# 評価理由の生成に基づく大規模言語モデルによる自動評価への取り組み

田中 杏　(指導教員：小林 一郎)

## 1　はじめに

近年，大規模言語モデルが生成した文章の評価の際，BLEU や ROUGE などの人間が作った正解文との一致度により評価する従来の評価指標の代わりに大規模言語モデルに文章を評価させる自動評価の研究が進んでいる．大規模言語モデルによる自動評価においては，人手の評価というコストはかからないが従来の指標と比べて人に類似した評価が可能であるという報告がされている [1]．しかし，大規模言語モデルが観点ごとに評価を行う際に他の観点についても誤って考慮してしまう問題があり，これが評価精度に悪影響を及ぼしている可能性が指摘されている [2]．本研究では，評価理由に基づいて評価を行うように学習させたモデルに評価をさせる．このことを通して，観点ごとの評価理由に基づいた評価を行うモデルの有効性の検討を行う．

## 2　研究概要

GPT-4o mini もしくは GPT-4o に事前に生成させた評価理由を評価指示の後につなげたプロンプトが与えられたときに回答を生成するモデルとして，LongT5-Local-Large[1] を学習させる．また，評価指示が与えられたときに評価理由自体を生成できるように同じモデルを学習させる．

### 2.1　データ作成

モデルの学習に必要である，質問，評価理由，回答が組となったデータセットを作成した．G-EVAL [1] で使用されているものに従い，テキスト要約タスクのデータとして SummEval [3] と対話生成タスクのデータとして USR [4] で提供されている Topical-Chat [5] を使用する．これらはタスクに対して大規模言語モデルが生成した文章とその文章に対する評価観点ごとの人手のスコアが含まれるデータであり，評価観点として，SummEval においてはそれぞれ 1 から 5 の整数の評価スコアが割り当てられる Coherence, Consistency, Fluency, Relevance，Topical-Chat においては 1 から 3 の整数のスコアが割り当てられる Naturalness, Coherence, Engagingness, 0 または 1 のスコアが割り当てられる Groundedness が存在する．

**質問部分**　質問部分はタスク説明，評価対象となる生成された文章などから構成される．指示として，SummEval，Topical-Chat で使用されている評価観点に GPT-4o に生成させたものを追加したリストから適切な評価観点を選びその観点に対してスコアを割り当てさせる．

**回答部分**　人手の評価スコアの平均値を四捨五入して整数にしたものを正解として回答に使用する．なお，評価スコアの取りうる範囲について，評価観点ごとに違いが生じないように全て 1 から 5 の整数となるように変換してから平均をとった．

[1] https://huggingface.co/google/long-t5-local-large

**評価理由部分**　評価理由部分については，Few-Shot の形式のプロンプトを GPT-4o mini もしくは GPT-4o に与えタスクごとに分けて作成する．Few-Shot のデータとしては GPT-4o に評価観点の選び方，スコアの割り当て方の理由を説明することなどを指示したプロンプトを与えて得られた回答をもとに作成したものをテキスト要約と対話生成タスクそれぞれについて 5 つ使用する．Few-Shot のデータは，出力形式を安定させる目的のみで作成したものと，プロンプトに正答を導くような理由を生成させる指示と評価観点リストに含まれる評価観点の説明を追加することで正答を導きやすくかつ観点の意味を考慮した理由になるように改善したものの 2 種類を採用する．GPT-4o mini や GPT-4o にプロンプトを与えて理由生成させる際の temperature は 1.0，top-p は 0.5，max-tokens は 200 に設定し，1 つの質問回答に対し 20 個の評価理由を生成させる．

### 2.2　学習

評価理由生成と推論の両方を行えるモデルに LongT5-Local-Large を調整するための学習を行う．損失関数として，式 (1)，(2) を用いる．なお，質問を $q$，評価理由を $r$，正答を $a^*$ と表す．

$$L_{QR}(\theta) = -\log p_{QR}(r|q;\theta) \tag{1}$$

$$L_{QRA}(\theta) = -\log p_{QRA}(a^*|q,r;\theta) \tag{2}$$

評価理由生成用の学習と推論用の学習を 500 ステップずつ交互に行い，合計 50,000 ステップの学習を行う．

## 3　実験

### 3.1　実験設定

データセットとして，2.1 節で紹介したデータを使用する．テキスト要約データは，学習データ数 1,196，検証データ数 239，対話生成データは，学習データ数 266，検証データ数 53 である．また，2.1 節で述べたように 1 つの質問回答に対してそれぞれ 20 個作成した評価理由を使用するため，学習に用いるデータ数は合計で 29,240 となる．

