ChatGPT

ELYZA「LLM-jp-4」ベース推論モデルの技術評価・反響調査と o1／o3 比較

エグゼクティブサマリ

2026年10月2日、KDDIグループのELYZAは、国立情報学研究所（NII）LLM-jpプロジェクトの `llm-jp-4` ベースモデルに独自の中間学習・事後学習を施した推論モデル **`ELYZA-Thinking-1.0-llm-jp-4-33b`** と **`ELYZA-Thinking-1.0-llm-jp-4-32b-a3b`** をHugging Faceで公開した。ITmediaの記事は同日15時20分に公開され、20時04分に「完全国産」という当初表現をELYZA側の説明を受けて訂正している。両モデルはApache License 2.0で、商用利用可能なオープンウェイトモデルである。 [1]

本件で技術的に最も重要なのは、「LLM-jp-4をそのまま公開した」のではなく、ELYZAが **Mid-training → 数百万件規模のSFT → RLVR（検証可能報酬による強化学習）** という3段階を追加した点である。日本語化した数学・STEM・コーディング推論、日本語の複雑な指示追従、Wikipedia/Wikidata由来の日本語知識、ツール呼び出しが強化されている。 [2]

公表された20ベンチマークの集約評価では、33B Denseモデルが **Overall 61.69**、32B-A3B MoEモデルが **58.46**。同一評価系の `llm-jp-4` に対して、それぞれ **+10.19点、+12.08点**、さらに新しい `llm-jp-4.1` に対しても **+1.05点、+4.57点** の総合平均差を出している。ただし全項目で4.1を上回るわけではなく、33BではGPQA、AIME、数学、function callingなど複数項目で4.1の方が強い。むしろELYZA版の特徴は、**日本語QA・厳密な指示追従・コーディングの一部を大きく改善し、総合点を押し上げたこと**にある。 [3]

一方、「o1/o3を超えたか」という問いについては、**現時点の証拠からは否定的に評価するのが妥当**である。ユーザー指定どおり「o1」「o3」の具体的定義は**未指定**なので、本稿では比較の便宜上、OpenAIのフルサイズ版スナップショット **`o1-2024-12-17`** と **`o3-2025-04-16`** を暫定参照する。o1は公式GPQA Diamondで75.7、ELYZA 33Bは同名ベンチマークで69.95、MoEは63.86であり、評価ハーネス差を差し引いても「ELYZAがo1を上回った」という証拠にはならない。OpenAIはさらにo3について「同一レイテンシ・コスト条件のo1より高性能」と公式に説明しているため、**汎用的な高度推論では o3 > o1 ≳ ELYZA 33B > ELYZA MoE と推定するのが現時点では最も保守的**である。 [4]

ただしこれは「ELYZAモデルの価値が低い」という意味ではない。ELYZAモデルは **重みを自社管理できること、日本語に強いこと、Apache 2.0、自社ネットワーク内で運用可能であること、ツール呼び出しを備えること**が価値の中心であり、API型フロンティアモデルとは競争軸が異なる。特に機密情報を外部APIに送れない日本企業、RAG、社内文書QA、定型業務エージェント、モデル改変・追加学習が必要な用途では、有力な評価候補である。 [5]

なお、公開から調査時点の2026年10月3日まで約1日しか経っておらず、**コミュニティによる独立再現評価はまだほぼ成立していない**。Hugging Faceでは33Bが約390 downloads/8 likes、MoEが約404 downloads/7 likesという初動が確認できるが、これは導入実績を意味しない。Qiita、Hacker News、Redditで対象モデル名に一致する有意な技術検証記事は公開Web検索では確認できず、企業による外部導入発表も確認できなかった。 [6]

総合判定

評価軸	判定	コメント
日本語オープンウェイトモデルとしての技術価値	**高い**	llm-jp-4からの改善が大きく、Apache 2.0、64K context、tool calling
LLM-jp-4からの性能向上	**明確**	Overall +10.19点（33B）、+12.08点（MoE）
LLM-jp-4.1との差	**総合ではELYZA優位、項目別は混在**	33B +1.05、MoE +4.57。ただしGPQA/AIME等では逆転あり
o1フルモデル超え	**証拠なし／下回る可能性が高い**	共通名称のGPQA Diamondで69.95 vs o1 75.7
o3フルモデル超え	**可能性はさらに低い**	OpenAI公式ではo3が同条件のo1より高性能
オンプレ・閉域導入適性	**高い**	重み公開、自前vLLMサーバが可能
即時の商用本番投入	**条件付き**	安全性評価、実機性能、業務固有E2E評価が不足
現時点でのコミュニティ評価成熟度	**低い／時期尚早**	公開翌日で第三者評価が少ない

記事内容と公式ソースの照合

ITmedia記事で何が報じられたか

対象のITmedia AI＋記事は **2026年10月2日15時20分公開、同20時04分更新**。記事によるとELYZAは同日、NIIの国産オープンLLM「LLM-jp-4」をベースに性能を強化した2モデルを無料公開した。両モデルともApache 2.0で、Hugging Faceから取得できる。記事はまた、同日に設立された「ELYZA RSI Research」の研究成果の一つであること、ただし現段階の成果はRSIそのものではなく前段階の要素研究であることも明記している。 [7]

特に重要なのはITmediaの**変更履歴**である。当初モデルカードの “Fully domestic foundation” を根拠に「完全国産AI」「完全国産基盤」と記述したが、ELYZAから「国内の企業で作成したという意図であり、『完全国産』という日本語のニュアンスは想定と異なる」と説明を受け、20時04分にタイトル・本文を修正した。したがって本モデルを「部材・データ・ソフトウェアスタックまで完全に国産」と理解するのは不適切である。 [8]

