国産推論モデル「ELYZA-Thinking-1.0」の現在地と戦略的示唆

o1/o3推論水準とのギャップ分析およびRSI（再帰的自己改善）がもたらす開発プロセスの革新

[DATE] 2026.10.03
[SOURCE] ELYZA-Thinking_LLM-jp-4_調査報告
[DOCUMENT TYPE] Strategic Intelligence Briefing

Gemini Notebook

Executive Summary: 3つの重要ポイント

国産・商用可の推論モデル誕生

NIIの「LLM-jp-4」をベースに、日本語の処理性能に特化した33B（Dense）と32B-A3B（MoE）をApache 2.0で公開。機密性の高い社内環境での推論タスクに新たな選択肢。

推論力は「o1-mini超、o1未満」

独自推論指標の厳密な補正分析の結果、33Bの推論力はo1-miniを上回るが、o1や同クラスの海外オープンモデル（Qwen/Gemma）には依然として明確な性能差が存在する。

真の革新は「性能」ではなく「プロセス」

最も注目すべきは、AIがAIを育てる「RSI（再帰的自己改善）」の実証。高度な学習パイプラインを「人手1〜2人日」で完了させた超高効率な開発自動化の証明である。

Gemini Notebook

[The Specs] ELYZA-Thinking-1.0 の輪郭

33B Dense Model	
Parameter	330億
Architecture	Dense
Base Model	llm-jp-4-33b-base

32B-A3B MoE Model	
Parameter	総320億 (アクティブ約30億)
Architecture	MoE (Mixture of Experts)
Base Model	llm-jp-4-32b-a3b-base

License: Apache 2.0 (商用利用可能)

Language: 日本語・英語対応

Tool Calling: OpenAI形式のtoolsに対応

Recommended Settings: Temp 0.6 / Top_p 0.95

Gemini Notebook

[The Pipeline] 推論能力を獲得する3段階の学習アプローチ

Stage 1: 中間学習 (Intermediate Training)

数学・コード・STEM問題を日本語に翻訳し推論過程を生成。エージェント系データも日本語化。

Stage 2: SFT (教師ありファインチューニング)

数百万件規模の合成データ。Wikipedia/Wikidataからの日本語知識合成と、ふりがな・漢数字などの複合制約付き指示追従データを投入。

Stage 3: 強化学習 (Reinforcement Learning)

