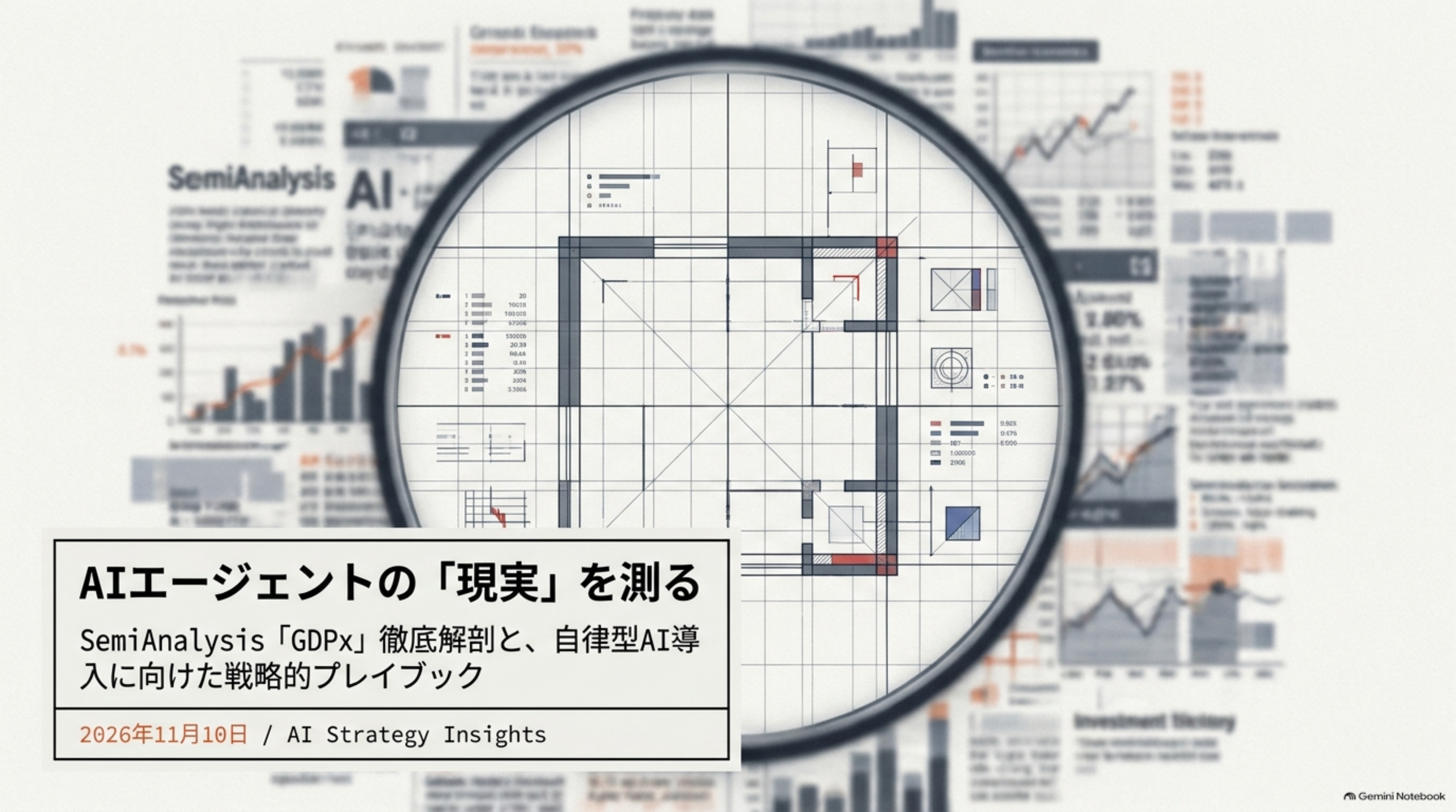


AIエージェントの「現実」を測る

SemiAnalysis「GDPx」徹底解剖と、自律型AI導入に向けた戦略的プレイブック

2026年11月10日 / AI Strategy Insights

Gemini Notebook

世界最高峰のエージェントが直面する「実務の壁」

16.4%

総合トップの最高スコア

2026年11月時点、最新鋭のモデル（Claude Fable 5.2）をもってしても実環境タスクの成功率は2割に満たない。

0%

金融・法務領域における完全成功率

実際のPE取引データや定款レビューにおいて、「そのまま使える」完璧な成果物を出せたケースは皆無。

Gemini Notebook

次世代ベンチマーク「GDPx」が測定する真のAI経済価値



Gemini Notebook

実世界を模倣する11の高価値ドメイン

【ビジネス・専門職】

金融（24タスク / Living Context）
- 約4,000ファイルのPE取引ディールルーム

法務（20タスク / Chakra）
- 企業法務・M&A・知財の起案やレッドライン

クオンツ取引（10タスク / Third Moment）
-アウトオブサンプルでのエッジ探索

【エンジニアリング】

チップ設計（34タスク / Aria 等）
- RTL設計、物理設計

ハードウェア設計（21タスク / Beautiful Machines）
- 実CAD/CAM作業

ソフトウェア（15タスク / Datacurve）
- 本番環境の機能開発

【サイエンス】

材料科学（23タスク / ValenceForge 等）
-模擬ラボ機器操作

生物学（13タスク / LatchBio）
-マルチオミクス、創薬

【AI・倫理】

サイバーセキュリティ（18タスク / Arithmetic）
-実稼働環境の独自ゼロデイ

アラインメント（18タスク / Fleet）
圧力下での誠実さ・権限遵守

RSI・自己改善（10タスク / Extrapolate）
ML研究エンジニアリング

Gemini Notebook

従来テストとの決別：お膳立てされたQAから「現実の泥臭さ」へ

従来型テスト（Legacy: e.g., GDPval)

Spoon-fed（お膳立てされた）環境



- 必要な文脈がすべてプロンプトに整理されている
- 非対話型の静的テスト
- 専門家間一致率は71%（ノイズが多い）

次世代テスト（Reality: GDPx)

