ELYZA-Thinking-1.0（LLM-jp-4版）の技術特性・市場評価およびフロンティア推論モデルとの比較分析

Gemini 3.8 Flash

記事の要旨とモデルの基本仕様

2026年10月2日、KDDIグループのAI開発企業である株式会社ELYZAは、国立情報学研究所（NII）の大規模言語モデル研究開発センターがスクラッチから開発した国産オープン基盤モデル「LLM-jp-4」をベースとする推論特化型言語モデル2種を公開した[1]。公開されたモデルは、全パラメータを稼働させるDense（稠密）構造の「ELYZA-Thinking-1.0-llm-jp-4-33b」と、Mixture-of-Experts（MoE）構造を採用した「ELYZA-Thinking-1.0-llm-jp-4-32b-a3b」の2系統である[1]。いずれも商用利用が可能なApache 2.0ライセンスのもと、Hugging Face上でオープンウェイトとして無償提供されている[2]。

本モデル群は、国内で約12兆トークンを用いてゼロから事前学習されたLLM-jp-4に対し、ELYZAが独自の中間学習（Mid-training）、教師あり微調整（SFT）、および検証可能な報酬に基づく強化学習（RLVR）を施したものである[1]。数学やプログラミング、STEM分野の論理推論データを日本語にローカライズして思考過程を生成させたほか、回答出力前に内部的な推論を展開する思考アーキテクチャが組み込まれている[1]。出力形式には思考プロセスと最終回答を別個のチャネルで制御するharmony形式が導入され、コンテキスト長は最大65,536トークン（約64K）を確保している[1]。

項目	**ELYZA-Thinking-1.0-llm-jp-4-33b**	**ELYZA-Thinking-1.0-llm-jp-4-32b-a3b**
開発元	株式会社ELYZA（KDDIグループ）	株式会社ELYZA（KDDIグループ）
ベースモデル	llm-jp-4-33b-base（NII）	llm-jp-4-32b-a3b-base（NII）
アーキテクチャ	Dense（全パラメータ常時稼働）	MoE（128エキスパート中8つを活性化）
総パラメータ数	約332億（33B）	約321億（活性化パラメータ約38億）
コンテキスト長	65,536 トークン	65,536 トークン

ライセンス	Apache 2.0（商用利用可能）	Apache 2.0（商用利用可能）
提供形態	Hugging Face（重み公開）	Hugging Face（重み公開）

本発表は、ELYZAが同日付で設立した新研究組織「ELYZA RSI Research」の研究成果の一環に位置付けられている[2]。同組織は、AIがAI自身を反復的に開発・改良する「再帰的自己改善（RSI: Recursive Self-Improvement）」の探求を目的として掲げている[2]。今回のモデル構築においては社内のLLM学習基盤を自動化・効率化し、中間学習から事後学習、評価に至る一連の工程を人間が介在する作業1〜2人日程度に抑えて完遂した点が強調されている[5]。なお、公式発表および報道では、本成果はRSIの前段階にあたる要素技術研究であり、RSIそのものを自律的に達成したわけではない旨が明記されている[2]。また、ITmediaの当初報道において「完全国産AI」「完全国産基盤」と見出しに掲げられていた表現は、モデルカード上の「Fully domestic foundation」が単に国内組織によるスクラッチ事前学習と追加学習を指す意図であり、使用した教師データ生成等のニュアンスと乖離があったことから、ELYZA側の補足説明を受けて記事タイトルおよび本文が修正された経緯を持つ[2]。

ベンチマーク性能と領域別の推論特性

ELYZAが提示した20種類のベンチマーク（7分野）による評価結果では、ベースモデルであるLLM-jp-4およびNII自身が改善を加えた「LLM-jp-4.1」シリーズに対する有意な性能向上が確認されている[2]。一方で、同時期のグローバルな最先端オープンウェイトモデルと比較した場合、得意分野と不得意分野が極めて明確に分かれる結果となっている[1]。

評価分野 / ベンチマーク	**ELYZA-Thinking-33b**	**llm-jp-4.1-33b**	**Qwen3.5-27B**	**Gemma 4 31B**
総合平均（**Overall Average**）	61.69	60.64	68.18	**70.54**
日本語平均（**Japanese Average**）	62.18	61.78	66.05	**68.88**
英語平均（**English Average**）	59.72	59.11	71.81	**73.09**
数理（**AIME 2024+2025**）	68.85	78.23	**93.33**	88.75

