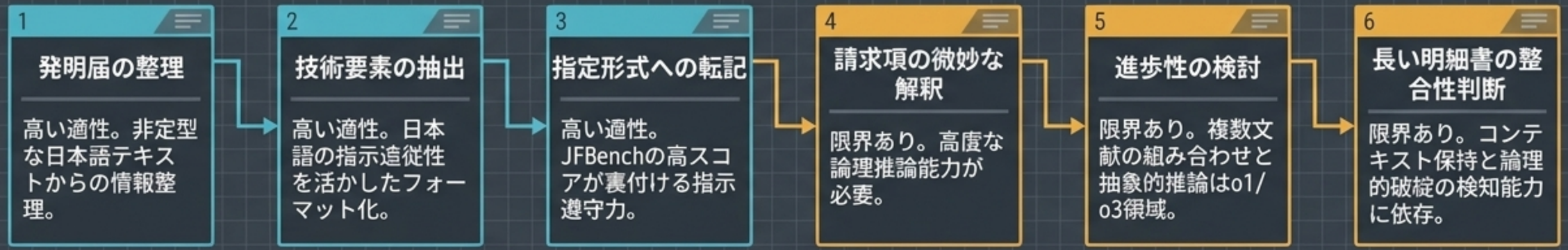
Gemini Notebook

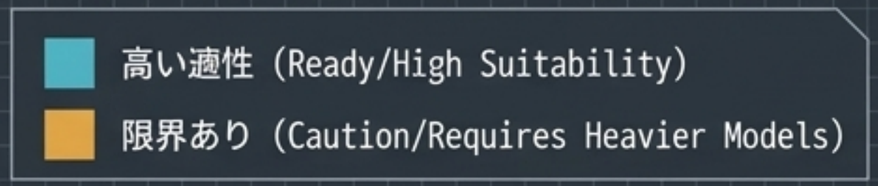
LLM進化の分岐点：絶対的推論力と局所的業務効率のトレードオフ



Gemini Notebook

エンタープライズ展開：知的財産（IP）業務における適用ヒートマップ





Gemini Notebook

自社導入のための評価フレームワーク（汎用ベンチマークからの脱却）

1. 比較条件の厳密な統制

- プロンプトの完全一致
- 提供する資料（RAG等のコンテキスト）の同一性
- 出力フォーマット制約の統一
- 外部ツールの有無の統一

2. 業務適合性の測定指標

- 根拠箇所の特定正確性
- 特許構成要件の見落とし率
- 否定表現・例外条件の誤読検知
- **専門家による最終修正時間（最重要指標）**

3. 量子化運用時の注意点

- Hugging FaceのGGUF/MLX 4bit版を検証する際は、必ず「本番導入予定の量化版」で実務テストを行うこと。
- 軽量化による日本語指示追従性の劣化度合いをモニタリングする。

Gemini Notebook

最終判定：戦略的ユースケース別 モデル選定ガイド

[ELYZA 33B（Dense）] を選ぶべきケース

自社専用のオンプレミス/VPC環境で、高セキュリティな日本語文書処理（要約、抽出、整形）を最高精度で行いたい場合。コストより性能優先。

[ELYZA 32B-A3B（MoE）] を選ぶべきケース

推論時の計算リソース（GPU）を抑制しつつ、並列処理や大量の定型的な日本語タスクを捌く必要がある場合。効率と速度優先。

[OpenAI o1 / o3] を選ぶべきケース

数学的証明、複雑なコーディング、複数特許文献の論理的矛盾の指摘など、高度な「推論・思考プロセス」が不可欠であり、クラウドAPIの利用が許容される場合。

Gemini Notebook