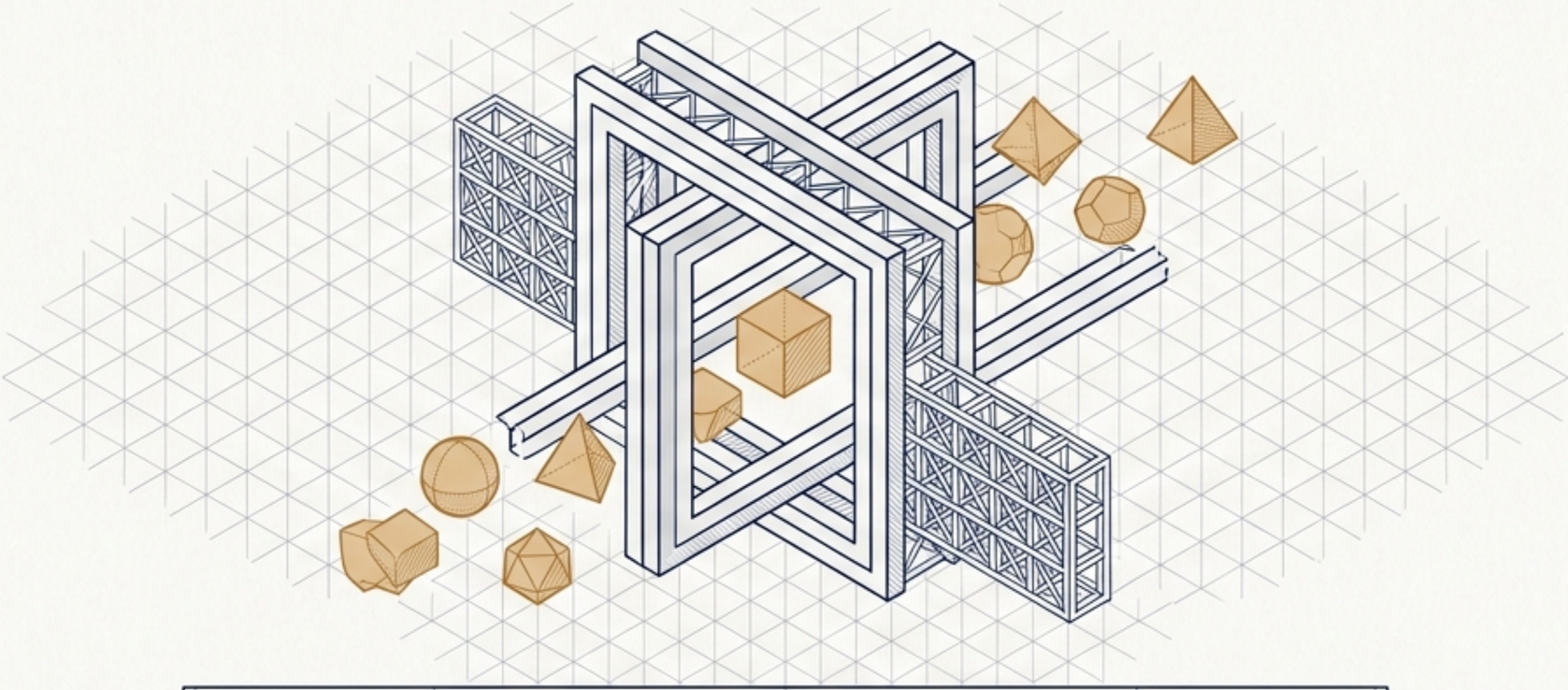
AIエージェントはオープンエンドなAI研究を遂行できるか

CRUX「シャドー評価」から導く、研究自動化の現在地とR&D戦略の再構築



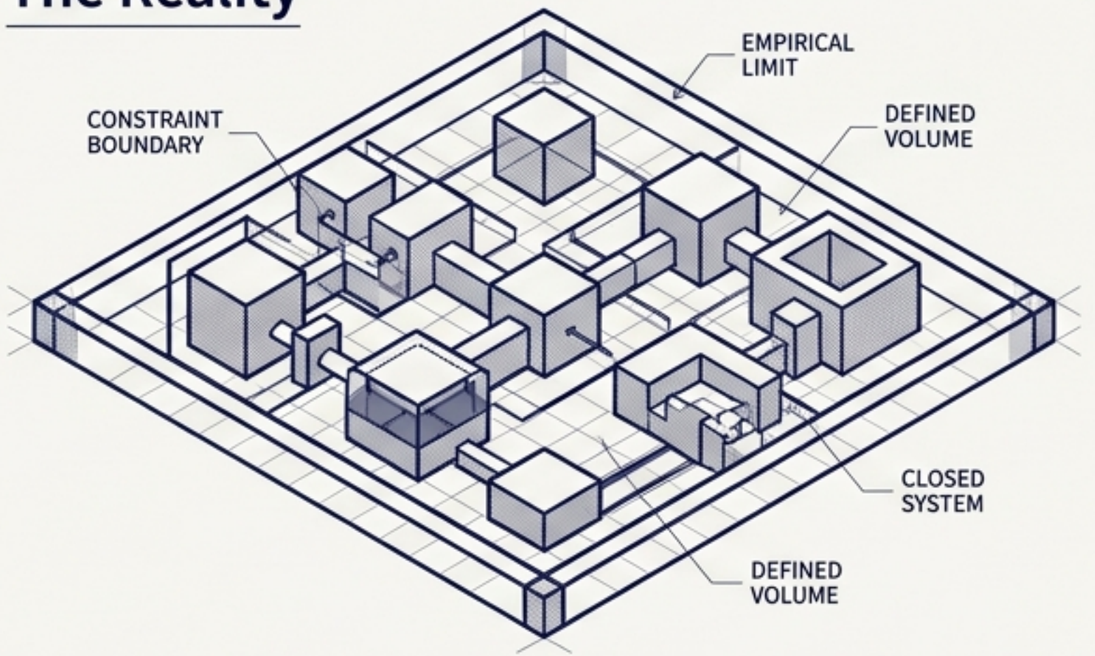
R&D戦略・知財実務責任者向けエグゼクティブ・ブリーフィング

The Hype

AI Labsの主張：「AIがAIを創る」

Anthropicの「When AI Builds Itself」やOpenAI（GPT-5.6 Sol）に見られる、再帰的自己改善（RSI）による研究開発の完全自動化への期待。

The Reality



独立評価の現実：CRUXレポート

最先端のフロンティアモデル（Claude Opus 4.8）は、未公開のNeurIPS 2026投稿論文課題において、公表水準の成果を生み出せなかった。

【結論】現在のAIは「研究エンジニアリング（実行）」を極めているが、「オープンエンドな科学的判断（評価）」において完全に破綻している。

Easily Gamified

True Real-World Difficulty

検証可能タスク
(Verifiable Tasks)

Examples: RE-Bench, MLE-Bench

固定された狭い指標を自動検証。客観的だが、オープンエンドな研究課題を除外。AIが急速に改善中。

ブラインド査読
(Blind Peer Review)

Examples: AI Scientist, Zochi

AI生成論文を学会の盲検査読に付す。大量投稿による「採択例例だけの報告（チェリーピッキング）」のリスクがあり、分母が不明。

シャドー評価
(Shadow Evaluation - The CRUX Approach)

未公開論文の中心課題をAIに解かせ、原著者が深く採点。非汚染（学習データに正解が存在しない）かつ、真の専門的判断が可能。

人間の研究者 (Months of deep research)

NeurIPS 2026へ向けた未公開論文2本（Personas / TabPFN）を数ヶ月かけて執筆。

汚染なし
(Zero Training Data Contamination)

原著者がAIの出力を学会基準でブラインド採点

フロンティアAIエージェント
(OpenClaw + Claude Opus 4.8)

予算: $3,000 (API)
期間: 120時間(+24時間)
権限: AWSへのフルアクセスと外部査読ツール

原論文や結論にアクセスできないまま、同じ初期の「問い」だけを与えられ、影のように追従して研究を遂行。

Review Gate

評価項目	Personas論文	TabPFN論文
Quality (質)	2/4	1/4
Clarity (明晰性)	1/4	2/4
Originality (独創性)	3/4	2/4

