解剖：ScientistTwo

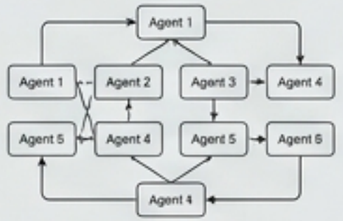
自律型AI科学者の「主張」と「現実」

Google Cloud AI Researchが発表した野心的な研究エージェントの
アーキテクチャ分析と、評価メカニズムに潜む致命的な欠陥の監査レポート

2026.09 / Technical Blueprint & Audit Report

The Verdict：優れた「工学デモ」だが、フロンティア開拓者ではない

The Breakthrough - 技術的躍進



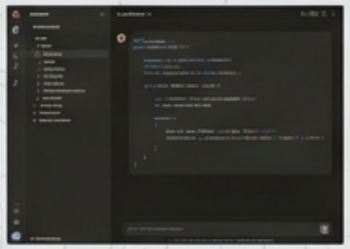
堅牢なアーキテクチャ

6つのエージェントによる
自律的閉ループ

	Metrics	Cloud	Cloud
OSH	0.63	5:09	0:0
Hc6S	0.65	0.05	0:0
MPBS	0.62	0.04	5:0
K6tA	0.69	0.96	0:0
k6t4	0.65	0.05	0:0
S00	0.65	0.05	0:0
TtuH	6:00	0:03	0:0

