ChatGPT

Gemini 3 Deep Think と ARC-AGI-2 ベンチマークに関する包括的・分析的報告書

Executive Summary

2026年2月12日、Google [1] は、Gemini 3の専門的推論モード「Gemini 3 Deep Think」を“科学・研究・工学の現実課題”向けに大幅更新したと発表した。公式発表では、研究者との共同改善（「不完全・ノイズを含むデータ」「正解が一意でない課題」などを想定）と、実運用（Geminiアプリ／Gemini API での提供）への拡張が強調されている。 [2]

この更新の象徴的指標が、ARC-AGI-2（v2 / Semi-Private）での **84.6%** である。これはARC Prize Foundationの「Verified」プロセスに基づく“ARC Prize Verified”として扱われ、評価はARC Prizeサイト（Semi-Private）由来と明記されている。 [3] 一方で、ARC-AGI-2の公式説明自体が「Semi-Privateは第三者（例：API）に限定的に露出した可能性がある」点を明示しており、**“Private（未露出）”と同一視はできない**。 [4]

ARC-AGI-2は、平均的なテスト参加者の平均スコアが **60%** とされ（一般公衆を中心に400+参加者で校正）、全タスクが「少なくとも2名が2試行以内（pass@2）で解ける」ように設計されている。 [5] 従って84.6%は、少なくとも“人間平均”を大きく上回る水準であり、同時に「効率（cost）を含めた知能」を重視するARC-AGI-2の思想に照らすと、コスト（$13.62/タスク）込みでの高得点は“推論系の進化（test-time computeと推論アルゴリズム）”の到達点を示唆する。 [6]

しかし、(a) 統計的ゆらぎ（テスト項目数が有限）、(b) Semi-Privateの露出可能性、(c) 「ARC形式／色マッピング」などフォーマット知識に起因する“汚染（contamination）”の兆候、(d) ベンチマーク領域の狭さ（2Dグリッド規則推論）により、**84.6%をAGI到達の証明とみなすのは不適切**である。ARC Prize側も「ARC PrizeベンチマークはAGIのリトマス試験ではない」と明記し、さらに2025年総括ではARC-AGI-3（探索・計画・記憶・目標獲得・アラインメント等のインタラクティブ推論）へ進む必要性を強調している。 [7]

一次情報に基づく事実関係と要約

本報告書が優先した一次情報（公式・査読・公式技術報告）は以下である（括弧内は本報告書での主な役割）。 [8]

- Google [1] 公式ブログ（2026-02-12）
  更新の狙い（科学・研究・工学の現実課題）、提供形態（Geminiアプリ／Gemini API早期アクセス）、事例（Rutgersの数学者、Dukeの材料研究、社内R&Dでのプロトタイプ）、主要ベンチ結果（ARC-AGI-2：84.6%、HLE：48.4%等）を提示。 [2]

- Google DeepMind [9] の評価方法PDF（2026-02）
  「ARC-AGI-2はARC Prizeサイト由来でVerified、v2 semi-private」と明記し、比較表（Gemini 3 Pro Preview、Claude Opus 4.6、GPT-5.2等）を提示。評価上の注意（他社値は原則自己申告、pass@1が基本、ベンチにより複数試行平均など）を明示。 [10]

- ARC-AGI-2公式ページ（ARC Prize Foundation）
  データセット構造（Training 1000 / Public Eval 120 / Semi-Private 120 / Private 120）、校正（400+参加者）、**平均人間スコア60%**、Semi-Privateの露出可能性、効率指標としてcostを導入する思想を明示。[11]

- ARC-AGI-2技術論文（arXiv, 2025-05版）
  人間実験のプロトコル・結果、タスクが複数test pairを持つ場合の扱い、選定基準（2試行以内に2人以上が解く）、ARC-AGI-2が難しい理由（独自性、情報量、合成的一般化）を詳細化。[12]

- ARC Prize Verified Testing Policy / Verified program
  Verifiedの考え方、Semi-Private／Privateの位置づけ、データ保持防止の合意（data retention）など、リーダーボード解釈の前提条件を規定。[13]