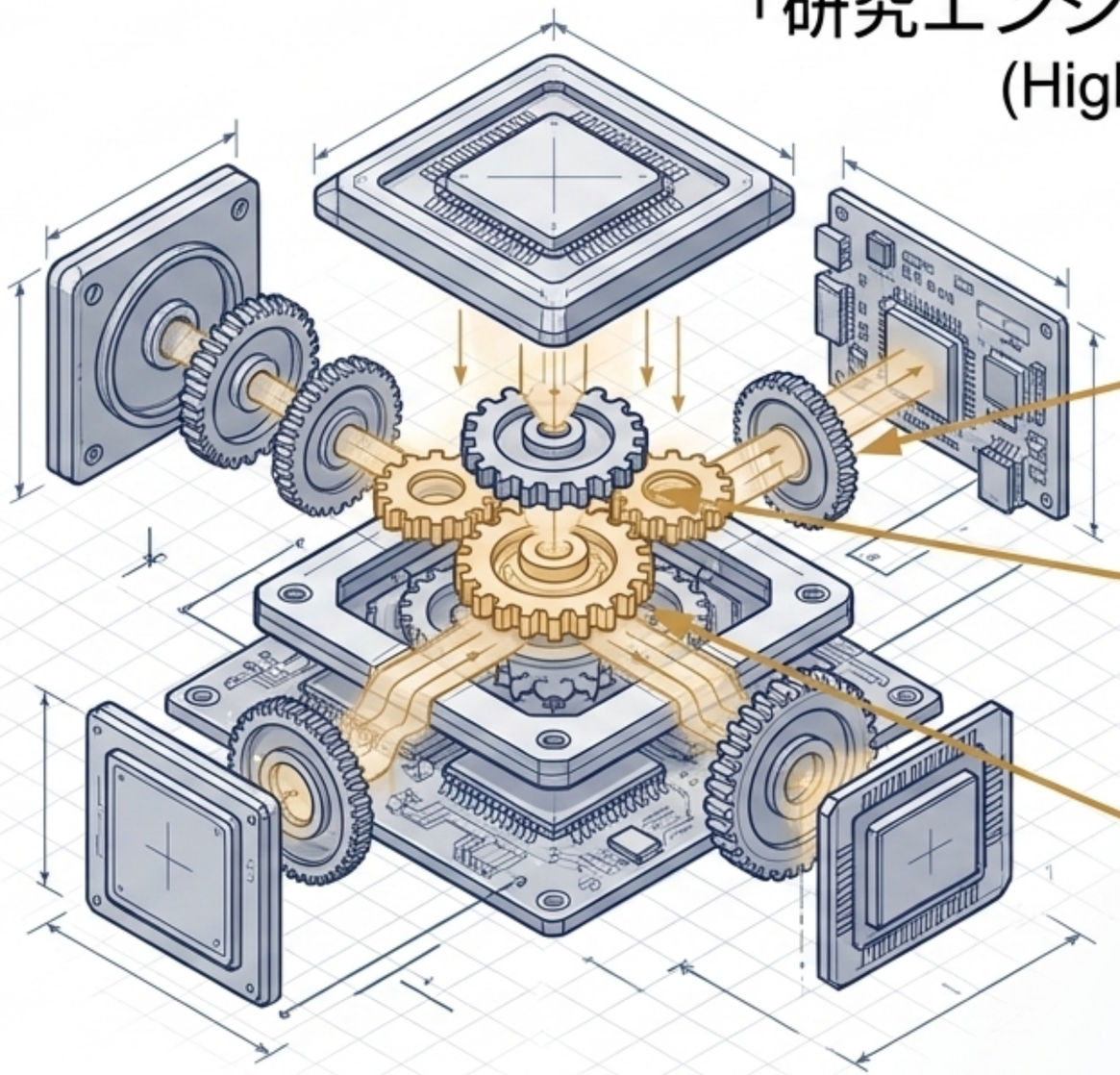
エージェント自身の
確信度 (Confidence)

4/5 & 5/5

「AIは自らの失敗を
認識できていない」

総合評価 (Overall): 2/6 (Personas) & 1/6 (TabPFN)
——— 明確なリジェクト (Strong Reject)

