

「見た目だけ」のAIはもう終わり？新指標MMGRが暴く生成AIの「賢さの死角」

これまでのAI評価は、生成される画像や動画の「見た目のリアルさ」に偏っていました。新ベンチマークMMGRは、AIが物理法則や論理を本当に理解しているかという「推論能力」を初めて測定。これで、その結果、最先端モデルが抱える意外な弱点と、今後の進化の方向性が明らかになりました。

なぜMMGRが必要か？ 見せかけの賢さを見抜く新基準 従来の評価では「推論の失敗」を見抜けない



従来の評価
(Conventional Evaluation)
見た目がリアルなら、物理法則を無視した映像でも高評価になってしまう問題があった。



MMGR: AIの「本当の賢さ」を測る初の包括的ベンチマーク
現実世界のルールを理解し、一貫したシミュレーションができるかを評価する。

測定する5つのコア推論能力

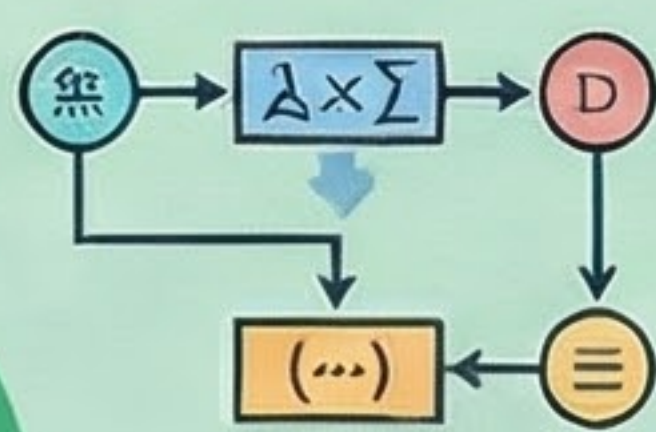
物理的推論
重力や衝突など、直感的な物理法則の理解度。



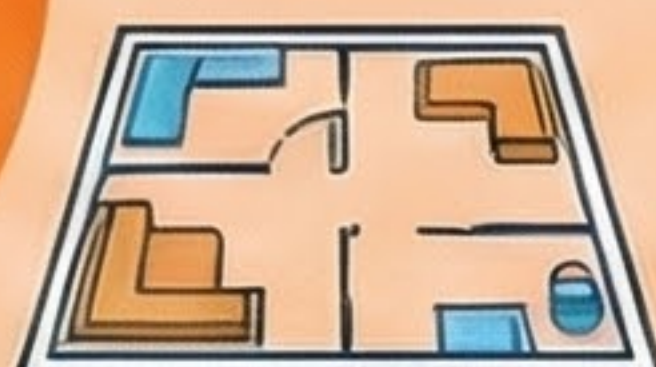
3D空間推論
3次元空間の関係やナビゲーション能力。



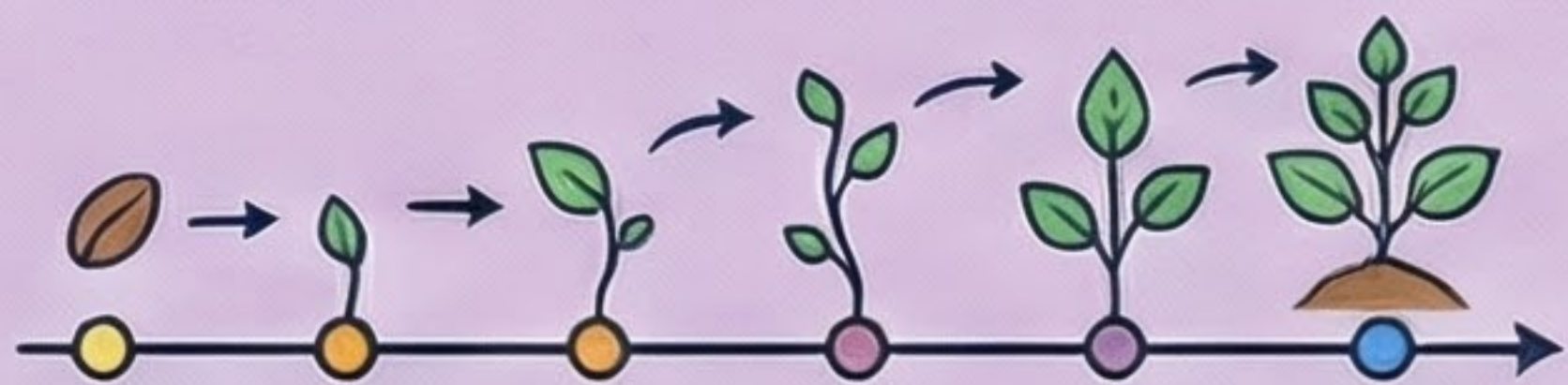
論理的推論
ルールに基づいた記号的・独創的な思考力。



2D空間推論
平面上のレイアウトや相対位置の正確な解釈。



時間的推論
因果関係やイベントの順序など、時間的な一貫性。



衝撃の結果：最先端AIが抱える「推論の錯覚」 「推論の錯覚」：答えは合うが、過程はデタラメ

問題：

入力

プロセス

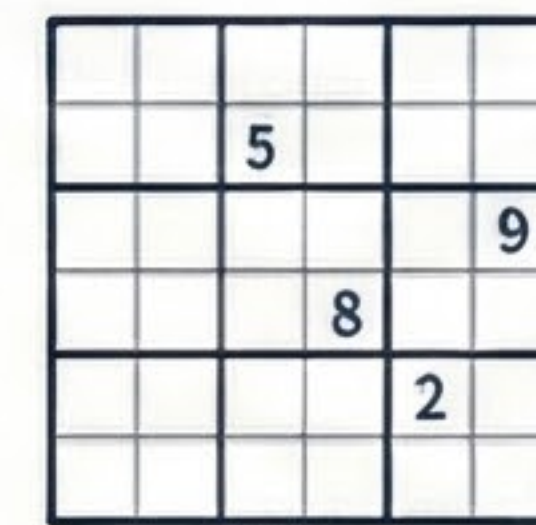
最終解答
17

最終解答
17

ある数学テストでは最終解答の正解率74%に対し、過中計算の正しさは値が12%だった。

12%
正しい過程

抽象的推論タスクで壊滅的な結果を記録



10%未満

数独やパズルの正解率は10%未満で、最先端モデルでも全く解けていない状態だった。

モデルが犯す典型的なエラー例



物理法則の違反
ビリヤードの球が衝突せずにすり抜ける。



空間的一貫性の欠如
ナビゲーション中に壁を通り抜けてワープする。



ルール違反
迷路の壁を自分で道してゴールへの道を作る。