国産 LLM-jp-4 ベースの推論モデル「ELYZA-Thinking-1.0」

― 公開内容・反響・評価と、o1／o3 推論水準との比較推定 ―

2026 年 10 月 3 日

Claude Opus 5.5

本書は 2026 年 10 月 3 日時点の公開情報に基づく。記述は「確認済み事実」と「分析・推定」を区別し、後者には（分析）（推定）と明記した。文中の上付き番号は文末の参考文献に対応する。

要旨

- ELYZA（KDDI 子会社）は 2026 年 10 月 2 日、国立情報学研究所（NII）の国産 LLM「LLM-jp-4」に追加学習を施した推論モデル 2 種（33B Dense／32B-A3B MoE）を、商用利用可能な Apache 2.0 ライセンスで Hugging Face に公開した[1,2,4,5]。
- 自社評価（20 種）では総合平均で NII の改良版 LLM-jp-4.1 を上回る。ただし 33B は数学・科学推論（AIME・GPQA）では LLM-jp-4.1 を下回り、優位はコード生成・指示追従にある。同サイズ帯の Qwen3.5・Gemma 4 とは総合平均で 6.5〜8.9 ポイントの差がある[4]。
- 反響は公開翌日時点で限定的である。報道の重心は同時発表の研究組織「ELYZA RSI Research」にあり、開発者コミュニティでは量子化版の公開や日本語 LLM 一覧への登録が進む一方、第三者による独立評価は確認できなかった[10,12,13,15]。
- **（推定）**33B は主要推論指標で o1-mini・o1-preview を上回るが、o1（正式版）には GPQA で約 6〜8 ポイント、AIME で約 8 ポイント及ばず、o3 とは 13〜21 ポイントの差がある。推論水準は「o1-mini と o1 の間」と推定する（第 3 章）。

1. 記事の概要（確認済み事実）

1.1 公開モデル

ITmedia AI+は、ELYZA が NII 開発の「LLM-jp-4」シリーズをベースに中間学習・事後学習を行い、「主に日本語の処理性能に焦点を当て」て性能を強化したモデルを無料公開したと報じた[1]。公開されたモデルの仕様は表 1 のとおりである。

表 1　公開モデルの仕様

項目	ELYZA-Thinking-1.0-llm-jp-4-33b	ELYZA-Thinking-1.0-llm-jp-4-32b-a3b
構造	Dense	MoE（Mixture of Experts）
パラメータ数	330 億	総 320 億（アクティブ約 30 億※）
ベースモデル	llm-jp/llm-jp-4-33b-base	llm-jp/llm-jp-4-32b-a3b-base
ライセンス	Apache 2.0（商用可）	Apache 2.0（商用可）
対応言語	日本語・英語	日本語・英語
推奨設定	temperature 0.6／top_p 0.95	temperature 0.6／top_p 0.95
ツール呼び出し	OpenAI 形式の tools に対応	OpenAI 形式の tools に対応

出典：[4,5]。※アクティブ約 30 億は名称「a3b」の慣例からの推定であり、モデルカード上の明記は確認していない。

1.2 学習手法

モデルカードによれば、学習は 3 段階で行われた[4,5]。

- **中間学習**：数学・コード・STEM 問題を日本語に翻訳して推論過程を生成し、エージェント系データも日本語化して投入。
- **教師ありファインチューニング（SFT）**：数百万件規模の合成データを使用。日本語 Wikipedia・Wikidata から日本語知識データを合成し、ふりがな・漢数字など日本語固有の制約を含む複合制約付きの指示追従データを作成。
- **強化学習**：数学・コード・ツール使用など、正誤を検証可能な難問を対象に最適化。

