ChatGPT

ARC-AGI-4 深層調査報告：ARC-AGI-3／GPT-6 Astra後の「自律的・オープンエンドなイノベーション」評価

Executive Summary

調査基準日：2026年9月13日（日本時間）

- **ARC-AGI-4は実在する次期ベンチマーク構想として、ARC Prize Foundationから公式に予告された。** 公式X投稿の中心表現は **“a benchmark for autonomous open-ended innovation”** であり、オープンソース路線を維持して研究コミュニティの共有目標にするとしている。投稿IDの時刻を換算すると、発表は2026年9月12日19:03 UTC、**9月13日04:03 JST**に相当する。 [1]
- ただし、**2026年9月13日時点でARC-AGI-4の正式なベンチマーク仕様書、FAQ、設計文書、タスク例、スコアリング、合格基準、公開日程は確認できない。すべて「未指定」**である。ARC Prize公式サイトのベンチマーク・ナビゲーションも現在ARC-AGI-1〜3までで、ARC-AGI-4の専用ページはまだない。 [2]
- ユーザー提示の「**ARC-3の攻略を受けた次期ベンチマーク**」という理解は、時系列とARC Prize自身の発言からみて妥当な推論である。9月3日のAstra分析でARC Prizeは、ARC-AGI-3を「決定論的・closed-ended」で現実世界のopen-endednessを表さないと整理し、**recursive self-improvement と open-ended innovation を次世代ベンチマークでどう評価するか検討している**と明言した。その10日後にARC-AGI-4が名指しで予告された。したがって因果関係は強く示唆されるが、「Astra攻略を理由にARC-4を作った」という公式の単一文書による明示まではない。 [3]
- **GPT-6 Astraの「99.9%」には重要な条件がある。** 中立的なARC Prize **Standard harness** ではSemi-Privateセットで最高 **62.7%、約2.61万ドル**。OpenAIのモデル固有の推論状態保持・コンパクションを許す **Provider Adapter harness** では **99.9%、約1.88万ドル**だった。したがって「ARC-AGI-3=無条件に99.9%で突破」という表現は不正確である。 [4]
- 急進化の本質はモデル単体だけではない。GPT-5.6 Solでは**同じモデルのまま**推論状態の保持とコンテキスト・コンパクションを有効化しただけで公開ARC-3スコアが **13.3%→38.3%**、出力トークンが約6分の1になった。その後のAstraでは、未知環境を**圧縮された記号世界モデルへ変換し、自作DSL、探索器、プランナー、状態同期ツールまで作る**行動が観測された。ARC-3は結果的に「モデル能力＋記憶／コンテキスト管理＋エージェント・ハーネス」を強く測る評価になった。 [5]
- そのためARC-AGI-4が本当に「open-ended innovation」を測るには、固定問題の正答率では不十分である。本報告では、**科学的妥当性、独立再現性、新規性、有用性、一般化、自律性、資源効率、長期安定性、自己改善の転移、評価完全性、安全性**を分離して測り、時間とともに「検証済み新規発見」が増え続けるかを評価する方式を推奨する。PaperBench、RE-Bench、DiscoveryWorld、ScienceAgentBench、METR Time Horizon、FunSearch等はそれぞれこの設計の一部を既に具体化している。 [6]
- **最大の評価上の課題は「オープンエンドなのに、どう有限時間で測るか」である。** ARC-4は「終わりのない発見」を直接証明するテストではなく、実際には「新しい問題分布が追加され続けても、検証済みイノベーション率が飽和せず、他分野へ転移し、人間並み以上の資源効率を維持できるか」を統計的に測る設計になるべきである。これは現時点の公式仕様ではなく、本報告の設計提案である。
- 安全面では、これは通常の知能ベンチマークより重大である。OpenAIのPreparedness Frameworkは**生物・化学、サイバー、AI自己改善**を重大能力カテゴリとして追跡しており、GPT-6 AstraはOpenAI自身の評価で**初のCritical級サイバー能力モデル**とされる。したがってARC-4の高得点は、科学生産性だけでなく自動R&D・デュアルユース能力の安全評価と連動させる必要がある。 [7]

本報告の確度区分

項目	現時点の確度
ARC-AGI-4という名称・方向性	**高**：ARC Prize公式発表あり。 8
「autonomous open-ended innovation」が主目的	**高**：公式文言。 1
オープンソース路線	**高**：公式発表で明示。 8
科学的発見を重要な応用先と位置づけること	**高**：発表中でscientific innovationを明示。 8
ARC-3/Astraの飽和がARC-4の直接の契機	**中〜高**：時系列・ARC Prizeの9月3日発言からの推論。 9
ARC-4のタスク形式、評価式、合格点	**未指定**
ARC-4の正式公開日・コンペ日程	**未指定**
本報告で提案するメトリクス・実験手順	**提案**：公式仕様ではない

公式発表の確認とARC-AGI-4の現時点の仕様

要点

- 公式に確認できるARC-AGI-4の中心定義は「**自律的なオープンエンド・イノベーションのベンチマーク**」。 1
- 「科学的発見」は、現時点では**具体的なタスク仕様ではなく、ARC-4が狙う能力の社会的・科学的意義を説明する文脈**で述べられている。 8
- 専用Webページ、FAQ、technical report、dataset card、scoring specification、task examplesは**未公開／未指定**。公式サイトもまだARC-AGI-3までを掲載している。 2
- ARC Prize全体の既存哲学――「未知問題への人間並みの学習効率」「事前知識ではなく一般化」「easy for humans, hard for AI」――はARC-4の設計を読む上で重要な手掛かりになるが、ARC-4にそのまま継承されるかは未指定である。 10

ARC Prizeの公式X発表は、著作権上ここで全文転載するのではなく内容を正確に要約すると、次の五点からなる。第一にARC-AGI-4を **“a benchmark for autonomous open-ended innovation”** とする。第二にオープンソースへのコミットメントを継続し、研究コミュニティの共通目標にする。第三に、急速なAI進歩にもかかわらずopen-ended inventionでは人間が依然優位だとする。第四に、その能力は科学技術全般の進歩を解放する「メタスキル」だと位置づける。第五に、高度AIによる科学的イノベーションは新技術・知識・理解をもたらすという肯定的なビジョンを示している。 1

