中国製LLMの「ベンチマーク・パラドックス」：なぜARC-AGI-2で苦戦するのか？

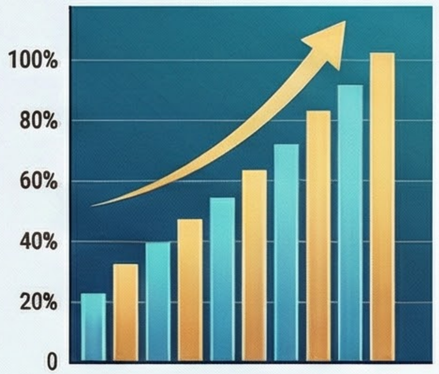
顕著な性能の乖離：知識の記憶 vs 真の推論

（既存知識の応用）　（未知の課題への適応）

MMLU / MATH

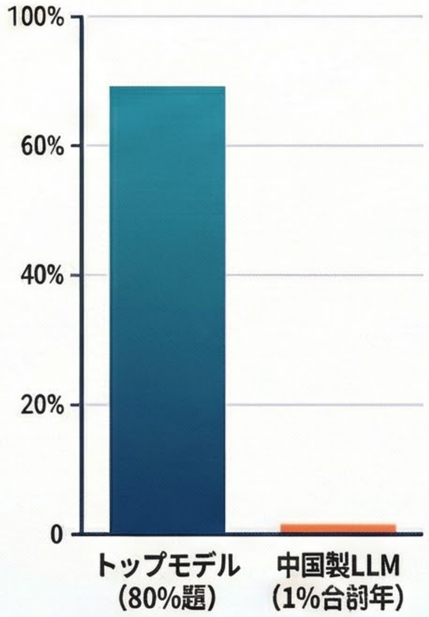
ARC-AGI-2

世界トップクラス（非常に高い）



世界トップクラス（非常に高い）
世界トップクラスの高にを記録する。

スコアの絶望的な格差



トップモデルが80%超を記録する中、多くの中国製LLMは1%台前年に篙まる。

完全新規で非公開

上位モデルは複数の推論経路を生成・自己評価する「推論ハーネス」を活用している。

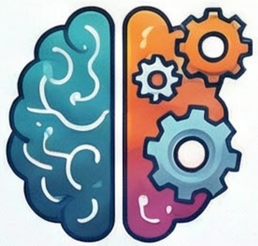
データ汚染の影響

学習データ（記憶）

既存テスト

データ汚染の影響
学習的な獄に、エデーデータを記惊を沢化している。

「結晶性知能」と「流動性知能」の差



結晶性：既存知識の応用
既存知識データな顳鞘が隅し、擷的テストルを連成している。

なぜ苦戦するのか？3つの構造的要因

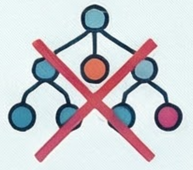
「検証可能なタスク」への過学習

数学やコードなど、正確を過せる。

数学やコードなど、正碥が閧題な領域に夥化した強化学習が汎化性能を助げている。

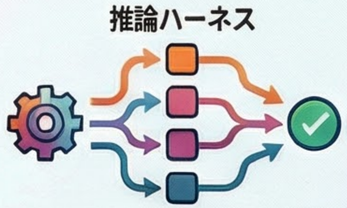
静的な推論アーキテクチャの限界

単一モデルの予測に依存し、動的な複験やプログラム合成を行う仕組みが不足している。



動的探索・合成

成功モデル（Gemini 3等）との差



上位モデルは複数の推論経路を生成・自己評価する「推論ハーネス」を活用している。