ELYZA は、人間が介在する作業を「1〜2 人日程度」に抑えながら、中間学習から事後学習・評価までのプロセスを実行したと説明している[2]。

1.3 性能に関する主張

ITmedia は、20 種類のベンチマークの総合スコア比較で、NII が 9 月 28 日に発表した LLM-jp-4.1 シリーズを「一部のベンチマークで上回るスコアを示した」と報じた[1]。モデルカードには、知識・STEM、数学、日本語 QA、指示追従、翻訳、コード、関数呼び出しの 7 分野・20 指標が掲載されている[4]。評価条件は、LLM-jp-4／4.1 と gpt-oss-20b を推論強度「high」、その他は思考モード ON で実行したとされる[4]。

表 2　33B モデルの主要ベンチマーク（モデルカード記載値）

指標	ELYZA 33b	LLM-jp-4 33b（参考）	LLM-jp-4.1 33b	Qwen3 32B	Qwen3.5 27B	Gemma 4 31B
MMLU-Pro（英）	77.53	75.74	78.82	78.75	85.42	85.50

指標	ELYZA 33b	LLM-jp-4 33b（参考）	LLM-jp-4.1 33b	Qwen3 32B	Qwen3.5 27B	Gemma 4 31B
GPQA-Diamond（英）	69.95	64.80	70.77	64.68	85.92	86.65
MATH-500（英）	96.00	95.00	96.80	97.80	90.60	99.20
AIME 2024+2025（英）	68.85	62.40	78.23	76.35	93.33	88.75
LiveCodeBench v6（英）	60.23	35.60	45.20	58.57	80.11	79.37
JMMLU（日）	84.69	85.22	85.49	86.09	90.11	90.17
JMATH-500（日）	90.00	90.00	92.40	92.00	93.60	94.00
IFEval（英）	95.32	87.95	93.59	88.61	91.64	96.22
Nejumi BFCL（日）	46.53	11.58	51.16	55.79	59.65	61.39
総合平均（20 指標）	61.69	51.50	60.64	58.11	68.18	70.54
日本語平均	62.18	53.04	61.78	58.97	66.05	68.88
英語平均	59.72	48.25	59.11	57.89	71.81	73.09

出典：[4]。LLM-jp-4／4.1 は「-thinking」版。表は 20 指標のうち主要指標を抜粋。

表 3　32B-A3B モデルの主要ベンチマーク（モデルカード記載値）

指標	ELYZA 32b-a3b	LLM-jp-4.1 32b-a3b	gpt-oss-20b	Qwen3.5 35B-A3B	Gemma 4 26B-A4B
GPQA-Diamond（英）	63.86	60.80	68.59	84.38	78.60
AIME 2024+2025（英）	62.29	57.71	88.23	92.92	90.62
LiveCodeBench v6（英）	53.66	32.40	59.94	74.91	77.09
JMMLU（日）	81.05	83.09	82.79	89.76	88.33
総合平均（20 指標）	58.46	53.89	57.26	66.37	68.58
日本語平均	59.61	56.65	55.04	64.24	66.68

出典：[5]。

（分析）総合平均では、ELYZA 33b は LLM-jp-4.1 33b を約 1 ポイント、32b-a3b は約 4.6 ポイント上回る。一方、33b は AIME（−9.4）・GPQA（−0.8）・MATH-500・MMLU-Pro・Nejumi BFCL などでは LLM-jp-4.1 を下回る。上回るのは LiveCodeBench（+15.0）、IFEval などであり、「一部で上回る」との表現は実態に即している[1,4]。

1.4 同時発表：ELYZA RSI Research

ELYZA は同日、AI が自身の能力や開発プロセスを反復的に改善する「再帰的自己改善（RSI）」を研究する組織「ELYZA RSI Research」の設立を発表し、安全性を重要な研究課題と位置づけた[2,3]。今

回のモデルはその成果の一つとされ、あわせて以下が公表された[2]。

- カスタマーセンター業務を音声・テキストで評価するベンチマーク「ELYZA Agent Tasks: Customer Service」の公開。
- 自律改善エージェントによるハーネス改善で、2 つのベンチマークの成功率がそれぞれ 14.4 ポイント、27.2 ポイント向上。問い合わせ応答業務では完遂率が 34.3%から 52.9%に改善し、推論コストを 67%削減。
- 研究成果を活用したバーティカル AI 事業の開始と、第一弾「ELYZA Voice Agent」の開発。