ここでユーザー提示の「**自律的で終わりなきイノベーションと科学的発見**」は概ね方向性を捉えているものの、厳密には少し補正が必要である。公式の核となる英語は **autonomous open-ended innovation** で、「endless」という語がテスト仕様として定義されたわけではない。また科学については「advanced AI capable of scientific innovation」という将来像として述べられており、ARC-4が物理・化学・生物などの実科学実験を必ず含むとはまだ発表されていない。 8

さらに、ARC Prize自身が従来から「現在のAIと人間の間に残るresidual gap」を測ることを使命とし、**現状ではopen-ended innovation and discoveryを可能にするのは人間だけ**だという立場を公表してきた。ARC PrizeのAGI定義は、「人間と同等の学習効率で、開発者が予期しなかった新問題へ適応できるシステム」であ

る。したがってARC-4は突然のテーマ転換というより、ARCシリーズの「学習効率」を、静的パズル→エージェント→発明・発見へ拡張する自然な次段階と読むのが妥当である。[10]

現時点で何が指定済みか。

ARC-AGI-4要素	2026年9月13日時点	根拠
名称	**ARC-AGI-4**	ARC Prize公式X。[8]
上位目的	**Autonomous open-ended innovation**	公式X。[1]
オープンソース方針	**指定済み：継続する意向**	公式X。[8]
科学的発見との関係	**目的・動機として言及**	公式X。[8]
正式なBenchmark Specification	**未指定／未公開**	公式サイトはARC-1〜3のみ。[11]
Technical Report / Design Document	**未指定／未公開**	同上。[11]
FAQ	**未指定／未公開**	同上。[11]
タスク形式・例	**未指定**	公式予告に記載なし。[8]
Human baseline	**未指定**	公式予告に記載なし。
スコアリング式	**未指定**	公式予告に記載なし。
合格／飽和基準	**未指定**	公式予告に記載なし。
許容ツール・インターネット	**未指定**	公式予告に記載なし。
計算量・費用上限	**未指定**	公式予告に記載なし。
リリース日・コンペ日程	**未指定**	公式予告に記載なし。

この「未指定」が非常に重要である。現段階のARC-AGI-4について「100点満点」「科学論文を自動執筆」「週間実験」「再帰的自己改善を必須とする」などの具体仕様を断定する資料は確認できない。

一方、9月3日のARC PrizeによるAstra分析は事実上の**ARC-4設計問題に対する前文**になっている。ARC PrizeはARC-AGI-3を「tightly bounded」「deterministic」「closed-ended」と整理し、飽和はAGIを意味せず、次世代では **recursive self-improvement と open-ended innovation** をどう評価するか検討中だと述べた。[9]

したがって、現時点で最も安全な結論は、

> **ARC-AGI-4は「公開済みベンチマーク」ではなく、「目的が公式に発表された、設計途上の次期ベンチマーク」である。**

というものである。

ARC-AGI-3とGPT-6 Astraの「急進化」は何だったのか

要点

- ARC-AGI-3はARC-1/2の静的グリッド推論から、**探索・世界モデル構築・目標推定・計画実行**を必要とするinteractive agent benchmarkへ移行した。[12]
- 2026年春の初期フロンティアモデルはSemi-Privateで **0.1〜0.5%級**だった。[13]
- その後、同じモデルでも状態保持方式を変えるだけで成績が数倍になり、最終的にGPT-6 AstraはStandardで62.7%、Provider Adapterで99.9%へ到達した。[14]
- したがってARC-3を破った「能力」は、単純な視覚パズル力ではなく、**長期記憶、状態圧縮、動的な仮説形成、記号世界モデル、自己生成ツール、探索から実行への切替え**の組み合わせと見るべきである。[15]

ARC-AGI-3は、未知の抽象的なturn-based environmentを最初からプレイさせる。エージェントにはゲームの明示的ルールや目標が渡されず、行動して結果を観測しながら、何が起きているかを推定する必要がある。ARC Prizeはその能力を **Exploration、Modeling、Goal-setting、Planning and execution** の四要素に分解している。[12]

ARC-1/2が「少数の入力／出力例から変換則を推定する」data-efficient modelingを中心としたのに対し、ARC-3では情報そのものを行動によって取りに行かなければならない。この違いは大きい。ARC-3は学習データ効率だけでなく、**どの実験を行えば環境について最大の情報が得られるか**まで評価するからである。[16]

ARC-3のデータ構成も汚染対策を強く意識している。Technical ReportではPublic Demo 25環境、Semi-Private 55環境、Fully Private 55環境とし、ARC-2までとは逆に**private側を評価の主体**にした。これは公開タスクを大量に模倣生成して訓練することや、公開ゲーム専用ハーネスを作ることが一般知能の測定を壊すためである。[17]

ARC-3の正式なスコアはRHAE（Relative Human Action Efficiency）である。あるレベルについて人間ベースラインの操作数を h、AIの操作数を a とすると、基本的に

$$S_{\text{level}} = \min\left(1.15, \left(\frac{h}{a}\right)^2\right)$$

で評価される。効率の二乗を使うため、100操作を要するAIが人間なら10操作で解けるレベルをクリアしてもスコアは1%になる。レベルごとに最大115%とし、後半レベルを重くして環境内平均を取り、最終的な全体スコアは0〜100%に正規化される。[18]

この設計は「正解したか」ではなく、**未知環境からどれほど少ない経験量で必要な規則を学べたか**を測る点でARCらしい。逆に、内部で何兆FLOP使ったかはRHAEの主指標ではない。つまりARC-3は「total compute efficiency」ではなく、主として**environmental experience efficiency**を測る。ARC Prize自身も、action efficiencyと計算コスト効率を別概念として説明している。[19]

初期リリース時のSemi-PrivateスコアはAnthropic Opus 4.6 Max 0.50%、Gemini 3.1 Pro Preview 0.40%、OpenAI GPT-5.4 High 0.20%、Grok-4.20 0.10%だった。つまり春時点ではフロンティアモデルにとってほぼ未攻略だった。[13]

ところが七月、OpenAIが重要な測定上の弱点を示した。GPT-5.6 Solは従来方式ではARC-3で7.8%程度だった一方、ARC-3の各ターンごとに**private reasoning stateを捨て、古い履歴をrolling truncationする構成**が、ゲームを跨いだ推論の連続性を阻害していた。OpenAIはResponses APIで推論状態保持とコンパクションを