実際、翌日までの一部二次記事にはなお「完全国産AI」とする表現が残っており、今回の発表では性能以上に**“国産”の定義をどう扱うかという情報発信上の注意点**が表面化した。 [9]

基本仕様の確認結果

項目	33B Dense	32B-A3B MoE	確認結果
正式名称	ELYZA-Thinking-1.0-llm-jp-4-33b	ELYZA-Thinking-1.0-llm-jp-4-32b-a3b	公式確認
公開日	2026-10-02	2026-10-02	公式確認

項目	33B Dense	32B-A3B MoE	確認結果
発表主体	ELYZA	ELYZA	公式確認
ベース	llm-jp-4-33b-base	llm-jp-4-32b-a3b-base	公式確認
構造	Dense	MoE、128 routed experts中8 activated	NII親モデル仕様
総パラメータ	33,219,548,160	32,139,028,992	NII親モデル仕様
1 token当たりactive params	約33.2B	約3.827B	NII親モデル仕様
Context	65,536 tokens	65,536 tokens	NII親モデル仕様
言語	日本語・英語	日本語・英語	ELYZA model card
主入力	テキスト	テキスト	model card
Reasoning	常時Thinking	常時Thinking	model card
Tool calling	対応	対応	model card
Parallel tool calling	評価系では非対応	同左	model card
ライセンス	Apache-2.0	Apache-2.0	公式確認
配布	Hugging Face重み	Hugging Face重み	公式確認
ELYZA運営の公開推論API	**未指定／確認できず**	同左	2026-10-03時点
HF Inference Provider	**未提供**	**未提供**	HFが明示
自前API化	vLLM等で可能	vLLM等で可能	公式Quickstart
ELYZAによるオンプレ製品提供	**未指定**	**未指定**	公表資料に記載なし

構造・パラメータ数について、NIIの親モデルカードではDense 33Bが64層、hidden size 5,120、40 heads、65,536 context、総33.2195B parameters。MoEは32層、hidden size 2,560、40 heads、128 experts中8 expertsをtokenごとに活性化し、総32.1390Bに対してactive parameterは3.8275Bとされる。 10

Hugging FaceのELYZA model cardは33Bを「Japanese and English language understanding and generation」のreasoning modelとし、日本語推論・指示追従、日本語知識を重点領域としている。ライセンス表示はいずれもApache-2.0である。 11

APIとオンプレ提供を分けて理解する必要がある

ここには誤解しやすい点がある。

「公開APIがない」ことと「APIとして使えない」ことは別である。

2026年10月3日時点で、Hugging Faceは両ELYZAモデルについて “This model isn't deployed by any Inference Provider” と表示しており、ELYZA公式発表にもELYZA運営の有料・無料推論APIは記載されていない。 [12]

しかしmodel cardにはvLLMを使って65,536 contextでサーバを起動する手順があり、OpenAI互換形式の `/v1/chat/completions` 相当で利用でき、標準的な `tools` フィールドを使ったfunction callingも可能である。したがって**重みを自社サーバ／クラウド／閉域環境に置くセルフホストは明確に可能**である。 [13]

一方、NIIの **LLM-jp-4.1** については別件として東京大学mdx-MaaS経由の研究者向けAPI・チャットサービスを2026年10月中旬頃から段階提供する計画が発表されている。これは今回のELYZAモデルのAPIではない。 [14]

公式ソース監査

ソース	確認	内容
ITmedia対象記事	**確認済み**	公開、ライセンス、モデル名、記事訂正
ELYZAプレスリリース	**確認済み**	RSI Research、モデル公開、開発プロセス
KDDI本体の本件専用プレスリリース	**確認できず**	調査時点ではELYZA発表が一次資料
Hugging Face 33B model card	**確認済み**	学習法、ベンチマーク、実行法、ライセンス
Hugging Face MoE model card	**確認済み**	同上
NII LLM-jp-4 base model card	**確認済み**	architecture、tokenizer、11.7T tokens等
NII LLM-jp-4.1発表	**確認済み**	比較対象、4.1の改良点、API予定
GitHub上の本モデル専用weight repository	**確認できず**	モデル本体はHugging Face配布
GitHubの関連benchmark	**確認済み**	ELYZA Agent Tasks: Customer Service
Apache 2.0正式ライセンス文書	**確認済み**	ASF正式文書

ELYZAの公式発表では、社内学習基盤にLLM-jp-4を適用した結果、人間の介在作業を約1〜2人日に抑えつつMid-training、Post-training、評価まで実行したとしている。これは**モデルの推論速度が高速という意味ではなく、「モデル開発工程の人的工数を削減した」という数字**なので混同すべきではない。 [15]

技術仕様、学習方法、ベンチマークの詳細評価

トークナイザーと事前学習

LLM-jp-4のtokenizerは、Hugging Face Tokenizersベースの **Unigram byte-fallback** 方式で、語彙は `llm-jp-tokenizer v4.0` から変換されている。OpenAI Harmony互換のresponse formatを意識したchat templateを持つ一方、`openai-harmony` が想定するtokenizer自体とは異なり、直接互換ではない。 [16]

親モデルのpre-training＋mid-trainingは合計 **11.7兆tokens**。コーパスの大部分は公開されているが、ライセンス上の制約で一部は公開対象外である。またNINJAL Web Japanese Corpus（whole-NWJC）を使用したこともNIIが明示している。 [17]

