

AIエージェントの評価入門

2026-08-14

自然な完了報告だけでは、AIエージェントの業務成功を判断できない。

● 事実 エージェントは複数手順でツールを使う

△ 注意 途中の誤りは後続処理へ伝わる

◆ 推論 回答と実際の最終状態は分けて確認する

長い仕事を任せる利用が拡大

AIの利用単位が、一問一答から複数手順の業務へ広がっている。

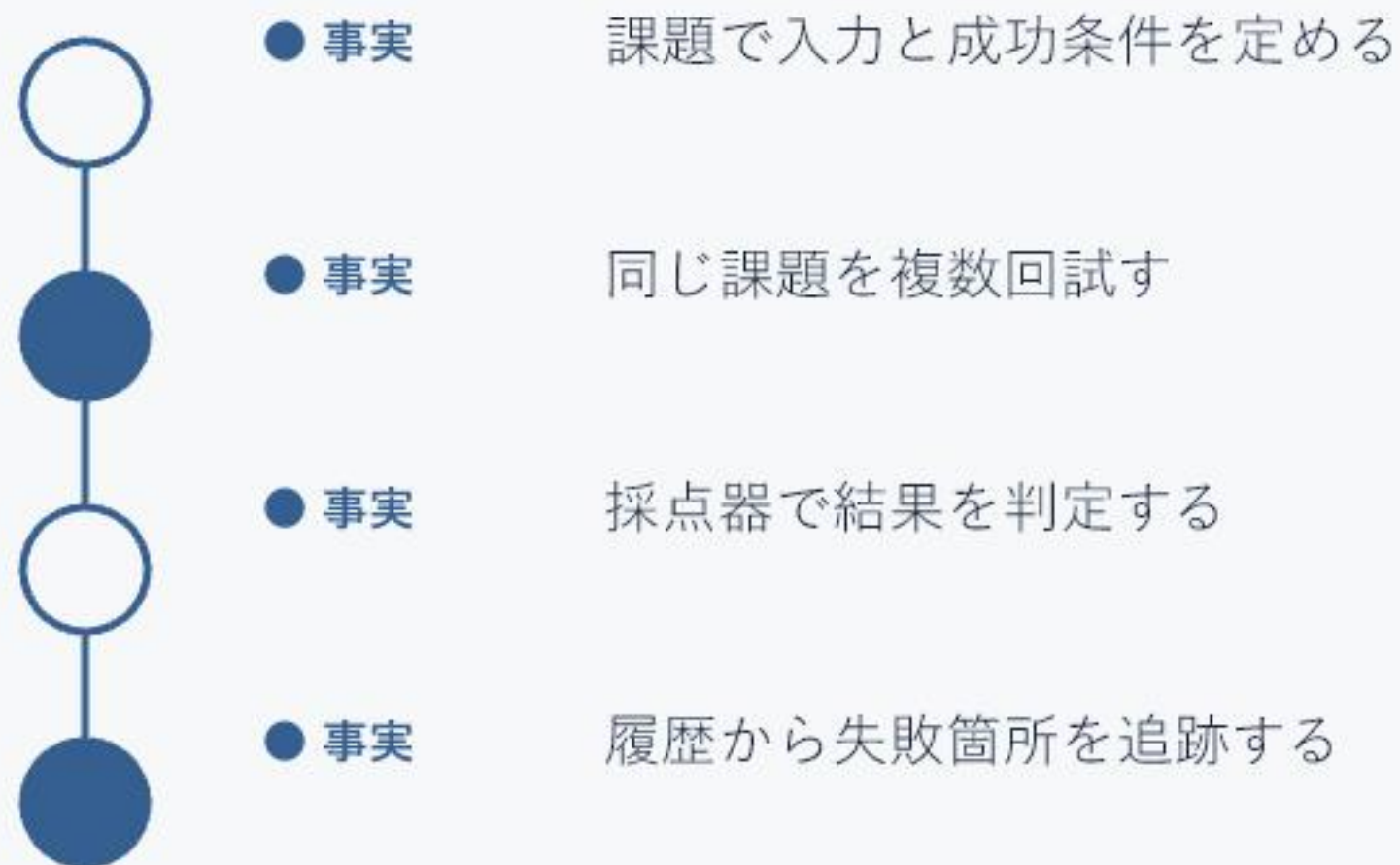
● 事実 ツールを組み合わせで作業

● 事実 1時間超の依頼が70%以上

△ 注意 同社内と同社製品の利用状況に基づく数値

4つの部品で評価する

課題・試行・採点・履歴をそろえ、偶然の成功と再現できる成功を分ける。



見るべきは最終状態

発言でなく、予約・データ・成果物の最終状態を確認する。

△ 注意 完了報告と実際の成功は別

◆ 推論 自動検査・AI採点・人で確認

◆ 推論 重大リスクは即時不合格

20件の評価から始める

代表業務を小さな評価セットにし、変更のたびに同じ条件で確かめる。

◆ 推論 通常例と例外・禁止依頼を含める

◆ 推論 成功率・重大失敗・時間・コストを記録

● 事実 モデルや指示の変更前後で再評価する