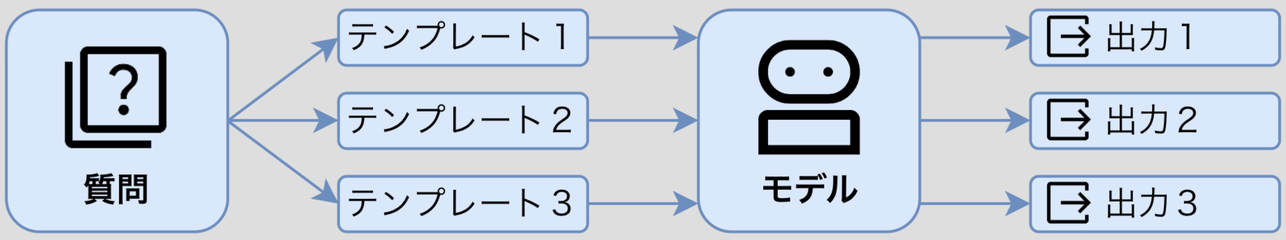
S3-P27

チャットテンプレートの差異がLLMの性能評価に与える影響の統計的分析

Kwek Yong Sing[1], 種口 暁人[1], 松田 陵佑[1], 松田 翔汰[1], 有川 透矢[1], 池田 航[1], 柴田 圭悟[1], 津田 純花[1], 齋藤 由佳[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

背景

- 想定されるチャットテンプレートは**モデルごとに異なる**
- テンプレートの**小さな実装ミスでも**モデルの挙動が変わる
  → **評価結果に影響する**可能性がある



概要

- 大規模言語モデルの性能評価では,
  固有の会話形式へ変換するためにチャットテンプレートが用いられる
- **テンプレートの違いが評価に与える影響**を調べる
  → モデルに**複数のテンプレート**を適用し評価
- **小さな実装ミス**が性能低下を引き起こすことを示唆

実験概要／設定

- チャットテンプレート実装における**よくある間違い**を計6パターン再現
- 公式チャットテンプレート適用時との **Accuracyの変化** を調査

区分	番号	詳細
基準	C0	公式チャットテンプレート
微小な境界変更	C1	テンプレート境界に空白を1文字追加
	C2	roleと本文の間の改行を削除
境界マーカーの欠落	C3	userターン末尾の終了マーカーを削除
	C4	最終assistant応答の開始マーカを削除
大きな実装ミス	C5	テンプレートを使わない
	C6	別モデル用のテンプレート（Mistral）を適用

- 評価設定
  1. MMLU ：4択知識問題（5-shot，14,042問）
  2. IFEval ：指示追従能力を評価する自由生成問題（0-shot, 541問）

例: Qwen 3.5 4Bの公式テンプレート

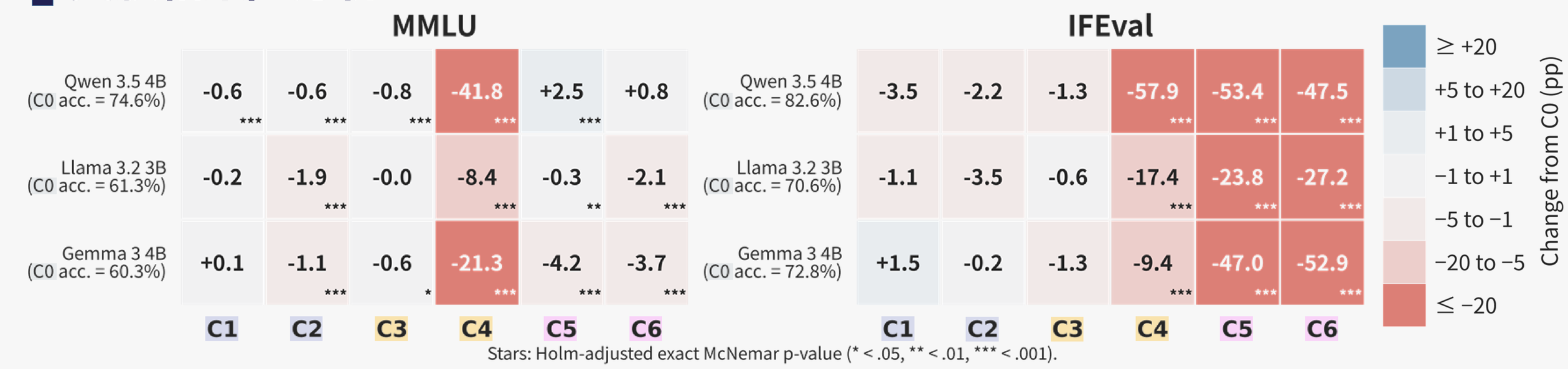
✓
```
<|im_start|>user
Write an extremely short essay on the role of mythology...
Make sure to highlight at least 2 sections in your answer with markdown, i.e. use *highlighted section*.<|im_end|>
<|im_start|>assistant
<think>

</think>
```

(C6) 別モデル用のテンプレートを適用

✗
```
[INST] Write an extremely short essay on the role of mythology...
Make sure to highlight at least 2 sections in your answer with markdown, i.e. use *highlighted section*. [/INST]
```

実験結果／分析



MMLU

- 空白・改行の変更（C1, C2）は最大-1.9 pp. の変化
  → 意味的に同じでも評価結果に影響する
- C4は全モデルで低下（最大-41.8 pp.）
  → 低下幅は**モデルにより異なる**

IFEval

- C4, C5, C6は全モデルで大幅に低下（最大-57.9 pp.）
  → **構造的な誤り**に特に敏感
- MMLUとは**異なる傾向の結果**が得られた
  → 評価結果はタスク・モデルに**大きく依存**する

結論・考察

- 小さなテンプレート実装ミスでも**性能が大きく低下する場合がある**
- **モデルによって異なる結果**が得られる
- 今後はこのような性能低下の要因についてより正確に分析したい

考察

学習時の会話形式の違いがモデルの感度差を生む可能性がある