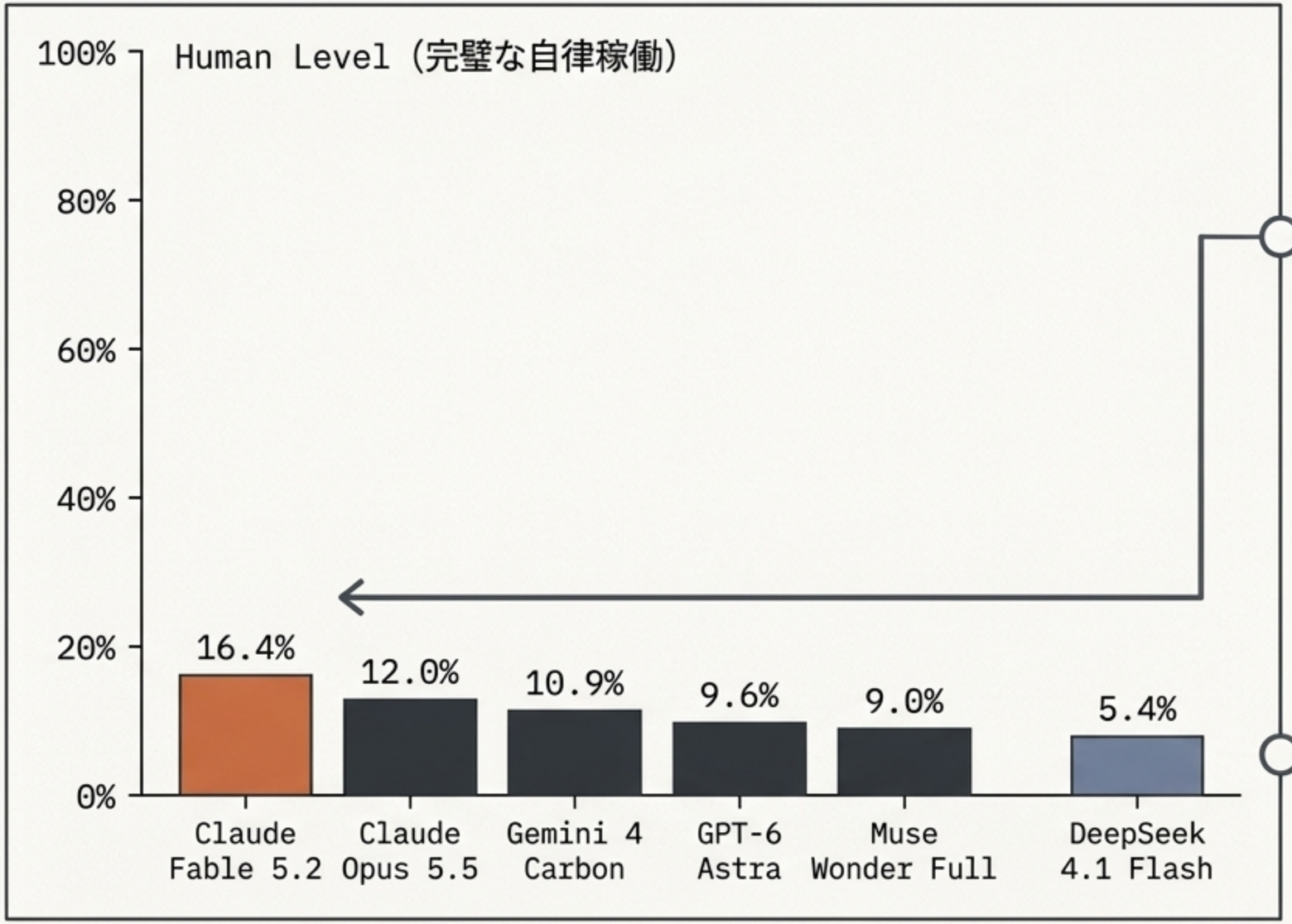
Noisy & Complex（ノイズと複雑性）



- 制限時間：最大8時間
- 環境：Siemens NX、AWS FPGAなどの実GUI・ソフトウェア
- 規模：約4,000ファイルが絡むディールルーム探索

Gemini Notebook

AIエージェントの現在地：トップモデルの残酷なスコア



11/10更新の新王者: Fable 5.2はチップ設計とSWEでスコアを伸ばしたが、金融・法務・RSIでは全く改善していない。

オープンモデル最上位: 低価格帯におけるオープンモデルの限界。