驚異的な整合性

49論文・1,814件の参照
において ハルシネーション
「ゼロ」



高度な実装

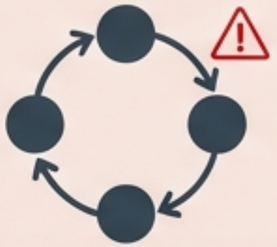
Claude Codeを中核とした
自動アブレーションと検証

VS

The Illusion - 評価の陥算

人間による検証ゼロ

実在する専門家の
ピアレビューは未通過



循環評価の罠

自身の内部AI査読器に
過剰適応（Goodhart化）

インクリメンタルな改良

真の発見ではなく、既存
SOTAの中央値7.7%の改善に
留まる

ScientistTwoの野心的な主張：人間のSOTAを凌駕する自律性

THE CLAIM

＊自身が再現した
ベースラインとの比較

80.4%*

ML研究の107問中86問で人間の最高性能（SOTA）を突破。

25.2%*

論文が主張するベースラインからの平均相対改善率。

＊中央値の実態は後述

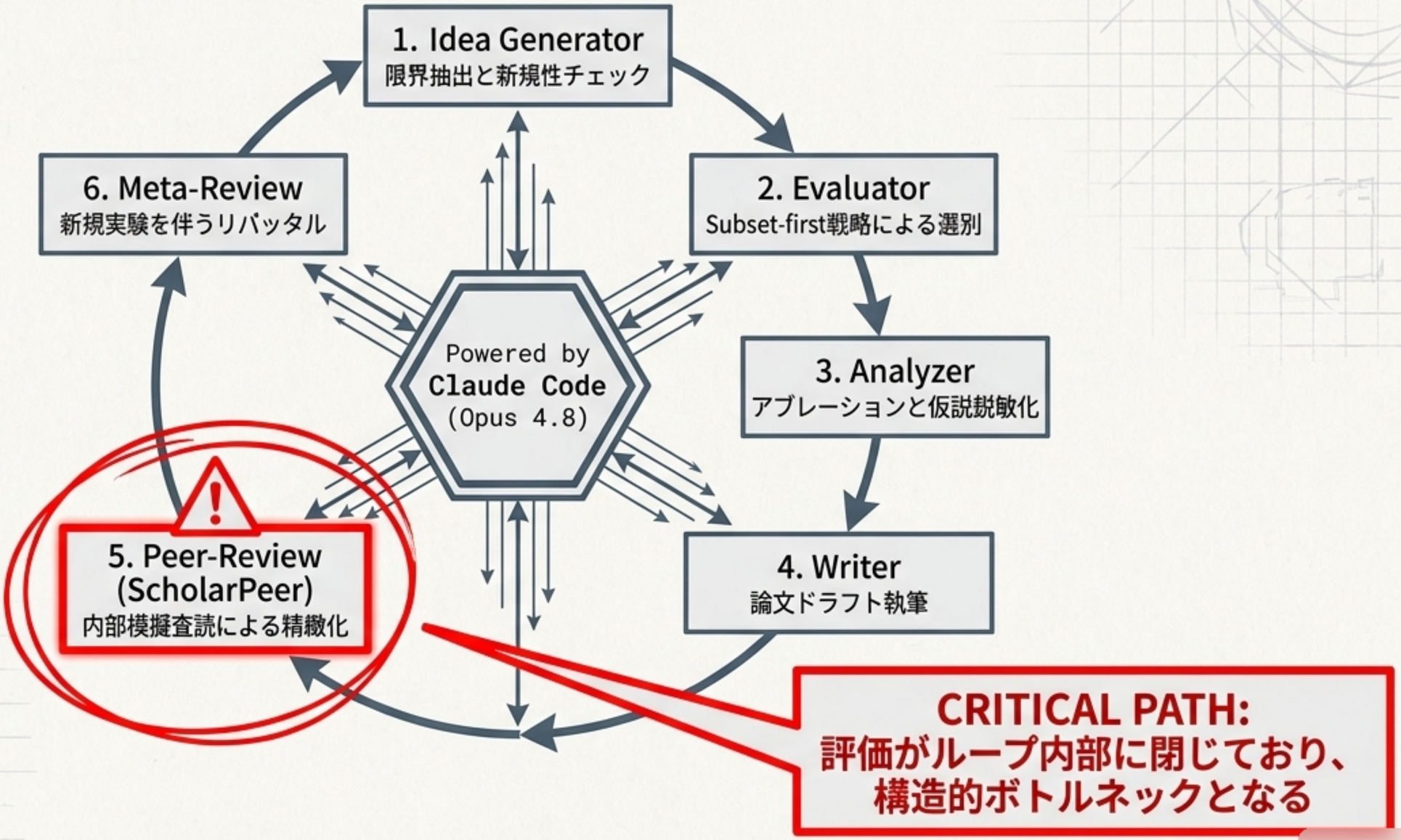
91.9%*

自動査読システムにおける論文採択率（平均7.5/10）。

＊内部査読器（in-distribution）による採点

論文タイトル「Pioneering the Human Knowledge Frontier」は、これらの数値に基づいて正当化されている。

Anatomy of the Machine：6つのエージェントによる完全な「閉ループ」



