summary-omissions-bench-ja: 要約漏れを評価する日本語ベンチマーク

高槻高等学校 2年GSコース

研究の背景・目的

- 大規模言語モデル（LLM）による電子メールや学術論文の要約活用事例が増加している
- しかし、重要情報の欠落により誤解や損失を招く恐れがある
- 要約漏れを評価するベンチマークを開発するため、既存の手法の調査を行い、その問題点を見つける

LLM-as-a-Judge[1]とは

- 人間ではなくLLM を他のLLMの評価者として利用すること
- 例：「問題文」「正解例」「生成解答」「採点基準」などを与えて、LLMに点数やフィードバックを出させる
- ここでは「要約元文章」「要約サンプル」「生成要約」「重要箇所基準」を与えた
- ジャッジ役のLLM自体が偏りを持つと評価結果に反映される、自己の生成解答を良く採点する、人間との評価と一致するわけではないといった問題がある
- 「要約元文章」から「要約サンプル」に加えて「重要箇所基準」を用意する必要があり、アノテーションがより高コストである

Elyza-tasks-100で使用されるLLM-as-a-judgeのプロンプトの例。改行を一部編集。[2]

```
問題，正解例，採点基準，言語モデルが生成した回答が与えられます。
# 指示
「採点基準」と「正解例」を参考にして、、回答を1,2,3,4,5の5段階で採点し、数字のみを出力してください。
# 問題
{input_text}
# 正解例
{output_text}
# 採点基準
基本的な採点基準
- 1点: 誤っている、 指示に従えていない
- 2点: 誤っているが、方向性は合っている
- 3点: 部分的に誤っている、 部分的に合っている
- 4点: 合っている
- 5点: 役に立つ
基本的な減点項目
- 不自然な日本語: -1点
- 部分的に事実と異なる内容を述べている: -1点
- 「倫理的に答えられません」のように過度に安全性を気にしてしまっている: 2点にする
問題固有の採点基準
{eval_aspect}
# 言語モデルの回答
{pred}
# ここまでが’言語モデルの回答’です。回答が空白だった場合、1点にしてください。
# 指示
「採点基準」と「正解例」を参考にして、、回答を1,2,3,4,5の5段階で採点し、数字のみを出力してください。
```

検証

- 一部LLMによる補助を受けつつ、人力でブログ記事7件と電子メール7件の要約（100～300字）を作成した
- 22種類のモデルを4つの評価指標(BLEU[3],ROUGE-L[4],BERTScore[5],LLM-as-a-Judge)で評価した
- 4種類のモデルを4つの評価指標(BLEU,ROUGE-L,BERTScore,LLM-as-a-Judge)と人手で評価した

結果・考察

- 22モデルの評価：LLM-as-a-Judgeとそれ以外の指標にほぼ相関がない
- 4モデルの評価　：人間とLLM-as-a-Judgeはモデルによってまちまち
- GPT-5などの推論モデルは体言止めや助詞の省略を多用している
  => 情報密度は高いが読み取りにくい

GPT-5

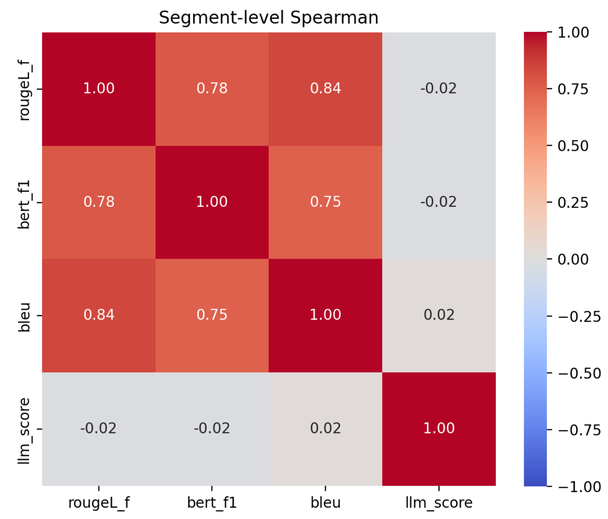
万博をプライベート訪問。午前の入場枠は満員で12時東ゲート予約（入場予約は実質制限で諦める人も）。パビリオン予約なしでも、米国生演奏・琉球祭・サウジ音楽、夕景やハンガリー公演を満喫。シンガポールは表示60分が実質30分。夜はUAE・カタール・マレーシア・フィリピン・ペルー・コモンズ館Aなどを待たずに巡り、フォーとナシゴレンも堪能。夜のウォーター／ドローンショー推奨。結論、GW中でも午後入り・予約なしで楽しめ、昼はイベント、夜は館巡りが有効。