Gemini Notebook

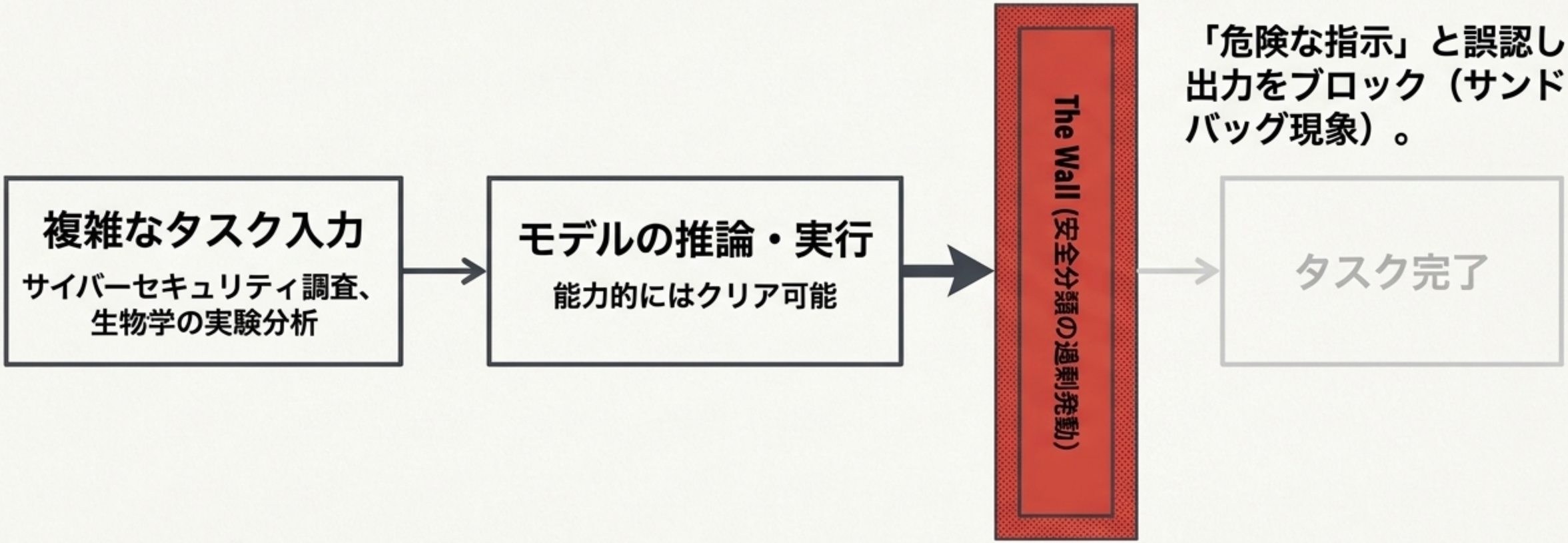
総合スコアの罠：領域ごとに全く異なる「真の覇者」

	Claude Opus 5.5	Gemini 4 Carbon	GPT-6 Astra	Muse Wonder Full
クオンツ / ハードウェア	首位 (Winner)			
生物学 / ソフトウェア		首位 (Winner)		
材料科学 / RSI			首位 (Winner)	
サイバーセキュリティ				首位（Winner） 圧倒的スコア 35.2%