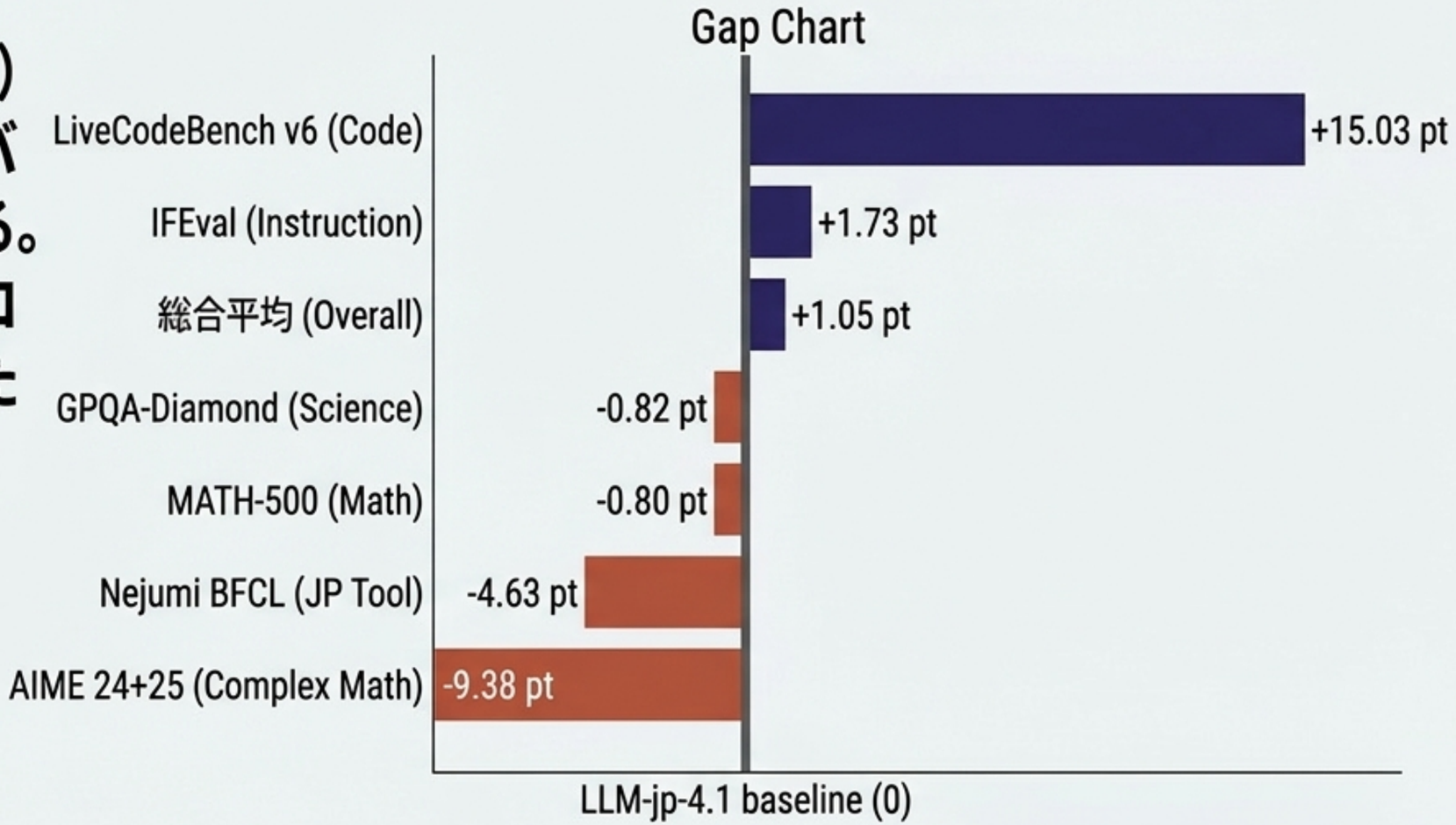
数学・コード・ツール使用など、正誤が客観的に検証可能な難問を対象にモデルを最適化。

驚異的な事実：この中間学習から事後学習・評価に至る全プロセスを、人間が介在する作業「わずか1〜2人日程度」で実行。

Gemini Notebook

[Gap Analysis] ベースモデル（LLM-jp-4.1）からの進化と局所的なトレードオフ

総合平均では微増（+1.05pt）にとどまるが、内部の能力バランスは劇的に変化している。純粋な理系推論を犠牲に、コードと実務能力に特化させたチューニングの痕跡が見える。



Gemini Notebook

[Market Reality] 30Bクラス・グローバルオープンモデルとの相対位置

同クラスの海外フロンティア・オープンモデル（Qwen/Gemma）と比較すると、総合平均で約6.5〜8.9ポイントの差が存在。
特に高度な推論領域でギャップが顕著である。