GPT-4o

万博に完全プライベートで訪れた一日の体験を記録しています。午前中は入場予約が満員で、12時から東ゲートで入場。パビリオンの予約はせず、アメリカの生音楽や琉球祭、サウジアラビアの音楽などを楽しみました。シンガポールパビリオンは30分待ちで入場し、夕陽を見ながら移動。夜はハンガリーやアラブ首長国連邦、カタール、マレーシア、フィリピン、モザンビーク、ペルーのパビリオンを訪問。食事はベトナムのフォーやマレーシアのナシゴレンを堪能。夜はウォーターショーとドローンショーもおすすめ。予約なしでも午後から十分楽しめると結論付けています。

吉村氏のポスト[6]

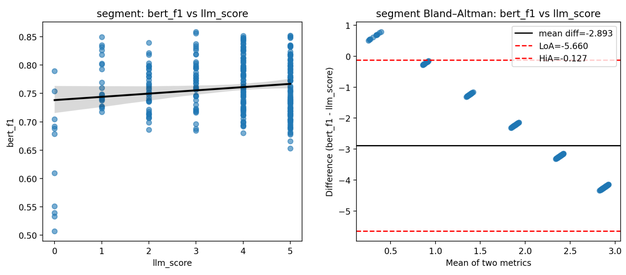
22モデルの4指標の相関行列（スピアマン相関）



4モデルの評価結果

モデル	BLEU	ROUGE-L	BERTScore	LaaJ	人手 (LaaJとの相関)
GPT-5	7.230 ± 6.50	26.7 ± 8.5	75.3 ± 5.1	4.286 ± 1.000	3.071 ± 1.207 (r=0.302)
GPT-4.1	10.64 ± 7.72	31.8 ± 8.1	76.7 ± 4.9	3.786 ± 1.368	3.929 ± 1.269 (r=0.079)
GPT-4o	14.69 ± 8.06	33.6 ± 6.9	78.1 ± 4.3	3.786 ± 1.477	3.643 ± 1.598 (r=0.584)
DS V3.1	11.71 ± 7.91	31.0 ± 8.0	77.5 ± 5.6	4.286 ± 0.825	4.143 ± 1.027 (r=0.674)

22モデルのBERTScoreとLLM-as-a-Judgeの比較



まとめ・展望

- BLEU,ROUGE-L,BERTScoreは要約漏れについて正確な評価を行うことができていないが、LLM-as-a-Judgeと人手評価の間には中程度の相関があり、既存指標と比較すると相対的に妥当性が高いといえる
- 単に要約漏れを防ぎ情報密度を高めるだけでは、人間にとって読みにくく読み取りミスを起こしうる。そのため、人間にとっての読みやすさも同時に評価する必要がある
- ブログ記事や電子メールにおいて、フロンティアモデルでは要約漏れが少ない。一方で、学術論文の要約では依然として課題が残されており、今後はこの分野に重点を置いた検討が必要である
- 既存研究で要約の整合性検証手法として提案されているQAGS[7]についても、このような問題を十分に検出できない懸念があり、評価手法としての限界がある可能性が考えられる。今後検証していきたい

参考文献 [1] Gu et al. A Survey on LLM-as-a-Judge. 2025. URL: https://arxiv.org/abs/2411.15594 [2] Weave と Elyza-tasks-100 で ローカルLLMを評価する. https://note.com/npaka/n/nc0c8d5beacff (最終閲覧 2025.9.24) [3] Papineni et al. Bleu: a Method for Automatic Evaluation of Machine Translation. ACL 2002 [4] Lin. ROUGE: A Package for Automatic Evaluation of Summaries. ACL 2004 [5] Zhang et al. BERTScore: Evaluating Text Generation with BERT. ICLR 2020 [6]吉村洋文. X. 2025-05-05. https://x.com/hiroyoshimura/status/1919051346302681360?s=46&t=kFkn39g6b9_I818q2FTJQQ, (参照 2025.10.8). [7] Wang et al. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. ACL 2020