数理（**MATH-500 en**）	96.00	96.80	90.60	**99.20**
数理（**JMATH-500 ja**）	90.00	92.40	93.60	**94.00**
知識・**STEM**（**GPQA-Diamond en**）	69.95	70.77	85.92	**86.65**
知識・**STEM**（**MMLU-Pro en**）	77.53	78.82	85.42	**85.50**
コード（**LiveCodeBench v6 en**）	60.23	45.20	**80.11**	79.37
日本語複数文書**QA**（**JEMHopQA ja**）	**73.85**	65.84	62.87	63.91
日本語指示追従（**JFBench ja**）	**39.02**	32.79	34.12	35.68
関数呼出（**BFCL v4 en**）	38.73	48.14	**69.37**	66.79

本モデル群が卓越した強みを発揮しているのは、日本語固有の知識想起および厳密な言語的制約の追従領域である[1]。複数の断片的な情報を統合して結論を導く日本語マルチホップQAベンチマーク「JEMHopQA」において、ELYZA-Thinking-33bは73.85点を記録し、海外の競合モデルであるQwen3.5-27B（62.87点）やGemma 4 31B（63.91点）を10ポイント近く上回った[4]。また、振り仮名の付与や漢数字の指定など、日本語固有の書式制限を厳密に評価する「JFBench」においても39.02点を記録し、比較対象となった全モデルの中でトップスコアを獲得している[4]。これらは、最新の日本語WikipediaやWikidataから合成した知識データの注入や、精緻な指示追従データによる事後学習が

的確に奏功した結果である[3]。
一方、記号処理や多段階アルゴリズムの探索を要する抽象的論理推論では大きな課題を残している[1]。難関数学競技レベルの推論を問う「AIME 2024+2025」において、本モデルは68.85点にとどまり、NIIのllm-jp-4.1-33b（78.23点）を下回っただけでなく、Qwen3.5-27B（93.33点）やGemma 4 31B（88.75点）に対して決定的な格差をつけられている[4]。プログラミング能力を測る「LiveCodeBench v6」（60.23点）やツール呼び出し能力を測る「BFCL v4」（38.73点）でもグローバル水準に遠く及ばず、言語化された知識処理には長けているものの、自律的なエージェント操作や極限的な数理推論では海外最先端モデルに劣後する構造が浮き彫りとなっている[4]。

産業界・技術コミュニティにおける反響と多面的評価

国内のAIコミュニティおよび産業界における反響は、技術主権（ソブリンAI）の確保や実務適用の容易さを歓迎する肯定的な論調と、グローバルな最先端モデルとの性能格差を冷静に指摘する分析の双方が存在する[6]。
実務者層やエンタープライズ導入の観点からは、完全な商用利用を担保するApache 2.0ライセンスでモデル重みが公開された点が極めて高く評価されている[8]。近年、海外の主要なオープンモデルでは商用利用にユーザー数制限を設けるなど独自の制約が増加しているが、本モデルは自社管理下のオンプレミス環境やプライベートクラウドに制約なく組み込むことができる[8]。金融機関、自治体、医療分野など、機密データの国外持ち出しが制限される業務領域において、国内でゼロから事前学習された基盤モデルが実用的な選択肢として提供された意義は大きい[8]。さらに、同時公開された窓口業務評価指標「ELYZA Agent Tasks: Customer Service」や、精度を維持したまま推論コストを67%削減したエージェントハーネス自動改善の実績など、実業務への社会実装を念頭に置いた包括的な公開設計も実務開発者から支持を集めている[5]。
他方で、AI研究者や技術アナリストの視座からは、客観的な性能指標に対する厳しい評価も下されている[6]。総合平均スコアにおいて半年以上前に登場した海外オープンモデルに6〜10ポイント及ばず、日本語評価の総合平均値においても海外モデルの後塵を拝している実態が指摘されている[6]。また、学習プロセスの人手工数を1〜2人日に抑えた効率性は評価されつつも、推論データの生成過程において海外の強力な教師モデルに依存している点がモデルカード等から示唆されており、真の意味での自立的な開発サイクルを確立するには依然として距離があるとの分析がなされている[6]。
あわせてコミュニティ内では、ELYZAが2025年5月に公開した「ELYZA-Thinking-1.0-Qwen-32B」との混同に対する注意喚起が広がった[1]。2025年版はAlibabaのQwen2.5-32B-Instructを基盤にモンテカルロ木探索（MCTS）データで微調整され、「OpenAIのo1-miniに匹敵する数学推論性能」として認知されていた[1]。しかし今回の2026年10月公開モデルは、土台がNIIの「LLM-jp-4」に刷新されており、アーキテクチャ、学習手法、性能バランスのすべてが異なる別系統のモデルである[1]。過去の「o1-mini並み」という評価を今回のLLM-jp-4版にそのまま適用することは技術的に誤りであるとの理解が共有されている[1]。

