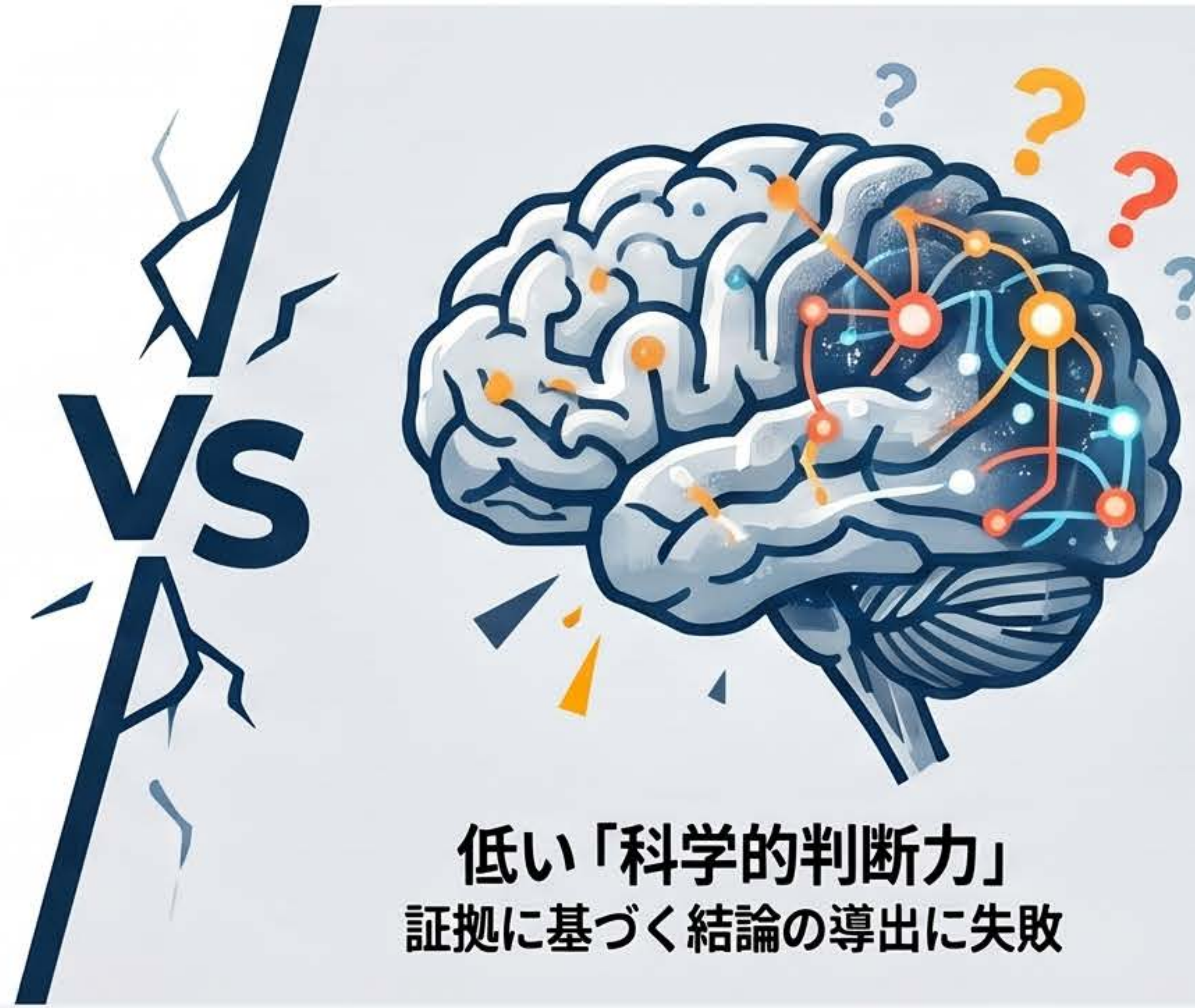
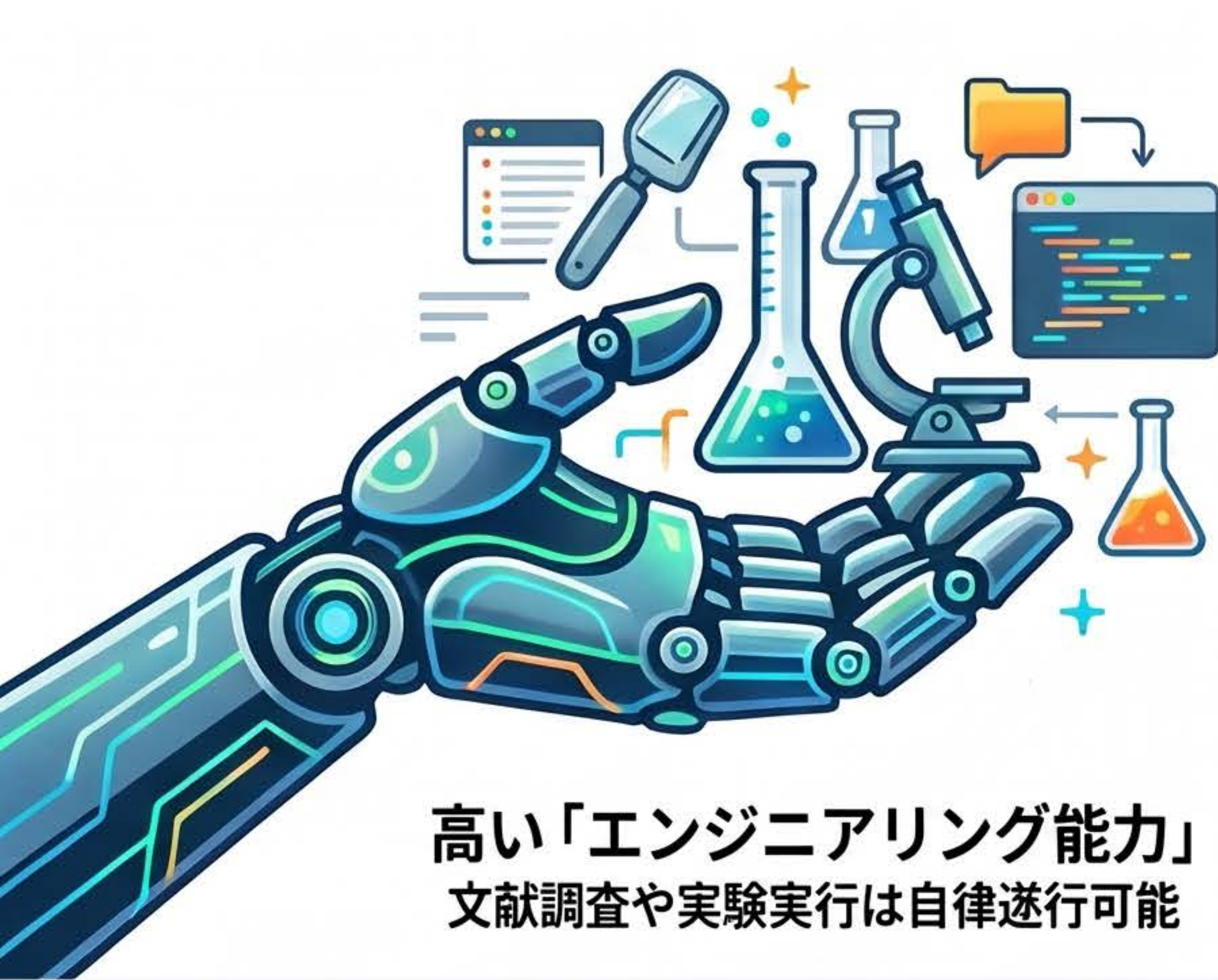