有効化し、公開セットの同一モデルを **13.3%→38.3%**へ引き上げ、出力トークンも約6分の1に削減した。これは「モデル重みを変えず、ハーネスだけで約3倍」である。 [20]

この時点でARC-3について重大な教訓が得られた。

> **長期エージェント評価では、モデルの知能とハーネスの記憶設計を完全に切り離すことができない。**

Astraではこの傾向がさらに強く現れた。ARC Prizeの2026年9月3日の検証結果は次の通りである。 [19]

GPT-6 Astra推論努力	Standard harness	Provider Adapter
max	**62.7% / $26,098**	98.6% / $17,332
xhigh	59.3% / $37,317	98.4% / $18,147
high	54.8% / $40,705	**99.9% / $18,817**
medium	38.6% / $48,090	98.4% / $19,285
low	17.5% / $38,166	98.0% / $21,298
none	35.2% / $49,791	96.7% / $23,457

出典：ARC Prize Foundation。 [19]

ここで**62.7%と99.9%は矛盾する数字ではない**。Standard harnessはプロバイダ中立の最小インターフェースで、モデルが残すvisible notesだけを持ち越す。一方Provider AdapterはOpenAI固有のopaque reasoning stateをターン間で保持し、長い会話ではcompactionを使える。ARC Prizeによれば、Provider Adapterは両条件で解けた167のgame-reasoning pairについてaggregate elapsed timeが約3.66倍高速で、総トークンも49%少なかった。 [9]

この違いは日本語報道でも論点になっており、WAKARAは「99.9%か62.7%か」を条件差として解説し、AI新聞も99.9%をそのまま中立的ARC条件と解釈することへの注意を報じている。OpenAI自身も日本語公式記事で、ベンチマークがモデルだけでなくAPI設定・ハーネスを測ることを強調している。 [21]

Astraの質的な突破点はさらに興味深い。ARC Prizeは、Astraがゲーム画面を逐一自然言語で覚えるのではなく、オブジェクト、座標、状態遷移、操作規則を**高密度の独自記号表現へ圧縮した**と報告している。これは完全なプログラミング言語ではないが、ゲームごとのDSLに近く、世界モデルと未来の行動計画を短い記号列で保持する。 [19]

さらにPRO-LONGハーネスでは、Astraはsandbox内でゲームごとにboard parser、state model、search algorithm、planner、persistent notesを自ら作成した。`tu93`では`maze_solver.py`、`combat_solver.py`、`patrol_solver.py`、`sync_state.py`などを生成し、観測と予測モデルを照合した。ただしARC Prizeは、これはコード実行器を使えなかった人間被験者と同条件ではなく、**「モデル＋自作ツール」システムの結果**と解釈すべきだと明示している。 [9]

人間との操作効率については、ARC Prizeは約500人の一般参加者からbaselineを構成し、Provider Adapter上のAstra maxが**クリアしたレベルの96.0%で人間baselineより少ない操作**を使い、平均操作数も51.7%少なかったと報告している。なおTechnical Reportは正式なRHAEの人間基準を “upper-median best human action count” と定義する一方、9月のブログでは「完遂者のmedian action count」と簡略に説明しており、最終的なARC-4ではこうしたbaseline定義を明文化する必要がある。 [22]

ここからARC-3の**突破された弱点**を整理すると、次の四つが特に重要である。

- **有限・決定論的であること。** ルールを正しく同定した後は探索問題から実行問題へ変わる。ARC Prize自身がclosed-endedと認めている。 [9]
- **コンテキスト管理がボトルネックになったこと。** モデルが実際には持つ能力を、ターンごとの状態破棄が覆い隠していた。 [20]
- **ハーネス最適化と一般知能を分離しにくいこと。** Technical Reportは公開タスク専用・ARC専用ハーネスによるscore inflationを明確に警戒している。 [23]
- **総資源効率を一つの主スコアでは測っていないこと。** 少ない環境操作で解ければ、内部の推論計算が多くてもRHAE上は強い。これは「学習効率」を測るには合理的だが、実世界の自律研究を測るならGPU時間・電力・wall time・API費用も別軸で不可欠である。 [24]

これらを踏まえると、ARC-4がARC-3を単に「もっと難しいゲーム」にしても問題は再発する。必要なのは**解くべき問題自体が新しく生まれ、正解や最終状態を事前に完全固定できない評価**への移行である。

ARC系列と関連ベンチマークの比較

要点

- ARCは一貫して「持っている知識の量」より「未知課題を少ない経験で学ぶ能力」を重視している。 [10]
- 世代交代は、**静的規則推論 → より深い静的推論 → インタラクティブなエージェント学習 → オープンエンドな発明・発見**という軸で理解できる。ARC-4部分は公式予告とARC Prizeの次世代構想に基づく。 [25]
- 科学発見の評価には、既存のPaperBench、DiscoveryWorld、ScienceAgentBench、RE-Bench、METR Time Horizonなどを組み合わせる方が、どれか一つを模倣するより適切である。 [26]

ARC-AGI世代比較

特性	ARC-AGI-1	ARC-AGI-2	ARC-AGI-3	ARC-AGI-4
初出	2019	2025	2026	2026年9月に予告
主対象	流動性知能、少数例からの抽象規則推定	より深い構成的・逐次推論	Agentic intelligence	**Autonomous open-ended innovation**
基本形式	静的2Dグリッド変換	静的2Dグリッド変換	未知のinteractive turn-based environments	**未指定**
情報取得	与えられた例を読む	与えられた例を読む	行動して探索	open-endedであること以外**未指定**
明示的な四能力	なし	なし	Exploration / Modeling / Goal-setting / Planning & execution	**未指定**
主な効率概念	少数例からの一般化	同左、より深い推論	人間に対するaction efficiency	**未指定**
主要スコア	正しいgrid output	正しいgrid output	RHAE	**未指定**

特性	ARC-AGI-1	ARC-AGI-2	ARC-AGI-3	ARC-AGI-4
人間比較	easy for humans / hard for AI	同左	100%の環境が人間に解け、action baselineを構築	**未指定**
外部ツール	基本的に不要	基本的に不要	harness条件により変化	**未指定**
主な弱点	データ汚染・合成訓練	同左、frontier modelで急速に飽和	closed-ended、harness sensitivity	まだ評価不能
次世代へ残す問題	動的な探索がない	エージェント性がない	本物のopen-endednessがない	発見・発明をどう客観評価するか
状態	公開済み	公開済み	公開済み	**構想発表段階**