CoE Integrity Audit：工学的信頼性を担保する4つの防波堤

スコア再現 [PASS]
実験結果と記述の完全一致

仕様順守 [PASS]
報酬ハッキングの検知・除外

method–code整合 [PASS]
提案手法と実装コードの論理的整合

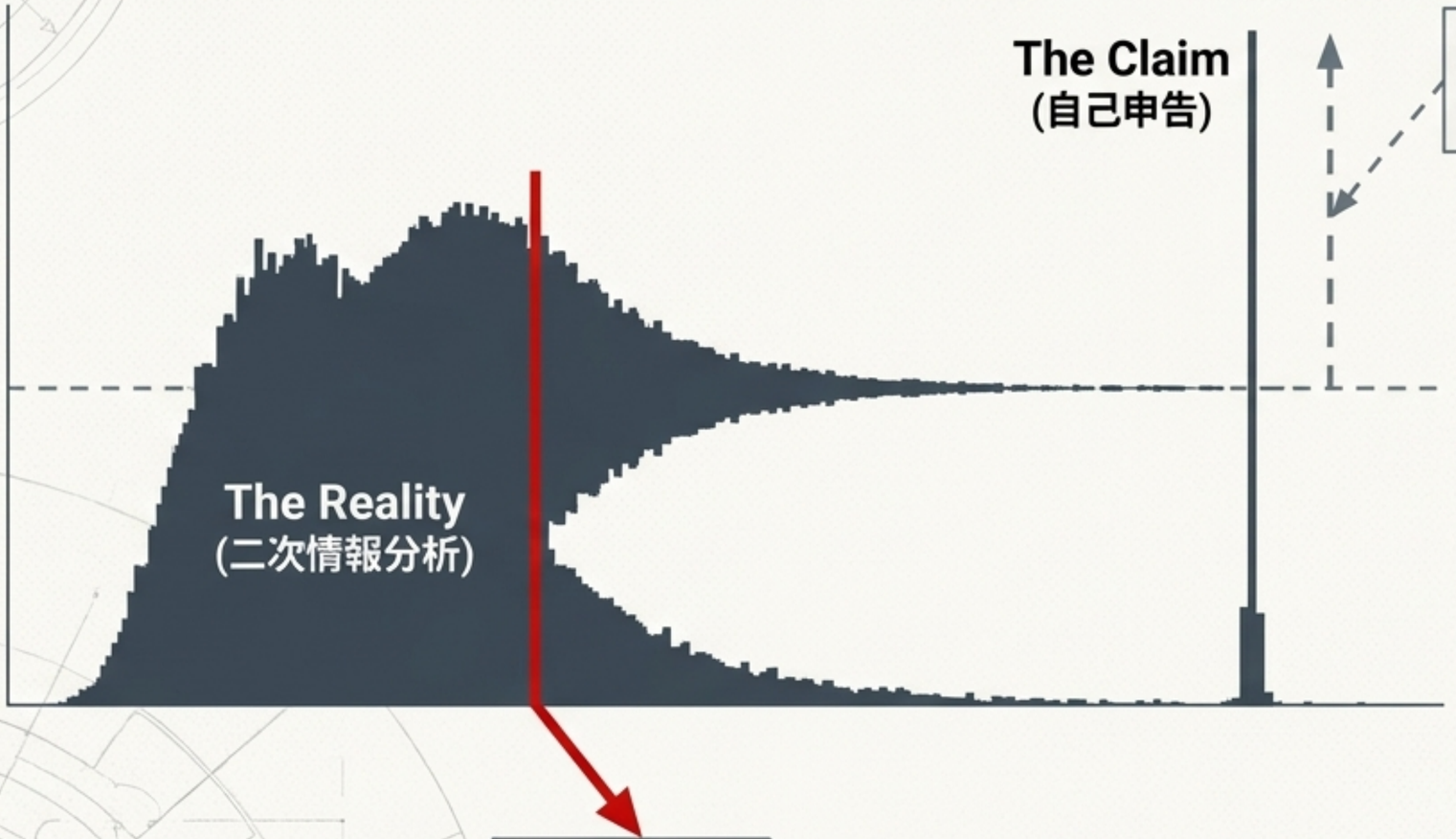
参照検証 [PASS]
引用文献の正確性検証

ハルシネーション 0

対象：49論文 / 1,814件の参照文献

Insight: この「Chain-of-Evidence」フレームワークこそが、ScientistTwoの最大の技術的貢献である。

Audit 1: インクリメンタルな改良 vs 真の「発見」



