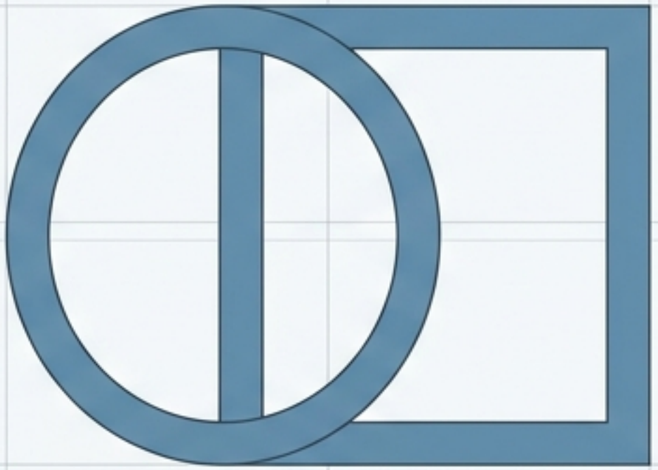


ScientistTwoの真実

自律型AIは本当に「科学者」になったのか？
― Google Cloud AIプレプリントの技術監査とデューデリジェンス

[STRICTLY CONFIDENTIAL / AUDIT REPORT // 2026]

論文の主張：「人類の知識フロンティアの開拓」

人間の介入なしで完全な科学的発見ループを実行。

トップ会議レベルの論文を自律生成し、人間の研究者を凌駕する。

監査の結論：「高度な自律型ML最適化エンジン」

未知の科学は発見していない。

既存の計算機科学課題に限定。

非常に強力な「閉ループ最適化」の実証だが、「万能の科学者」という表現は誇張。

【結論】ScientistTwoは画期的な「研究OS」であるが、ゼロからの科学的発見者ではない。

2025-04

AI Scientist-v2 (生成重視)

論文自動生成の実証。

2026-04

AutoSOTA (スケール重視)

大規模なベースライン性能の自動改善。

2026-05

ScientistOne (監査重視)

引用・コードの整合性チェック。

2026-09

ScientistTwo (閉ループOS)

【本論文】実験・アブレーション・査読反論を統合。

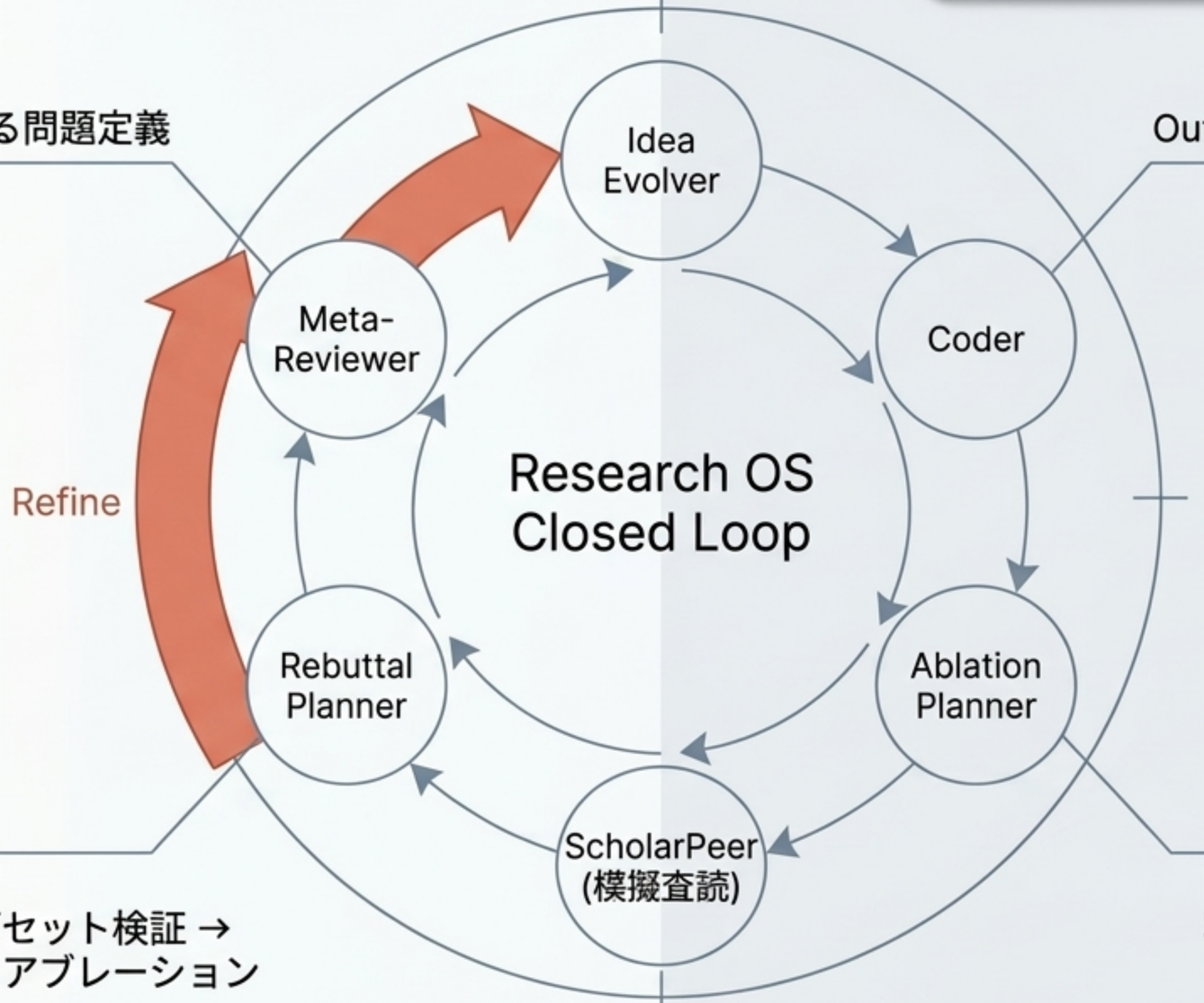
最も重要な新規性は「LLMに研究させること」ではない。科学研究を「状態を持った反復最適化プロセス」としてシステム実装した点にある。