- ARC Prize 2025 Results & Analysis（2025-12）
  「refinement loop」潮流、ベンチ汚染（knowledge-dependent overfitting）の懸念、Geminiの出力がARC色マッピングを用いる例などを提示し、**“AGIではない”**と明言。[14]

- Gemini 3 Proモデルカード（Google DeepMind, 2025-11）
  Gemini 3 Proのアーキテクチャ（sparse MoE Transformer、ネイティブ・マルチモーダル、1M context等）と訓練データ概観、Deep Think mode（推論時の任意設定）、安全評価枠組みを記述。[15]

- Google DeepMind研究ブログ（2026-02-11）
  Deep Thinkモードを用いた研究エージェント（Aletheia：生成→検証→修正の反復、失敗宣言、検索・ブラウジング併用）など、より“科学研究の現場”に寄せたワークフローを記述。[16]

YouTube動画について（一次情報の扱い）：ARC Prize公式サイトはYouTube（チュートリアル等）への導線を提供しているが、本環境ではYouTube本体の取得が不安定で、公式動画の全文テキスト（字幕）を一次として精査できないものがある。そのため、本報告書は **公式ページ・論文・技術報告に含まれる同等内容**を一次根拠とし、動画固有の追加主張は「未指定（未検証）」として扱った。[17]

Gemini 3 Deep Thinkの技術分析

Gemini 3 Deep Thinkは、Google DeepMind[9]が「専門的推論モード」と位置づける“思考時間や探索を厚く使う”系統のモードであり、2026年2月の更新は「科学・研究・工学」向けの実課題適用（不完全データ、明確なガードレールや単一正解がない問題）を前面に出している。[2]

アーキテクチャ的特徴（公開範囲）

Gemini 3 Deep Thinkそのものの内部アーキテクチャ仕様（層数、総パラメータ数、エキスパート数、ルーティング方式の詳細、推論時探索アルゴリズムの形式化など）は、一次情報では詳細に公開されていない（未指定）。[18]
ただしベースとなるGemini 3 Proについては、**sparse Mixture-of-Experts（MoE）Transformer**で、入力トークンごとに一部エキスパートを動的に活性化し、総容量と提供コスト（計算量）を分離する狙いが明記されている。[15]

加えて、Gemini 3 Proはテキスト・画像・音声・動画入力をネイティブに扱い、**最大1M tokenのコンテキスト**、最大64K tokenの出力とされる。これらはDeep Thinkなど“長い思考／検証”を必要とするワークフロー（長文・複数モダリティ・コード混在）に整合的な仕様である。 15

訓練データと学習手法（公開範囲）

Gemini 3 Proモデルカードによれば、事前学習データは「公開Web文書、テキスト、コード、画像、音声（音声含む）、動画」を含む大規模・多領域・多モダリティであり、事後学習に instruction tuning、強化学習データ、人間の選好データが含まれる。さらに「多段推論・問題解決・定理証明データを活用可能な強化学習技術」で訓練される旨が記載されている。 15
一方、Deep Think固有の追加学習（例：specific RL objective、思考トレースの蒸留、探索木の訓練、自己検証器の併学習など）がどの程度行われたかは、Gemini 3 Deep Think（2026年2月更新）については未指定である。 18

推論モード「Deep Think」の技術的特徴（公開範囲）

Deep Thinkの“動作原理”を一次情報の範囲でまとめると、概ね以下の3層に整理できる。

1) **推論時間（inference-time compute）の拡張と並列探索**
Google公式（2025-08のDeep Think解説）では、Deep Thinkは「parallel thinking techniques」により“多数のアイデアを同時生成し、同時に検討し、必要に応じて改稿・統合して最良解へ収束する”と説明されている。また、思考時間を延ばして複数仮説を探索すること、さらにその“長い推論経路”を使うよう促す新しい強化学習技術を開発したとしている。 19
Gemini 3 Deep Thinkの2025-12提供時にも「advanced parallel reasoning」「複数仮説の同時探索」など、同系統の説明がなされている。 20