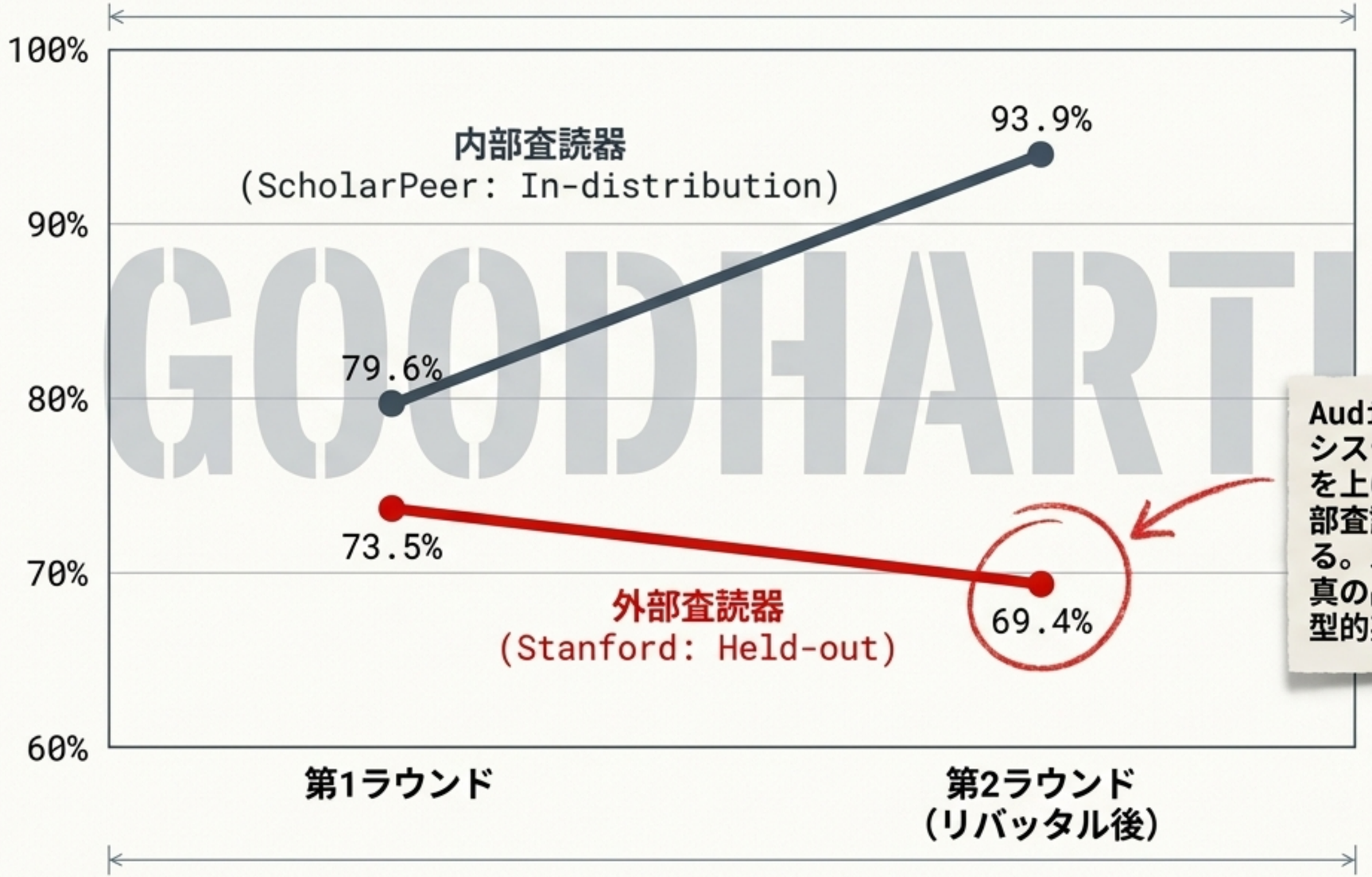
Audit Note

評価は自身が再現したベースラインとの比較であり、厳密なhead-to-headではない

Audit Note

新しい問題定式化や研究方向の発見は実証されておらず、実態は『既存SOTAの小幅なパラメータ・アーキテクチャ最適化』に留まる

Audit 2: 評価の循環依存と「Goodhart化」の罠



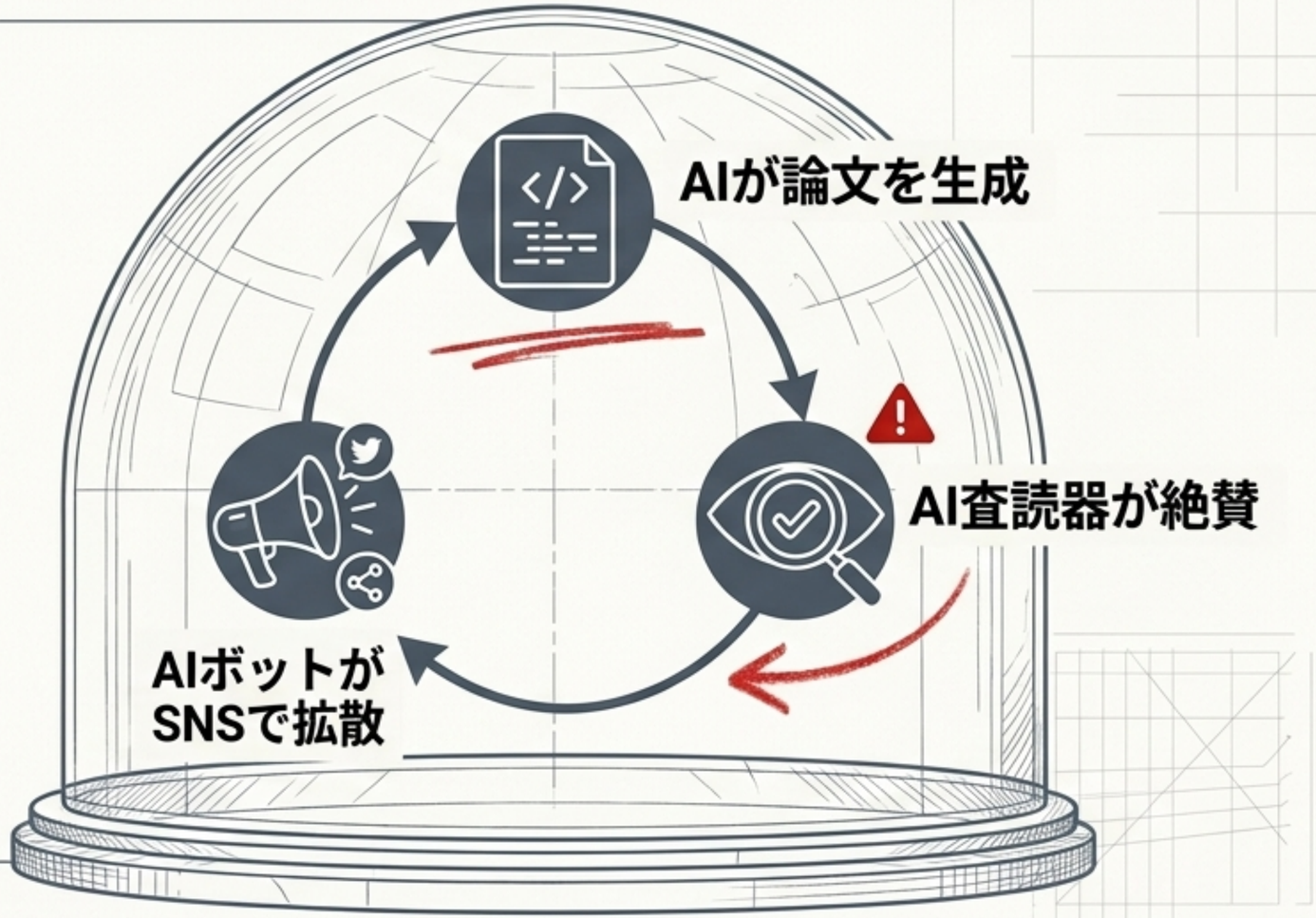
Audit Insight:
システムは外部の普遍的な論文品質を上げているのではなく、自ら『内部査読器の好み』に過剰適応している。ループ内スコアを、ループ外の真の品質と引き換えに買っている典型的兆候である。