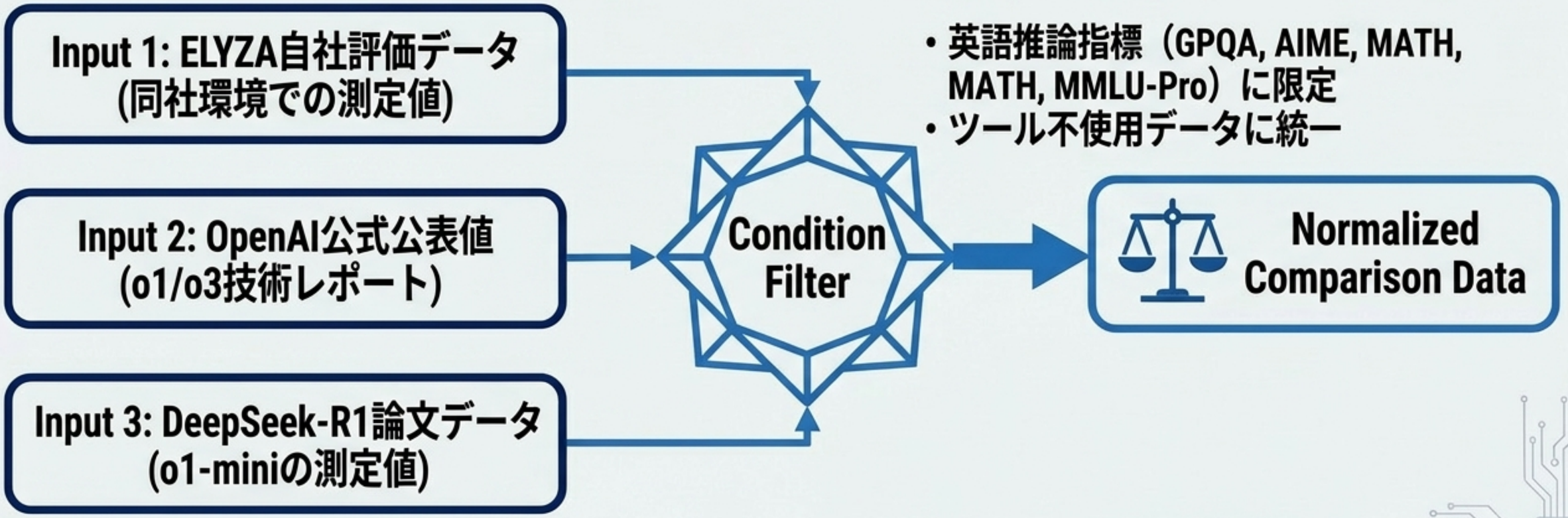
	ELYZA 33B	Qwen3.5 27B	Gemma 4 31B
総合平均 (Overall)	61.69	68.18	70.54
GPQA-Diamond (Science)	69.95	85.92	86.65
AIME (Math)	68.85	93.33	88.75
LiveCodeBench (Code)	60.23	80.11	79.37

結論：純粋な推論ベンチマークにおいて、現時点でグローバルモデルに対する優位性はない。

Gemini Notebook

[Methodology] フロンティアモデル（o1/o3）との比較方法論

評価環境（試行回数、プロンプト、推論予算）が異なるため、単純比較は危険。数ポイントの差は「誤差」としてして扱い、次スライドで評価環境の「辛さ」の補正を行う。



[Calibration] 測定環境のバイアス補正：「gpt-oss-20b」をアンカーとした分析

両者が共通して測定している「gpt-oss-20b」のスコアを比較すると、ELYZAの自社測定はOpenAIの公表値よりも「約3〜4ポイント辛め」に出る傾向がある。これを補正値として適用する。