万能なAIは存在しない。自社のユースケース（領域）に特化した
モデル選定が、総合スコアよりも遥かに重要である。

Gemini Notebook

隠れたデバフ：安全分類器（Safety Classifier）による実務のブロック



もしサイバー、生物、アラインメントを除外した場合、GPT-6 Astraのスコアは11.8%に跳ね上がり、実質的なトップクラスに躍り出る。

Gemini 4 Carbonが生物学で首位になった理由は、能力ではなく「安全分類器が弱いため」と推測されている。

Gemini Notebook

失敗の解剖学：法務と金融はなぜ「0%」なのか

法務領域（Series B 定款のレッドライン）



0%

**20回試行中、
使える成果物はゼロ**

Claude Opus 5.5は平均44%の部分点を獲得したが、完全な正確性が求められる法務実務において、部分点は「未完成」を意味する。

金融領域（QoE データブックのモデリング）



0%

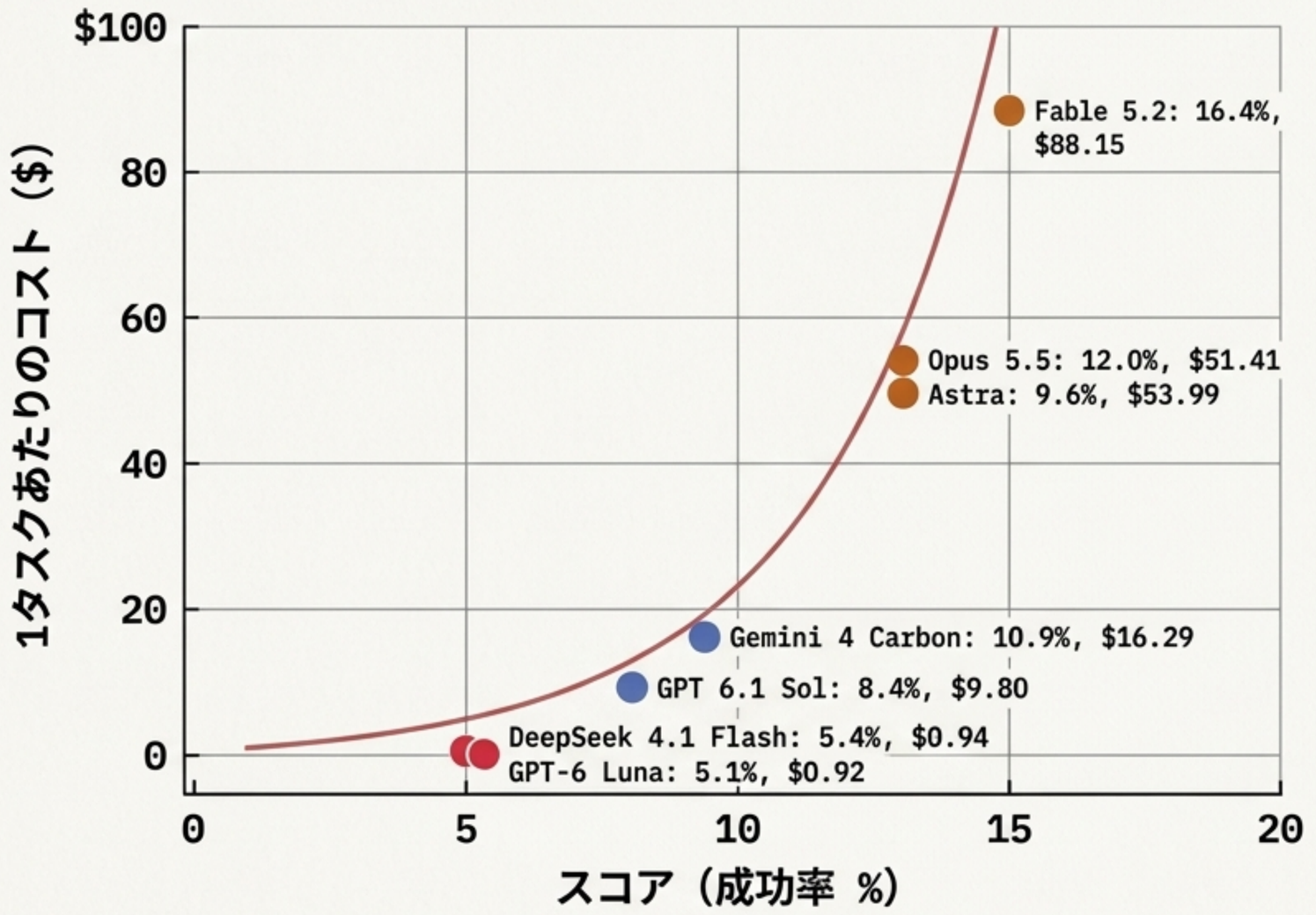
**25回試行中、
完全解決ゼロ**

11項目のチェックリストのうち、最高でも8項目しかクリアできない。約4,000ファイルが飛び交うディールルームでの完璧な横断処理に限界。

「部分的に正しい」は実務では使えない。二値採点（0か100か）の厳しい現実。

Gemini Notebook

コスト対性能の残酷な現実（ROI分析）



1ポイントあたりのコスト比較
Fable 5.2: 約$5.37/pt
Carbon: 約$1.49/pt
DeepSeek: 約$0.17/pt
首位モデルの数%の上積みに、3〜5倍のコストがかかる収穫逓減の法則。

