S2-P24

LLM-as-a-Judgeにおけるリッカート評価とランキング評価の比較分析

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

TOHOKU UNIVERSITY

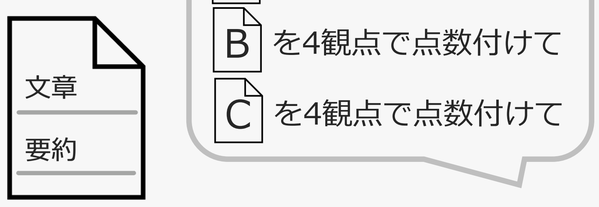
概要

- 同一ベンチマークを**リッカート・フルランク・ペアワイズ**で評価
- ペアワイズでリッカートと同等の人手との相関を得るには**より多くの件数を評価する必要性**
  →ペアワイズによってリッカート評価を代替し，**コストを削減することは難しい**
- GPT-4oとGPT-5.5でフルランクにおける人手評価との相関に**大きな差**
  →同じベンチマークでも評価形式を変えることで**分解能を変えられる**

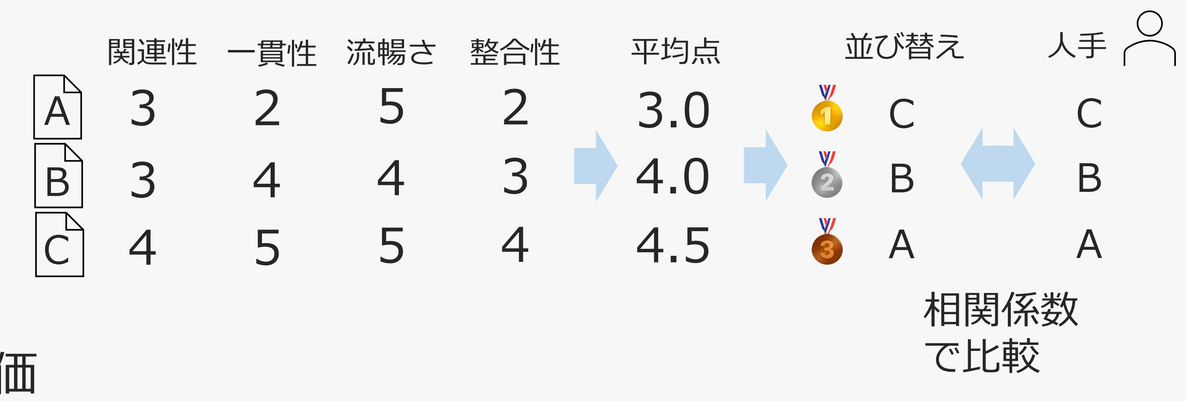
ベンチマークの評価形式

リッカート評価 （ベースライン）

- 各観点における点数の平均値
- 人手評価とLLM評価の相関係数で性能を評価
  (Spearmans'ρ を使用)



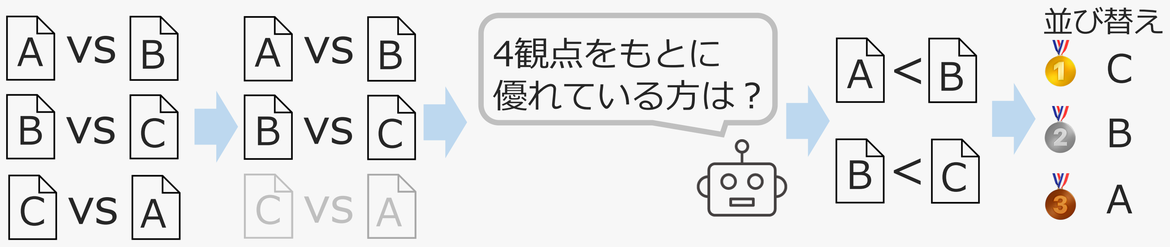
SummEval（要約の評価）[Fabbri+TACL'21]