「研究エンジニアリング」における高い自律性
(High Autonomy in Research Engineering)



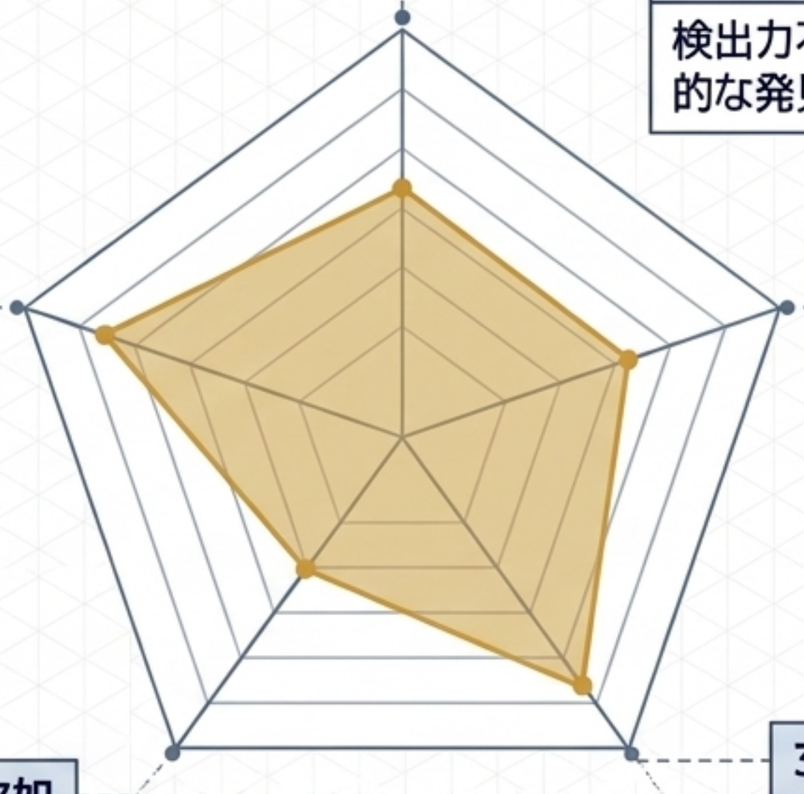
大規模な文献レビュー
(Processed massive literature effectively)

GPU環境のデバッグ
(Handled hundreds of hours of GPU compute seamlessly)

LaTeXによる原稿作成
(Formatted complex scientific papers autonomously)