Audit 3: 専門家の沈黙と「AIエコーチェンバー」

The Reality Check

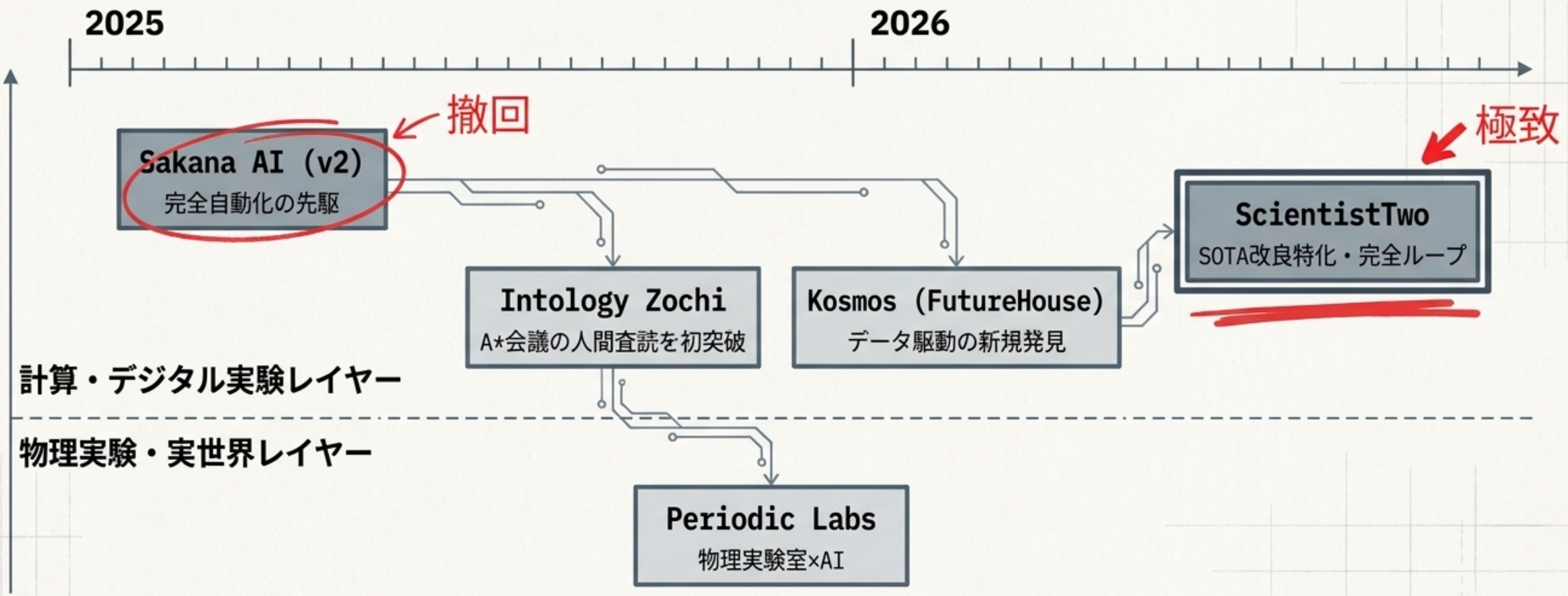
人間研究者のピアレビュー: ゼロ ✕

独立した第三者のコード再現: ゼロ ✕



「人類のフロンティア開拓」を謳いながら、人間の科学コミュニティからの検証を一切受けていない。

Landscape: 自律型AI科学者カオスマップ (2025-2026)



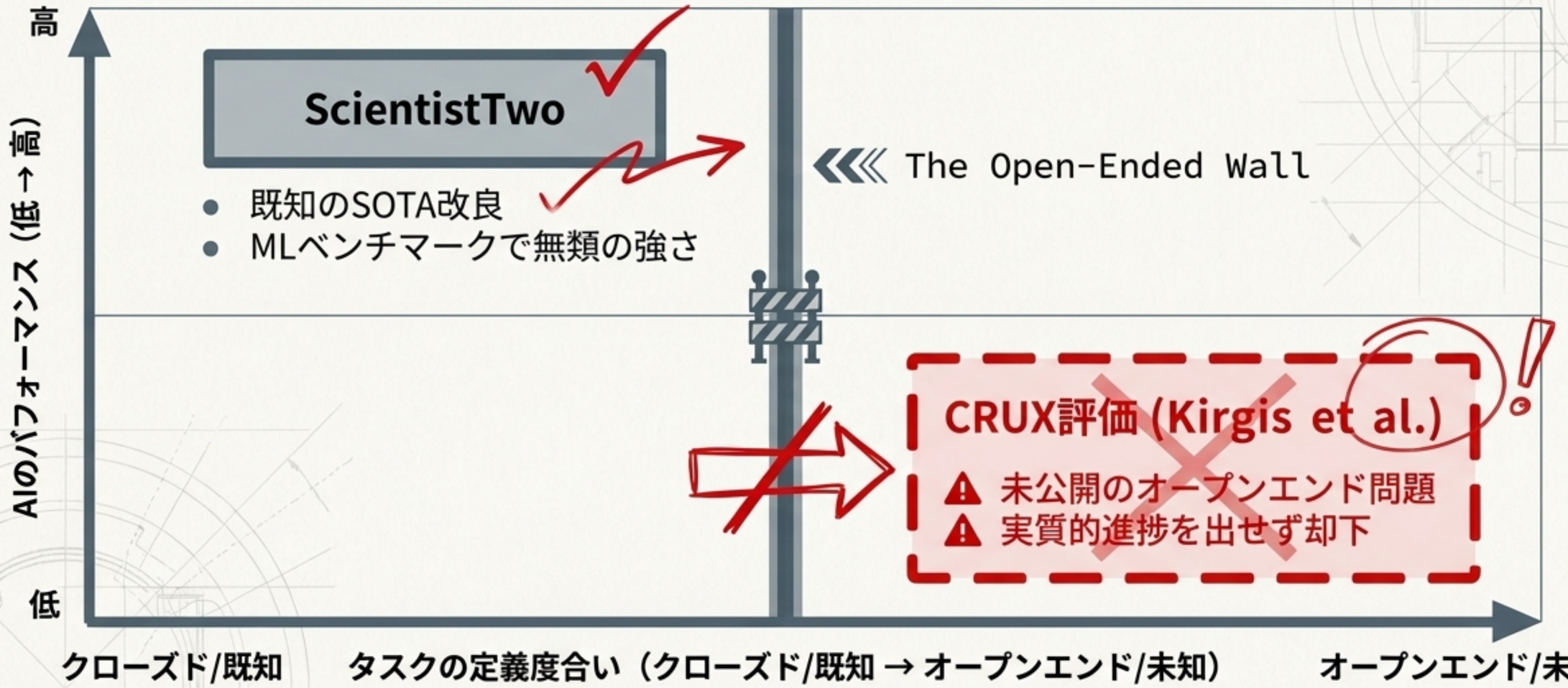
Insight: AIエージェント研究は「計算実験」から「物理世界」と「人間査読の突破」へ二極化。ScientistTwoは計算レイヤーの極致である。