2) **検証・修正ループ（verifier / reviser）を含むエージェント的推論**
研究ブログでは、Deep Thinkを中核にした数学研究エージェント（Aletheia）が「生成→候補解→検証→（軽微修正なら修正器へ）→再検証／致命的欠陥なら生成へ戻る」という反復構造を持つと説明している。さらに「失敗を認める」ことが研究効率を改善した、と明記される。 16
この構造は“単発出力で当て切る”よりも、探索と検証の回転数で精度を上げる設計思想であり、ARC-AGI-2のような“少数例から規則を導出する”課題にも概念的に整合する。 5

3) **実運用向けのツール併用（検索・コード実行）と汚染対策**
2026年2月の評価方法PDFは、「検索＋コード実行」の設定を扱い、ベンチ結果の混入を避けるため“huggingface.com等を避けるブロックリスト”を実装していると明記する。 21
Deep Thinkの価値を研究・工学に結び付ける上で、検索や計算（コード）を推論ループに組み込み、出力の妥当性（引用・数値）を補強しようとする方針が示唆される。 22

性能向上の要因（推定と、一次情報で確定できること）

一次情報から“確定的”に言える要因は、**推論時の探索（並列思考＋長い思考時間＋RLによる誘導）**と、**研究・工学の現実ワークフロー（検証・修正、場合により検索・コード）**の組み合わせである。 23
さらにARC Prize側の分析では、同一タスクで「Gemini 3 Proは96 reasoning tokens、Gemini 3 Deep Thinkは138,000 reasoning tokens」という対比が紹介され、推論モードが“より長いプログラム（推論）”に強く相関するという観察が述べられている。 24
ただし、(a) どの程度がモデル重みの改善か、(b) どの程度が推論時探索（scaffolding / harness）か、(c) どの程度がデータ汚染・フォーマット既知性か、の分解は未指定であり、外部再現での因子解析が必要である。 25

ARC-AGI-2ベンチマークの設計とリーダーボード解釈

ARC-AGI-2は、ARC-AGI-1（2019）と同形式の2Dグリッド推論だが、推論系モデルを“より強く”試すために、タスクの独自性・情報量・合成的一般化を強化し、さらに人間側難度で校正した設計である。 [26]

データセット構成とタスク種別

公式ページでは、ARC-AGI-2が次の4区分から構成されるとされる（いずれも120タスクへ増加した旨の変更点が明示）。 [11]

- Training：1000タスク（非校正、公開）
- Public Eval：120タスク（校正、公開、pass@2で少なくとも2人が解ける）
- Semi-Private Eval：120タスク（校正、非公開、Kaggleの途中順位＆公開リーダーボードに使用。**第三者へ限定的露出の可能性**がある）
- Private Eval：120タスク（校正、非公開、Kaggle最終順位に使用。**第三者へ未露出**を意図） [27]

ARC-AGI-2が狙う能力面の代表例として、公式launch記事・技術論文は、(a) 記号的解釈（シンボルに意味を付与）、(b) 複数ルールの合成、(c) 文脈依存のルール適用（制御フロー的要素）、(d) インコンテキストのシンボル定義、といった“単一変換では解けない”方向性を強調する。 [28]

校正（calibration）と人間ベースライン

ARC-AGI-2は「Public / Semi-Private / Privateのevalセットは統計的に同等（IDD）で、過学習がなければスコアは<1pp程度で比較可能」と説明される。これは“semi-privateで高得点→privateでも近い”という期待を生む一方、**その前提が「no overfitting」**である点が重要である。 [11]

人間側の難度は、400+人の一般公衆を用いた現地実験で校正され、各タスクは「少なくとも2名が2試行以内（pass@2）で解ける」ことが要件となる。また公式ページは「平均テスト参加者スコアは60%」と明示する。 [5]
技術論文では、タスクの多くが1 test pair（68%）だが、2以上を持つタスクもあり、完全正解は“全test pairを解く”こと、部分正解は“一部のみ解く”こと、といった定義が使われている。 [29]

採点方法とリーダーボード解釈上の注意点

ARC Prizeの公式ガイドでは、コンペ採点として「各test input gridに対し2つの出力（attempt_1/2）を提出し、いずれかが完全一致なら1、外れなら0。複数test inputがあるタスクはそれぞれを採点し、平均する」という方式が説明される。 [30]
この設計は、(a)“2試行（pass@2）”が与えられる、(b)“タスク≒test outputの総数”が必ずしも一致しない（test inputが複数あり得る）という点で、スコアの統計解釈（標本サイズn）に影響する。 [31]

