Gemini 3 Deep Think が示す 推論 AI の到達点と限界

ARC-AGI-2 ベンチマーク 84.6%達成の技術的背景と AGI への含意

2026 年 2 月 13 日

■ **エグゼクティブサマリー**　Gemini 3 Deep Think は、2026 年 2 月 12 日に Google DeepMind が発表した推論特化モードの大幅アップグレードであり、ARC-AGI-2 ベンチマークで人間平均（約 60%）を大きく上回る 84.6%を達成した[1][2]。Claude Opus 4.6（68.8%）や GPT-5.2（52.9%）を大差で引き離す一方、1 タスクあたり$13.62 という推論コストは、ARC Prize が定める人間レベル効率閾値（$0.42/タスク）の 32 倍にのぼる[3][4]。

アーキテクチャの核心は「考える時間」への投資

Gemini 3 Deep Think は、Gemini 3 ファミリーの中で速度や汎用性ではなく「推論の深さ」を最大化する専用モードとして設計されている。技術的基盤は「テストタイム・コンピュート」（推論時計算）と呼ばれるアプローチで、応答生成前に大量の計算資源を投入して「考える時間」を確保する[1][5]。具体的には、高度な並列推論により複数の仮説を同時に探索し、内部検証によって誤った推論経路を剪定してから回答を生成する。

その計算規模は桁違いだ。ある ARC-AGI タスクにおいて、通常の Gemini 3 Pro が 96 推論トークンで回答したのに対し、Deep Think は同じタスクに 138,000 推論トークンを費やした[6]。約 1,400 倍の「思考量」を投入して精度を担保する設計思想である。

2026 年 2 月 12 日のアップグレード版は、2025 年 11 月の初期版（ARC-AGI-2 で 45.1%）から 84.6%へとほぼ倍増する飛躍を遂げた[2][7]。Google AI Ultra サブスクリプション（月額$249.99）のユーザーに提供されるほか、Gemini API を通じた研究者・企業向けの早期アクセスプログラムも初めて開設された。

ARC-AGI-2 で 84.6%が意味すること

ベンチマークの設計思想

ARC-AGI-2 は、Keras の開発者として知られる François Chollet が率いる ARC Prize Foundation が開発した抽象推論ベンチマークである。色付きグリッド上のパターン変換タスクで構成され、1〜2 組の入出力例を観察した後、新しい入力に対する正しい出力グリッドを生成する必要がある。完全一致のみ正解というルールで、1 セルでも誤れば不正解となる[8][9]。

ARC-AGI-1（初代）が 2024 年末には AI モデルによって概ね攻略されたことを受け、ARC-AGI-2 は「力業による攻略への耐性」を強化して 2025 年 3 月に公開された。2025 年 5 月時点では最先端モデルでも 5%未満しか解けず、純粋な LLM はスコアほぼ 0%だった[9][10]。

スコアの推移

下表は、ARC-AGI-2 におけるスコアの急速な進展を示している。

時期	モデル	スコア	タスク単価
2025 年 5 月	最先端 LLM 群	0〜5%	—
2025 年 11 月	Gemini 3 Pro	31.1%	$0.81
2025 年 11 月	Deep Think（初期版）	45.1%	$77.16
2025 年 11 月末	Poetiq	54.0%	$30.57
2025 年 12 月	GPT-5.2	52.9%	~$1.90
2025 年 12 月	Claude Opus 4.5	37.6%	$2.20
2026 年 2 月	**Deep Think（更新版）**	**84.6%**	**$13.62**

人間の平均スコアが約 60%であることを踏まえると、AI が初めて一般人の平均を大幅に上回った歴史的マイルストーンと言える[3][4]。

$13.62 という「知能の代償」

Deep Think の 84.6%という精度は、ARC Prize Grand Prize（賞金$700K 以上）の精度閾値 85%にわずか 0.4 ポイント及ばない。しかし仮に精度を満たしたとしても、Grand Prize のもう一つの要件であるコスト閾値$0.42/タスク以下は到底クリアできない。$13.62 は閾値の 32 倍以上であり、初期版の$77.16 からは大幅に改善されたものの、依然として実用的な効率性とは程遠い[3][4]。

Chollet は「知能は問題を解決する能力だけでなく、その能力の獲得と展開の効率性こそが決定的に重要」と主張する。コスト効率の観点で比較すると、GPT-5.2 は 52.9%を約$1.90/タスクで、Claude Opus 4.5 は 37.6%を$2.20/タスクで達成しており、精度あたりのコスト効率では Deep Think は劣位にある[4][11]。

競合モデルとの多面的な比較

推論・抽象知能では Deep Think が圧倒

ベンチマーク	Deep Think	Claude Opus 4.6	GPT-5.2	Gemini 3 Pro
ARC-AGI-2	84.6%	68.8%	52.9%	31.1%
Humanity's Last Exam	48.4%	40.0%	34.5%	37.5%
GPQA Diamond	93.8%	—	93.2%	91.9%
Codeforces Elo	3,455	2,352	—	2,512
MMMU-Pro	81.5%	73.9%	79.5%	81.0%