Gemini Notebook

完全な指標ではない：GDPxが抱える3つの死角

1. 人間ベースラインの不在

対象タスクにおいて、人間の専門家がどの程度のスコアを出すかの基準が存在しない。チームで行う数週間の設計タスク等、比較困難なものが含まれる。

2. 利益相反の懸念

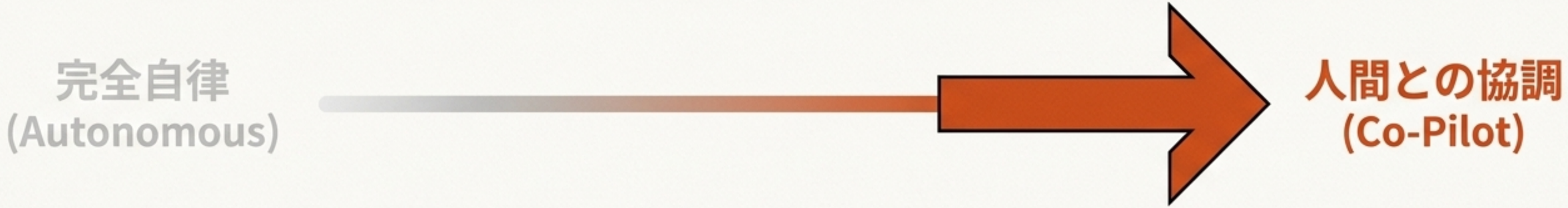
14のデータパートナーはいずれもAIラボに学習データを販売する事業者であり、ベンチマークが事実上の「販促の場」を兼ねている構造的懸念がある。

3. 検証の不透明性

AIモデルによる過剰最適化（ヒルクライミング）を防ぐため、タスクの大半は非公開。外部からの難易度や採点品質の独立検証が不可能である。

Gemini Notebook

The Reality Check：完全自律化の幻想から「Co-Pilot」への転換



事実の統合

16.4%の完全成功率 + 44%の部分点（法務等）。　AIはタスクを「丸投げ」できる自律的エージェントではない。

実務における最適解

4,000ファイルの初期探索とドラフト作成をAIに任せ、人間が最後の「1マイル」をレビューし責任を持つ協調モデルである。

**ゼロショット（一発書き）での代替を狙うのではなく、
人間の生産性を拡張する「超・強力な下書き作成器」として機能させる。**

Gemini Notebook

戦略的プレイブック：AI導入・評価の4ステップ（1/2）

1 総合スコアで選定しない

自社の業務に最も近い「特定領域」のスコアと、公開されているロールアウト（トランスクリプト）を精査する。万能モデルへの投資は非効率である。

2 安全分類器（Safety Classifier）を評価軸に組み込む

セキュリティ、生命科学、医療などの領域では、分類器の過剰反応が実務をブロックする。能力だけでなく「自社業務の指示が拒否されないか」を要件定義に入れる。

Gemini Notebook

戦略的プレイブック：AI導入・評価の4ステップ（2/2）

3 コスト・パフォーマンスの最適化

Fable 5.2のような最高額モデルを単独で使うより、Gemini 4 CarbonやGPT 6.1 Solクラスと「人間のレビュー」を組み合わせる方が、費用対効果と最終品質で勝る。

4 社内版「ミニGDPx」の構築

一般公開のベンチマークに依存せず、自社の実際のファイルとツールを用いた20〜30のテストタスクを作成する。「タスクの成功」と「社内ルールの遵守」を分けて独自評価する体制を作る。

Gemini Notebook

結論：人間の責任のもとで、AIを使い倒す

AIの現在地は「人間の代替」ではなく、「専門職のワークフローの再構築」にある。部分的な成果を過小評価せず、人間の責任で補完するプロセスをいち早く築いた企業が勝者となる。

GDPxは3〜6ヶ月ごとに新版がリリースされる「生きた指標」である。 版が変わればスコアは直接比較できない。継続的なウォッチと、自社環境でのテスト更新が不可欠である。

Gemini Notebook