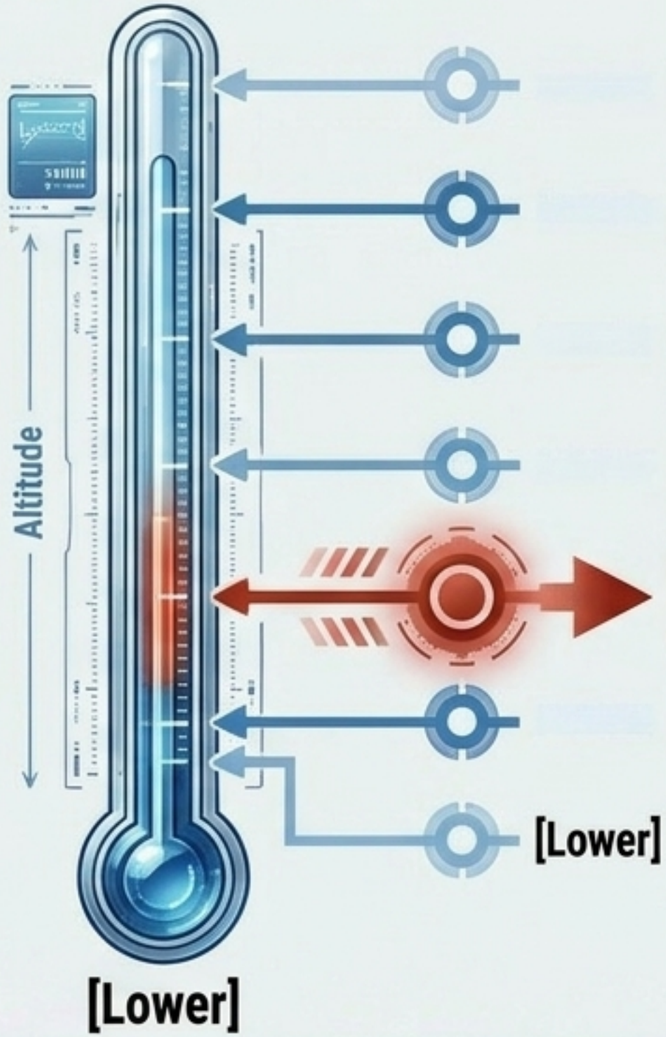


推論バイアス（+3〜4pt）をELYZA 33Bのスコアに上乗せして補正しても、o1（GPQA 75.7, AIME 76.8）には届かない。結果の信頼性は担保された。

Gemini Notebook

[The Verdict] o1/o3シリーズ推論能力スケールにおける現在地

[Higher Capability]



[Apex] o3: GPQA 83.3 / AIME 90.3

[High] o3-mini (high):
GPQA 77.0 / AIME ~85

