ChatGPT

ARC-AGI-2 がリリース後に 70% に到達するまでの時間と、ARC-AGI-3 の 70% 到達予測

エグゼクティブサマリー

ARC-AGI-2（ARC-AGI v2）は ARC Prize [1] が **2025-03-24 に「本日ローンチ」**として公開した静的（入出力ペア）推論ベンチマークで、当初は「純粋なLLMは 0%、公開されている“推論AI”でも一桁%」と説明されました。 [2] その後、公式の半非公開（Semi-Private）評価での最高スコアは、2025-05-14 時点で約 3%（上位モデル群）だったところから、2025-12-05 に Poetiq [3] の refinement（ハーネス）で 54% が検証・公表され、さらに 2026-02-19 に Google [4] が Gemini 3.1 Pro の **検証済み 77.1%** を発表したことで、**70% 超えが確認**できます。 [5]

したがって、「ARC-AGI-2 がリリース後に ≥70% に到達するまでの時間」は、**（検証済み 77.1% が公表された）2026-02-19 を到達点**とみなすと、**2025-03-24 → 2026-02-19 の 332 日（約 11 か月）**です。 [6] ただし、この“70%”が **どの評価セット（Public / Semi-Private / Private）**で、**何の指標（pass@2 など）**かが曖昧なまま語られることが多いため、本レポートでは **「70%」の定義を複数提示し、定義ごとに結論が変わる**点を明確にします。 [7]

ARC-AGI-3 は 2026-03-25 に公開された **インタラクティブ環境（ゲーム）型**の新ベンチマークで、公開時点のフロンティアAIのスコアは **0.26% 前後（Semi-Private）**と報告されています。 [8] この初期点（0.26%）から 70% までの距離は非常に大きく、ARC-AGI-2 の“急伸”を単純に外挿するのは危険です。ARC-AGI-1（2019→2024 で 70% 超）と SWE-bench Verified（2023→2025〜2026 で 70% 超）という **少なくとも 2 つの類似「ベンチマーク進捗曲線」**を参照すると、ARC-AGI-3 の 70% 到達は **最短でも 1.5〜2 年、中央値で 3〜4 年、遅い場合は 6〜8 年**程度のレンジが合理的（ただし不確実性は極めて大）という結論になります。 [9]

ベンチマークと「70%」の定義

ARC-AGI-2 の公式データ構造と評価セット

ARC-AGI-2 の公開リポジトリ（タスクデータ）では、以下が明示されています。

- 公開トレーニング 1000 問、公開評価 120 問。 [10]
- さらに **2つの非公開テストセット**：
- **Semi-Private**：商用APIモデル等を（低リークリスクで）遠隔評価するため。
- **Fully-Private**：コンペの最終順位決定等、リークリスクを最小化するため。 [10]
- 「解けた（solve）」の基準は、初見でテスト入力すべての正解出力を作り、**各テスト入力につき 2 回の試行**が許される、という形で記述されています。 [11]

この「2 回試行」ルールから、ARC-AGI 系でしばしば用いられる **pass@2 相当の“正解率（%）”**がスコアとして扱われている、と解釈するのが自然です（ただし “pass@2” という表記自体は、ソースによっては明示されません）。 [11]

「70%」が意味しうる複数の定義

本件は、ソース側が「70%」の定義を固定していないため、少なくとも次の3通りが現実的です。

- **定義A：ARC-AGI-2 Semi-Private の“検証済み（Verified）”スコアが ≥70%**
  最も“公式・比較可能”になりやすい定義。Google [4] が 2026-02-19 に ARC-AGI-2 で **検証済み 77.1%**を公表しており、この定義では **到達済み**です。 [12]
- **定義B：ARC-AGI-2 Fully-Private（コンペ最終）スコアが ≥70%**
  2025 年大会の最終（Private）SOTA は **約 24%**と報告されており、この定義では **未到達**（少なくとも 2025 年末時点）です。 [13]
- **定義C：Public Eval（公開評価セット）で ≥70%**
  公開セットは反復検証による過適合リスクが高く、“科学的な進捗”として扱うには注意が必要ですが、Poetiq の 2025-11-20 の記事は Public Eval 上での優位（平均的人間を超えた旨など）を主張し、Semi-Private への性能低下も想定しています。 [14]

以降の「ARC-AGI-2 が ≥70% に到達するまでの時間」は、ユーザー要件（公式・一次ソース優先）に合わせ、**定義A（Semi-Private / Verified）を主軸**にしつつ、定義B・C も併記します。 [15]

ARC-AGI-3 のスコア定義

ARC-AGI-3 は「インタラクティブ環境を探索し、ルールも目標も明示されない中で勝利条件を発見し、上位レベルへ一般化する」ことを要求するベンチマークとして、2026-03-25 に公開されました。 [8]