Technical Auditor Dossier

Core Engines: Gemini 3.6 Flash & Claude Code Opus 4.8

Input: 人間による問題定義

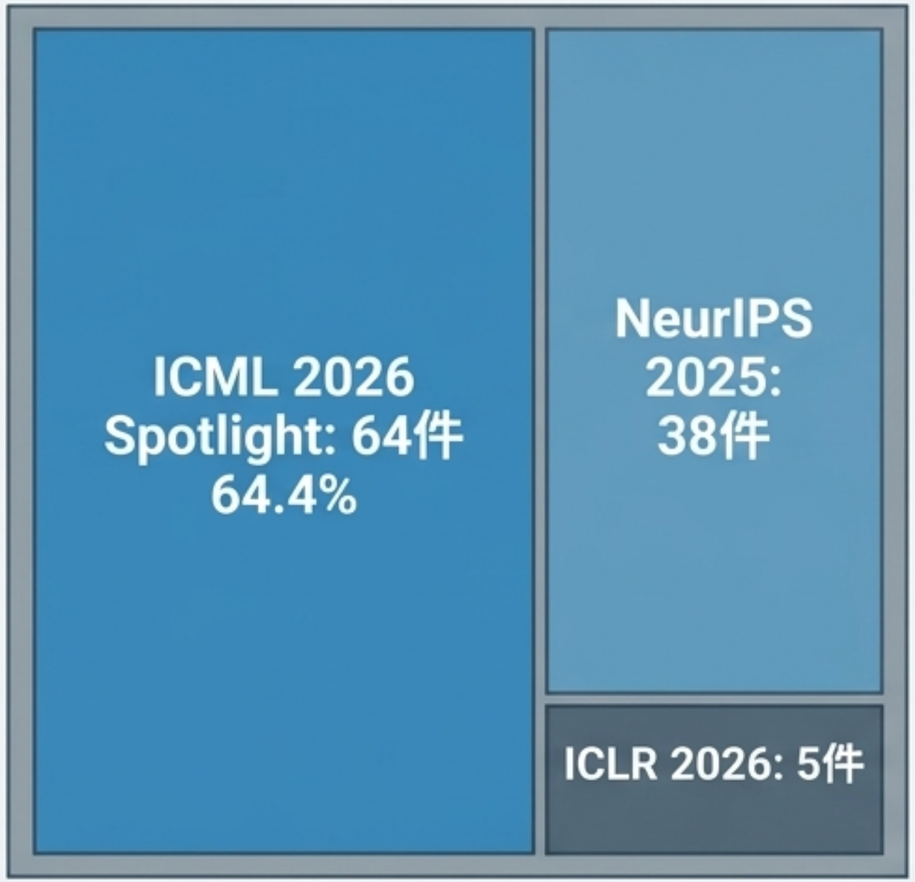
Output: 論文 + 実行可能コード

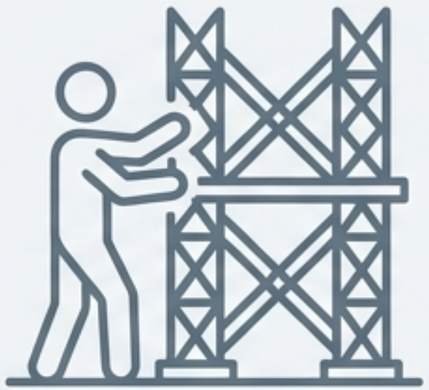


Process:
仮説生成 → サブセット検証 →
全データ拡張 → アブレーション

Feedback:
模擬査読 (ScholarPeer)
→ 追加実験 (Rebuttal)