[Mid-High] o1 (正式版):
GPQA 75.7-78.0 / AIME ~76.8
(※ELYZAはここに届かず)

[Mid] o3-mini (low):
(o1-miniと同等クラス)

[Target Highlight] ELYZA 33B (補正後):
GPQA ~73.0 / AIME ~72.5

[Target] ELYZA 32B-A3B: 概ね o1-miniと同等

[Lower] o1-mini: GPQA 60.0 / AIME 63.6

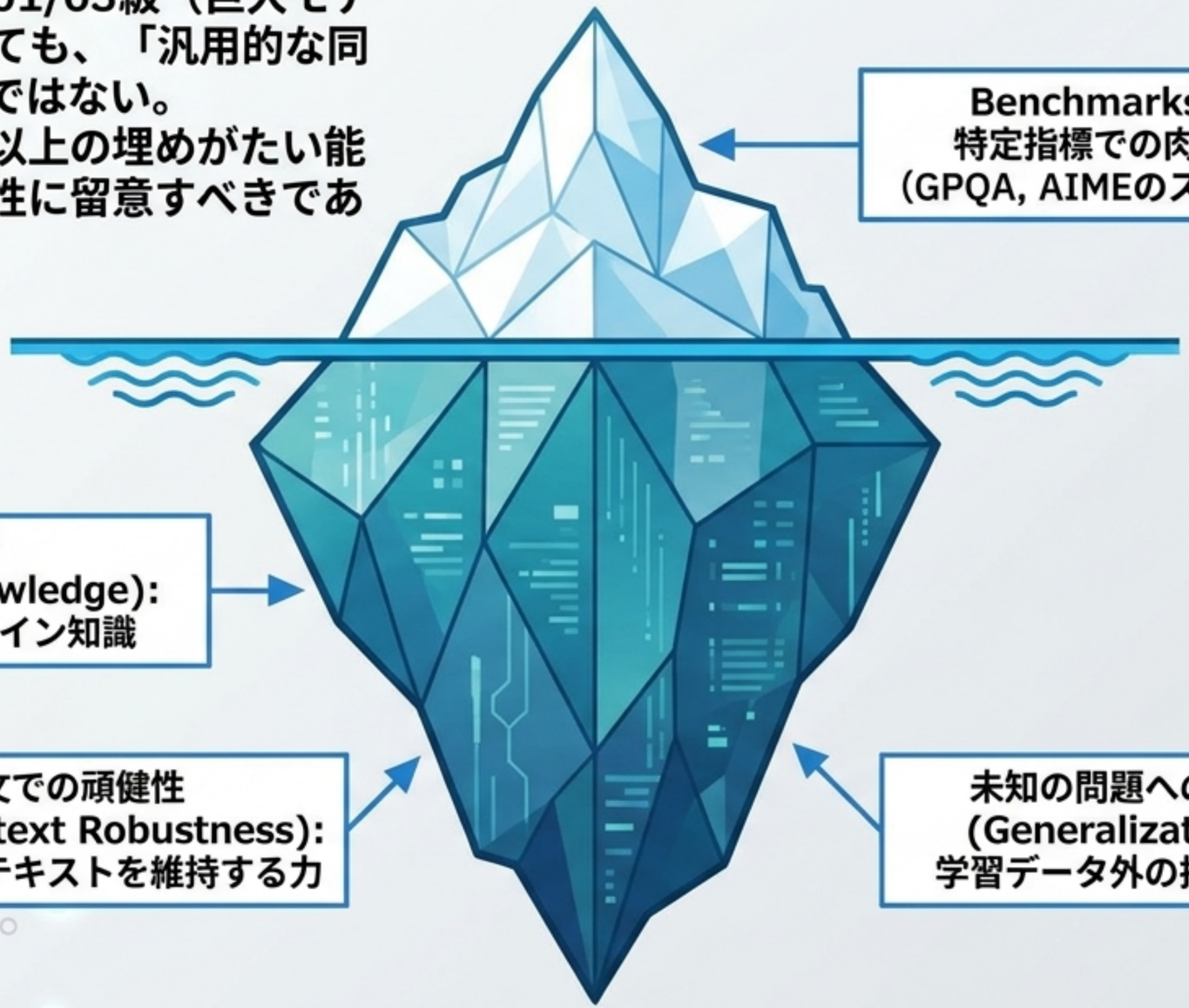
Executive Takeaway:

