S5-P27

年代別評価データにおける大規模言語モデルの性能差の検出力分析

有川 透矢[1], 池田 航[1], 齋藤 由佳[1], 松田 翔太[1], Kwek Yong Sing[1], 柴田 圭悟[1], 種口 暁人[1], 津田 純花[1], 松田 陵佑[1], 赤間 怜奈[1, 2, 3], 鈴木 潤[1, 2, 4]

[1]東北大学, [2]理化学研究所, [3]国立国語研究所, [4]国立情報学研究所 LLMC

is-failab-research@grp.tohoku.ac.jp

概要

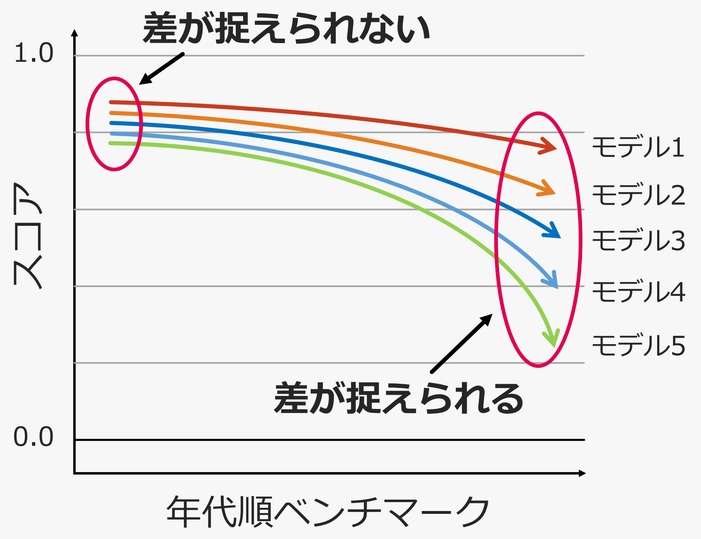
- LLMの性能向上に対応して，**高難度化・細分化**された評価データが提案されている
- ❓ 評価データのモデル識別能力が**年代とともにどのように変化したのか？**
- **評価データが捉えるモデル性能差**を**年代の観点**から調査 → 新しいタスクほど優れた分解能を示すと予想
- ❗ 評価データの年代とモデル間の性能差は，**ドメインごとに異なる傾向を示した**
  - Reasoningではスコア差が拡大傾向 → 近年の推論能力への注目により，他ドメインより変化が顕著だったか

評価データの年代別整理に関する仮説

- 年代が新しくなるほど，性能の高いモデルに対応するために
  - スコアの飽和を防ぐための**高難度化**
  - 能力差を測るために**モデル間スコア差の拡大**

  が生じていると予想

- 性能水準の異なるモデルで評価
  → **幅広い性能帯に対する評価データの識別力を検証**
- ❓ 年代の違いによるスコアの傾向は確認できるのか



実験設定

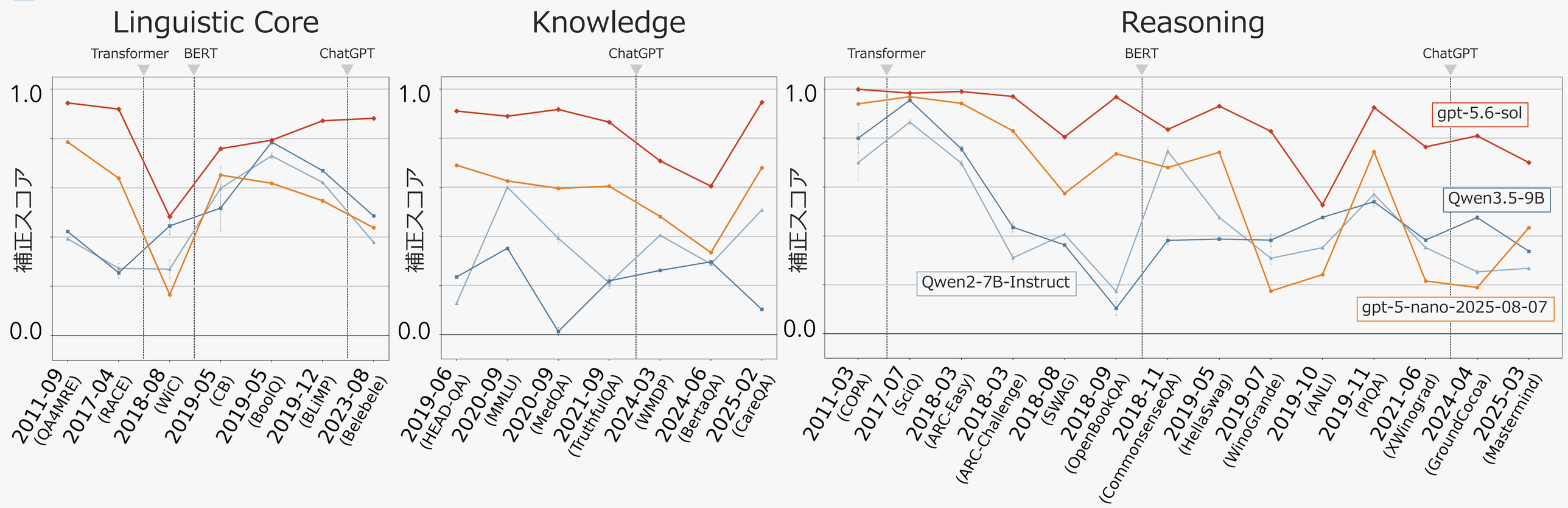
- 多肢選択式タスクを対象に**公開年代順に比較**
- チャンスレート補正スコアを使用

$$S_{m,t} = \frac{A_{m,t} - C_t}{1 - C_t} \quad \begin{cases} S_{m,t} & : \text{補正スコア} \\ A_{m,t} & : \text{モデル}m\text{のAccuracy} \\ C_t & : \text{タスク}t\text{のチャンスレート} \end{cases}$$

対象モデル

- GPT系列モデル
  - ◆ gpt-5.6-sol
  - ● gpt-5-nano-2025-08-07
  - 回答ラベルを生成して判定
- Qwen系列モデル
  - ■ Qwen3.5-9B
  - ▲ Qwen2-7B-Instruct
  - 選択肢ごとの尤度で判定

実験結果



- Linguistic Core, Knowledgeドメインは年代による明確な**スコア傾向は見られず**
- ❗ タスクの専門領域や問題形式のような**年代以外の要因が支配的**だったか
- Reasoningドメインのみ，**スコアが減少傾向**
- Reasoningドメインにおいて，新しいタスクほどGPT系列モデル間の**スコア差が拡大傾向**
- ❗ **推論能力への注目**に伴い，近年のタスクが**高難度化・識別力向上**した可能性

分析 ❓ 仮説と異なる結果になった要因は？

- ドメイン分類の粒度が荒い
- ベンチマーク**作成意図の混在**が年代傾向を不明瞭にした
- モデル系列間での**評価方式の違い**
- **対象タスク数・モデル数が不十分**
- ❗ **タスク数・モデル数の拡大**により，**個別差を平均化した傾向**の確認が期待される