ELYZAが追加学習に使ったデータについては、model cardから次の範囲まで確認できる。

Mid-trainingでは、数学・コーディング・STEM問題を日本語化し、新しいreasoningを生成しつつ元の英語データも維持。agentic dataも、tool calls、function names、arguments、JSON構造を壊さず日本語化している。 [2]

SFTでは **数百万件（several million）のsynthetic examples with thinking traces** を利用し、その約半分を高品質日本語データとした。teacher modelが回答を生成し、数学的同値性、unit tests、rule checkersなどでプログラム検証に合格した回答を残した。また、振り仮名や漢数字など日本語固有条件を含む多制約instructionを作り、Wikipedia/Wikidataから日本語知識をsynthetic生成している。 [2]

最後のRLVRでは、数学、coding、日本語知識、複雑なinstruction following、tool useについて、正解判定可能な問題をまとめて学習し、回答一致・test execution・constraint checker・tool call matching等からプログラム的にrewardを計算する。reward hackingにつながる抜け道も訓練前に探して閉じた、とELYZAは説明している。 [18]

ただし、**teacher modelの正体、Mid-trainingの正確なtoken数、SFTの正確な件数、RLVRの問題数・training compute・GPU数・学習時間は今回確認した公式資料では未指定**である。

ベンチマークの主要結果

ELYZA model cardのOverallは、20個を単純平均した値ではなく、複数ベンチマークを7能力カテゴリ内でまとめた上で集約した値である。このため「61.69 = 全20問系の単純平均」と読むべきではない。 [19]

Dense 33B

Benchmark	ELYZA 33B	LLM-jp-4 33B	LLM-jp-4.1 33B	ELYZA vs 4.1
MMLU-Pro en	77.53	75.74	**78.82**	-1.29
JMMLU ja	84.69	85.22	**85.49**	-0.80
GPQA-Diamond en	69.95	64.80	**70.77**	-0.82
GPQA ja	60.97	59.54	**64.19**	-3.22
MATH-500	96.00	95.00	**96.80**	-0.80
JMATH-500	90.00	90.00	**92.40**	-2.40
AIME 2024+2025	68.85	62.40	**78.23**	-9.38
JamC-QA ja	**66.35**	57.30	57.69	+8.66
JEMHopQA ja	**73.85**	63.19	65.84	+8.01
IFEval en	**95.32**	87.95	93.59	+1.73
IFBench en	**69.40**	54.58	61.63	+7.77
M-IFEval-ja	**85.84**	67.48	80.53	+5.31
JFBench ja	**39.02**	30.29	32.79	+6.23
LiveCodeBench v6	**60.23**	35.60	45.20	+15.03

Benchmark	ELYZA 33B	LLM-jp-4 33B	LLM-jp-4.1 33B	ELYZA vs 4.1
JHumanEval ja	**95.73**	95.18	95.61	+0.12
BFCL v4	38.73	12.45	**48.14**	-9.41
Nejumi BFCL ja	46.53	11.58	**51.16**	-4.63
Overall	**61.69**	51.50	60.64	**+1.05**
Japanese	**62.18**	53.04	61.78	**+0.40**
English	**59.72**	48.25	59.11	**+0.61**

出典はELYZA公式model card。 [20]

この表から見えるのは、「ELYZA版が4.1を全面的に凌駕した」のではないということだ。4.1はMMLU-Pro、JMMLU、GPQA、数学、AIME、function calling等で優勢。それでもELYZAは、**日本語QA、instruction following、LiveCodeBenchを大幅に改善したため、カテゴリ集約後のOverallで僅差ながら逆転している。**これはELYZAの追加学習の狙いと整合する。 [21]

MoE 32B-A3B

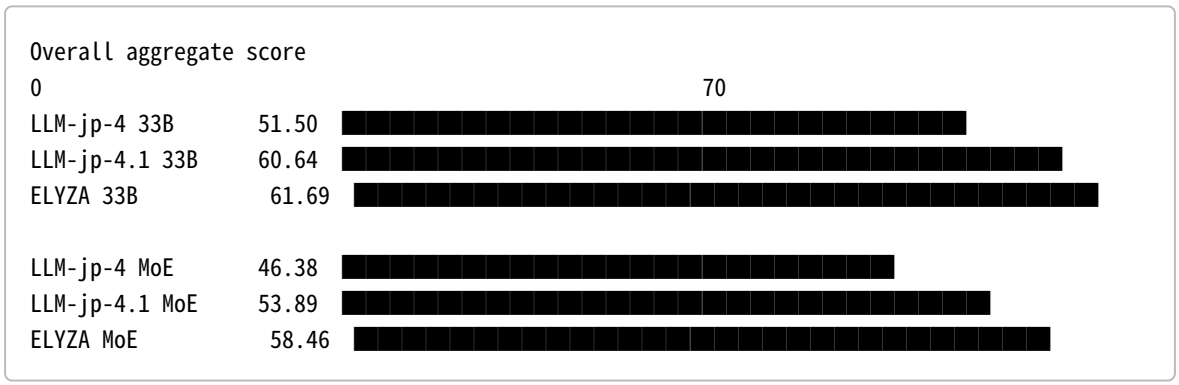
Benchmark	ELYZA MoE	LLM-jp-4 MoE	LLM-jp-4.1 MoE	ELYZA vs 4.1
MMLU-Pro	73.14	69.70	**73.90**	-0.76
JMMLU	81.05	81.32	**83.09**	-2.04
GPQA-Diamond	**63.86**	51.93	60.80	+3.06
GPQA ja	**57.24**	51.03	57.06	+0.18
MATH-500	**95.20**	84.20	93.80	+1.40
JMATH-500	**88.20**	83.20	87.40	+0.80
AIME 2024+2025	**62.29**	34.48	57.71	+4.58
JamC-QA	**59.51**	53.75	53.79	+5.72
JEMHopQA	**71.59**	62.71	65.51	+6.08
M-IFEval-ja	**81.75**	63.05	77.32	+4.43
LiveCodeBench v6	**53.66**	26.86	32.40	+21.26
BFCL v4	**36.31**	13.29	27.96	+8.35
Overall	**58.46**	46.38	53.89	**+4.57**
Japanese	**59.61**	49.16	56.65	**+2.96**
English	**55.94**	41.16	49.72	**+6.22**