Codeforces の 3,455 Elo は「Legendary Grandmaster」ティアに相当し、これを上回るアクティブな人間プログラマーは世界でわずか 7 名しかいない[12]。

Deep Think が劣位にある領域

一方、すべての領域で Deep Think が最強というわけではない。ソフトウェア工学においては

Claude Opus 4.5 が SWE-bench Verified で 80.9%を達成し、Gemini 3 Pro の 76.2%を上回る。数学競技の一部では GPT-5.2 が AIME 2025 でツールなし 100%を達成した[13][14]。推論の深化が抽象推論に偏重して効果を発揮していることを示唆する。

科学オリンピックから結晶成長まで——具体的な実績

学術競技での達成

Deep Think は 2025 年の主要国際科学オリンピック 3 大会すべてで金メダル水準の成績を記録した。国際数学オリンピック（IMO 2025）、国際物理オリンピック（IPhO 2025）、国際化学オリンピック（IChO 2025）の筆記部門でいずれも金メダルレベルを達成[1][5]。また、MathArena Apex では 23.4%を記録し、競合の GPT-5.1（1.0%）や Claude Sonnet 4.5（1.6%）に対して 20 倍以上の差をつけた。

実世界の研究応用

ラトガース大学の数学者 Lisa Carbone は、アインシュタインの重力理論と量子力学を橋渡しする研究に Deep Think を活用し、人間の査読を通過していた微妙な論理的欠陥を Deep Think が発見した[1][15]。デューク大学の Wang Lab は半導体材料探索のための結晶成長最適化に使用し、DeepMind の Aletheia エージェントは 18 の研究課題のボトルネックを解消した。

AGI 到達の「シグナル」か「マイルストーン」か

専門家の反応は慎重な楽観

ARC-AGI-2 の設計者である Chollet 自身は、「新しい Gemini Deep Think は ARC-AGI-2 で本当に驚異的な数値を達成している」と評価した[16]。ただし彼の一貫した立場は、ARC 攻略は「AGI への必要条件ではあるが、AGI そのものではない」というものだ。ARC Prize Foundation の Mike Knoop も「AGI は達成されていない。新しいアイデアが依然として必要」と明言している[4]。

AGI タイムラインをめぐる専門家の見解は大きく分かれる。Anthropic CEO Dario Amodei は 2026 年中の実現を予測する一方、DeepMind CEO Demis Hassabis は 2030 年末までに約 50%の確率を見込む[17]。懐疑的な立場の Gary Marcus は最近の数ヶ月が AGI 楽観論にとって「壊滅的」だったと主張し、元 OpenAI の Andrej Karpathy は AGI を「10 年先」と見積もっている。

ベンチマーク飽和のパラドックス

Deep Think の成果は、皮肉にもベンチマーク自体の限界を浮き彫りにした。ARC Prize Foundation は既に ARC-AGI-3（インタラクティブ環境での「エージェンシー」を試すベンチマーク）を 2026 年 3 月 25 日にリリース予定と発表している[4][18]。

さらに、ARC Prize Foundation はオーバーフィッティングの懸念も指摘している。Deep Think の推論プロセスにおいて、ARC 固有のカラーマッピングが使われている証拠が検出され、「ARC デ

ータがモデルの学習データに十分に含まれていることを強く示唆する」と述べた[4]。ベンチマークスコアの一部が汎用的適応力ではなく、学習データの記憶に由来する可能性が否定できない。

残された課題とベンチマークの先にあるもの

コストと効率のスケーリング問題。1 タスク$13.62 のコストは人間の効率性基準の 32 倍であり、大量のタスク処理が求められる実用シナリオでは依然として非現実的だ[3][4]。

能力の非均一性。抽象推論やコーディングでは圧倒的でも、実世界のソフトウェア工学、マルチモーダル処理、創造的文章執筆では競合に対する優位性が限定的か劣位にある[13][14]。

ハルシネーションの持続。推論の深化は技術的なハルシネーションを減少させるが、完全には排除できない。コード変更の提案、金融判断、研究上の主張においては人間の監視が不可欠とされる[19]。

ベンチマークの信頼性。Hacker News のコミュニティでは「ベンチマークをまだ信頼できるのか」「学習データに含まれていないとどう保証できるのか」という根本的な疑問が提起されている[19]。

結論——「考える機械」の新段階とその代償

Gemini 3 Deep Think は、テストタイム・コンピュートの大規模投入により、抽象推論・数学・科学・競技プログラミングにおいて AI の能力フロンティアを明確に押し広げた。人間平均を大幅に超える ARC-AGI-2 スコア、世界トップ 7 に迫る Codeforces Elo、3 つの国際科学オリンピック金メダル水準は、「推論の深さ」が正面から評価される領域では既に AI が人間の専門家に匹敵あるいは凌駕し得ることを実証している。