ARC-AGI-3 の技術レポートでは、リリース時点の公式（Semi-Private）リーダーボードとして、例として **0.37% / 0.26% / 0.25% / 0.00%**といった極小スコアが掲載されています。 [16]
この “Score” は、ARC Prize 側の説明では「人間が 100%」で、フロンティアAIは 0.26% 程度と述べられています。 [8]
また、ARC-AGI-3 Preview の位置づけ記事では「**行動効率（action efficiency）を人間ベースラインに対して正規化し、ゲームごとに 0–100% にし、全ゲーム平均で 0–100%**」という枠組みが明示されています。 [17]

従って ARC-AGI-3 の「70%」は、**“人間の行動効率に対する正規化スコアで 70%”**を意味し、ARC-AGI-2 の“正答率 70%”とは **同じ「%」でも意味が異なる**点が重要です。 [18]

ARC-AGI-2 の ≥70% 到達タイムライン

公式リリース日と初期状況

ARC Prize は 2025-03-24 の告知で「本日 ARC-AGI-2 をローンチ」とし、純粋 LLM は 0%、公開推論システムでも一桁%と述べています。 [2]
同日付の ARC-AGI-2 技術報告（HTML版）でも、ARC Prize 2025 が 2025-03-24 にローンチし、ARC-AGI-2 を中心に据えると記載されています。 [19]

タイムライン表（定義A：Semi-Private / Verified）

下表は、「70%」を **定義A（Semi-Private の Verified/公式スコア）**としたときの、時系列に沿った主要データ点です。スコアは原則 “%” 表記に統一し、ソースが小数（0.771 等）で表す場合は % に換算しています。 [20]

日付	スコア（%）	評価セット / 指標の読み	測定主体・方法（一次ソース）	根拠
2025-03-24	0（定性的）	ARC-AGI-2 公開直後の概況（純粋LLMは 0%）	ARC Prize のローンチ告知（定性的説明）	2
2025-05-14	約 3.0	Semi-Private（ARC-AGI-2 技術報告の “as of May 14, 2025” 表）	ARC-AGI-2 技術報告 Table 1（Semi-Private）	19
2025-06-05	8.6	Semi-Private（主要“推論AI”横並びテスト）	ARC Prize の比較記事に掲載された表（ARC-AGI-2 Scores）	21
2025-12-05	54.0	Semi-Private（Verified / 公式検証）	ARC Prize 側が Poetiq の OSS solver を Semi-Private で評価し 54% と公表、また ARC Prize 年次分析でも 54% と言及	22
2026-02-19	77.1	「verified score」＝公式検証スコア（文脈上、ARC Prize の Verified 指標）	Google の公式記事が ARC-AGI-2 で “verified score 77.1%” を公表	23

≥70% 到達時点（定義A）：2026-02-19 に 77.1% が公表されており、ここで 70% を超えています。 23

70% 到達までに要した日数

ARC-AGI-2 の公開日（2025-03-24）から、検証済み 77.1% が公表された日（2026-02-19）までの差は **332 日（約 11 か月）**です（計算は日付差分）。 6

同様に、参考として 54%（2025-12-05）までは公開日から **256 日**です。 24

定義B・C の場合の到達状況

- **定義B（Fully-Private / コンペ最終）**
  2025 年（ARC Prize 2025）では Private Eval の上位スコアが約 24% と報告されており、≥70% は未到達です。 13
- **定義C（Public Eval）**
  Public Eval は 2025-11-20 の Poetiq 記事が優位を主張しますが、同記事自身が Semi-Private への性能低下（Public→Semi-Private のドロップ）を想定しており、Public 上の ≥70% を “公式到達” とみなすのは慎重であるべきです。 14

結論として、「ARC-AGI-2 が ≥70% に到達したか？」は **どの定義かで答えが変わる**ものの、ユーザー要件の「公式/一次ソース」優先に沿う **定義A（Semi-Private / Verified）では 2026-02-19 時点で到達が一次ソースで確認可能**です。 12

```
timeline
    title ARC-AGI-2 と ARC-AGI-3：主要マイルストーン（%は各ベンチマークの定義に従う）
    2025-03-24 : ARC-AGI-2 公開（純粋LLMは0%と説明）
    2025-05-14 : ARC-AGI-2（Semi-Private）上位モデル ~3%
    2025-12-05 : ARC-AGI-2（Semi-Private, Verified）54%
```

```
2026-02-19 : ARC-AGI-2（Verified）77.1%（≥70% 到達）
2026-03-25 : ARC-AGI-3 公開（Frontier ~0.26%）
```

