# ELYZA 公開推論モデルの内容と評価

OpenAI o1 と o3 との推論性能比較

調査基準日 2026 年 10 月 3 日 ｜ 公開資料に基づく調査と暫定評価

GPT-6 Astra

**ELYZA が公開した LLM-jp-4 ベースの推論モデルは、日本語の業務処理に適した国産の公開モデルとして意義がある。ただし、正式版 o1 や o3 の推論能力を上回ったと判断できる証拠はない。正式版 o1 に対しては分野によって接近しているものの総合的な優位性は未確認であり、o3 に対しては汎用的な難問推論で下回る可能性が高い。これは公表値からの推定であり、同条件の実機比較による結論ではない。[3][4][9][10]**

## 1 公開されたモデルと技術的特徴

KDDI グループの ELYZA は 2026 年 10 月 2 日、国立情報学研究所（NII）が開発した LLM-jp-4 を基盤とする 2 つの推論モデルを公開した。いずれも Apache 2.0 で、商用利用が可能である。無料公開はモデルの配布を指し、実行用の機器や運用まで無料になるという意味ではない。[1][2]

| モデル | 構成 | 位置付け |
|---|---|---|
| ELYZA-Thinking-1.0-llm-jp-4-33b | 約 330 億<br>Dense 型 | 今回の 2 モデルでは<br>性能重視 |
| ELYZA-Thinking-1.0-llm-jp-4-32b-a3b | 約 320 億<br>MoE 型 | 一部のパラメータを使い<br>計算量を抑える構成 |

出典 [3]–[5]。MoE 版の基盤モデルは総パラメータ約 321 億、稼働パラメータ約 38 億。稼働パラメータが少ないことと、保存・実行に必要なメモリ量が同じ比率で減ることは別である。

ELYZA は中間学習、教師ありファインチューニング（SFT）、検証可能な報酬による強化学習（RLVR）を実施した。数学・科学・プログラミングの問題を日本語化するとともに、日本語の複雑な指示や日本固有の知識を重点的に学習している。RLVR では正答、プログラムのテスト結果、条件遵守などを機械的に検証して報酬を与える。[3][4]

同社は、整備済みの学習基盤を使い、中間学習・事後学習・評価に人間が介在する作業を 1〜2 人日程度に抑えたと説明する。基盤モデルや学習設備を含む全開発を 1〜2 日で完了したという意味ではない。同時に設立した ELYZA RSI Research についても、再帰的自己改善（RSI）の前段階の研究であり、RSI そのものを実現したとは発表していない。[2]

## 2 公開直後の反響

公開翌日の段階では、反響は報道・紹介と初期の利用準備が中心であり、性能に関する第三者評価はまだ固まっていない。以下の動きは確認できるが、モデルの汎用推論能力が世界の上位モデルを上回ったことを示すものではない。

ローカル利用への対応として、Hugging Face には 33B 版の GGUF 版と MLX 4bit 版が既に公開されている。利用への関心を示す動きだが、量子化後の性能や実用上の安定性は別途評価する必要がある。[7]

AI Watch の発表会報道では、AI による AI 開発について、安全性への質問が集中したとされる。隔離環境や停止の仕組みなどが問われており、公開モデルの性能そのものとは別に、RSI 研究への関心が高いことがうかがえる。[8]

ITmedia は 10 月 2 日 20 時 04 分、ELYZA から説明を受け、当初の「完全国産」という表現を修正した。国内主体が開発したという説明から、学習データ・教師モデル・技術要素まで全て国内由来だと広げて解釈することはできない。[1]

調査時点で、独立した評価者が今回のモデルと o1・o3 を同条件で比較し、優位性を実証した報告は確認できていない。公開直後のため、評判と再現可能な性能検証を分けて扱う必要がある。