[22]

MoE版は33B以上に改善幅が目立つ。特に親LLM-jp-4 MoEからOverall **+12.08 points**、4.1からも **+4.57 points**。計算時のactive parametersが約3.8Bであることも考えると、**「絶対性能では33B、推論効率候補ではMoE」という使い分け**が自然である。ただしMoEだからメモリ使用量も3.8B相当になるわけではない。総weightは約32.1Bあり、後述のとおり格納メモリはDense 33Bと大差ない。 [23]

ベンチマークをグラフ化すると

同一ELYZA評価ハーネス内のOverall比較である。



[19]

ここでは同一model cardの評価条件を使っているため比較価値が比較的高い。一方、後述するo1/o3との比較では評価環境が異なるため、このようなOverall値を直接並べることはしない。

「AIによるAI開発」の数字をモデル性能と混同しない

ELYZAは別成果として日本語カスタマーセンター向け `ELYZA Agent Tasks: Customer Service` を公開している。これは音声・テキスト、SOP選択・手順遵守・tool use・最終処理結果を評価するagent benchmarkである。 [24]

プレスリリースでは、agent harnessの改善により開発用benchmarkで+14.4/+27.2 points、customer support testのtask completionを34.3%から52.9%へ改善し、別のaccuracy/cost最適化ではaccuracyをほぼ維持してinference costを67%削減したとする結果も掲載されている。これらは**今回公開された2モデル単体のtoken throughputや推論コストのbenchmarkではなく、ELYZAのagent/harness研究の結果**である。したがって「ELYZA 33Bは67%安い」などと読み替えるのは誤りである。 [25]

コミュニティ、メディア、導入面での反響

反響は「肯定的だが、まだ評価というより紹介」の段階

公開翌日である2026年10月3日という調査時点を考える必要がある。検索できた日本語報道・技術解説の中心は、「Apache 2.0で公開された」「LLM-jp-4/4.1より一部指標が改善」「自社ホスト可能」という一次資料の整理であり、長期間の実運用や第三者による大規模再現実験ではない。 [26]

FIXITの技術解説は、33B Overall 61.69、MoE 58.46、MoE active約3.8Bという仕様を整理しつつ、「重み公開であり自社ホスト前提」「customer-service関連の数値はELYZA自身の実験で第三者検証ではない」と慎重な読み方をしている。この点は本調査の評価とも一致する。 [27]

AI-Revolutionも、33Bは総合では4.1を上回るが、数学・tool callingなどでは4.1に負ける項目があり、Qwen3.5/Gemma 4など他の新しい公開モデルが総合平均でさらに上位にいることを指摘している。実際、ELYZA自身の表でも33B Overall 61.69に対しQwen3.5-27B 68.18、Gemma 4 31B 70.54であり、**世界のオープンウェイト最先端を総合性能で更新したモデルではない**。 [28]

チャネル別の調査結果

チャネル	2026-10-03時点の確認結果	評価
ITmedia	詳細報道、当日「完全国産」表現を訂正	重要な一次報道
SBクリエイティブ系メディア	33B 61.7、MoE 58.5、1〜2人日等を紹介	主として発表内容の報道
日本語技術ブログ	FIXIT、AI-Revolution等が早期解説	有用だが大部分は一次資料分析
X/Twitter	紹介・歓迎系の投稿を少数確認	まだ定量評価には使えない
Qiita	exact model名の有意な検証記事を公開検索で確認できず	**未確認**
Hacker News	exact model名の有意なdiscussionを確認できず	**未確認**
Reddit	exact model名の有意なdiscussionを確認できず	**未確認**
Hugging Face	33B 約390 downloads/8 likes、MoE 約404/7	初動関心のみ
外部企業の本番採用発表	確認できず	**未指定／未確認**

Hugging Faceの数字は公開から約1日後という極端に短い期間のため、「390社が導入した」などとは解釈できない。download countは試用、CI、再download等を含みうるプラットフォーム指標であり、**企業採用の証拠ではない**。 [6]

また、検索エンジンによるX、Reddit等のindexは完全ではないため、「投稿が存在しない」のではなく、**本調査の公開Web検索で有意な議論を確認できなかった**という意味である。

「ELYZA Voice Agent」を採用実績と扱うべきではない

ELYZAは同日の発表で、自律研究基盤の研究成果をバーティカルAI事業に活用し、その第一弾として「ELYZA Voice Agent」を開発したとする。 [25]

しかしプレスリリースは、Voice Agentが今回公開した **この2つのweightを直接production modelとして使用している**とは明記していない。したがって、Voice Agentを「ELYZA-Thinking-llm-jp-4の商用導入事例」とカウントすることはできない。これは現時点では**未指定**である。

現状の評価を一言で表すなら

「技術的な初期評価は良好。しかし“評判が確立したモデル”ではまだない」。

特に公開翌日という時間軸から、現在目立つ記事数やSNS言及量より、今後数週間で出てくる以下のデータの方がはるかに重要になる。

第三者による同一promptでの日本語生成品質比較、実GPUでのtokens/secとTTFT、量子化後の品質劣化、長文64Kでの実効性能、RAG hallucination率、tool-call success rate、安全性benchmark、業務データでのend-to-end成功率が揃って初めて、本格的な評価が可能になる。