OpenAI o1およびo3との推論能力比較・推定

OpenAIが開発した推論特化型モデル「OpenAI o1」および後継のフロンティアモデル「OpenAI o3」と比較した場合、ELYZA-Thinking-1.0（LLM-jp-4版）がこれらの推論水準を上回っているかについて、各種ベンチマークとシステム設計の観点から検証を行う[4]。

ベンチマーク指	**ELYZA-Thinkin**	**ELYZA-Thinkin**	**OpenAI o1**	**OpenAI o3**

標	g-33b	g-32B-A3B	(High/Full)	(Full)
AIME 2024 / 2025	68.85%	62.29%	83.3%	**96.7%**
GPQA Diamond	69.95%	63.86%	78.0%	**87.7%**
MATH / MATH-500	96.00%	95.20%	94.8%〜96.4%	**97.0%+**
MMLU-Pro	77.53%	73.14%	**91.0%**	85.0%〜87.0%
コーディング評価	LiveCodeBench: 60.23%	LiveCodeBench: 53.66%	LiveCodeBench: 約75〜80%	**SWE-bench: 71.7%**

客観的なベンチマークデータの比較から、ELYZA-Thinking-1.0（LLM-jp-4版）は、OpenAIのo1モデルおよびo3モデルの推論レベルを明確に下回っていると推定される[4]。
数学オリンピック予選水準の高度な論理展開を要するAIMEにおいて、ELYZA-Thinking-33bの正答率は68.85%にとどまる[4]。対してOpenAI o1は83.3%を達成しており、さらにo3に至っては96.7%という数学の専門家をも上回る極めて高い推論精度に到達している[15]。大学院・博士レベルの理系学術知識を問うGPQA Diamondにおいても、ELYZA-Thinking-33bの69.95%に対し、o1は78.0%、o3は87.7%を記録しており、抽象的な専門科学推論において有意な実力差が存在する[4]。高校生レベルの数学問題集であるMATH-500では96.00%を記録してo1に匹敵する数値を示しているが、このベンチマークは近年の思考型モデル全般においてスコアが飽和（天井効果）に達しつつある指標であり、極限的な推論深度を測る指標としてはAIMEやGPQAにおけるスコア乖離が両者の決定的な推論格差を物語っている[4]。
プログラミングおよび汎用論理タスクの領域でも同様の傾向が確認される[4]。LiveCodeBench v6におけるELYZA-Thinking-33bのスコアは60.23%であり、80%前後に達する海外の主要オープンモデルに及ばないばかりか、競技プログラミングの難問を突破しSWE-bench Verifiedで71.7%を記録するOpenAI o3の領域とは依然として隔絶した距離がある[4]。学術総合テストであるMMLU-Proにおいても、o1の91.0%に対してELYZA-Thinking-33bは77.53%にとどまっている[4]。
この推論格差が生じる根本的な背景には、モデル規模と計算パラダイムの設計思想における相違が存在する。OpenAIのo1およびo3は、数千億パラメータ規模のフロンティア基盤に対し、強化学習を通じて推論時計算量（Test-time Compute）を動的にスケーリングさせる大規模探索メカニズムを実装している[14]。対してELYZA-Thinking-1.0は、企業が単一ノードや手頃なGPU環境でオンプレミス運用可能な33B（MoE版では活性化約38億）というコンパクトなサイズに設計されている[1]。限られたパラメータ容量の多くを日本語固有の文脈や知識の再現、厳密な指示制約への適合に割り振っているため、膨大な計算資源を背景に記号論理や数学的証明を極限まで追求するグローバルなフロ