成長率の分析と類似ベンチマークとの比較

ARC-AGI-2 の成長率をどう捉えるべきか

ARC-AGI-2　のスコア推移は、単純な「毎月一定割合で伸びる」タイプではなく、**（1）評価手法（ハーネス/推論ループ）と（2）基盤モデルの世代更新**が組み合わさって「段差（ステップ）」として現れています。
2025-12-05 時点で ARC Prize 自身が“refinement loop”による進展（Poetiq の 54% など）を大きく取り上げていることからも、改善の主因が単純なスケール則だけではないことが示唆されます。 [25]

上の折れ線図（本回答内のチャート）に対応する区間平均（概算）をとると、例えば 8.6%→54% は約 6 か月で +45.4pt 程度であり、平均すると月 +7〜8pt に相当します。しかしこの“平均”は、Poetiq の Verified のような **新方式の投入**に強く依存しており、外挿には不向きです。 [26]

類似進捗曲線の比較

ユーザー要件に合わせ、「あるベンチマークが 70% に到達するまでの時間」という観点で、少なくとも 2 つの類似ケースを並べます。

ARC-AGI-1

ARC-AGI-1 は 2019 年に François Chollet [27] により導入され、長期にわたり進捗が遅かった一方、2024 年末に o3 系のテストタイム計算（test-time adaptation / compute）で 70% 超の領域に入った、と ARC-AGI-2 技術報告が総括しています。 [28]
このケースは、**「静的ARC形式」でも“新しい推論パラダイム”が出現するまでは 70% に到達しない**ことを示す前例です。 [19]

SWE-bench Verified

SWE-bench は 2023 年 10 月にリリースされ、初期ベースラインは 1.96% と説明されています。 [29]
一方で SWE-bench の公式サイトは、SWE-bench Verified で **mini-SWE-agent が最大 74%**に達している旨を掲示し、2025 年 7 月に 65% 到達などの節目も併記しています。 [30]
このケースは、**“エージェント化（ツール・実行環境・反復）”が入ると、ベンチマークが比較的短期間で 70% のラインに到達し得る**ことを示唆します。 [30]

比較表

ベンチマーク	リリース（一次/公式）	≥70% 到達の観測（一次/公式または公式サイト掲示）	概算所要時間	注記
ARC-AGI-2（Semi-Private, Verified）	2025-03-24 [2]	2026-02-19 に 77.1%（verified） [23]	332 日	“70%”=Verifiedスコア（定義A）
ARC-AGI-1（Semi-Private など）	2019（導入） [19]	2024 年末に 70%超（o3-preview 等） [28]	約 5 年	パラダイム転換まで遅い

ベンチマーク	リリース（一次/公式）	≥70% 到達の観測（一次/公式または公式サイト掲示）	概算所要時間	注記
SWE-bench Verified	2023-10（リリース）[29]	公式サイト掲示で up to 74% [31]	≤約 2–2.5 年（上限推定）	≥70% の“正確な初到達日”はサイト上からは確定できず

この比較から、ARC-AGI-2 の 332 日は「速い部類」ですが、**ARC-AGI-1 のように長い停滞 → 急伸**という形もあり得るため、ARC-AGI-3 予測では“急伸前提の短期外挿”を避けます。[32]

ARC-AGI-3 が ≥70% に到達するまでの予測

出発点と到達目標の定量化

ARC-AGI-3 は 2026-03-25 公開で、公開時点のフロンティアAIは **0.26% 前後**と報告されています。[8]
この 0.26%→70% は、単純比で **約 270 倍**の改善が必要です（70 / 0.26 ≈ 269）。[33]

また ARC-AGI-3 は「ルールも目標も明示されず、探索して学習し、レベルを跨いで適応」する形式で、静的 ARC とは必要能力（探索・記憶・方策学習・長期計画）が異なります。[8]
この差は、ARC-AGI-2 の“推論ループと基盤モデル刷新”が効いた構図が、そのまま ARC-AGI-3 に当てはまらない可能性を意味します。[34]

シナリオ別予測（楽観・中央値・悲観）

ここでは「70%」を ARC-AGI-3 の **正規化スコア 70%**として、3シナリオを提示します（確率は主観的ベイズ事前であり、後述の感度要因にかなり敏感です）。

楽観シナリオ（到達：2027 年後半〜2028 年前半、確率 20%）

- 予測到達：**公開から約 18〜24 か月**（2027-09〜2028-03 頃）
- 主要仮定
- 大規模な“インタラクティブ学習”訓練（合成環境やゲーム様タスク）と、推論モデル（LRM）を統合した **探索→仮説→検証→方策更新**の標準レシピが確立する。[35]
- ARC Prize 側の評価条件が大きく変わらず（ゴールポスト移動が小さく）、手法改良がスコアに直結する。[36]
- 根拠にした類推
- ARC-AGI-2 が 11 か月で 70% 超に到達した“急伸局面”が、ARC-AGI-3 でも早期に再現される場合。[37]