セキュリティ、倫理、ライセンス上の評価

Apache 2.0はかなり使いやすいが「無条件」ではない

ELYZAの両model cardはApache-2.0を明記している。Apache Software Foundationの正式文書では、使用・複製・派生物作成・配布等について世界的・無償・非独占的なcopyright licenseを付与し、一定範囲のpatent licenseも含む。再配布時にはライセンス文書、変更表示、関連する著作権・特許・商標・attribution notice、NOTICEが存在する場合はそのattributionを保持する義務がある。 [29]

同時にApache 2.0は商標利用権を与えず、成果物を **AS IS** で提供し、非侵害性・商品性・特定目的適合性等の保証をしない。したがって「Apache 2.0だから出力内容まで著作権・特許リスクが完全免責になる」という理解は誤りである。 [30]

Training-data provenanceは相当透明だが完全ではない

LLM-jp-4は11.7T tokenのpre/mid-training corpusの大部分を公開している点で透明性が高い一方、NII自身が「ライセンス制約で公開できない部分がある」と明記する。 [16]

ELYZA追加学習も手法・データ種類はかなり具体的だが、今回確認したmodel cardからは、teacher model名、全datasetの出典一覧、正確なdataset構成比、全生成データのprovenanceを完全には追跡できない。したがって、金融・医療・公共調達などprovenance要件の厳しい用途では、Apache 2.0表示だけで法務審査を終えるべきではない。

Safety評価には大きな空白がある

親LLM-jp-4のNII model cardは、自身のモデル群について、研究開発の早期段階であり、**human intentやsafety considerationsへのalignmentを保証する調整はしていない**と明示している。 [31]

ELYZA版にはSFT/RLVRが追加されているため、NIIの注意書きをそのまま「ELYZA版も安全でない」と転用することはできない。しかし、今回確認したELYZA model cardには、AnswerCarefully、jailbreak rate、harmful request refusal、CBRN、cyber misuse、bias/toxicity等の**独立した詳細安全性スコアは掲載されていない**。よって「追加学習で安全性も改善済み」と推定する根拠も不足している。

ELYZA自身はRSI Researchについて、想定外の挙動と悪用リスクを明示し、安全性を主要研究課題とするとしている。これはガバナンス面では好材料だが、個別モデルの安全認証とは別である。 [32]

Tool callingは攻撃面を広げる

両モデルはtool callingをサポートする。agentとして外部DB、メール、社内API、shell、browser等につなげれば、単なるtext-generation modelよりprompt injection・権限昇格・誤操作の影響が大きくなる。model card自身もfunction/tool callを主要機能に含めている。 [2]

実運用ではモデルに直接権限を与えるのではなく、

Model → policy/permission layer → schema validation → human approval for high-impact action → tool

という分離が望ましい。

さらにELYZAの推奨vLLM起動例では `--trust-remote-code` と、repositoryからdownloadしたreasoning/tool parser pluginを使用する。 [33]

これは便利である一方、**モデル重みだけでなくremote repository由来の実行コードも信頼境界に入る**ということを意味する。企業環境ではrepository revision/commitを固定し、plugin codeをreview・vendor化し、hashを固定してからproductionに持ち込むべきである。これはmodel card記載の実行方法から導かれるセキュリティ上の推論である。

reasoning traceの扱いにも注意

公式Quickstartでは最終回答を `content`、reasoningを別の `reasoning` fieldとして返す。 [2]

そのため社内システムでraw reasoningを無条件にユーザー表示・ログ保管すると、promptの機密要素、内部処理、tool結果等を不要に蓄積する可能性がある。productionでは「最終回答」と「内部推論ログ」を別データクラスとして扱い、保存期間・アクセス権・監査方針を設定する方が安全である。

o1・o3との推論性能、速度、メモリ、コスト比較

比較対象の定義

ユーザー指定に従い、**「o1モデル」「o3モデル」の具体的定義自体は未指定**とする。

ただし数値比較を成立させるため、本節だけ次を**暫定参照モデル**とする。

呼称	本稿の暫定参照	注意
o1	`o1-2024-12-17`	o1-mini / o1-proではない
o3	`o3-2025-04-16`	o3-mini / o3-proではない
ELYZA	33B Dense / 32B-A3B	2026-10-02公開weights

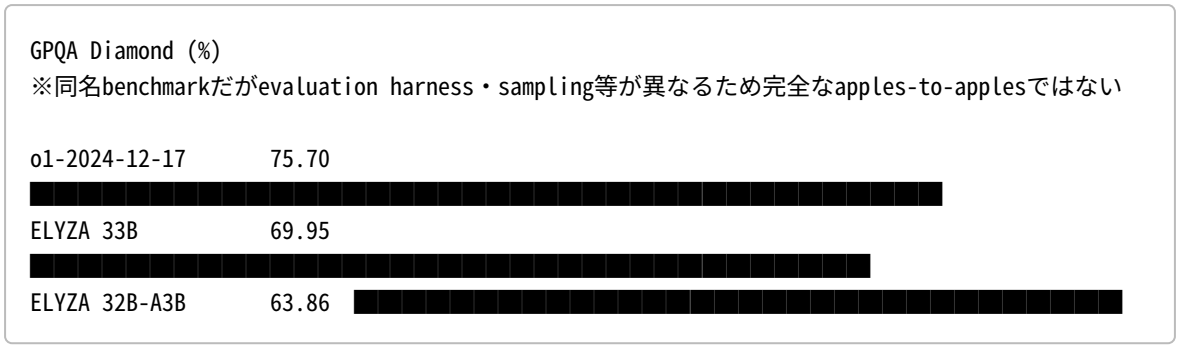
OpenAIの現行ドキュメントでは、2026年10月時点でo1とo3はいずれもdeprecated表示で、o3はGPT-5にsuccededされたモデルとして扱われている。したがってここでの比較は**現行OpenAI最高性能モデルとの比較ではなく、ユーザーが指定した歴史的o-seriesとの比較**である。 [34]