比較手法 **同一ベンチマーク**を異なる評価形式で評価

Q1. ベンチマークの評価コストの削減は可能？

評価難易度：ペアワイズ評価 < リッカート評価



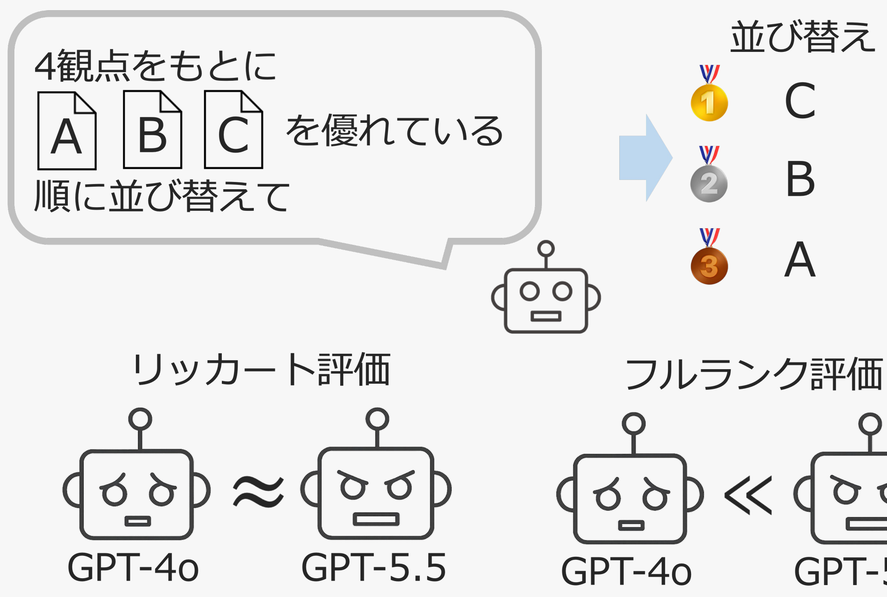
ペア ${}_{N}C_{2}$　nサンプル

- リッカート評価の評価件数 N (=100)
- ペアワイズ評価の評価件数 n
  例１）n = 1だとコスト大幅削減だが，精度が悪い
  例２）全部サンプルすると ($n = {}_{100}C_{2} = 4950$)
  精度は高いが，リッカートより件数が多い

性能が飽和するサンプル率の評価件数で比較

Q2. ベンチマークの分解能の向上は可能？

評価難易度 ：フルランク評価 > リッカート評価

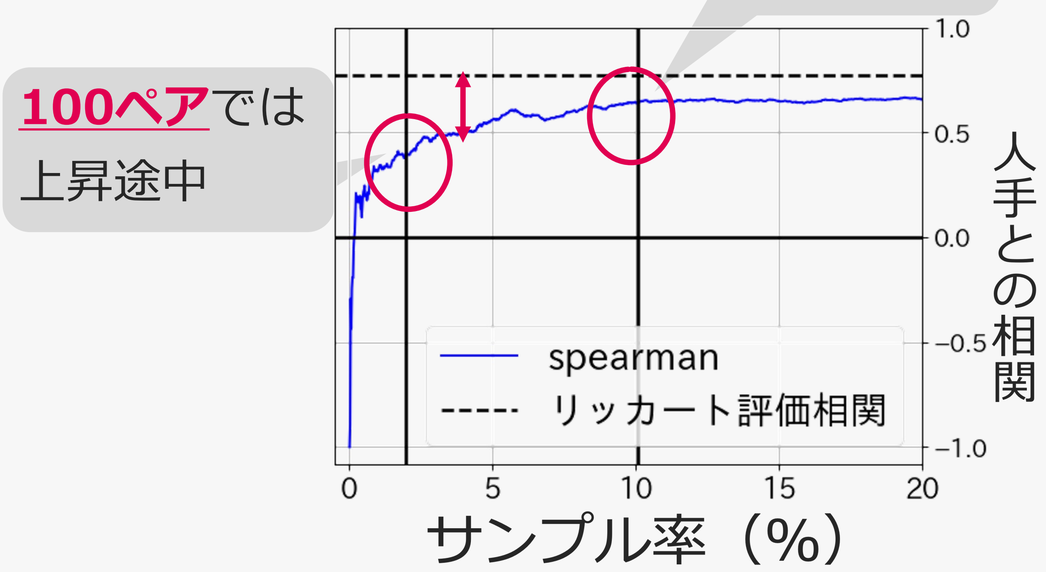


リッカート評価では見えない性能差がフルランク評価では見えるか？

実験結果

ペアワイズ評価

サンプル率と人手との相関の関係



リッカート評価件数 **100件** < ペアワイズ飽和ペア数 **500ペア**

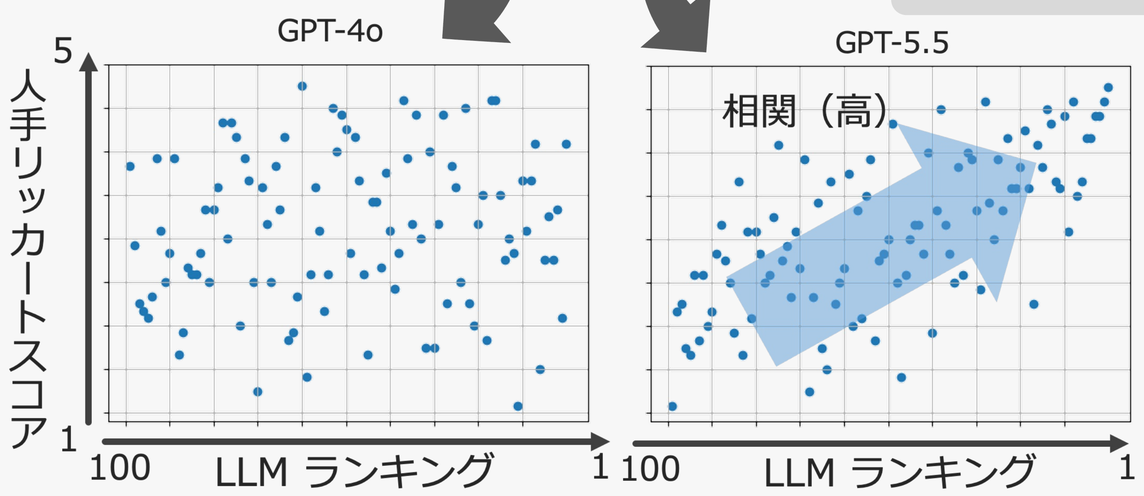
！ペアワイズによるコスト削減は難しい

フルランク評価

評価形式	GPT-4o	GPT-5.5	相関差
リッカート	0.77	0.72	0.05
フルランク	0.095	0.66	**0.565**

差が見えない

差が見える



！リッカートで見られない性能差をフルランクで検出可能