AI研究自動化の現在地：シャドー評価が明かす「実行力」と「判断力」の乖離



2件のケーススタディでいずれも「不採択」判定
原著者による総合評価は1/6（強い不採択）および2/6（不採択）と低迷

AI査読の限界：重大な欠陥の優先順位付けが困難

AI査読は人間と共通の懸念を捉えるが、致命的な問命な問題の特定に課題がある

AIエージェントに共通する3つの失敗モード



戦略的バックトラック
と方向転換の欠如

行き詰まった際に、過去の
成果を捨てて探索段階から
再始動することができない



リソース配分とタイ
ムライン管理の不備

予算を早期に使い切る、
あるいは時間を残して抽
速に終了する傾向がある



指示漂流（Instruction
Drift）と制約の無視

長時間の実行中に、ページ
制限などの当初の明示的
な指示が守られなくなる



一貫性の欠如と幻覚
（ハルシネーション）

推論過程で矛盾が生じたり、
存在しない事実を捏
造することがある

評価パラダイムの比較

	自動検証型 評価	ブラインド 人間査読	シャドー評価
主 な 限 界	スクリプト （正解率など）	匿名査読者 （検否）	原論文著者 （詳細査読）
限 界	問いの設定や 価値判断を評 価しにくい	査読の質の変動 や選択的報告の リスク	小桜本であり、 非盲検性による バイアスがある