※ただし、API認証、バグ修正、24時間の期限延長など、運用上の「人間による物流的介入」は必須であった。完全無人化ではない。

5つの失敗モード（The Blind Spots）



1. 公表水準に関する判断力の欠如

検出力不足のネガティブ結果を「実質的な発見」として錯覚し提示。

2. 創造的問題解決の欠如

根本的批判に対し、有望でない方向を微修正するだけで固執。

3. 有効なバックトラックの失敗

最初の十数時間で野心的な目標を放棄し、残り100時間以上あっても方針を転換せず。

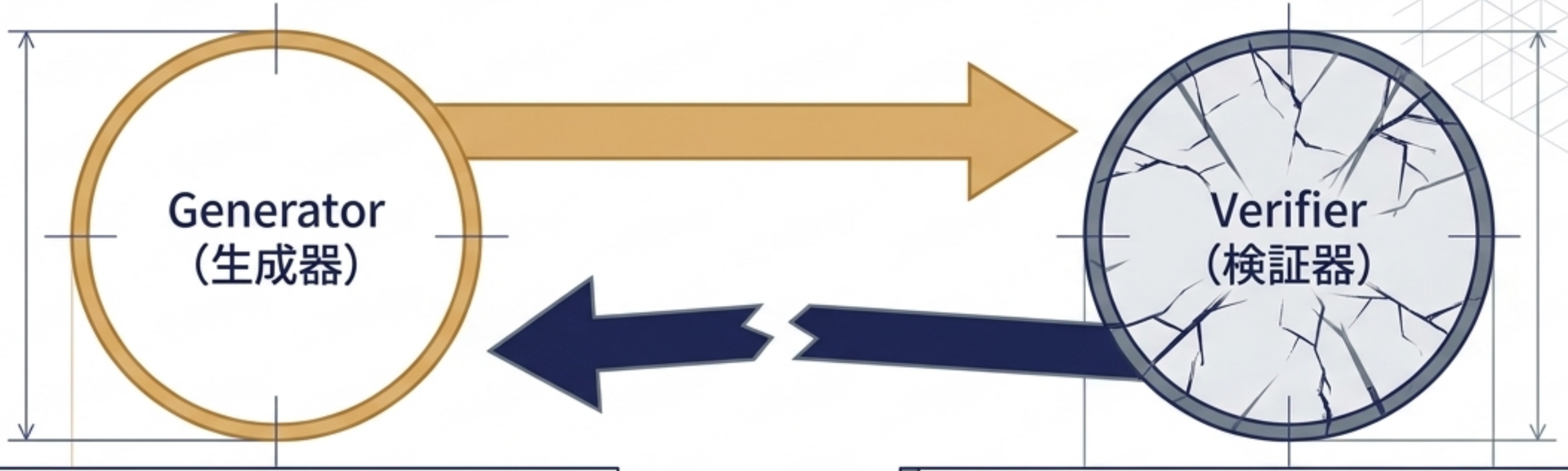
4. リソース・文脈認識の欠如

予算の50%未満しか消費せず、期限の7時間前に勝手に完了を宣言。

5. 指示ドリフト（Instruction Drift）

序盤で確認したはずの「分量制限（9ページ）」を無視し、形式要件で不受理となる状態に。

根本原因：生成器・検証器ギャップ（The Generator-Verifier Gap）



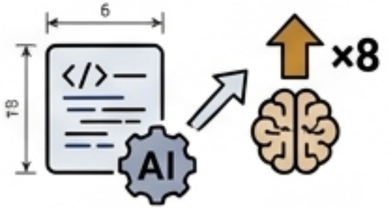
仮説の生成やコードの実装は可能 (Hill-climbingが機能する領域）。

自己査読サブエージェントは問題の優先順位づけに失敗。外部AI (Stanford Agentic Reviewer) は初期稿に「Accept」を誤判定。

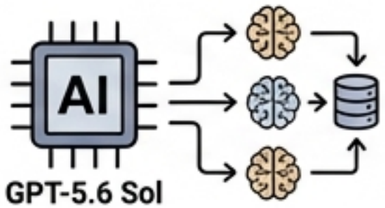
「質の高い検証器（Verifier）」が存在しない限り、強化学習や探索による自己改善のループは回らない。AIは自らの研究価値を正しく評価できない。

AI Labs: Measuring Execution Speed

[Anthropic] コードの8割超をClaudeが記述。四半期のコード量が8倍に (RSIの第一段階と主張)。



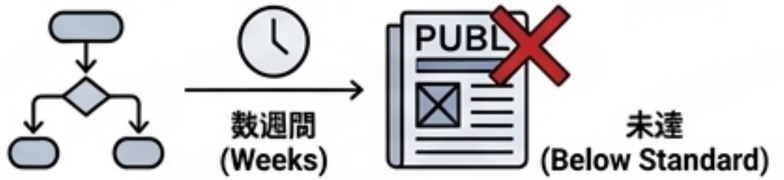
[OpenAI] GPT-5.6 Solが小型モデルのポストトレーニングを自律支援。