## 3 公表ベンチマークから見える強みと限界

ELYZA が公開した 33B 版の評価表から、特徴の出る項目を抜粋した。以下は同社の公表値であり、第三者の共通ランキングではない。各指標とも高い方が良好である。[3]

| 評価項目 | ELYZA 33B | NII LLM-jp-4.1 33B | Qwen3.5 27B |
|---|---|---|---|
| 総合スコア | 61.69 | 60.64 | 68.18 |
| GPQA Diamond<br>難しい科学問題 | 69.95 | 70.77 | 85.92 |
| AIME 2024＋2025<br>競技数学 | 68.85 | 78.23 | 93.33 |
| JEMHopQA<br>日本語の複数段階の質問応答 | 73.85 | 65.84 | 62.87 |
| JFBench<br>日本語の指示追従 | 39.02 | 32.79 | 34.12 |
| BFCL v4<br>外部ツールの呼び出し | 38.73 | 48.14 | 69.37 |

出典 [3]。総合スコアは 20 のベンチマークを 7 つの能力群に分け、群ごとの平均をさらに平均した値。BFCL の比較では全モデルの並列呼び出しを除外している。推論設定や出力上限も完全に同一ではない。

結果は、日本語の質問応答や細かな指示への対応に強みがある一方、数学・科学・ツール利用には改善の余地があることを示している。NII 版より総合スコアが高くても、難問推論まで一律に上回ったとはいえない。Qwen3.5-27B との比較でも、得意分野と不得意分野が明確である。[3]

NII も 9 月 28 日の LLM-jp-4.1 公開時に、STEM 分野、ツール呼び出し、回答の簡潔さを改善したと説明している。ELYZA 版と NII 版は、同じ基盤から異なる方向に調整されたモデルとして捉えるのが適切である。[6]

## 4 o1 と o3 との推論性能比較

比較では、2024 年 9 月に公開された o1-preview と、後に公開された正式版 o1 を区別する必要がある。ELYZA のモデルカードと OpenAI 公式資料の値を並べると、次のようになる。[3][4][9][10]

| モデル | GPQA Diamond<br>科学推論 | AIME<br>競技数学 | AIME の対象 |
|---|---|---|---|
| ELYZA 33B | 69.95 | 68.85 | 2024＋2025 |
| ELYZA 32B-A3B | 63.86 | 62.29 | 2024＋2025 |
| o1-preview | 73.3 | 42.0 | 2024 |
| 正式版 o1<br>2024-12-17 | 75.7 | 79.2 | 2024 |
| o3<br>推論設定 high | 83.3 | 88.9 | 2025 |

出典 ELYZA は[3][4]、o1 系は[9]、o3 は[10]。o1 系の AIME は pass@1、o3 の掲載値はツールなし。AIME は出題年が異なるため、この表を直接の順位表として使用できない。

同一環境での比較ではないことを前提とした暫定評価は、以下のとおりである。判断の確かさは統計的な信頼区間ではなく、公開根拠の強さについての定性的な評価である。

| 比較対象 | 推定 | 判断の確かさ |
|---|---|---|
| o1-preview | 数学などで上回る可能性はあるが、<br>総合的な超過は未確認 | 中程度以下 |
| 正式版 o1 | 一部能力は接近。汎用的な難問推論で<br>上回ったとする根拠は不足 | 中程度 |
| o3 | 今回の両モデルが汎用的な難問推論で<br>上回っている可能性は低い | 中程度から<br>比較的高い |

比較しやすい GPQA Diamond では、ELYZA 33B は正式版 o1 より約 5.8 ポイント、o3 より約 13.4 ポイント低い。ただし、プロンプト、推論に使うトークン量、試行回数などがそろっておらず、この差をそのまま統計的に確定した能力差とは扱えない。[3][9][10]

AIME は対象年が異なるため、数値差を直接の勝敗には使えない。正式版 o1 を超えたと主張するには、年度別の結果と評価条件をそろえる必要がある。また、日本語の指示追従の高さと未知の難問を解く推論力は別の評価軸であり、日本語の文書処理で使いやすいことだけから o3 以上の推論力を推定することはできない。

## 5 知財業務への適用評価

本モデルは、社内で動かす日本語文書処理モデルの候補として検証する価値がある。発明届の整理、技術要素の抽出、指定形式への転記などは試行対象にしやすい。一方、請求項の微妙な解釈、複数文献を組み合わせた進歩性の検討、長い明細書の整合性判断については、今回の一般ベンチマークから実務性能を推定する限界がある。以下は業務への適用に関する提案であり、公開された実証結果ではない。

