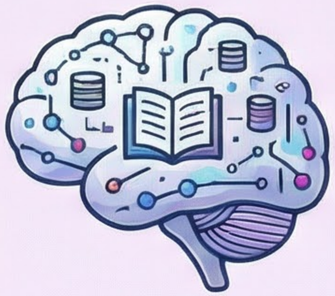
なぜ中国製LLMはARC-AGI-2で苦戦するのか？：知識依存型モデルと抽象推論の壁

多くの中国製LLMはMMLUやHumanEvalで世界トップクラスの成績を収めていますが、ARC-AGI-2という「視覚的な抽象規則を解く」テストでは、人間や一部の最新推論システムに大きく引き離されています。これはモデルの「知能」の欠如ではなく、ベンチマークの設計とLLMが得意とする学習の方向性のミスマッチによるものです。

ベンチマーク設計の決定的な違い

「知識の検索」か「未知の規則の発見」か

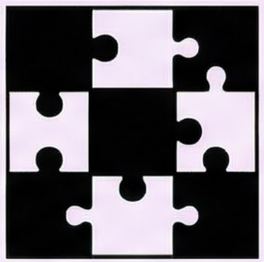


一般的なベンチマーク (MMLU, HumanEval)

膨大な学習コーパスに依存するが、過去問の混入で点数を稼ぎやすい。

ARC-AGI-2

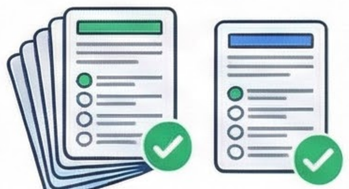
初見の図形規則をその場で推論する力を問う。非公開セットには転移が効かない。



採点と探索枠

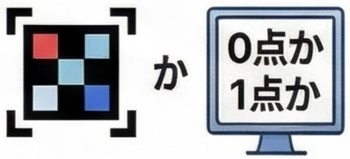
MMLU / C-Eval & HumanEval

多分野の知識・言語推論・プログラム合成



選択肢一致（多肢選択）
・pass@k（多数試行が可能）

ARC-AGI-2



完全一致（0/1）・pass@2
（極めて狭い探索枠）

0点か1点か：部分点なしの厳格採点

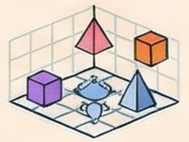
ピクセル単位の完全一致のみを正解とし、生成AIが得意な「もっともらしい回答」を許容しない。

pass@2という極めて狭い探索枠

提出チャンスが2回しかないため、多数の候補から正解を当てる総当たりが適用しにくい。

性能を阻む5つの技術的仮説

2D空間表現（Encoding）のミスマッチ



グリッドをテキストとして言葉自空間認識(必要)

グリッドをテキストとして扱う既存手法では、回転や反転などの空間的直感（帰納バイアス）が働きにくい。

「探索・検証・修正」ループの欠如

探索 (Search)　修証 (Verify)　修正 (Refine)

ARCでの高得点には、LLM単体ではなく、外部ツールで候補を機証し自己修正するシステムが必要。

ベンチマーク特化型学習の限界

既知のパターン

未知の規則

知識ベースの試験は過去問の混入で点数を稼ぎやすいが、ARCの非公開セットにはその転移が効かない。