ELYZA 33Bは明確に「o1-miniを上回り、o1には届かない」水準。OpenAIの基準に照らせば、o3-miniのLowとMediumの間に位置する。

Gemini Notebook

[Diagnostic] ベンチマーク上の「近さ」が意味しないもの

33Bクラスのモデルがo1/o3級（巨大モデル群）に数値上肉薄しても、「汎用的な同等性」を意味するわけではない。
実務においては、数値以上の埋めがたい能力力差が存在する可能性に留意すべきである。



Benchmarks:
特定指標での肉薄
(GPQA, AIMEのスコア)

知識の幅
(Breadth of Knowledge):
世界の広範なドメイン知識

長文での頑健性
(Long-context Robustness):
複雑なコンテキストを維持する力

未知の問題への汎化
(Generalization):
学習データ外の推論能力

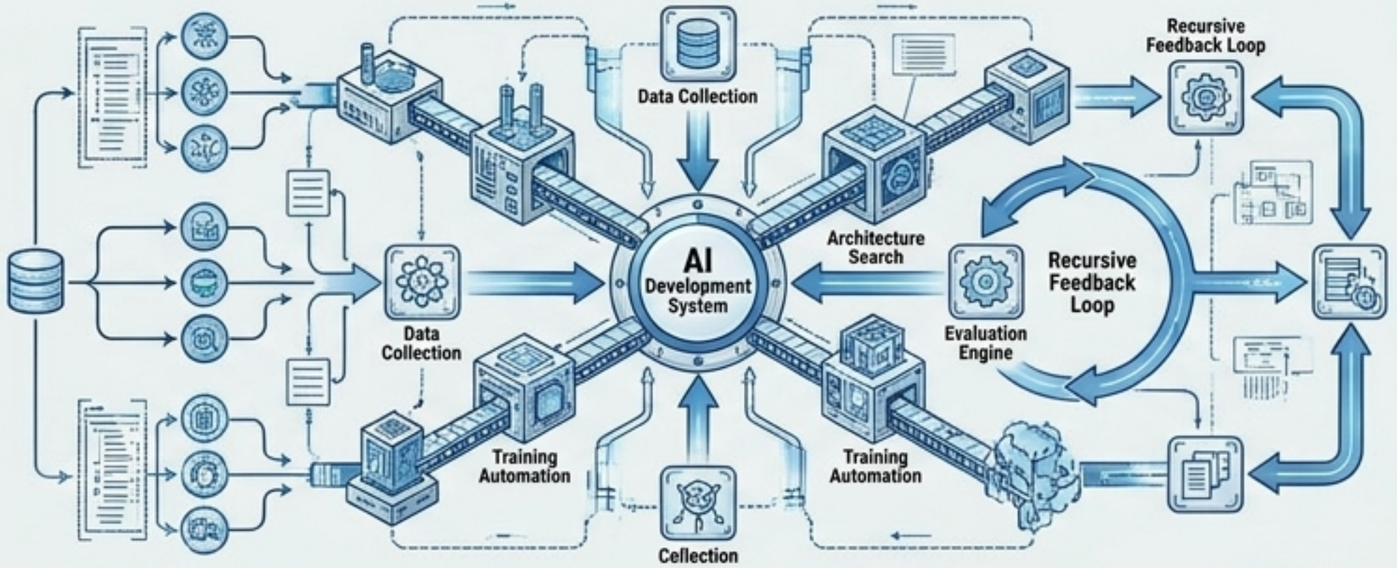
Gemini Notebook

[The Big Picture] 成果物（モデル）ではなく、製造プロセス（RSI）の証明

真のブレイクスルーは、o1に肉薄したことではない。「AIを用いてAIを開発する」自己改善プロセスの実証にある。



Model (成果物)



Automated Factory Pipeline (製造プロセス)

ELYZA RSI Researchの設立: 再帰的自己改善（Recursive Self-Improvement）を専門とする研究組織の同時立ち上げ。

開発の自動化: 単なるファインチューニングではなく、ハーネス（周辺の仕組み）から評価基盤までをAI自身が自律的に改善。

前段階の要素研究: 今回のモデルは完全なRSIに至る前段階（ステージ1〜3）の成果物としての位置づけ。

Gemini Notebook

[The Impact] RSI要素技術がもたらす超高効率化のインパクト

人的コストの破壊

1~2人日

中間学習から事後学習・評価までの全プロセスを、人間が介在する作業をわずか1~2人日に圧縮して実行完了。

自律改善による性能向上

+27.2 pt

自律改善エージェントによる評価ハーネス改善のみで、ベンチマーク成功率が最大27.2ポイント劇的向上。

実務タスクの最適化

コスト67% 削減

問い合わせ応答業務において、業務完遂率を34.3%から52.9%へ引き上げつつ、推論コストを2/3カット。

Gemini Notebook

[Market Reaction] コミュニティの初動とエコシステムの課題

公開初動のPR効果は「モデル」よりも「RSI研究」に集中。
一方で、ローカル実行に向けた開発者コミュニティの動きは極めて速い。



初動の反響 (Moderate)

公開翌日時点での反響は限定的（Hugging Faceでの初期いいね数は0）。独立検証の報告待ち状態。



量子化版の即時登場 (Fast)