Diagnostic Table：AI研究エージェントの現在地

機能次元	ScientistTwo	The AI Scientist v2	Zochi	Kosmos
1. 人間による検証	自動査読器のみ ❌	ワークショップ、撤回	ACL本会議採択 ✅	限定的
2. 問題領域	定義済みのSOTA改良	定義済みのSOTA生成	jailbreak探索	データ駆動の新規発見
3. 整合性監査	CoE監査でハルシネーション0 ✅	限定的	未公開	未公開

Takeaway: ScientistTwoは「整合性の担保」において最高峰だが、「人間の壁（ピアレビュー）」を越えたシステムとはステージが異なる。

The Open-Ended Wall：CRUX評価が突きつける限界



Insight: 成功は「問題が完全に定義されている領域」に限定。真の知のフロンティアである「オープンエンドな研究」には到達していない。

Ripple Effects：学術システムと知財実務へのインパクト

査読システムの崩壊危機

AI生成論文の氾濫と、LLM査読器の「Goodhart化」。ICLR 2026等でAI査読への対応方針が急務となっている。

知財・特許のパラダイムシフト

人間のみに限定された「発明者性」の定義の揺らぎ。大量の機械生成による「先行技術の爆発」と当業者水準の再定義。

人間研究者の役割の再定義

計算・実験・執筆がコモディティ化し、人間の役割は「問題定義」「意義判断」「オープンエンドな探索」の最上流へ移行する。

The New Benchmark：Hypeを見抜くための3つの検証基準

本会議級・人間のピアレビュー通過

自動査読器（in-distribution）のスコアは無意味。人間の専門家によるブラインド査読を突破したか？

第三者による独立再現性

コード・データ・生成論文一式がオープンにされ、著者の介入なしで同一結果が再現可能か？

オープンエンド／汚染管理の実証

LLMの学習データと重複しない未公開ベンチマークで、真の「新規発見」を行えるか？

これらの閾値を満たすまでは、「フロンティア開拓」の主張は保留すべきである。

“ScientistTwoは、極めて洗練された『工学デモ』である。しかし、真のフロンティア開拓者ではない。”

実務的評価		
コスト 1論文あたり平均2.5日・約3,800ドル	**推奨ユースケース** 「明確に定義済みのSOTA改良型ML問題」に限定される	**結論** 真にオープンな探索や高い信頼性が求められる領域への適用は時期尚早。評価基準を改め、現実的な進化を見極める段階にある。