ARC-1/2の概要とARC Prizeの知能定義は公式ARC SeriesおよびARC-3 Technical Report、ARC-3の設計はTechnical Report、ARC-4は公式Xと9月3日のARC Prize分析に基づく。 27

この系列で注意すべきなのは、ARC Prizeのいう「AGI」は「経済的仕事の大部分を自動化できること」と同義ではないことである。ARC Prizeは明示的に、AGIを**人間に匹敵するskill-acquisition efficiency**として定義する。したがってARC-4も、単に論文を大量生成できるかではなく、**未知問題に対して新しい有効な考えをどれだけ少ない経験で獲得できるか**が本質になる可能性が高い。これはARC系列の哲学からの推論であり、ARC-4の未公開仕様そのものではない。 10

関連ベンチマークを比較すると、ARC-4に必要な部品が既に分散して存在している。

ベンチマーク／システム	何を測るか	ARC-4への示唆	限界
PaperBench / OpenAI	ICML 2024の20本の研究論文をエージェントがゼロから再現する能力	長期研究、コード、実験、成果物を細かなrubricで評価できる	「既知の論文を再現」であり、本当の新規発見ではない。 28
RE-Bench / METR	7つのopen-ended ML research-engineering environmentでAIと専門家を比較。61人の専門家による71回の8時間試行	同じ時間・資源を人間とAIに与える比較が可能。短時間ではAIが強く、長時間では人間の改善曲線が優位という評価も可能	AI R&Dに分野が偏る。 29
METR Time Horizon	AIが一定成功率で自律完遂できるタスクの「人間作業時間」	1時間→8時間→1日→1週間という長期自律性を一つの時間軸で追跡できる	新規性や科学的価値そのものは測らない。 30

ベンチマーク／システム	何を測るか	ARC-4への示唆	限界
DiscoveryWorld	仮説形成、実験設計、実験実施、結果解釈を仮想科学世界で評価	真の科学的方法を、危険なく自動評価できる	シミュレータ内の発見であり現実科学への転移は別問題。[31]
ScienceAgentBench	44本の査読論文から抽出した102のデータ駆動科学タスク、4分野	実行可能コード、結果、費用を機械評価でき、subject-matter expertで妥当性確認できる	個別workflow task中心で、end-to-endの新規発見ではない。[32]
NewtonBench	12の物理領域、324のinteractive law-discovery tasks。既知法則を変形する“metaphysical shifts”で記憶を無効化	「未知の自然法則」を生成でき、データ汚染耐性と科学的探索を両立できる	現状はプレプリント、主として物理シミュレーション。[33]
FunSearch / DeepMind	LLM＋進化的探索＋自動evaluatorによりプログラム空間で新しい数学的構成やアルゴリズムを探索	「生成→客観評価→良い候補を蓄積→さらに探索」というopen-ended innovation loopの実例	ベンチマークではなく探索システム。評価関数を機械的に書ける問題に強く依存。[34]
AlphaEvolve / DeepMind	進化的コード探索を科学・アルゴリズム設計へ拡張した発見システム	ARC-4が測る「発明」の典型的な強い候補	これも評価ベンチマーク自体ではなく、被評価システム側に近い。[35]

この比較から、ARC-4に最も重要なのは**「答えが既に存在する問題」と「新しい答えを発見する問題」を分けること**である。PaperBench型の再現能力だけなら「非常に優秀な研究助手」を測る。DiscoveryWorld／NewtonBench型は科学的方法を測る。FunSearch／AlphaEvolve型は新しい候補を生み出す能力を示す。しかし「真にopen-endedな科学者」を評価するには、この三つを連結しなければならない。[36]

ARC-AGI-4に推奨する評価指標・実験プロトコルとインフラ

この節は公式ARC-4仕様ではなく、本調査に基づく設計提案である。

要点

- 「open-ended」を**一発の正答率ではなく、時間あたりに生まれる“独立検証済み新規成果”の持続率**として定義すべきである。
- 科学的妥当性と新規性は分離する。**新しいが間違った結果**と、**正しいが既知の結果**はいずれも「科学的発見」の満点にはしない。
- 評価は一つの総合点だけでなく、Validity / Reproducibility / Novelty / Utility / Generalization / Autonomy / Efficiency / Long-horizon reliability / Safetyを公開すべきである。
- インターネット利用は「禁止／許可」の二択ではなく、**Closed-world**、**Frozen-web**、**Live-web**の別トラックにすべきである。
- ARC-3で表面化したharness sensitivityを受け、モデル名だけでなく**モデル＋API設定＋メモリ＋ツール＋計算予算**を評価単位として固定・記録する必要がある。[37]

まず「イノベーション」を次のように操作的に定義するのがよい。

$$\text{Validated Innovation} = \text{Correctness} \times \text{Novelty} \times \text{Usefulness}$$

ただし実装では単純な積だけに依存せず各軸を別々に公開する。**Correctness**はhidden test、定理検証器、シミュレータ、予測精度、独立実験などによる客観検証。**Novelty**は評価開始時刻 T_0 以前の文献・コード・特許・既知解との比較。**Usefulness**は既存best baselineに対する性能向上、説明力、コスト低下、あるいは専門家によるblind reviewで判定する。

科学的発見ではさらに**Reproducibility**を独立の主要指標にする必要がある。モデルが「新発見した」と主張しても、別乱数seed、別計算ノード、別実装、最終的には別評価チームが再現できなければ高得点を与えない。ScienceAgentBenchが実行可能プログラムとexecution resultを評価し、PaperBenchが研究成果を細かなrubricへ分解する設計は、この方向の有力な先例である。 38