さらにリーダーボードは、単純な得点比較ではなく「スコア × コスト（cost-per-task）」の散布図で“効率”を可視化する思想を明確にしている。ARC Prize側は「知能は解けるかだけでなく、資源を最小にして効率的に解くこと」と位置づけ、costを比較軸に採用した。 [32]
一方で、ARC Prize Verified Testing Policyは「このリーダーボード（Semi-Private）には計算やインターネット利用の制限がない」と明記しており、**“経済的に有用な利用”をそのまま測る設計ではない**点にも注意が必要である。 [33]

84.6%スコアの意味と限界

84.6%が指すもの（評価セットと比較対象）

一次情報の整合する記述は次の通りである。

- 84.6%は「ARC-AGI-2（v2 / semi-private）」で、ARC Prizeサイト由来かつ“ARC Prize Verified”として取り扱われる。 [34]
- 同一の比較表では、Gemini 3 Pro Preview（Thinking High）31.1%、Claude Opus 4.6（Thinking Max）68.8%、GPT-5.2（Thinking xhigh）52.9%が並ぶ。 [35]
- 人間側は公式に「平均テスト参加者60%」で、全タスクは少なくとも2人がpass@2で解ける。 [11]

ここから、“同一ベンチ上での相対位置”としては **人間平均（60%）を大きく上回り、直近の主要推論モデルの上位設定を超える**ことが読み取れる。 [36]

統計的有意性（不確実性の源泉を明示）

ARC-AGI-2のSemi-Private Evalは「120タスク」と明記されているが、採点は“タスク内のtest output数”に依存し得る（複数test inputがあるため）。 [37]
そのため、厳密な信頼区間（CI）には **(A) 採点の分母（総test output数）** と **(B) 84.6%が小数第1位までの丸めか** が必要だが、一次情報では未指定である。 [38]

それでも“オーダー感”として、もし分母を120（タスク単位）と仮定して二項分布近似を置けば、84.6%の標準誤差は概ね3.3%程度となり、95%CIは±6–7ポイント規模になる（試行が独立同分布である等の仮定付き）。一方、分母がtest output総数（例：≈160前後）ならCIはやや狭くなる。これにより、**「68.8%（Claude）との差」や「52.9%（GPT-5.2）との差」は、通常の統計的不確実性を大きく上回る可能性が高い**が、厳密なCIは未指定である。 [39]

難度・分布・“解けた”の意味

ARC-AGI-2は、ARC-AGI-1より「独自性が高い」「情報量が多い」「合成的一般化が深い」ことが難度増の理由として挙げられる。 [40]
また、一般参加者の中央値は「test pairあたり2.3分程度」と報告され、ARC-AGI-1より“即答できる”問題が減っていることが示唆される。 [29]
したがって84.6%は、（少なくともARCが狙う）“短い例示からの規則導出”で高いカバレッジを達成したことを意味する。 [41]

過学習・リーク可能性（「semi-private」ゆえの構造的リスク）

84.6%の最大の解釈上リスクは、**評価セットそのものがSemi-Privateである**点である。公式説明では、Semi-Privateは「API等を通じて限定的に第三者へ露出した可能性がある」ため“semi”とされ、Privateは第三者へ未露出を意図するとされる。 [11]

ARC Prize側は、Verified Testing Policyで、モデル提供者と機密保持・データ保持防止（data retention）を含む合意を結び、Semi-Privateデータが保持されないよう協力すると述べる。ただし同時に、閉源モデル評価において“完全な秘匿を永久に保証できない”という構造は残る。 [42]

さらに重要なのは、ARC Prize自身が2025年総括で「ARC形式が公開データに十分含まれていることによる新しいタイプの“overfitting / contamination”が起きている可能性」を指摘している点である。具体例として、ARCタスクや色形式を明示していない検証ハーネスにも関わらず、Gemini 3 Deep Thinkが“ARC色マッピング