導入判断では、同じ特許文献、同じ指示、同じ出力条件を使い、根拠箇所の正確さ、構成要件の見落とし、否定・例外条件の誤読、専門家による修正時間を比較する。量子化して利用する場合は、導入予定の量子化版で評価する必要がある。

推論能力の比較と業務適合性の比較は、目的に応じて条件を分ける。推論能力を見る場合は、資料の与え方や外部ツールの有無をそろえる。業務適合性を見る場合は、同じ作業時間や費用の範囲で、どこまで正確に完了できるかを測る。これらの実務評価が良好なら、汎用推論で o1・o3 に届かなくても、担当業務を明確にした活用価値はあり得る。

## 6 評価の範囲と留意点

本稿は 2026 年 10 月 3 日時点の公開資料に基づく暫定評価である。モデルを実際に動かした独自比較、知財分野の専用ベンチマーク、統計的有意差の検定は実施していない。公表ベンチマークの違い、モデルの版、推論設定、量子化、ツール利用の有無によって結果は変わり得る。第三者による同条件の検証が蓄積した段階で、特に正式版 o1 との位置関係を再評価することが望ましい。

## 参考文献

各 URL の参照日 2026 年 10 月 3 日。本文中の番号と対応する。更新されるモデルカードや一覧ページの数値は、参照時点の掲載内容による。

[1] ITmedia．国産 AI「LLM-jp-4」ベースの AI モデル、KDDI 傘下の ELYZA が無料公開．ITmedia AI＋。2026 年 10 月 2 日公開、同日 20 時 04 分更新。
https://www.itmedia.co.jp/aiplus/article/2610/02/2000001965/

[2] 株式会社 ELYZA．ELYZA、「AI による AI 開発」の研究を推進する「ELYZA RSI Research」を設立。基盤モデル・評価・自律改善に関する研究成果を公開．プレスリリース。2026 年 10 月 2 日。
https://elyza.ai/news/2026/10/02/rsiresearch

[3] ELYZA, Inc.．ELYZA-Thinking-1.0-llm-jp-4-33b．Hugging Face モデルカード。モデル概要、学習方法、Dense 型モデルの比較評価。
https://huggingface.co/elyza/ELYZA-Thinking-1.0-llm-jp-4-33b

[4] ELYZA, Inc.．ELYZA-Thinking-1.0-llm-jp-4-32b-a3b．Hugging Face モデルカード。モデル概要、学習方法、MoE 型モデルの比較評価。
https://huggingface.co/elyza/ELYZA-Thinking-1.0-llm-jp-4-32b-a3b

[5] LLM-jp．llm-jp-4-32b-a3b-base．Hugging Face モデルカード。基盤モデルの構成、総パラメータ数、稼働パラメータ数。
https://huggingface.co/llm-jp/llm-jp-4-32b-a3b-base

[6] 国立情報学研究所 大規模言語モデル研究開発センター．大規模言語モデル「LLM-jp-4.1」シリーズを公開．研究成果発表。2026 年 9 月 28 日。
https://llmc.nii.ac.jp/topics/llm-jp-4-1/

[7] Hugging Face．Quantized Models for elyza/ELYZA-Thinking-1.0-llm-jp-4-33b．量子化モデル一覧。GGUF 版と MLX 4bit 版の公開状況。
https://huggingface.co/models?other=base_model:quantized:elyza/ELYZA-Thinking-1.0-llm-jp-4-33b

[8] 船津稔．再帰的自己改善で AI による AI 開発を目指す「ELYZA RSI Research」設立．AI Watch。2026 年 10 月 2 日 16 時 36 分。発表会および質疑応答の報道。
https://ai.watch.impress.co.jp/docs/news/2145227.html

[9] OpenAI．OpenAI o1 and new tools for developers．2024 年 12 月 17 日。o1-2024-12-17 と o1-preview の公表評価。
https://openai.com/index/o1-and-new-tools-for-developers/

[10] OpenAI．Introducing GPT-5 for developers．2025 年 8 月 7 日。Detailed benchmarks 内の OpenAI o3 high の公表値を参照。
https://openai.com/index/introducing-gpt-5-for-developers/