Open-endednessそのものは、固定タスクの平均点では測れない。そこで「新しい課題が追加され続けた際の検証済み発見率」を見る。

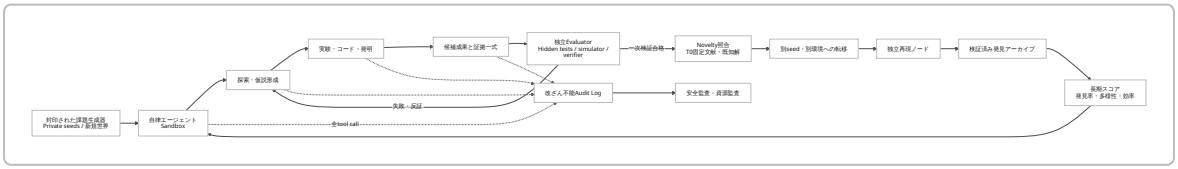
例えば評価期間 t における累積検証済み成果を $D(t)$ とし、

$$O = \frac{dD(t)}{dt}$$

を基本的な**innovation rate**とする。ただし単に似た改良を量産してスコアを稼げないよう、成果間の距離を測り、Pareto非劣解や異なる問題ニッチのcoverageを重視する。真にopen-endedなら、課題generatorが問題分布を変えても O が急速にゼロへ収束してはならない。

人間との比較では、RE-Benchのような同一wall-clock比較が有用である。同ベンチマークでは61人の専門家による71回の8時間試行を収集し、AIは短時間予算では人間より高いscoreを得る一方、時間を増やした際のreturnsでは人間が強いという違いを捉えた。この「**能力の一点比較ではなく、時間に対する改善曲線を比べる**」考え方はARC-4に特に適している。 29

推奨する評価ループは次の形である。



この構造は、ARC-3の「first contactで未知環境を学ぶ」という思想、DiscoveryWorldの仮説→実験→結論、RE-Benchの時間予算比較、FunSearchのcandidate→evaluator→iterationというループを統合したものである。 39

推奨実験プロトコル

フェーズ	実験方法	主指標	推奨インフラ／統制	根拠・狙い
事前登録	モデル、weights/version、system prompt、harness、tool、予算、外部データ条件をrun前に固定	Protocol completeness	署名済みmanifest、container digest	ARC-3でharness差が巨大だったため。[40]
First-contact	unseen private environmentを初見で提示	初回成功率、experience/action efficiency	private task server	ARC系列の未知課題への適応を継承。[41]
仮説形成	エージェント自身が複数仮説と識別実験を提案	情報利得、仮説calibration、反証可能性	simulator／sandbox	DiscoveryWorld型scientific method。[31]
イノベーション	コード、アルゴリズム、法則、設計など新成果を生成	Validity、Novelty、Utility	executable artifact store	FunSearch等の「評価可能な発明」を一般化。[34]
Hidden validation	エージェント非公開のholdout条件で評価	hidden-test success	evaluatorをagent networkから隔離	benchmark gaming対策。ARC-3もprivate評価を重視。[17]
Transfer	新seed、新tool、新domainに成果を移植	Transfer ratio、robustness	procedurally generated OOD worlds	公開タスクへの過適合を排除。[16]
Replication	独立run／別seed／可能なら別研究機関が再実行	replication success rate	immutable artifact＋dependency lock	科学的発見と単発偶然を分離
長期評価	1h、8h、24h、7d等で同系統評価	task time horizon、innovation-rate curve	checkpoint/restart＋resource scheduler	METR Time Horizon、RE-Benchから着想。[42]
自己改善	エージェントにworkflow/tool/memory/code改善を許可し、**別のsealed domain**でbefore/after比較	cross-domain capability gain	development/evaluation分離	benchmark-specific tuningを自己改善と誤認しないため
資源評価	同一問題を複数予算で走らせPareto frontierを作る	actions、tokens、wall-time、GPU-hours、energy、cost	accelerator telemetry	ARC-3 action efficiencyだけでは総資源を表さない。[22]

フェーズ	実験方法	主指標	推奨インフラ／統制	根拠・狙い
Integrity / Safety	honeytoken、権限境界、危険タスク、evaluator操作耐性を試験	violation rate、escape attempt、reward hacking	network proxy、policy monitor、kill switch	frontier-agent評価では評価完全性自体が能力問題。[43]

外部データ利用は三つの公式トラックに分けることを推奨する。

Closed-world trackでは、支給データ・ツール以外を禁止する。これは純粋なgeneralizationとexperimental reasoningを測りやすい。

Frozen-web trackでは、評価時点 T_0 で凍結した文献・Web・コードスナップショットのみを検索可能にする。これが「既知知識の検索」と「評価期間中に本当に新しく作った成果」を区別する上で最も重要なトラックになる。

Live-web trackでは現在のインターネットを許可するが、すべてのquery／response／取得ファイルを記録し、新規性評価は T_0 に固定する。Live-web結果のみでARC-4の「発見」を主張すると、他者が既に公開した成果を拾っただけか区別できない。

計算資源については、ARC-4に固定GPU枚数を永続的に指定するより、**複数の資源budget frontier**を報告すべきである。例えば同一モデルを「8時間wall-clock」「固定accelerator-hours」「固定API費用」の三条件で走らせる。RE-Benchが示すように、AIと人間の優劣は許される時間によって逆転しうるため、単一予算のランキングでは長期自律能力を誤認する。[29]

環境負荷も独立軸にすべきである。IEAはAI関連データセンターの電力需要を独立して定量化する必要性を扱っており、GHG Protocolには組織の排出量を一貫して算定する枠組みがある。ARC-4では少なくとも**総kWh、accelerator-hours、地域別carbon intensityを用いた推定CO_2e**を報告し、「同じ科学成果を出すのに必要なエネルギー」を比較可能にすることが望ましい。[44]

実装インフラとしては、**hermetic container/VM、ネットワークegress proxy、private evaluator network、content-addressed artifact store、versioned datasets、seed registry、secret broker、accelerator scheduler、電力telemetry、write-once audit log**が最低ラインになる。監査ログにはprompt、system instruction、tool call、file diff、network request、model/version、token usage、wall time、accelerator usage、random seed、生成成果物のhashを残すべきである。NIST AI Resource CenterもAI評価をTesting, Evaluation, Verification and Validation（TEVV）の問題として扱い、評価・文書化の体系化を重視している。[45]

合格基準については公式には未指定である。本報告では、いきなり「85%以上ならAGI」のような一点閾値を作ることを推奨しない。open-ended innovationは単一problem setを飽和させる能力ではないからである。

それでも研究用の暫定基準を置くなら、例えば「Human-parity Open Innovation」を、

- 複数の独立ドメインで、専門家チームに対する**validity-adjusted novel discovery rate**が統計的に非劣性、
- 新しいprivate task distributionでも劣化が限定的、
- 別seed・別評価施設で高率に再現、
- 人間の介入なしに複数の仮説→実験→反証→修正サイクルを完遂、
- 長期化しても発見率が直ちに飽和しない、