（例：Green=3, Magenta=6等）”を用いて推論しているログが示される。 14
この現象は、(a) ARC-AGIの“問題そのもの”が混入していなくても、(b) ARCに類似した形式・記法・色番号慣習が学習済みで、推論を有利にする可能性を示唆する。ARC-AGIの狙いは“内容の暗記”耐性だが、**“フォーマット既知性”は別の汚染チャネル**になり得る。 43

コスト（$13.62/タスク）とスケーラビリティ

ARC Prize（X投稿）では、Gemini 3 Deep Think（Semi-Private）について「ARC-AGI-2: 84.6% / $13.62 per task」、ARC-AGI-1では「96.0% / $7.17 per task」とされる。 44
ARC-AGI-2は120タスクなので、単純積ならフル評価一回あたり約$1,634程度のオーダーになる（外部要因で変動し得る）。 45

スケーラビリティを考えると、$13.62/タスクは「研究者の高単価な時間」「試作・実験コスト」「失敗の外部費用」を削減できる局面では十分正当化され得る。実際、Google公式は（a）数学論文の微妙な論理欠陥検出、(b)　半導体材料探索のための結晶成長レシピ設計、(c)　スケッチ→3Dプリント用ファイル生成、といった“高コスト領域での補助”を示している。 2
一方、同じARC Prizeの枠内でも、たとえば（年度・条件が異なるが）Opus 4.5（Thinking 64k）が37.6%で$2.20/タスク、TRMがARC-AGI-2で6.2%・$2.10/タスクなど、**「低コスト低精度」から「高コスト高精度」までの幅**が存在する。 46
したがって実務では、(1) “84.6%が必要な領域”か、(2) “より安いモード＋追加検証（人間／ツール）”で十分か、をタスク単位で判断する設計が重要となる。 47

AGI到達の主張を批判的に検証し、産業・研究・社会影響を見立てる

「AGI」と呼ぶための定義問題

ARC　Prizeは、ARC-AGIを「人間に易しくAIに難しいタスク集合とのギャップ」と捉え、このギャップがゼロになった時にAGI到達とみなす、という“ベンチマーク中心の定義”を提示している。 48
ただし同時に、Verified Testing Policyは「ARC Prizeのベンチマークは AI進歩を測るためであり、AGIのリトマス試験ではない」と明記している。 33
この二重性は、ARC-AGIが“AGIに向けた北極星”でありつつも、「単一ベンチでAGIを断定できない」ことを示す。 49

Gemini 3 Deep Thinkは何を満たし、何を満たさないか

満たしている可能性が高い点（一次情報で支持される範囲）：　- **“少数例からの規則獲得”の領域で、人間平均を超える適応能力**（ARC-AGI-2: 84.6%、人間平均60%）。 36
- **高度STEM領域での推論支援**（HLE、オリンピアド、コードフォース等の指標、および研究現場での事例）。 50
- **推論のプロセスを“探索→検証→修正”に寄せるワークフロー**（Aletheia等）と、それを支える“長い思考時間＋並列探索＋RL誘導”という設計思想。 51

満たしていない／未確定な点（本質的ギャップ）：- **実世界一般化**：ARC-AGI-2は2Dグリッドの抽象推論に強く、現実世界の長期計画、物理行為、社会的相互作用、環境探索などの多様な一般知能を直接測らない。
ARC　Prize自身が、探索・計画・記憶・目標獲得・アラインメント等を測るARC-AGI-3へ移る必要性を述べている。 52
- **安全性・信頼性**：Gemini 3 Proモデルカードは幻覚やタイムアウト等の一般的限界を認め、フロンティア安全評価を実施したとするが、Deep　Thinkが“より長い推論”を行うほど、(a)　誤推論の長文化、(b)　ツール誤用、(c) 説得・操作などの副作用リスクが上がり得る。現時点でDeep Think更新版の包括的リスク評価詳細は

未指定である（Gemini 3 Pro向けFSFは別資料）。 [53]
- **ベンチ汚染耐性**：ARC Prize側が“ARC形式（色マッピング等）”の内在化を汚染の兆候として示しており、84.6%が純粋な流動性（fluid intelligence）だけで説明できるかは未確定である。 [54]