なお、ITmedia は現段階を RSI そのものではなく「前段階の要素研究」と伝えている[1]。

1.5 背景：LLM-jp-4 系列の経緯

- 2026 年 4 月 3 日：NII が LLM-jp-4 の 8B・32B-A3B を公開。約 12 兆トークンで学習し、一部ベンチマークで GPT-4o や Qwen3-8B を上回るとした[8]。
- 2026 年 8 月 18 日：33B（base／thinking）を公開[9]。
- 2026 年 9 月 28 日：事後学習を改良した LLM-jp-4.1（8B／32B-A3B／33B）を公開。STEM 性能の強化、ツール呼び出し対応、回答の簡潔化を図り、学習データも公開した[6,7]。NII は、4.1-33b は日本語・英語の総合評価で gpt-oss-120b に比肩する一方、Qwen3.8-27B や Gemma-4-31B とは差が残ると説明している[6]。

ELYZA は 2025 年 5 月にも Qwen ベースの推論モデル「ELYZA-Thinking-1.0-Qwen-32B」を公開し、「o1-mini に近い性能」と説明していた[16]。今回は、ベースを国内でゼロから学習した LLM-jp-4 に切り替えた点が異なる。

2. 反響・評価

2.1 報道

ITmedia AI+はモデル公開を主題として報じた[1]。AI Watch とビジネス＋IT（Yahoo!ニュース転載）は、RSI 研究組織の設立を見出しに据えた[10,11]。AI Watch によれば、ELYZA は AI による開発自動化の到達段階を、1000 億パラメータ未満の LLM 開発で「ステージ 1」、ハーネス（周辺の仕組み）で「ステージ 3」、評価基盤で「ステージ 2」と位置づけている。研究段階では隔離環境や緊急停止の仕組みで安全性に配慮するという[10]。

AI ニュースのキュレーション「SIGNAL10」は、10 月 3 日付の朝刊で本件を 10 本中 1 位（重要度 80/100）に選んだ。一方で「性能値は発表元の評価であり、独立検証とは限らない」と注記している[15]。

2.2 開発者コミュニティ

- LLM-jp の GitHub 組織が公開する日本語 LLM 一覧「awesome-japanese-llm」に、公開当日、両モデルを追加するプルリクエストが提出された[12]。
- 第三者によるローカル実行向けの量子化版が、公開翌日までに登場した。GGUF 版[13]と Mac 向け MLX 4bit 版[14]である。
- GGUF 版の作成者は、10 月 3 日時点で「標準の llama.cpp ではチャット API が失敗する」とし、パッチ適用と再ビルドが必要だと記している[13]。
- Hugging Face 上の「いいね」は、閲覧時点で両モデルとも 0 だった[4,5]。初動の注目度は控えめである。
- 今回の検索範囲では、第三者による独立評価や詳細な試用記事は確認できなかった（公開翌日のため）。

2.3 評価（分析）

肯定的な評価点

- 国内でゼロから学習したベースモデルに推論能力を付与し、商用可能なライセンスで公開した。LLM-jp-4 エコシステムから民間の派生モデルが生まれた具体例といえる[2,4]。
- 前作（Qwen ベース）は「o1-mini に近い」とされていた[16]。今回は国産ベースで o1-mini 超の水準に達したと推定される（第 3 章）。
- 人手 1〜2 人日で学習パイプラインを回した点は、RSI 研究の実証として、モデル性能そのもの以上に注目に値する[2]。
- 元の LLM-jp-4-33b-thinking から総合平均を約 10 ポイント引き上げた。とくにコード生成（LiveCodeBench +24.6）と関数呼び出し（Nejumi BFCL +35.0）の改善幅が大きい[4]。