107課題中タスクマルグッドの実成





「後知恵の優位（Hindsight Bias）」

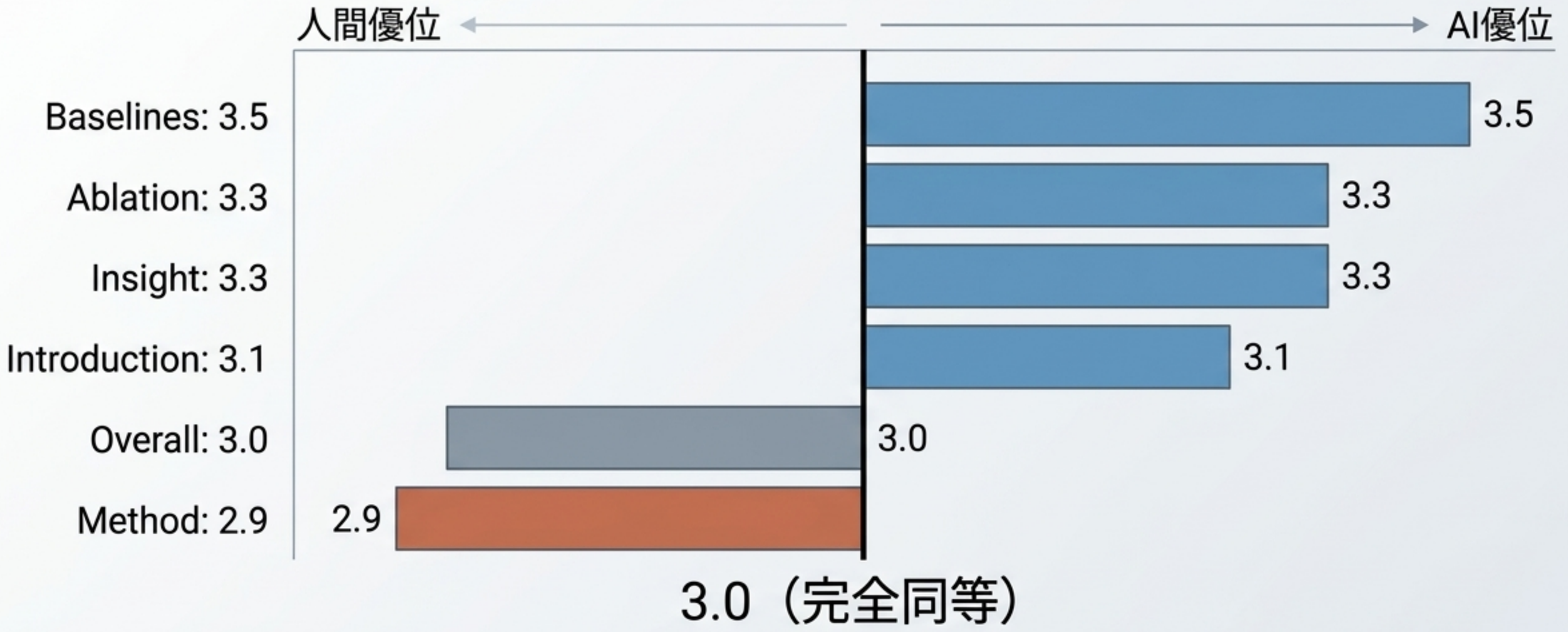
これらは「白紙からの発見」ではない。

既に人間が問題として定式化し、評価指標を定義し、公開コードが存在する既知のML課題である。

AIは「何を研究すべきか」を見つける能力は評価されていない。

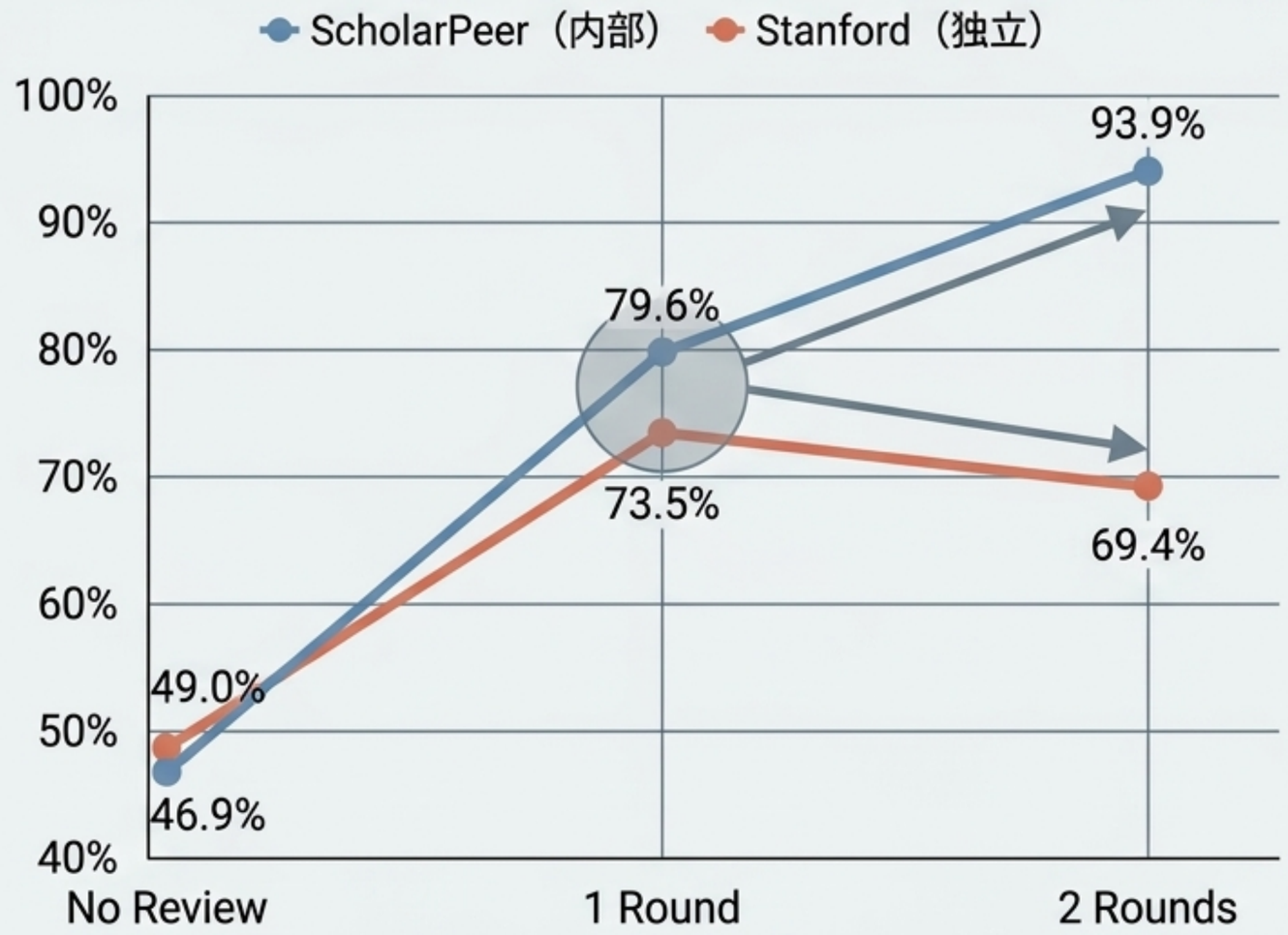
107課題中 86件（80.4%）で改善案生成に成功。

AI Output vs. Human Parity Audit



AIは実験・アブレーションの網羅性で勝るが、方法論の厳密さや総合評価は「トップ会議の人間と同等」に留まる。全面的超越ではない。

査読プロセスにおけるスコアの収束と発散
(Score Convergence & Divergence in Review Process)