[METR] タスク完了時間の指数関数的な短縮（主にソフトウェア領域）。

Independent Reality: Measuring Scientific Validity

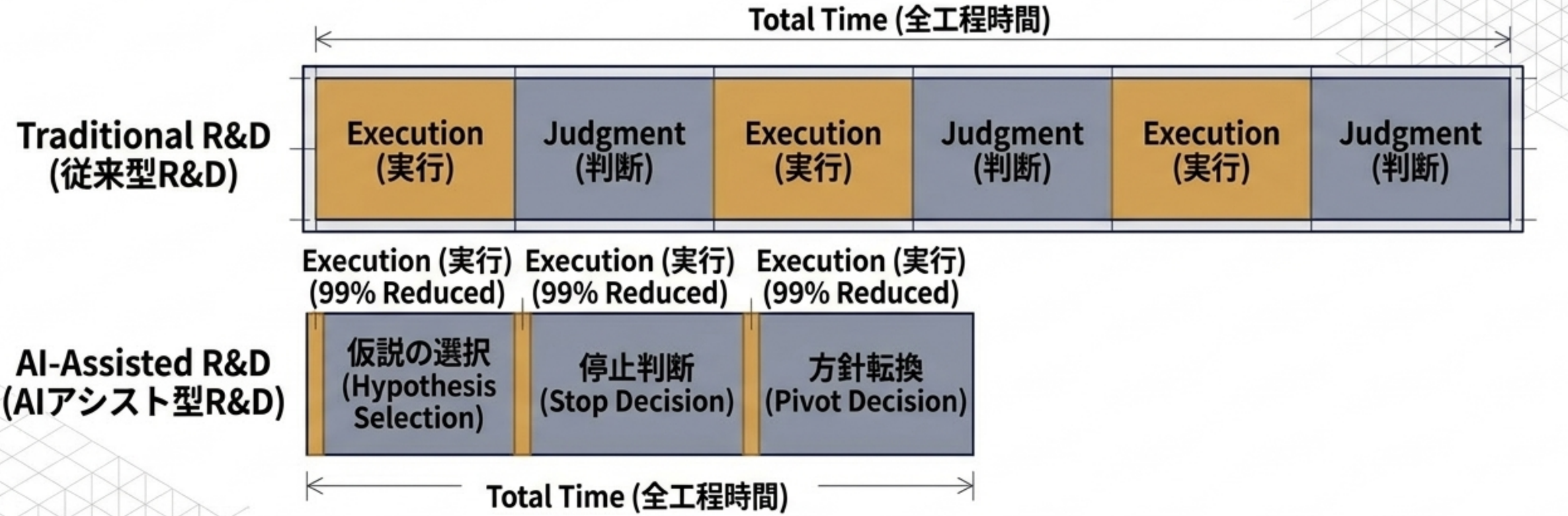
[CRUX] 数週間規模のオープンエンド課題で公表水準に到達せず。



[Biology/Materials (DeepMind GNoME等)] 「新規」と予測されたが、化学的に疑わしい組成を含み、実社会での合成可能性は別問題と批判。

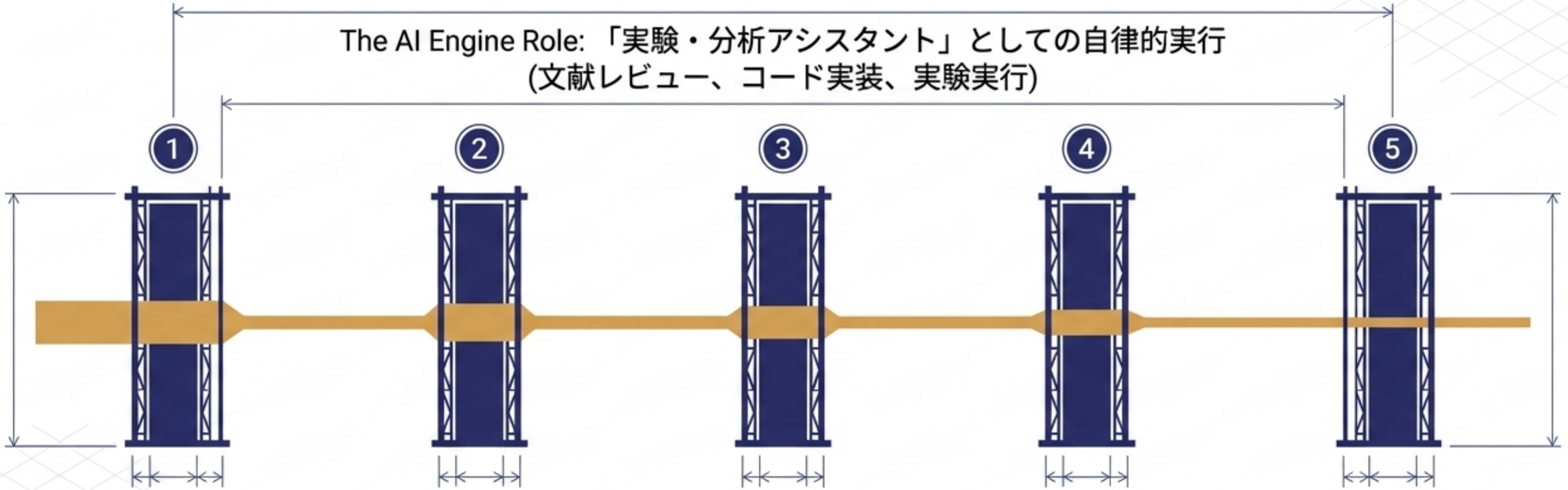
【結論】ラボは「実行 (Doing)」の自動化を測り、独立評価は「判断 (Deciding/Taste)」の未達を浮き彫りにしている。