しかしその代償は大きい。32 倍のコスト超過という事実は、現在の推論 AI が「正解に到達する能力」と「効率的に到達する知能」の間にまだ深い溝を抱えていることを示す。ARC Prize Foundation が 2026 年 3 月に ARC-AGI-3 をリリースする予定であることは、現行ベンチマークの飽和速度がいかに速く、真の汎用知能の測定がいかに困難であるかを物語っている。Deep Think が切り拓いた地平は確かに印象的だが、AGI への道程はまだ「知能の効率化」という次の険しい峠を残している。

引用文献

[1] Google, "Gemini 3 Deep Think: Advancing science, research and engineering," Google Blog, Feb 12, 2026. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-deep-think/

[2] OfficeChai, "Google Releases Gemini 3 Deep Think, Tops ARC-AGI 2 Benchmark With 84.6%," Feb 2026. https://officechai.com/ai/gemini-3-deep-think-benchmarks-arc-agi/

[3] ARC Prize, "ARC Prize 2025 Results and Analysis," ARC Prize Foundation, 2026. https://arcprize.org/blog/arc-prize-2025-results-analysis

[4] ARC Prize Foundation, "ARC-AGI-2 Leaderboard," 2026. https://arcprize.org/arc-agi/2/

[5] Google DeepMind, "Gemini Deep Think: Redefining the Future of Scientific Research," Feb 2026. https://deepmind.google/blog/accelerating-mathematical-and-scientific-discovery-with-gemini-deep-think/

[6] The Decoder, "Google Deepmind upgrades Gemini 3 Deep Think for complex science and engineering tasks," Feb 2026. https://the-decoder.com/google-deepmind-upgrades-gemini-3-deep-think-for-complex-science-and-engineering-tasks/

[7] 9to5Google, "Gemini 3 Deep Think gets 'major upgrade' aimed at practical applications," Feb 12, 2026. https://9to5google.com/2026/02/12/gemini-3-deep-think-upgrade/

[8] ARC Prize Foundation, "ARC-AGI-2 A New Challenge for Frontier AI Reasoning Systems," Technical Report. https://arcprize.org/blog/arc-agi-2-technical-report

[9] Adaline Labs, "ARC-AGI In 2026: Why Frontier Models Still Don't Generalize," 2026. https://labs.adaline.ai/p/what-is-the-arc-agi-benchmark-and

[10] Medium (Artificial Synapse Media), "AI Models Struggle with New ARC-AGI-2 Benchmark, Raising Doubts About AGI Progress," 2025. https://medium.com/artificial-synapse-media/ai-models-struggle-with-new-arc-agi-2-benchmark-raising-doubts-about-agi-progress-1c5b2fcb9cf6

[11] Poetiq, "Poetiq Shatters ARC-AGI-2 State of the Art at Half the Cost," 2025. https://poetiq.ai/posts/arcagi_verified/

[12] OfficeChai, "Google Gemini 3 Deep Think Scores 3455 On Codeforces, Is Now Better Than All But 7 Human Programmers," Feb 2026. https://officechai.com/ai/google-gemini-3-deep-think-scores-3455-on-codeforces-is-now-better-than-all-but-7-human-programmers/

[13] R&D World, "How GPT-5.2 stacks up against Gemini 3.0 and Claude Opus 4.5," 2026. https://www.rdworldonline.com/how-gpt-5-2-stacks-up-against-gemini-3-0-and-claude-opus-4-5/

[14] MarkTechPost, "Is This AGI? Google's Gemini 3 Deep Think Shatters Humanity's Last Exam And Hits 84.6% On ARC-AGI-2 Performance Today," Feb 12, 2026. https://www.marktechpost.com/2026/02/12/is-this-agi-googles-gemini-3-deep-think-shatters-humanitys-last-exam-and-hits-84-6-on-arc-agi-2-performance-today/

[15] Google, "Gemini 3 Deep Think is now available in the Gemini app," Google Keyword Blog, Feb 2026. https://blog.google/products/gemini/gemini-3-deep-think/

[16] François Chollet (@fchollet), X post, Feb 13, 2026. https://x.com/fchollet/status/2021983310541729894

[17] AIMultiple, "AGI/Singularity: 9,300 Predictions Analyzed," 2026. https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/

[18] ARC Prize Foundation, "Announcing ARC-AGI-2 and ARC Prize 2025," 2025. https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025

[19] GitHub Gist, "HN community discussion summary on Gemini 3 post (814 comments)," 2025. https://gist.github.com/primaprashant/3786f3833043d8dcccae4bfd4ff9f4a7