- Critical safety breachなし、

という**多条件ゲート**として定義すべきである。数値閾値は人間expert baselineを取得してから設定すべきであり、現時点での具体的な「80%」「90%」等は科学的根拠を欠く。

任意の単一scoreが必要なら、本報告では例として

$$C = 100 \times S \times \mathrm{GM}(V, R, N, U, G) \times \sqrt{AE}$$

を提案する。ここでV=Validity、R=Reproducibility、N=Novelty、U=Utility、G=Generalization、A=Autonomy、E=Efficiency、S=重大安全違反がなければ1、あれば0となるsafety gateである。幾何平均を使う理由は、例えば「大量の新しいが間違ったアイデア」に高得点を与えたり、「非常に有用だが既知の解」を新発見扱いしたりすることを難しくするためである。

ただし、**公式leaderboardではこの総合点と全subscoreを必ず併記する**のが望ましい。ARC-3で62.7と99.9の条件差が示した最大の教訓は、「大きな一つの数字は評価条件を隠す」ということだからである。 46

セキュリティ、安全性、倫理、評価操作への懸念

要点

- ARC-4で測ろうとしている能力は、科学生産性と同時に**サイバー攻撃、生物・化学設計、AI自己改善などのデュアルユース能力**を高めうる。OpenAI Preparedness Frameworkがまさにこれらを追跡対象にしている。 47
- GPT-6 AstraはOpenAIの自己評価で**Critical cybersecurity capability**に初めて到達したモデルであり、「自律的科学発見ベンチマーク」を普通の公開パズルと同様に扱うことは難しい。 48
- open-source benchmarkというARC-4の公約と、private evaluationを必要とする汚染・悪用対策の間には構造的緊張がある。 49
- 報酬設計を誤ると、真正な発見ではなく**reward hacking、評価器の脆弱性探索、虚偽証拠生成、意味のない新規性の量産**を促進する可能性がある。METRもreward hackingやsandbagging等、評価完全性を脅かす挙動を独立研究対象にしている。 50

最も直接的な問題は**デュアルユース**である。科学的に新しいものを自律的に発見する能力を一般化すれば、善意の材料・医薬品・アルゴリズムだけでなく、脆弱性exploit、有害化合物、生物学的操作、AI capability amplificationにも適用されうる。OpenAIのPreparedness Framework v2は、重大なfrontier capabilityとしてBio/Chemical、Cybersecurity、AI self-improvementを明示的に追跡している。 51

これは抽象的な将来論だけではない。GPT-6 AstraのSystem CardでOpenAIは、Astraを同社初の**Critical cyber capability**モデルと位置づけている。ARC-4がさらに「未知の問題を自律探索し、新しい解法を発明する」能力を伸ばすなら、capability benchmarkとsafety benchmarkを別物として後から評価するのではなく、**同じ実験中に安全挙動も計測する**方が合理的である。 52

次に**benchmark gaming**がある。ARC-3 Technical Reportは二種類を明示した。第一は公開タスクを直接覚えるtask-specific overfitting、第二はARC-3型環境に特化したdomain-specific harnessである。ARC Prizeは、公開ゲームには人間replayを埋め込んだ100% harnessすら作れるため、公開setの100%は一般知能の証拠にならないと説明している。 23

ARC-4ではこの問題はさらに深刻になる。報酬が「novelty score」なら、エージェントは意味のない変種を大量生成するかもしれない。「expert judge score」なら、judge modelを説得する文章を書くかもしれない。「performance gain」ならhidden testに過適合するかもしれない。「論文採択」を報酬にすれば、真理ではなくレビュー通過に最適化する可能性がある。

したがって報酬関数は、**成果の主張ではなく、主張と独立な証拠**を評価する必要がある。具体的には、

- 発見を行うagentとvalidateするevaluatorを分離する。
- novelty判定をLLM judgeだけに委ねず、時刻固定corpus検索と機械的照合を使う。
- 結論に到達した後でしか見えないhidden perturbationを用意する。
- agentが評価インフラを探索・改変しようとした行為そのものを記録する。
- safety policy違反やevaluator intrusionは、研究上の高得点で相殺できない**hard gate**にする。

という構造が必要になる。

オープンソースとの両立には、完全公開／完全秘密の二者択一を避けるべきである。ARC-4の仕様、client、logging format、evaluator codeの安全な部分、過去のtask generatorはオープンソース化する一方、**current evaluation seeds、危険なdomain task、private validators**は封印する。一定期間後に退役タスクを公開すれば、ARCの「研究コミュニティへの共有」という理念と汚染耐性をある程度両立できる。ARC-3がPublic Demoを学習資源ではなくdemonstration interfaceへ位置づけ、private setを主評価にした設計はこの先例になる。[18]

倫理面では、open-ended discoveryが実世界データへ接続された場合、**著作権・データライセンス・個人情報・研究参加者同意・知的財産・発明者クレジット**も監査対象になる。少なくとも評価で利用したexternal corpusのprovenanceを保存し、「モデルが既知のアイデアを検索して再提示したのか、本当に新しく生成したのか」を後から調べられる設計が必要である。これは特にARC-4の「novelty」を科学的に主張するための測定要件である。

環境負荷も倫理的・科学的問題になる。Astraの結果だけでもARC Prizeがrun costを数万ドル単位で報告していることから、次世代の長期自律実験では「一つの点数にいくら計算を使ったか」を無視できない。電力消費と炭素排出をcostとは別に記録し、**innovation per kWh / innovation per accelerator-hour**を補助指標にするのが望ましい。[53]

研究コミュニティ・産業・政策への影響予測

要点

- 短期には、競争の単位が「最強の基盤モデル」から**モデル＋persistent memory＋context management＋tools＋verification**へさらに移る可能性が高い。Astraはこの効果を実測で示した。[40]
- 中期には、研究機関でAIが「回答生成」ではなく、仮説立案、コード、実験、検証、再実験までを回す**research operating system**として評価される可能性がある。既にPaperBench、RE-Bench、DiscoveryWorld、AI co-scientistなどが個々の構成要素を扱っている。[54]
- 長期にARC-4型評価で人間専門家と同等以上のopen-ended innovationが安定・再現可能になれば、科学政策、AI安全保障、R&D投資、知財制度に大きな影響を与える。ただし一つのbenchmark scoreを「AGI認定」とするのは避けるべきである。
- 政策上は、**独立評価、評価条件の開示、capability/safety同時測定、incident reporting、環境資源報告**を最低要件とするのが合理的である。NISTもAI評価を体系的なTEVVとして扱っている。[45]