慎重に見るべき点

- 33B の数学・科学推論は NII 自身の LLM-jp-4.1 を下回る。「上回る」のは総合平均と一部指標に限られる[4]。
- 同サイズ帯の海外オープンモデル（Qwen3.5、Gemma 4）とは総合平均で 6.5〜8.9 ポイント、GPQA や AIME では 15〜25 ポイントの差がある[4]。
- 数値はすべて開発者自身の評価であり、第三者の検証を待つ必要がある。
- llama.cpp への対応など、利用環境の整備はこれからである[13]。

3. o1／o3 の推論水準を上回るか（推定）

3.1 比較の方法と前提

- 比較指標は、ELYZA と OpenAI 系の双方で公表値がある英語の推論系ベンチマーク（GPQA-Diamond、AIME、MATH-500、MMLU-Pro）とし、いずれもツール不使用の値を用いた。
- ELYZA の値は同社の評価環境による測定である[4,5]。o1／o3 系の値は OpenAI の公表値[17,18,19,20]、または DeepSeek-R1 論文での測定値[22]を用いた。評価環境（試行回数、プロンプト、推論予算）が異なるため、数ポイントの差は誤差の範囲となり得る。
- AIME について、ELYZA は 2024 年と 2025 年の合算値を示している。OpenAI は年度別に公表しているため、o1／o3 は両年の単純平均で比較した（分析上の処理）。
- LiveCodeBench は版が異なる（ELYZA は v6、o1 系は DeepSeek-R1 論文時点の版）ため、比較から除外した。日本語推論ベンチマーク（JMATH-500 等）は o1／o3 の公表値がなく、比較できない。

3.2 比較表

表 4　ELYZA-Thinking と o1／o3 系モデルの推論ベンチマーク比較

指標	ELYZA 33b	ELYZA 32b-a3b	o1-preview	o1-mini	o1	o3-mini (high)	o3
GPQA-Diamond	69.95	63.86	73.3	60.0	75.7〜78.0	77.0	83.3
AIME	68.85	62.29	44.6	63.6	76.8	83.6〜87.3	90.3
（AIME の内訳）	24+25 合算	24+25 合算	2024	2024	74.3／79.2	2024	91.6／88.9
MATH-500	96.0	95.2	(85.5)	90.0	96.4	—	—
MMLU-Pro	77.5	73.1	—	80.3	—	—	—

出典：ELYZA[4,5]。o1-preview：OpenAI[17]（MATH は MATH-500 ではなく MATH 全体の値のため括弧書き）。o1-mini：DeepSeek-R1 論文[22]。o1：GPQA は DeepSeek-R1 論文の 75.7[22]と OpenAI の 78.0[19,23]、AIME は 2024 年 74.3・2025 年

79.2 の平均[19,23]、MATH-500 は[22]。o3-mini（high）：GPQA[18,20]、AIME 2024 は資料により 83.6[18]と 87.3[23]。o3：GPQA[20]、AIME は 2024 年 91.6・2025 年 88.9 の平均[19,23]。

3.3 測定条件の差の補正（分析）

ELYZA の評価表には、OpenAI の gpt-oss-20b も含まれている[5]。同モデルについて ELYZA の測定値と OpenAI の公表値を比べると、ELYZA の測定はやや辛めに出ている（表 5）。

表 5　gpt-oss-20b における測定値の差

指標	ELYZA 測定値	OpenAI 公表値（ツールなし）	差
GPQA-Diamond	68.59	71.5	−2.9
AIME（2024・2025）	88.23（合算）	92.1／91.7（平均 91.9）	−3.7

出典：[5,20,21]。

（推定）この差（約 3〜4 ポイント）を ELYZA 33b に上乗せしても、GPQA は約 73、AIME は約 72.5 にとどまる。o1（GPQA 75.7〜78.0、AIME 76.8）には届かず、結論は変わらない。