精度比較で最も有効なのはGPQA Diamond

o1公式発表は `o1-2024-12-17` について、GPQA Diamond **75.7**、MMLU **91.8**、MATH **96.4**、AIME 2024 **79.2**、SWE-bench Verified **48.9** 等を公表している。 [35]

ELYZAではMMLUではなくMMLU-Pro、MATHではMATH-500、AIMEでは2024+2025、codingではLiveCodeBench v6を使うため、それらをo1の数字と直接比較することはできない。

比較可能性が比較的高い共通名称benchmarkがGPQA Diamondである。



OpenAI o1: [36]
ELYZA: [37]

差は、

- ELYZA 33B − o1 = **−5.75 percentage points**
- ELYZA MoE − o1 = **−11.84 points**

となる。

ただしこれは**同一ハーネスの対照試験ではない**ので「o1が正確に5.75点上」と断定することはできない。それでも「ELYZA 33Bがo1を超えた」という主張を支える証拠にはならず、むしろ逆方向の証拠である。

o3について

OpenAIはo3について、coding、math、science、visual reasoningなどで強力なreasoning modelと位置づけ、o3/o4-mini世代ではtoolをagentically組み合わせる能力も追加したとしている。 [38]

さらに重要なのは、OpenAIが、

同じlatencyとcostのo1との比較でもo3のperformanceは高く、より長く考えさせるとさらに向上する

という検証結果を公表している点である。 [39]

従って、

1. ELYZA 33Bは少なくとも共通GPQA Diamondではo1を下回る。
2. OpenAIの内部/公開評価ではo3は同条件のo1より高性能。
3. ELYZA側にはo3を直接破った共通benchmarkは提示されていない。

という3点から、**「ELYZAモデルの汎用推論能力がフルo3を超えている」と推定するのは合理的でない**。

能力別の比較推定

能力	ELYZA 33B vs o1	ELYZA 33B vs o3	確信度
一般的な高度推論	**下回る可能性が高い**	**下回る可能性が非常に高い**	中〜高

能力	ELYZA 33B vs o1	ELYZA 33B vs o3	確信度
科学・難問推論	下回る可能性が高い	下回る可能性が高い	中〜高
数学	o1優勢と推定	o3優勢と推定	中
coding	o1/o3優勢の可能性	o3優勢が有力	中
日本語固有知識	**不明**	**不明**	低
日本語の細かい制約追従	競争力あり得るが共通評価なし	不明	低
Tool calling/agentic capability	対応するがELYZAには制約あり	o3優勢と推定	中
Vision reasoning	**ELYZA対象外**	o1/o3対応	高
重みの自己保有・改変	**ELYZAが圧倒的に有利**	**ELYZAが圧倒的に有利**	高
閉域・オンプレ	**ELYZA優位**	OpenAI APIとは性質が異なる	高

o1はvision、function calling、structured outputsをサポートし、o3はtext/codeだけでなくimage reasoningとagentic tool useを強化している。一方、ELYZA今回モデルはtext generationを中心とした公開weightであり、画像reasoningの同等機能はmodel cardで示されていない。 [40]

したがって「言語理解／生成」という狭い日本語タスクではELYZAに勝機があっても、**multimodal・agentic・scientific reasoningを含めた総合モデル能力ではo1/o3を超えたとは評価できない**。

推論速度

ここは公開情報だけでは正確な順位を付けられない。

ELYZAはmodel architectureとweightsが公開されているが、公式model cardには特定GPU上の `tokens/s`、TTFT、request/secといったproduction throughput benchmarkが掲載されていない。o1/o3側は逆にAPIとして提供されるため、parameter countやGPU topologyが非公開である。従って「ELYZA 33Bはo3の何倍速い」のような比較はできない。

OpenAIはo1についてo1-previewより平均60%少ないreasoning tokensで処理するとしており、o3世代についてはChatGPTで複雑なtool-using questionにも典型的には1分未満で回答すると説明するが、これは固定hardware上のtokens/sec benchmarkではない。 [41]

ELYZAについては次の定性的推定ができる。

33B Denseは各tokenでほぼ全weightを使うため、一ユーザーあたりdecode throughputはMoE版より重いと予想される。

32B-A3B MoEは総32.1B weightsを持つ一方、token当たりactive paramsは約3.83Bなので、expert演算量を大きく削減できる。ただしattention、routing、memory transfer等は残るため、「3.8B dense modelと同じ速度」にはならない。 [42]

さらに両ELYZAモデルはreasoning effortを切り替える設定がなく、公式評価では**thinking always on**である。簡単な問い合わせでもreasoning tokenを生成する場合、non-reasoning modelよりend-to-end latencyが大きくなる可能性がある。 [43]

したがって速度の現時点の判定は、

MoE > Dense の可能性が高いが、o1/o3との絶対速度比較は未指定・測定不能

である。

メモリ使用量

ELYZAについてはparameter countから最低限のweight memoryを推計できる。

計算式は、

`BF16 weight memory ≈ parameters × 2 bytes`

である。

Model	Params	BF16純weight量・概算	補足
ELYZA 33B Dense	33.2195B	**約66.44 GB**（約61.88 GiB）	全weight active
ELYZA 32B-A3B	32.1390B	**約64.28 GB**（約59.86 GiB）	全weight自体は保持必要
MoE active params相当	3.8275B	約7.65 GB相当	**総VRAMではない**
o1	非公開	**未指定**	server-side
o3	非公開	**未指定**	server-side