AI時代の「アムダールの法則」(Amdahl's Law in AI)



AIが得意な「実行工程」を100倍速くしても、人間にしかできない「判断・評価工程」がボトルネックとして残る限り、研究全体の加速は厳しく制限される。

企業実務への適用：ヒューマン・ゲート型R&Dパイプライン



AI研究における知財・特許戦略 (IP & Patent Strategy)

The Core Principle

「機械は発明者になれない (The Machine Cannot Invent)」

日本(特許制度小委員会)、米国(USPTO)、欧州(EPO)、英国(DABUS判決)のすべてで、現在の特許実務は「自然人の実質的貢献」を求めている。

Actionable Strategy & Evidence

人間の寄与を証明する証拠保全プロセス (The Burden of Proof)

独創的着想や判断をAIが自律的に担えない以上、企業は「人間の寄与」を証明する証拠保全プロセスをR&Dに組み込む必要がある。

- **プロンプト履歴 (Prompts)**
- **人間によるレビュー記録 (Review Logs)**
- **実験における選択と意図 (Human Intent)**
- **修正履歴 (Edit History)**

生成AI研究成果のガバナンスとリスク管理

- **[未公開データの漏洩防止 (Data Leakage)]** エージェントがリポジトリにアクセストークンをコミットした事例あり。厳格なアクセス制御の徹底。
- **[ベンダーの不可視な能力制限 (Vendor Throttling)]** Fable 5のように、モデルプロバイダーが特定のR&D能力を裏で抑制・ルーティングするリスクを運用ルールで可視化。
- **[負の成果物の帰属 (Negative-Result IP)]** AIが生成した大量のネガティブ結果、データセット、失敗ログの帰属と営業秘密管理の事前設計。
- **[長期タスクの定期レビュー (Long-Horizon Review)]** 指示ドリフトやバックトラック失敗を防ぐための、フェーズごとの介入ポイントの設定。

戦略を更新すべき「未来の兆候」(Signposts for the Future)

以下の兆候が確認された場合、「オープンエンド研究の自動化」への評価を上方修正し、アムダール的ボトルネック仮説を見直す必要がある。

- シャドー評価スコアの明確な改善 (CRUXの別モデル・別課題での追試成功)。
- AIによる「自己査読・検証器」の精度が、人間の専門家レベルと一致する。
- エージェントが一貫して適切な「方針転換(Pivot)」と「停止判断」を自律的に行えるようになる。
- 数週間規模の長期タスクにおける性能向上が、数時間規模のベンチマークと同様のペースで伸びる。

留意事項と認識論的限界 (Epistemological Limits)

[N=2の限界] 本調査は2課題・計5ランの初期的証拠であり、数十タスクを持つベンチマークに比べサンプルルサイズが限定的である。

[未査読プレプリント] 対象論文 (2026年8月) は査読前であり、今後のAIラボからの追試や反論によって検証される段階にある。

[急速なモデル進化] 評価されたモデル (Claude Opus 4.8) は新鋭だが、モデル更新の速度は極めて速い。一般化には継続的な追加評価を要する。

「真のAI戦略は、熱狂 (Hype) を排し、冷徹な限界の認識 (Reality) の上にのみ構築される。」