公開翌日にはローカル実行向けの「GGUF版」およびMac向け「MLX 4bit版」がコミュニティから公開完了。

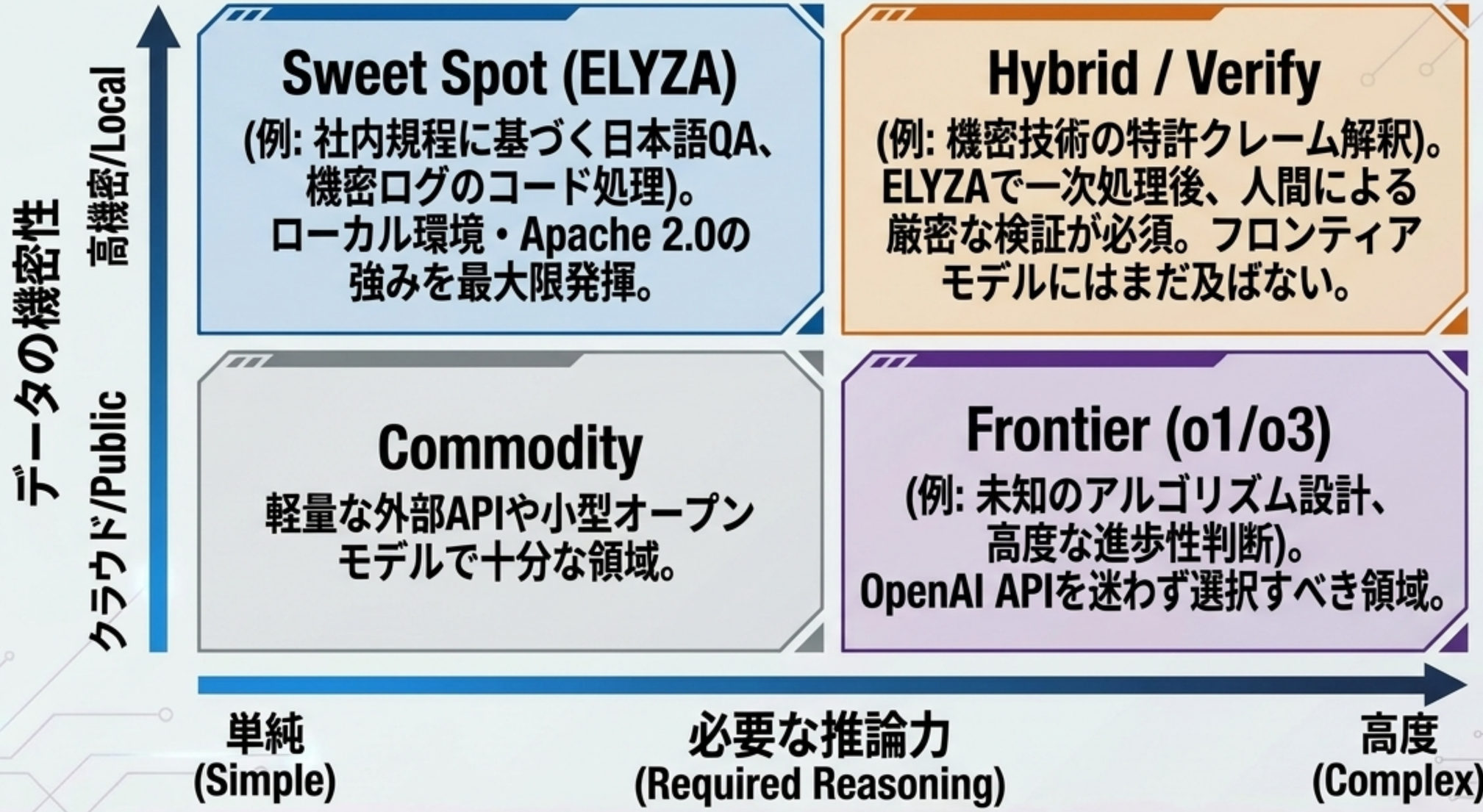


インフラ課題 (Ongoing)

標準のllama.cppではチャットAPIが失敗する現象が確認されており（10/3時点）、パッチ適用などの環境整備が今後の課題。

Gemini Notebook

[Deployment Matrix] 国産推論モデルの「実務上の使いどころ」



2023-06-01 15:36.54.293
Gemini Notebook

[Strategic Conclusion] トレードオフを理解したハイブリッド戦略の構築

Q1: 学習データや開発体制の透明性・国産性が必須か?

NO

グローバルモデル
(o1/o3/Qwen) を採用。

YES

Q2: そのタスクは「進歩性判断」など、o1レベルの高度な汎用推論を要するか?

YES

現時点では人間による
プロセス介入、またはクラウドAPI
の特例利用を検討。

NO

**ELYZA-Thinking-1.0を
ローカル環境で積極採用。**
(日本語特有の制約、コード生成、定型推論)

「国産性・ライセンス・日本語特化」という独自価値と、「純粋な推論力」の限界。
このトレードオフを正確に把握し、RSIによる将来
の爆発的進化を見据えながら「適材適所」のアーキテクチャを設計せよ。

Gemini Notebook