パラメータ数の一次情報: [10]

これは純weightsだけなので、実際にはKV cache、activations、CUDA/runtime、vLLM memory pool等が上乗せされる。特に65,536-token contextを大きく使う場合はKV cacheが無視できない。

理論上のweight-only量子化なら、

- 8-bit: 33B ≈ 33 GB、MoE ≈32 GB
- 4-bit: 33B ≈16.6 GB、MoE ≈16.1 GB

まで縮められるが、これは単純なbit換算であり、quantization metadataやruntime overheadを除いた**理論値**である。品質低下も別途検証が必要である。

この点で32B-A3Bの重要な特徴は、「計算量はactive 3.8Bに近づけられるが、**重み保存量は32Bのまま**」という点である。MoEを「3Bなので8GB GPUでそのまま動く」と理解するのは誤りである。

コスト

OpenAIの2026年10月時点のmodel documentationでは、

モデル	Input / 1M tokens	Output / 1M tokens	重み取得
o1	**$15**	**$60**	不可
o3	**$2**	**$8**	不可
ELYZA 33B/MoE	API課金なし	API課金なし	Apache 2.0で取得可能

OpenAI料金: [44]
ELYZA license/distribution: [45]

ここで「ELYZAは無料だから推論コストゼロ」と考えてはいけない。無料なのはweights/licenseであり、GPU費用と運用費用が発生する。

セルフホストのoutput-generation costは概算で、

$$\text{Cost per 1M generated tokens} \approx \frac{1,000,000}{3,600R} \times N \times C$$

ここで、

- R = aggregate decode throughput [tokens/s]
- N = 使用GPU台数
- C = 1 GPUあたり1時間の実効費用 [$ / GPU-hour]

である。

つまり、

$$\approx \frac{277.78NC}{R}$$

ドル／100万generated tokens。

仮に**実測ではなく計算例**として、

仮定throughput	1M output tokensに必要なGPU時間
10 tok/s	27.78時間
20 tok/s	13.89時間
40 tok/s	6.94時間
80 tok/s	3.47時間

となる。

40 tok/s、1 GPUなら **約6.94 × GPU-hour単価**、2 GPUなら **約13.89 × GPU-hour単価** が純generation costの概算になる。実際にはprompt prefill、idle time、冗長化、ストレージ、network、engineer運用費等が加わる。

この式から分かるのは、現在のo3の **$8/1M output tokens** はかなり低い価格であり、低〜中利用率のセルフホストが必ずしもo3より安くなるとは限らないことである。一方、o1は$60なので、高稼働率でGPUを回せる環境ではELYZAが経済的になる可能性がはるかに高い。 [34]

そして自社ホストの最大のメリットはtoken単価そのものより、

データを外に出さない、rate limitを自分で制御できる、weightをfine-tuneできる、モデルを固定versionで保持できる

という運用上の主権にある。

o1/o3超えに関する最終推定

私の判定は次の通りである。

ELYZA-Thinking-1.0-llm-jp-4-33b がフルo1の「総合的な推論レベル」を上回っている可能性は低い。

根拠は、比較可能性の比較的高いGPQA Diamondでo1 75.7に対しELYZA 69.95であること、o1がMATH/AIME/codingなどでも当時のfrontier reasoning modelとして非常に高い公式値を持つこと、ELYZA側にo1を共通harnessで破った試験がないことによる。 [46]

o3を超えている可能性はさらに低い。

OpenAI自身が同latency/cost条件でo3がo1より高性能としており、ELYZAがo1超えを立証できていない以上、o3超えを主張する根拠はない。 [39]

ただし、

「日本語固有知識」「日本語の複雑な形式制約」「閉域環境でRAG＋tool callingを行う」という限定された業務タスクで、ELYZAがo1/o3より良い実用結果を出す可能性は否定できない。

なぜならELYZAはまさにそれらを意図して追加学習されており、公開benchmarkでもJamC-QA、JEMHopQA、M-IFEval-ja、JFBench等に明確な改善が出ているからである。 [47]

しかしo1/o3を同じ日本語benchmark・同じprompt・同じjudgeで走らせた一次データがない以上、そこは現時点では**未指定／未検証**である。

結論と導入推奨

結論

今回のELYZAモデルについて最も適切な評価は、

> **「o1/o3を超える新しいfrontier model」ではなく、「LLM-jp-4を日本語reasoning・instruction following・coding/tool use向けに大幅強化し、企業が自社保有できる実用的な国産系open-weight reasoning model」**

である。

技術的成果は十分に実質的である。33Bは親LLM-jp-4からOverall +10.19 points、MoEは+12.08 points。新しいLLM-jp-4.1に対してもカテゴリ集約Overallではそれぞれ+1.05、+4.57であり、追加post-trainingの効果は明白である。 19

しかも追加学習は単なる日本語instruction tuningではなく、Mid-training、数百万synthetic SFT、programmatically verified data、RLVRまで行い、math/coding/STEM、日本語知識、複雑な指示、tool useをまとめて改善している。 2

一方で絶対性能を見ると、ELYZA自身の比較表においてもQwen3.5やGemma 4などが総合平均では上にいる。さらにフルo1との共通GPQA DiamondではELYZA 33Bが約5.75 points下回るため、「o1/o3級を凌駕」という評価は現在のデータからは支持できない。 48