産業・研究への影響と競合比較

研究・工学の現場への影響は、少なくとも「高価な専門家の思考プロセス（探索・検証・反例探し）を支援する“知的増幅器”」として強い。Google公式が提示する事例（論理欠陥検出、材料設計、3Dプリント生成）は、価値が“タスク正解率”ではなく「研究サイクル短縮」「失敗コスト削減」「探索範囲拡大」にある領域を狙っている。 [55]

競合比較は、DeepMindの表（ARC-AGI-2、HLE、MMMU-Pro等）により“少なくともこの設定ではDeep Thinkが上回る”という主張が構成されているが、同PDFは「非Geminiモデルの数値は原則プロバイダ自己申告」と明記し、比較公平性には留保が必要である。 [10]
このため、比較の最も堅い読み方は「**ARC Prize Verifiedのsemi-privateで84.6%**」という点（第三者機関の評価枠）に重心を置き、他ベンチ比較は“参考情報”として扱うことだろう。 [56]

倫理・社会的含意（ディストピア的懸念を含む）

本件の社会的含意は、単なる“チャット精度”ではなく、**高難度推論が（コストを払えば）広範に利用可能になる**点にある。ARC Prizeが強調する通り、推論系では「モデル＋推論時計算量＋ハーネス（反復）」が性能を決めるため、資本（計算資源・API費用）と性能の結びつきが強まる可能性がある。 [57]
その帰結として、(1) 研究・開発の速度格差（資本を持つ組織が優位）、(2) 労働市場の再編（分析・設計・レビューの自動化）、(3) デュアルユース（科学・工学の加速が悪用にも転用され得る）などが懸念される。
安全面では、Gemini 3 Proが危険能力に関する評価枠組み（FSF）に基づき評価された旨は示されるが、Deep Think更新（2026-02）固有の詳細は未指定であり、継続的監査が必要である。 [58]

今後の実験・検証提案

以下は、再現性・妥当性・コスト効率の3軸で、今後の検証を“研究として成立させる”ための提案である。
ARC Prize側が強調する「効率（cost）込みでの性能評価」「semi-privateの取り扱い」「ベンチ汚染への適応」を前提にする。 [59]

再現性テスト（第三者再現と分解可能性）

- **ARC Prize Verifiedの再試験**：同一モデルID・同一設定で、(a) 複数日実行、(b) 温度・サンプリング差、(c) 失敗タスクの再現性（同一タスクでの揺らぎ）を記録し、得点の分散を推定する。DeepMind側も“小規模ベンチは複数trial平均で分散低減”を述べており、ARC-AGI-2でも同様の分散解析が望ましい。 [60]
- **推論時資源の計測標準化**：ARC Prizeが“cost-per-task”を採用する一方、推論トークンや推論時間が実質的な内部資源であることが示唆されている（例：Deep Thinkのreasoning tokenが桁違い）。スコアとともに、(1) 生成トークン、(2) 推論トークン、(3) wall-clock、(4) APIコストを同時に公開するフォーマットを標準化する。 [61]

ストレステスト（汚染・フォーマット依存・分布外一般化）

- **色マッピング撹乱テスト**：ARC Prizeが示した“色マッピング（Green=3等）”依存の兆候に対し、(a) 色番号の置換、(b) 色数の変更、(c) 背景色のランダム化を行い、性能がどれだけ落ちるかを測る。これは“汚染”と“真の規則推論”を分離する強い診断になる。 [62]

- **フォーマット撹乱テスト**：JSON配列ではなく別表現（例：ランレングス表現、座標リスト、画像入力のみ）に変換し、性能低下を測る。ARC-AGIが“言語知識に依存しない”ことを目指す一方、モデルが特定表現に最適化される可能性があるため。 [63]
- **インタラクティブ能力の代替評価**：ARC-AGI-2は静的問題だが、ARC Prizeは探索・計画・記憶・目標獲得・アラインメント等を測る次世代（ARC-AGI-3）を示している。静的高得点が“行動を伴う一般知能”へ転移するかを検証する。 [64]