中央値シナリオ（到達：2029 年頃、確率 50%）

- 予測到達：**約 3〜4 年**（2029-03〜2030-03 頃）
- 主要仮定
- エージェントの探索と学習は改善するが、
  - 長期計画、
  - レベル間の転移、
  - 人間に近い行動効率（無駄手を減らす）
    の同時達成が難しく、70% に近づくほど“最後の数十ポイント”が重くなる（性能飽和）。[38]

- 2026〜2027 にかけては 1〜5% 程度→10〜20% 程度への改善は起こるが、70% は複数回のパラダイム改良を要する。 16
- 根拠にした類推
- ARC-AGI-1 のように“新形式のベンチマーク”は、急伸の前に一定期間の試行錯誤が続く可能性。 39

悲観シナリオ（到達：2032〜2034 年頃、確率 30%）

- 予測到達：**約 6〜8 年**
- 主要仮定
- エージェントが「勝たせる」こと自体はできても、**人間並みの行動効率**（探索コストを抑えつつ一般化）を要求するスコア設計が、LLM+外部ループの延長では伸びにくい。 40
- ベンチマーク側がリーク耐性・過適合耐性を高めるために構成や評価を継続的に調整し、実質的に“移動ターゲット”になる。 41
- 根拠にした類推
- 静的 ARC でも、長期的には手法が“買い増し（計算量）”に寄りがちで、効率化が難しい局面があるという ARC Prize 側の問題意識。 42

感度分析（ETA を大きく動かす要因）

ARC-AGI-3 の 70% 到達予測は、以下 4 変数に特に敏感です。

- **データ（学習環境）の供給**：多様なインタラクティブ環境で“探索→学習→転移”を鍛えられるか。 43
- **計算資源とコスト構造**：ARC-AGI-3 は“効率”が指標に入るため、単純な計算量増加がスコアに反映されにくい可能性。 44
- **アーキテクチャ変化**：世界モデル、階層型方策、メモリ、探索計画など、LLM推論の外側にある中核要素の成熟。 43
- **評価の変動（ゴールポスト）**：Semi-Private の更新頻度や、反復提出・学習の扱い、ハーネスの扱い（公式かコミュニティか）。 36

不確実性・欠損データ・前提の明示

本調査には、構造的に避けにくい不確実性があります。

第一に、「70%」という数値は、ARC-AGI-2 では **評価セット（Public/Semi-Private/Private）**と**検証状態（Verified/非検証）**で意味が変わります。特に Public Eval は反復利用しやすく、リーク・過適合の議論が避けられません。 45

第二に、ARC-AGI-2 の“いつ 70% を初めて超えたか”について、現時点で一次ソースとして確実に日付が固定できるのは **Google の 2026-02-19 の 77.1% 公表**であり、それ以前に 70% を超えた可能性は否定できません（例：他社が先に検証で超えていたが未公表、など）。このため本レポートの “332 日” は「最遅でも（no later than）」の意味合いを含みます。 46

第三に、ARC-AGI-3 は公開直後で、時系列データ点がほぼ存在しません。したがって予測は、ARC-AGI-1/2 や SWE-bench Verified のような外部曲線からの類推に強く依存し、分布の裾（特に悲観側）が厚い推定になります。 47

---

1 3 12 23 37 46 https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/

2 6 24 27 https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025

4 7 10 11 15 45 https://github.com/arcprize/ARC-AGI-2
https://github.com/arcprize/ARC-AGI-2

5 9 19 20 28 32 39 https://ar5iv.org/abs/2505.11831
https://ar5iv.org/abs/2505.11831

8 35 43 https://arcprize.org/blog/arc-agi-3-launch
https://arcprize.org/blog/arc-agi-3-launch

13 25 https://arcprize.org/blog/arc-prize-2025-results-analysis
https://arcprize.org/blog/arc-prize-2025-results-analysis

14 https://poetiq.ai/posts/arcagi_announcement/
https://poetiq.ai/posts/arcagi_announcement/

16 33 34 36 41 47 https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf
https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf

17 18 38 40 44 https://arcprize.org/blog/arc-agi-3-preview-30-day-learnings
https://arcprize.org/blog/arc-agi-3-preview-30-day-learnings

21 26 42 https://arcprize.org/blog/which-ai-reasoning-model-is-best
https://arcprize.org/blog/which-ai-reasoning-model-is-best

22 https://poetiq.ai/posts/arcagi_verified/
https://poetiq.ai/posts/arcagi_verified/

29 https://www.swebench.com/original.html
https://www.swebench.com/original.html

30 31 https://www.swebench.com/
https://www.swebench.com/