導入判断

PoC導入は強く推奨、本番採用は条件付き推奨とする。

最も相性が良いのは、日本語社内文書RAG、規程・FAQ・manual QA、customer support agent、Japanese coding assistant、複雑な書式制約を伴う文書生成、自社APIを操作するtool-calling agent、外部LLM APIに機密データを送信できない環境、独自fine-tuningが必要な企業システムである。

逆に、最先端の数学・科学研究支援、最高性能coding agent、visionを含むmultimodal reasoning、hardware運用を抱えたくない小規模システム、低利用率でo3のAPI単価を下回りにくいワークロードでは、今回のELYZAモデルを選ぶ理由は弱くなる。o3は現在の公式価格でinput $2/M、output $8/Mであり、GPUのidle costまで考えると少量利用では外部APIの方が合理的になり得る。 49

モデル選択だけなら、現状は次の使い分けが妥当である。

要件	第一候補
ELYZA系で最高精度を優先	**33B Dense**
inference compute効率を重視	**32B-A3B MoE**
最先端の汎用推論能力	o3等のfrontier APIも比較対象に残す
機密性・閉域・モデル所有	**ELYZA**
日本語固有task	**ELYZAを必ずPoC候補に入れる**
Multimodal vision	今回のELYZAモデルは不適
GPU運用なし・低volume	API型モデルが有力

本番採用前に必須の追加検証

優先順位が最も高いのは、**自社taskでo1/o3および他のopen-weight modelと完全に同条件でblind evaluationすること**である。

最低でも日本語50〜500件程度の実業務test setを固定し、正答率だけでなくhallucination、instruction adherence、citation correctness、tool-call success、schema validity、unsafe action rate、latency、reasoning token量、GPU costを測るべきである。特に「日本語ではELYZAの方が良い」という仮説は、公開benchmarkから十分あり得るが、現状ではo1/o3との直接データがない。 48

次に必要なのは実機benchmarkである。33BとMoEを同一GPU、同一vLLM version、同一context長で動かし、TTFT、decode tok/s、concurrent throughput、peak VRAM、8K/32K/64K context degradationを測定する必要がある。公式QuickstartはvLLM 0.30.0と65,536-token contextを示しているが、production性能値そのものは公表していない。 33

三つ目は安全性である。NII親モデルにはalignment上の注意書きがあり、ELYZA版には今回確認できる範囲で包括的なsafety benchmarkがない。prompt injection、jailbreak、PII leakage、Japanese harmful content、tool misuse、RAG poisoningについて自社red-team評価を行うべきである。 50

四つ目はsoftware supply-chain reviewである。`--trust-remote-code` と独自parser pluginをproductionで直接remote取得する運用は避け、commit固定、code review、container image固定、hash verificationを行うことが望ましい。 33

最後に法務・ガバナンス面では、Apache 2.0のLICENSE/NOTICE義務、商標、学習データprovenance、生成物に関する自社ポリシーを確認する必要がある。Apache 2.0は非常に寛容なライセンスだが、non-infringement等を保証しているわけではない。 30

したがって、2026年10月3日時点の最終判断は「日本企業向けopen-weight reasoning modelとしてかなり有望でありPoC価値は高い。しかし“o1/o3超え”とは評価しない。本番導入の判断材料としては、公開ベンチマークよりも、今後の第三者評価と自社E2E benchmarkを重視すべき」である。

---

1 7 8 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＋
https://www.itmedia.co.jp/aiplus/article/2610/02/2000001965/

2 3 5 6 11 12 13 19 20 21 29 33 37 43 47 48 elyza/ELYZA-Thinking-1.0-llm-jp-4-33b · Hugging Face
https://huggingface.co/elyza/ELYZA-Thinking-1.0-llm-jp-4-33b

4 35 36 40 41 46 OpenAI o1 and new tools for developers | OpenAI
https://openai.com/index/o1-and-new-tools-for-developers/

9 ELYZA、完全国産AIを無料公開 - コプラス株式会社
https://coplus-one.jp/elyza-domestic-llm-free-release/

10 16 17 23 31 42 50 llm-jp/llm-jp-4-33b-base · Hugging Face
https://huggingface.co/llm-jp/llm-jp-4-33b-base

14 大規模言語モデル「LLM-jp-4.1」シリーズを公開 - LLM-jp
https://llm-jp.nii.ac.jp/news/llm-jp-4-1/

15 24 25 32 45 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　〜研究成果を活用した「バーティカルAI事業」も開始、第一弾として「ELYZA Voice Agent」を開発〜 | 株式会社ELYZA | 株式会社ELYZA
https://elyza.ai/news/2026/10/02/rsiresearch

18 22 elyza/ELYZA-Thinking-1.0-llm-jp-4-32b-a3b · Hugging Face
https://huggingface.co/elyza/ELYZA-Thinking-1.0-llm-jp-4-32b-a3b

26 27 ELYZA LLM-jp-4 とは｜Apache 2.0 で商用利用を検証 - FIXIT
https://fixit.co.jp/tips/elyza-thinking-llm-jp-4/

28 ELYZA-Thinking-1.0とは？LLM-jp-4ベースの国産推論モデル・33B ...
https://ai-revolution.co.jp/media/what-is-elyza-thinking/

30 Apache License, Version 2.0 | Apache Software Foundation
https://www.apache.org/licenses/LICENSE-2.0

34 44 49 o3 Model | OpenAI API
https://developers.openai.com/api/docs/models/o3

38 39 Introducing OpenAI o3 and o4-mini | OpenAI
https://openai.com/index/introducing-o3-and-o4-mini/