ンティア推論モデルとは、そもそもの最適化目標と運用ターゲットが異なっている[3]。

総括

ITmediaが報じた「ELYZA-Thinking-1.0（LLM-jp-4版）」は、国内学術主導の基盤モデル「LLM-jp-4」に対して民間企業の自動化された事後学習技術を適用し、実務的な日本語推論モデルへと昇華させた重要なマイルストーンである[1]。日本語特有の多段階情報検索（JEMHopQA）や厳格な書式指示（JFBench）においてはグローバルモデルを凌駕する実用性能を示し、商用利用可能なApache 2.0ライセンスのもとでローカル運用を可能にした点は、国内産業界のデータガバナンスとソブリンAI確立の観点から高く評価できる[4]。

一方で、純粋な論理推論能力の比較において、本モデルがOpenAIのo1モデルおよびo3モデルの推論水準を上回ることはなく、競技数学、博士レベルのSTEM推論、自律的コーディングの各領域において依然として大きな性能ギャップが存在する[4]。したがって本モデルの価値は、クローズドな最先端フロンティアモデルと推論競争を行う点にあるのではなく、企業が自社の管理下でセキュアかつ低コストに運用できる実用的な計算規模において、高度な日本語対話・指示追従能力をオープンソースとして確立した点に見出すべきである[8]。

引用文献

1. 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/
2. 国産AI「LLM-jp-4」ベースのAIモデル、KDDI傘下のELYZAが無料公開, https://www.itmedia.co.jp/aiplus/article/2610/02/2000001965/
3. 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
4. elyza/ELYZA-Thinking-1.0-llm-jp-4-33b - Hugging Face, https://huggingface.co/elyza/ELYZA-Thinking-1.0-llm-jp-4-33b
5. ELYZA、「AIによるAI開発」の研究を推進する「ELYZA RSI, https://elyza.ai/news/2026/10/02/rsiresearch
6. ELYZA-Thinking-1.0-llm-jp-4とは 国産モデルの性能と ... - AI時短ラボ, https://www.ai-jitan-hub.com/news/elyza-thinking-llm-jp-4-rsi-research
7. ELYZA、国産基盤のAIモデル2種を公開 商用利用可能で人間の作業を, https://aipicks.jp/news/20261002-elyza-ai-2-1-2
8. 日本語性能を高めた国産ベースのAIモデルを無料公開 | Yoshi_it - note, https://note.com/natty_toucan386/n/na06105ceab03
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. ELYZA-Thinking-1.0-llm-jp-4-32b-a3b とは #十02 | DopingConsomme, https://note.com/dopingconsomme/n/n2f6916197fc5
11. ELYZAが日本語での論理的思考能力を強化した ... - Web担当者Forum, https://webtan.impress.co.jp/n/2025/05/02/49117
12. ELYZA、論理的思考能力を強化した「Reasoning Model」を開発し, https://news.mynavi.jp/techplus/article/20250501-3271709/
13. ローカルLLM のリリース年表 | npaka - note, https://note.com/npaka/n/n1d99253ae2cf

14. LLM で論理的思考を学ぶ - OpenAI, https://openai.com/ja-JP/index/learning-to-reason-with-llms/
15. OpenAI o3 release: AI model one step closer to AGI - dida.do, https://dida.do/highlights-from-openai-s-o3-model-one-step-closer-to-agi-in-2025
16. OpenAI o3/o3-mini: 人間を超える推論力と安全性を備えた最新AI - note, https://note.com/kakeyang/n/n3d2dbfd3bec8
17. OpenAI o1とは？GPT-4oとの違い・6つの特徴・Claude比較で見る, https://genai-ai.co.jp/ai-kanri/blog/cc-openai-o1-guide/
18. OpenAI o1 Benchmark Results & Rankings | DataLearnerAI, https://www.datalearner.com/en/ai-models/pretrained-models/openai-o1/analysis
19. OpenAI Unveils o3 Model and Becomes First to Crack the ARC-AGI, https://beebom.com/openai-unveils-o3-model-cracks-arc-agi-benchmark/
20. OpenAI o1/o3 vs DeepSeek R1 vs QwQ-32B - Prem AI, https://www.premai.io/blog/reasoning-models-explained-openai-o1-o3-vs-deepseek-r1-vs-qwq-32b/