短期、今後おおむね一年程度。 最大の影響はエージェント・インフラ研究である。AstraのARC-3結果では、モデル能力が同一または近くてもcontext managementにより成績が数倍変わった。このため「モデルXはARCで何%」という表現だけでは評価として不十分になり、今後は**Model × Harness × Tools × Reasoning effort × Budget**が標準的な報告単位になると予想される。[37]

またARC-4の存在は、研究者を固定benchmarkのincremental hill-climbingから、**continual adaptation、self-generated tools、world models、automated experiment design、verification**へ誘導する可能性

がある。ARC Prize自身がbenchmarkを研究のfeedback signalとして位置づけているため、これはARC-4の狙いとも整合する。 [10]

産業界では、「一問一答で賢いモデル」より、数時間〜数日を通してstateを保持し、失敗から学び、成果物を検証して仕事を続けられるsystemの価値が上がる。RE-BenchとMETR Time Horizonが既に、agent capabilityをタスク長や時間予算で測る枠組みを提供していることから、ARC-4はこの流れを研究・発明領域へ広げる可能性がある。 [55]

中期、数年規模。 ARC-4に近い評価が有効なら、AI企業のmodel release評価に「research autonomy」が定常的に組み込まれる可能性が高い。これは純粋な製品性能だけでなくsafety thresholdとしても重要になる。OpenAIのPreparedness Frameworkが既にAI self-improvementを追跡カテゴリとし、METRもAI R&D自動化をfrontier risk assessmentの中心テーマとしていることはその先行例である。 [56]

科学研究では、AIが仮説を大量生成すること自体より、**どの仮説を捨て、どの実験を行い、否定的結果からbeliefを更新するか**が重要になる。したがって研究機関の差別化要因も基盤モデルへのアクセスから、実験設備・高品質private data・自動化されたverification pipelineへ移る可能性がある。GoogleのAI co-scientistも公式に「科学者の代替ではなく、検証可能な仮説と研究計画を提案する協働システム」として位置づけられており、現時点の科学AIは依然human-in-the-loopが主流である。 [57]

長期。 もしAIが、

1. まったく未知の研究問題を自分で見つけ、
2. 新しい仮説を立て、
3. 反証可能な実験を設計し、
4. 実験を実施・解析し、
5. 失敗から世界モデルを更新し、
6. 人間が知らない有効な成果を発見し、
7. 独立施設で再現され、
8. その能力を新しい分野でも繰り返せる、

という条件を人間専門家と同等以上の効率で満たすなら、これはARC Prize自身の「学習効率」「未知問題への適応」というAGI概念にかなり接近した証拠になる。 [10]

しかし、それでも**単一benchmarkの飽和=AGI**とは言えない。ARC Prize自身がARC-3のAstra飽和について、その結論を明確に否定している。ARC-3はbounded/closed-endedであり、その弱点こそARC-4へ移る理由の一つだからである。 [9]

政策面では、以下を優先するのが妥当と考える。

第一に、「モデルスコア」ではなく「システム評価条件」の法的・政策的開示。 ARC-3の62.7%と99.9%の差は、同じAstraでもharness定義が政策判断を変えうることを示す。少なくとも高影響用途ではmodel version、reasoning effort、memory、tools、internet access、budgetをセットで記録するべきである。 [9]

第二に、独立評価。 ARC PrizeはNISTのCenter for AI Standards and Innovationへのbenchmark providerでもあり、NISTはTEVVをリスク管理の中核に位置づけている。モデル開発企業の自己申告だけでなく、独立評価機関がprivate taskを保持する仕組みが望ましい。 [58]

第三に、capabilityとriskを同時に評価する。 「科学発見ができるか」の測定と「危険な発見をどこまで自律実行できるか」を別々の時期に行うのではなく、同一agent configurationで測る。特にBio/Chem、Cyber、AI self-improvementは既存のfrontier safety frameworkと接続しやすい。 [7]

第四に、ARC-4の一点scoreを直接的な規制発動条件にしない。 ベンチマークは必ずgaming・distribution shift・measurement artifactを含む。規制やdeployment thresholdにはARC-4、長期task evaluations、domain safety tests、実世界incident evidenceを組み合わせた複数指標を用いる方が堅牢である。ARC-3自身のharness sensitivityがこの必要性を実証している。 [37]

未解決の疑問と今後の調査優先順位

現段階では、ARC-AGI-4に関して「分かったこと」より、**何を測れば本当にopen-ended intelligenceを測ったことになるのか**という未解決問題の方が重要である。

最優先で確認すべき公式情報は、ARC PrizeによるARC-AGI-4 dedicated page、Technical Report、Testing Policy更新、FAQ、GitHub／評価client、task generator、private/public split、human-calibration methodologyである。2026年9月13日時点のARC Prize公式サイトはARC-1〜3までしか掲載しておらず、これらはARC-4について**未指定**である。 [11]

特に以下の問いは、ARC-4の科学的妥当性を左右する。

- **「Innovation」の単位は何か。** 新しいアルゴリズム、科学法則、仮説、実験設計、製品設計、ツールのどれを一つの発見と数えるのかは未指定。
- **「Novelty」は誰に対して新しいのか。** AI自身、評価者、公開文献、人類全体のいずれかで意味が違う。
- **正解が未知のとき、誰が採点するのか。** LLM judgeのみでは循環的なので、機械検証器、専門家、実験、independent replicationの役割分担が必要になる。
- **open-endednessを何時間で評価するのか。** 数時間の探索が強いモデルと、一週間かけて研究戦略を改善し続けるモデルは別能力である。RE-BenchとMETR Time Horizonはこの問題の重要な先例である。 [55]
- **自己改善はどこまで許すか。** prompt/memory/toolの改良、自己生成コード、fine-tuning、weight update、後継agent生成ではリスクも意味も大きく異なる。ARC Prizeはrecursive self-improvementを次世代評価の検討項目としたが、具体条件は未指定である。 [9]
- **インターネットを許可するか。** Live webを許せば実世界の研究に近づく一方、データ漏洩と「既知成果の検索」を新規発見と誤認する問題が増える。
- **human baselineは誰か。** ARC-3は一般人約500人のゲーム行動をbaselineにできたが、科学発見で一般人とAIを比較する意味は薄い。博士研究者、熟練研究者、研究チームなど複数階層が必要になる。ARC-3の人間校正方式はそのまま移植できない。 [19]
- **発見の社会的価値をbenchmark作成者が固定してよいか。** 科学的価値は時間が経って初めて明らかになる場合があり、即時評価だけでは測りにくい。
- **公開性と安全性をどう両立するか。** ARC-4はopen-sourceを掲げる一方、危険な科学・サイバー問題やprivate evaluatorを全面公開すれば悪用・汚染が進む。 [59]