3.4 推定結論

表 6　比較対象ごとの判定（推定）

比較対象	判定	根拠（ELYZA 33b との差、ポイント）
o1-preview（2024 年 9 月）	同等以上	AIME +24.3（ただし年度構成が異なる）、GPQA −3.4
o1-mini	上回る（33b）／概ね同等（32b-a3b）	33b：GPQA +10.0、AIME +5.3、MATH-500 +6.0、MMLU-Pro −2.8。32b-a3b：GPQA +3.9、AIME −1.3、MATH-500 +5.2、MMLU-Pro −7.2
o1（正式版）	下回る	GPQA −5.8〜−8.1、AIME −7.9、MATH-500 −0.4（同等）
o3-mini（high）	下回る	GPQA −7.1、AIME −14.8〜−18.5
o3	大きく下回る	GPQA −13.4、AIME −21.4

（推定）以上から、ELYZA-Thinking-1.0-llm-jp-4-33b の英語推論水準は「**o1-mini を上回り、o1 には届かない**」と推定する。OpenAI は、o3-mini の推論強度 low が o1-mini 並み、medium が o1 並みと説明している[18]。この物差しで見れば、33b は o3-mini の low と medium の間に位置する。32b-a3b は概ね o1-mini 並みである。

（分析）数学に限れば、NII 自身の LLM-jp-4.1-33b（AIME 78.23）が o1（76.8）並みに達している[4]。ELYZA 版の強みは推論そのものより、コード生成・指示追従・日本語特有タスクにある。また、ベンチマーク上の近さは総合力の同等を意味しない。知識の幅、長文での頑健性、未知の問題への汎

化では、33B 級と o1／o3 級の差は数値以上に大きい可能性がある。

4. 留意点・限界

- ELYZA の数値はすべて開発者自身の評価であり、第三者による検証はまだない。
- Hugging Face のモデルカードの数値は、元ファイルを直接取得できず、ページ読み取りで取得した。主要値は 2 回取得して一致を確認したが、原表での再確認を推奨する[4,5]。
- o1／o3 の AIME の年度別値（74.3／79.2、91.6／88.9）は、OpenAI 公式ブログの図表を転記した二次資料[23]で確認したもので、公式ページ本文からは数値を直接確認できなかった[19]。GPQA（o3 83.3、o3-mini 77.0）は OpenAI の資料[20]で確認した。
- o3-mini（high）の AIME 2024 は、資料により 83.6[18]と 87.3[23]が併存する。
- ビジネス＋IT 転載記事[11]は本文を取得できず、見出しのみ確認した。MLX 版[14]は検索結果で存在のみ確認した。

5. 実務上の示唆（分析）

- 国内でゼロから学習したベースと Apache 2.0 の組合せにより、機密性の高い発明情報や技術文書を社内環境で扱う推論モデルの選択肢となる。学習データや開発主体の説明可能性（AI のサプライチェーン）を重視する場面で意味を持つ。
- 推論力は o1 未満と推定されるため、進歩性判断や複雑なクレーム解釈など高度な推論を要する業務では、フロンティアモデルとの使い分けと人による検証が前提となる。
- 同サイズでは海外オープンモデルが総合性能で上回る。選定は「国産性・ライセンス・日本語特有タスク」と「純粋な性能」のトレードオフで判断すべきである。

参考文献

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

[2] 株式会社 ELYZA ニュースリリース「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」2026 年 10 月 2 日. https://elyza.ai/news/2026/10/02/rsiresearch

[3] PR TIMES（株式会社 ELYZA）同上リリース, 2026 年 10 月 2 日.

https://prtimes.jp/main/html/rd/p/000000082.000047565.html

[4] Hugging Face「elyza/ELYZA-Thinking-1.0-llm-jp-4-33b」モデルカード（2026 年 10 月 3 日閲覧）. https://huggingface.co/elyza/ELYZA-Thinking-1.0-llm-jp-4-33b

[5] Hugging Face「elyza/ELYZA-Thinking-1.0-llm-jp-4-32b-a3b」モデルカード（2026 年 10 月 3 日閲覧）. https://huggingface.co/elyza/ELYZA-Thinking-1.0-llm-jp-4-32b-a3b

[6] LLM-jp「LLM-jp-4.1 モデルの公開」2026 年 9 月 28 日. https://llm-jp.nii.ac.jp/blog/llm-jp-4-1/