コスト削減のための評価設計（“スコアを買う”問題への対処）

- **段階的・逐次検定**：ARC Prize自身が、性能報告に効率指標（cost）を導入した背景には“計算で押し切る”ことへの懸念がある。実務では、(1) 低コスト設定で解けるか、(2) 解けないときのみ高コスト設定へ、というカスケード評価が有効である。 [65]
- **“コストあたり正解”指標の導入**：単純な$ / taskでは、難タスクと易タスクが混在する場合の解釈が難しい。`expected cost per correct`（例：$13.62 / 0.846）や、パレートフロンティア（同スコアで最安、同コストで最高）を併記することで、実務に近い判断が可能になる。 [66]

参考図表・比較表

リーダーボード比較表（主要ベンチの“同時提示”）

下表は、DeepMindの評価表（2026-02）に基づく主要指標の比較である（非Geminiモデルは原則プロバイダ自己申告、ARC-AGI-2はARC Prize Verifiedのsemi-private由来）。 [10]

ベンチマーク	Gemini 3 Deep Think（Feb 2026）	Gemini 3 Pro Preview（Thinking High）	Claude Opus 4.6（Thinking Max）	GPT-5.2（Thinking xhigh）
ARC-AGI-2	84.6%	31.1%	68.8%	52.9%
Humanity’s Last Exam（No tools）	48.4%	37.5%	40.0%	34.5%
Humanity’s Last Exam（Search + code）	53.4%	45.8%	53.1%	45.5%
MMMU-Pro（No tools）	81.5%	81.0%	73.9%	79.5%
Codeforces（Elo）	3455	2512	2352	—
IPhO 2025（theory）	87.7%	76.3%	71.6%	70.5%
CMT-Benchmark（pass@8）	50.5%	39.5%	17.1%	41.0%
IChO 2025（theory）	82.8%	69.6%	—	72.0%

性能対コスト表（ARC-AGI-2を中心に“効率”を見る）

ARC-AGI-2は“効率も知能の一部”という思想を公式に掲げる。 [67]
ただし、条件（semi-private / private、制約の有無、ハーネスの有無）が混ざると誤解が生じるため、下表は **一次情報でコストが明示されたもののみ**を載せ、条件差を明記する。

対象	セット/条件	スコア	コスト（$/task）	注記
Gemini 3 Deep Think	semi-private（2/26 実行とされる）	84.6%	13.62	ARC Prize X投稿で明示（詳細内訳は未指定）。 [68]
Claude Opus 4.6	semi-private（とされる）	68.8%	3.64	ARC Prizeの投稿一覧に含まれる旨が検索スニペットに示される（一次のページ取得は不安定）。 [69]
Opus 4.5（Thinking 64k）	semi-private（verified commercial model, 2025-12時点）	37.6%	2.20	ARC Prize年次分析で明示。 [24]
TRM（Tiny Recursion Model）	ARC-AGI-2（評価条件詳細は要確認）	6.2%	2.10	ARC Prize公式HFレポジトリの“replication results”。 [70]

読み取りの要点：84.6%は“性能”として突出する一方、ARC-AGI-2の思想では「性能を買っただけでは知能ではない」ため、今後は“同等性能をより低コストで達成する”方向（推論アルゴリズムの改良、検証ループの効率化、表現汚染耐性の改善）が主要研究課題になる。 [71]

Mermaidタイムライン（主要出来事）

（出典は各年の公式発表・論文・ブログに基づく。） [72]

```
timeline
    title ARC-AGI と Deep Think の流れ（公開一次情報ベース）
    2019 : ARC-AGI-1（ARC）提唱
    2024 : ARC Prize運営で検証・競技・リーダーボード強化（semi-privateの位置づけ強化）
    2025-03 : ARC-AGI-2公開・ARC Prize 2025開始（効率=costを重視）
    2025-08 : Deep Think（Gemini 2.5）で「並列思考＋推論時間拡張＋RL誘導」を公式説明
    2025-12 : Gemini 3世代でDeep Think提供（並列推論の説明）
    2026-02 : Gemini 3 Deep Thinkアップグレード（ARC-AGI-2 semi-privateで84.6%）
```