⚠

「査読者への過適合（Goodhart’s Law）」

内部査読器（ScholarPeer）を報酬モデルとして反復すると、独立評価器（Stanford）でのスコアが低下する。
「優れた科学」ではなく「その査読者が好む文章」へ最適化されるリスクの実証。

AutoSOTA vs. ScientistTwo: 診断比較マトリックス (Diagnostic Comparison Matrix)

AutoSOTA（速度とスケール）	ScientistTwo（深さと厳密性）
Focus: 高速性能改善	Focus: 研究全体の閉ループ化
Success Rate: 105 SOTA	Success Rate: 86改善 / 107課題
Cost/Time: 平均 約5時間 / 論文	Cost/Time: 平均 約2.5日 / 論文 (約 $3,765)
Average Improvement: 7.5% (Median: 2.7%)	Average Improvement: 25.2% (Median: 7.7%)

結論 (CONCLUSION)

AutoSOTAが単なるパラメータ調整器であるのに対し、ScientistTwoは計算資源を「深さ（アブレーション・追加実験）」へ投資するアーキテクチャ。

評価指標の錯覚 (Evaluation Bias)

相対改善率の平均(25.2%)は外れ値に牽引されている。中央値は7.7%。また、Stanford評価器も最初の15ページしか解析できず、査読エラーの可能性がある。

成功確率の再解釈 (Selection Effect)

「生成成功した論文の72.1%が採択水準」だが、失敗例を含めたエンドツーエンドの実質採択相当率は約58.0%に下がる。

新規性検証の欠陥 (Novelty Failure)

新規性チェックで検索する先行研究はわずか「2本」。これでは既知のアイデアの再発見・再包装を排除できない。

検証チェックリスト比較: 内部整合性 vs. システム再現性

内部監査（極めて優秀）	
✓	Score Verification (49/49)
✓	Method-Code Alignment (49/49)
✓	幻覚引用 (Hallucinations) (0 / 1,814件)

メタシステムの再現性（未確立）	
✗	公式コード・リポジトリの公開 (未確認)
✗	実行プロンプトの完全公開
✗	モデル依存の排除 (変動するGemini/Claude APIに依存)

論文自体の整合性はScientistOneから劇的に改善したが、
システム全体を第三者が再現する土台は欠如している。

システム・セキュリティ (System Risk)

LLMが外部コードを読み、自律実行する性質上、悪意あるパッケージ依存やSandboxエスケープの脅威があるが、論文内に包括的なセキュリティ評価はない。

科学エコシステムの汚染 (Ecosystem Risk)

1課題2.5日で論文を量産可能。査読システムの飽和、AIの自己参照的引用ループ、検索不足による既存技術の「新発見」化など、深刻なスケーラビリティ問題を引き起こす。