[7] PC Watch「NII、ツール呼び出し対応・STEM 強化の『LLM-jp-4.1』公開」2026 年 9 月. https://pc.watch.impress.co.jp/docs/news/2144150.html

[8] gihyo.jp「NII、約 12 兆トークンのコーパスで学習した『LLM-jp-4 8B モデル』『LLM-jp-4 32B-A3B モデル』をオープンソースライセンスで公開」2026 年 4 月. https://gihyo.jp/article/2026/04/llm-jp-4

[9] カレントアウェアネス・ポータル「国立情報学研究所（NII）、新たな大規模言語モデル（LLM）『LLM-jp-4 33B』を公開」2026 年 8 月. https://current.ndl.go.jp/car/283665

[10] AI Watch「再帰的自己改善で AI による AI 開発を目指す『ELYZA RSI Research』設立 開発には安全性に配慮」2026 年 10 月. https://ai.watch.impress.co.jp/docs/news/2145227.html

[11] ビジネス＋IT（Yahoo!ニュース転載）「KDDI 傘下 ELYZA、AI の『再帰的自己改善（RSI）』研究組織『ELYZA RSI Research』を設立」2026 年 10 月. https://news.yahoo.co.jp/articles/45be3bcda4c5edf64cb07dbc6bdde89e87587524 ※本文未確認（見出しのみ）

[12] GitHub llm-jp/awesome-japanese-llm, Pull Request #726「ELYZA-Thinking-1.0-llm-jp-4-33b と 32b-a3b を追加」2026 年 10 月 2 日. https://github.com/llm-jp/awesome-japanese-llm/pull/726

[13] Hugging Face「na0x2c6/ELYZA-Thinking-1.0-llm-jp-4-32b-a3b-GGUF」（2026 年 10 月 3 日閲覧）. https://huggingface.co/na0x2c6/ELYZA-Thinking-1.0-llm-jp-4-32b-a3b-GGUF

[14] Hugging Face「hjmr/ELYZA-Thinking-1.0-llm-jp-4-33b-MLX-4bit」. https://huggingface.co/hjmr/ELYZA-Thinking-1.0-llm-jp-4-33b-MLX-4bit ※検索結果で存在のみ確認（未閲覧）

[15] SIGNAL10「AI 朝刊 2026.10.03」GitHub Takuya-ops/signal10, Issue #31. https://github.com/Takuya-ops/signal10/issues/31

[16] Web 担当者 Forum「ELYZA が日本語での論理的思考能力を強化した LLM『ELYZA-Thinking-1.0-Qwen-32B』を開発」2025 年 5 月 2 日. https://webtan.impress.co.jp/n/2025/05/02/49117

[17] OpenAI「Learning to Reason with LLMs」2024 年 9 月 12 日. https://openai.com/index/learning-to-reason-with-llms/

[18] OpenAI「OpenAI o3-mini」2025 年 1 月 31 日. https://openai.com/index/openai-o3-mini/

[19] OpenAI「Introducing OpenAI o3 and o4-mini」2025 年 4 月 16 日. https://openai.com/index/introducing-o3-and-o4-mini/

[20] OpenAI「gpt-oss main capabilities」（gpt-oss モデルカード付属図）. https://deploymentsafety.openai.com/data/eval-sets/gpt-oss/assets/capabilities_plots/gptoss_main_capabilities_1.pdf

[21] OpenAI「gpt-oss-120b & gpt-oss-20b Model Card」arXiv:2508.10925, 2025 年 8 月. https://arxiv.org/abs/2508.10925

[22] DeepSeek-AI「DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning」arXiv:2501.12948, 2025 年 1 月. https://arxiv.org/abs/2501.12948

[23] Presidentlin「llm-pricing-calculator」src/lib/benchmarks/openai.ts（Hugging Face Spaces）.

https://huggingface.co/spaces/Presidentlin/llm-pricing-calculator/blob/7dfbc6e3b487affada22cc89b7af5611324f36b1/src/lib/benchmarks/openai.ts ※OpenAI 公式ブログ図表を転記した二次資料