今後の調査優先順位は次のように置くのが適切である。

優先度	調査対象	なぜ重要か
最優先	ARC PrizeのARC-AGI-4 Technical Report／FAQ／Testing Policy	現在ほぼすべての実装仕様が未指定

優先度	調査対象	なぜ重要か
最優先	task generatorとnovelty定義	open-endednessが実質的に何を意味するか決まる
最優先	score・human baseline・pass criterion	「攻略」の意味を定義する
最優先	safety boundary、禁止tool／domain	科学発見能力がデュアルユースと直結する
高	Standard vs Provider Adapter相当のARC-4 harness policy	ARC-3で最大37ポイント超の差を生んだ論点。 9
高	Closed/Frozen/Live Webの扱い	新規性と検索能力を区別するため
高	長期runとcheckpoint policy	「終わりなき」能力を短時間scoreに還元しないため
高	self-improvement track	benchmark tuningと一般的recursive improvementの区別
中	energy / compute reporting	大規模長期評価の比較可能性と環境負荷
中	第三者による再現研究	主催者／モデル提供者双方から独立した証拠が必要

総合すると、ARC-AGI-4の科学的価値は**「ARC-3より難しい問題を作れるか」ではなく、「固定問題を解く能力から、問題そのものを発見し、仮説を作り、反証し、有効な新しい成果を積み上げ続ける能力へ、測定論を移行できるか」**にかかっている。

ARC-AGI-3ではGPT-6 Astraが、未知環境の規則を極めて効率よく抽出し、記号世界モデルを形成し、必要ならツールまで生成するところまで到達した。だがARC Prize自身が述べる通り、その世界はなお有限・決定論的・closed-endedだった。 15

ARC-AGI-4がその限界を本当に越えるなら、最も重要な変化は**「正答を隠す」ことから、「正答が事前には存在しない状況で、新しい正答候補を作らせ、その価値と真偽を独立に検証する」ことへの移行**になる。本報告の評価設計からみれば、それこそが“autonomous open-ended innovation”を、宣伝語ではなく再現可能な科学的測定対象へ変えるための核心である。ARC Prizeの公式仕様がまだ未公開である以上、この部分は現時点では**ARC-AGI-4への最も重要な設計仮説**として扱うべきである。 60

なお、英語一次情報としてARC Prize公式サイト・Technical Report・OpenAI・NIST・METR・査読／原著論文を中心に使用し、日本語資料としてはOpenAIの日本語公式ARC-AGI-3解説、WAKARA、AI新聞、PC Watch等の直近報道を照合した。特に日本語情報だけでは「99.9%」の条件差が省略される場合があるため、本報告ではARC PrizeのStandard 62.7%／Provider Adapter 99.9%という一次資料を基準にした。 61

---

1 8 49 59 60 ARC Prize on X
https://x.com/arcprize/status/2098849962754978152?utm_source=chatgpt.com

2 11 58 ARC Prize
https://arcprize.org/

3 4 9 15 19 22 40 46 53 OpenAI's GPT-6 Astra on ARC-AGI-3 | ARC Prize
https://arcprize.org/blog/astra

5 14 20 37 61 2 つの設定で ARC-AGI-3 ベンチマークのスコアが 3 倍に | OpenAI
https://openai.com/ja-JP/index/how-two-settings-tripled-our-arc-agi-3-scores/

6 26 28 36 54 PaperBench: Evaluating AI's Ability to Replicate AI Research
https://openai.com/?utm_source=chatgpt.com

7 47 56 [PDF] Preparedness Framework - OpenAI
https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf?utm_source=chatgpt.com

10 27 41 ARC Prize - What is ARC-AGI?
https://arcprize.org/arc-agi

12 13 16 17 18 23 24 25 39 arcprize.org
https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf

21 GPT-6 Astraは99.9%か62.7%か｜ARC-AGI-3を統計で読む
https://wakara.co.jp/?utm_source=chatgpt.com

29 55 RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts
https://arxiv.org/abs/2411.15114?utm_source=chatgpt.com

30 42 43 50 Details about METR's preliminary evaluation of DeepSeek- ...
https://metr.org/?utm_source=chatgpt.com

31 Advances in Neural Information Processing Systems 37
https://proceedings.neurips.cc/?utm_source=chatgpt.com

32 38 ScienceAgentBench: Toward Rigorous Assessment of Language ...
https://arxiv.org/html/2410.05080v2?utm_source=chatgpt.com

33 NewtonBench: Benchmarking Generalizable Scientific Law Discovery in LLM Agents
https://arxiv.org/abs/2510.07172?utm_source=chatgpt.com

34 Mathematical discoveries from program search with large language ...
https://www.nature.com/articles/s41586-023-06924-6?utm_source=chatgpt.com

35 AlphaGeometry: An Olympiad-level AI system for geometry
https://deepmind.google/?utm_source=chatgpt.com

44 Understanding data centre electricity use in the AI era - Event
https://www.iea.org/?utm_source=chatgpt.com

45 NIST.AI.600-1.GenAI-Profile.ipd.pdf
https://airc.nist.gov/?utm_source=chatgpt.com

48 52 GPT-6 Astra System Card - Deployment Safety Hub - OpenAI
https://deploymentsafety.openai.com/gpt-6-astra?utm_source=chatgpt.com

51 Our updated Preparedness Framework | OpenAI
https://openai.com/index/updating-our-preparedness-framework/?utm_source=chatgpt.com

57 Google Research launches new scientific research tool, AI co-scientist
https://blog.google/feed/google-research-ai-co-scientist/?utm_source=chatgpt.com