Mermaid推論フロー図（Deep Think型の“探索→検証→修正”）

（Aletheiaの記述と、Deep Thinkの公式説明を抽象化。） [73]

```
flowchart TD
    A[入力: 問題/データ/制約] --> B[並列仮説生成<br/>（複数案を同時探索）]
    B --> C[候補解（複数）]
    C --> D{検証器<br/>（整合性/反例/計算チェック）}
```

```
    D -->|合格| E[最終出力]
    D -->|軽微修正| F[修正器（reviser）]
    F --> C
    D -->|致命的欠陥| B
    D -->|解けないと判断| G[失敗宣言/要追加情報]
    G --> H[人間・追加ツール・追加実験]
```

Mermaid ER図（評価・コスト・スコアの関係）

（ARC-AGI-2が“score×cost”で評価する思想を、情報モデルとして整理。） [74]

```
erDiagram
    MODEL ||--o{ EVALUATION_RUN : "is tested in"
    BENCHMARK ||--o{ EVALUATION_RUN : "defines"
    DATASET_SET ||--o{ EVALUATION_RUN : "uses"
    EVALUATION_RUN ||--|| SCORE : "produces"
    EVALUATION_RUN ||--|| COST_METRIC : "produces"

    MODEL {
      string name
      string provider
      string mode  "e.g., base/Deep Think"
    }
    BENCHMARK {
      string name  "e.g., ARC-AGI-2"
      string metric "e.g., pass@2-like"
    }
    DATASET_SET {
      string visibility "public/semi-private/private"
      int tasks
    }
    EVALUATION_RUN {
      string date
      string protocol "prompt+harness+tools"
      string verification "verified/unverified"
    }
    SCORE {
      float value
      string unit "fraction/percent"
    }
    COST_METRIC {
      float dollars_per_task
      string pricing_basis "retail token pricing etc."
    }
```

---

[1] [4] [5] [9] [11] [27] [32] [37] [41] [45] [47] [59] [65] [66] [67] [71] [74] https://arcprize.org/arc-agi/2/

https://arcprize.org/arc-agi/2/

2 8 18 55 https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-deep-think/
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-deep-think/

3 10 21 34 35 36 38 39 50 56 60 https://storage.googleapis.com/deepmind-media/gemini/gemini_3_deep_think_model_evaluation.pdf
https://storage.googleapis.com/deepmind-media/gemini/gemini_3_deep_think_model_evaluation.pdf

6 44 68 https://x.com/arcprize/status/2021985585066652039
https://x.com/arcprize/status/2021985585066652039

7 13 33 42 https://arcprize.org/policy
https://arcprize.org/policy

12 26 28 29 31 40 63 72 https://arxiv.org/html/2505.11831v2
https://arxiv.org/html/2505.11831v2

14 24 25 43 46 49 52 54 61 64 https://arcprize.org/blog/arc-prize-2025-results-analysis
https://arcprize.org/blog/arc-prize-2025-results-analysis

15 53 58 https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf
https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf

16 22 51 73 https://deepmind.google/blog/accelerating-mathematical-and-scientific-discovery-with-gemini-deep-think/
https://deepmind.google/blog/accelerating-mathematical-and-scientific-discovery-with-gemini-deep-think/

17 30 https://arcprize.org/guide
https://arcprize.org/guide

19 23 https://blog.google/products-and-platforms/products/gemini/gemini-2-5-deep-think/
https://blog.google/products-and-platforms/products/gemini/gemini-2-5-deep-think/

20 https://blog.google/products-and-platforms/products/gemini/gemini-3-deep-think/
https://blog.google/products-and-platforms/products/gemini/gemini-3-deep-think/

48 https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025

57 https://home.mlops.community/public/videos/greg-kamradt-benchmarking-intelligence-or-arc-prize
https://home.mlops.community/public/videos/greg-kamradt-benchmarking-intelligence-or-arc-prize

62 https://huggingface.co/datasets/arcprize/arc_agi_v2_public_eval/blob/main/gemini-3-deep-think-preview/8698868d.json
https://huggingface.co/datasets/arcprize/arc_agi_v2_public_eval/blob/main/gemini-3-deep-think-preview/8698868d.json

69 https://x.com/arcprize?lang=en
https://x.com/arcprize?lang=en

70 https://huggingface.co/arcprize/trm_arc_prize_verification
https://huggingface.co/arcprize/trm_arc_prize_verification