コスト構造（Cost/ROI）

1論文あたり約$3,765。大学向けには高価だが、企業R&Dで人間数週間分の検証を代替するなら十分なROIが見込める。

著作権（Copyright - JP）

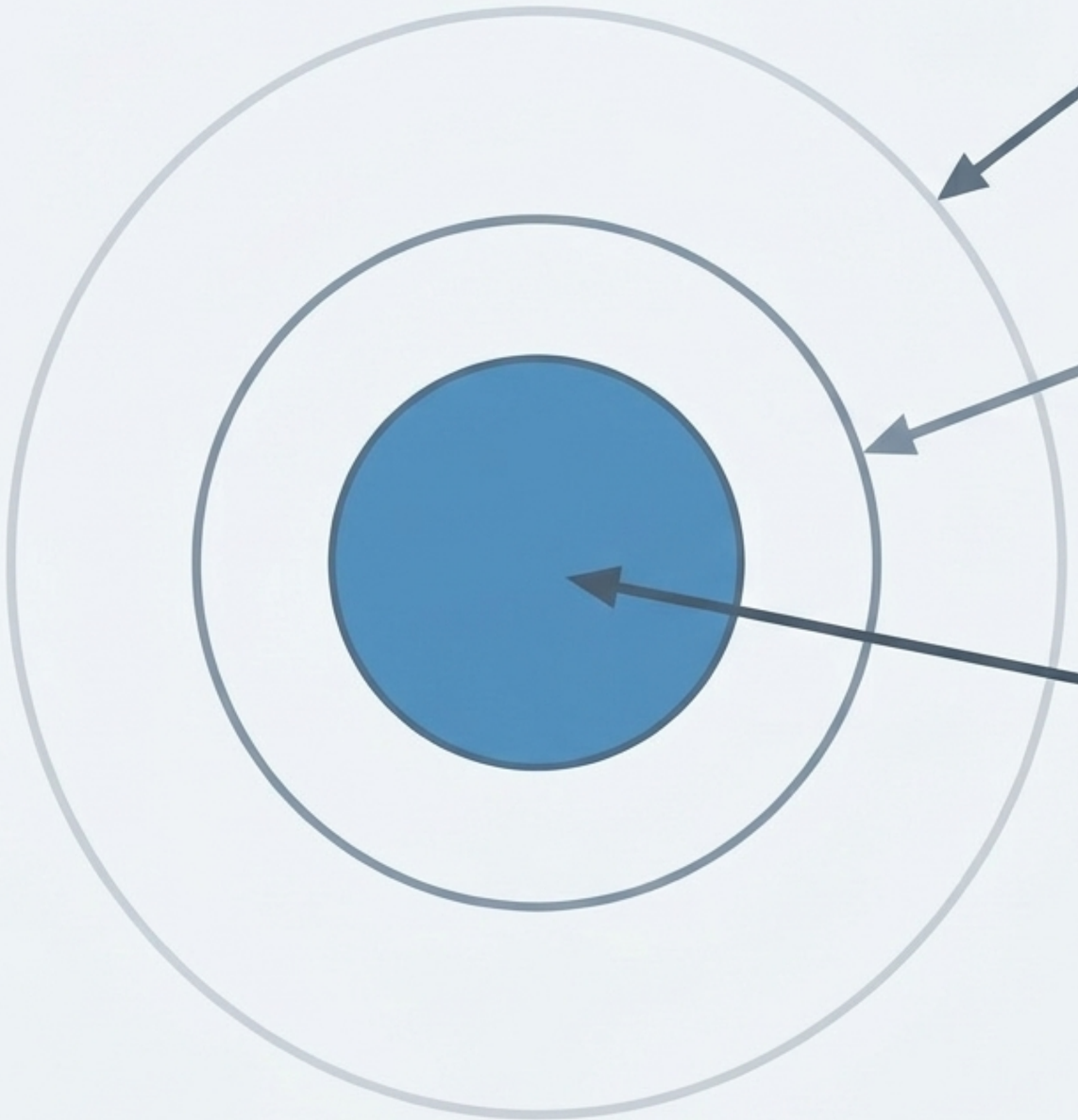
日本の文化庁見解では、人間の「創作意図・寄与」がなければ著作物性は認められない。完全自律生成論文の権利保護は極めて不透明。

特許権（Patents - US）

USPTO規則によりAIは発明者になれない。自律的に「発明」した新アルゴリズムの特許化には、自然人の寄与の証明が壁となる。

規制（Regulations - EU AI Act）

純粋な科学R&Dには適用除外があるが、商用製品として市場投入すれば透明性要件等の対象となるリスク。



未知の基礎科学・理論発見

評価関数が未定義な領域では機能しない。

長期的な階層的探索・物理実験

ロボットラボや生化学統合は未知数。

企業のMLパイプライン最適化 (Autonomous R&D Platform)

課題と評価関数が明確な環境での、SOTA改善、アブレーション自動化、モデルチューニングに絶大な威力を発揮する。

未来課題の事前登録

過去の受理論文（後知恵）ではなく、学習データに存在しない新規課題を事前登録し、人間チームと対決させる。

人間によるブラインド実査読

AI評価器への過適合を排除するため、AI生成論文であることを伏せて実際のトップ会議へ投稿する。

オープンソースと新規性監査

システム全体の公開と、単なる2本の検索ではない、大規模特許・文献グラフ検索による「真の新規性」の証明。

「問うべきは『AIは科学者になったか』ではなく、『研究のどの部分を自動化すれば、科学的信頼性を保ちながら速度を最大化できるか』である。」