2.2 節で説明した方法で学習を終えたモデルに対して，2 種類のデータセットの検証データの質問部分に事前に生成した評価理由部分を結合したものを入力として与え，評価結果を出力として得る．なお，処理効率を考慮して，入力として与える評価理由は複数生成したもののうち最初の 1 つ目を使用する．

提案手法を用い，事前に生成する評価理由が以下の 3 パターンに対して評価を行う．GPT-4o mini に対して 1 種類目の Few-Shot を与えて生成された評価理由を用いる **fsv1-mini**，GPT-4o mini に対して 2 種類目の Few-Shot を与えて生成された評価理由を用いる **fsv2-mini**，GPT-4o に 2 種類目の Few-Shot を与えて生成された評価理由を用いる **fsv2-4o** とする．

### 3.2　比較手法

**評価理由なしファインチューニング**　評価理由を使用せず以下の損失関数を採用することで質問から直接正

表 1: テキスト要約データでの実験結果 ($\gamma$: ピアソンの積率相関係数，$\rho$: スピアマンの順位相関係数，$\tau$: ケンドールの順位相関係数)

| 手法 | Coherence | | | Consistency | | | Fluency | | | Relevance | | | AVG | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\gamma$ | $\rho$ | $\tau$ | $\gamma$ | $\rho$ | $\tau$ | $\gamma$ | $\rho$ | $\tau$ | $\gamma$ | $\rho$ | $\tau$ | $\gamma$ | $\rho$ | $\tau$ |
| G-EVAL | 0.035 | 0.487 | 0.378 | 0.029 | 0.483 | 0.407 | 0.456 | **0.489** | 0.403 | 0.153 | 0.506 | 0.401 | 0.168 | 0.491 | 0.397 |
| directQA | 0.619 | 0.599 | 0.531 | 0.458 | 0.374 | 0.363 | **0.575** | 0.482 | **0.466** | 0.493 | 0.449 | 0.408 | 0.536 | 0.476 | 0.442 |
| fsv1-mini | 0.516 | 0.490 | 0.435 | 0.511 | 0.444 | 0.431 | 0.314 | 0.249 | 0.243 | 0.576 | 0.523 | 0.483 | 0.479 | 0.426 | 0.398 |
| fsv2-mini | **0.653** | **0.636** | **0.560** | 0.517 | 0.494 | 0.477 | 0.410 | 0.276 | 0.265 | 0.631 | 0.592 | 0.545 | 0.553 | 0.500 | 0.462 |
| fsv2-4o | 0.635 | 0.614 | 0.551 | **0.647** | **0.575** | **0.555** | 0.484 | 0.413 | 0.398 | **0.661** | **0.631** | **0.579** | **0.606** | **0.558** | **0.521** |

答を導くように学習させたモデルに対して，提案手法と同じ検証データを入力として評価結果を得る．この手法を **directQA** と呼ぶことにする．

$$L_{QA}(\theta) = -\log p_{QA}(a^*|q) \tag{3}$$

**G-EVAL**　G-EVAL [1] に対して，提案手法と同じ検証データを入力として評価結果を得る．G-EVAL で使用するモデルとして GPT-4 を採用し，プロンプトについては GitHub[2]に載っているものを参考に作成した．

### 3.3　評価指標

**相関係数**　ピアソンの積率相関係数，スピアマンの順位相関係数，ケンドールの順位相関係数を，予測評価結果と人手評価結果との観点ごとの関係を見るために算出する．観点ごとの相関係数の他に，それぞれのタスクの 4 つの評価観点における相関係数の平均をとったものを AVG として算出する．

### 3.4　実験結果

表 1 にテキスト要約データでの実験結果，付録 B に対話生成データでの実験結果を示す．

テキスト要約タスクで生成される文章の評価については，Fluency を除いて提案手法である **fsv2-mini**, **fsv2-4o** のいずれかが最も良い結果となった．提案手法の 3 パターンそれぞれについて見ると，**fsv2-mini**, **fsv2-4o** の 2 つは比較手法である G-EVAL, **directQA** と比べて Fluency を除いた全ての評価指標においてより良い結果を記録した．形式を揃える目的のみで評価理由生成の際の Few-Shot データを作成した **fsv1-mini** においても，Relevance については比較手法より高い精度を記録した．提案手法 3 パターンの精度を比較すると，Coherence において **fsv2-mini** が **fsv2-4o** より高い精度を記録している点を除くと，**fsv2-4o**, **fsv2-mini**, **fsv1-mini** の順に精度が高くなった．

対話生成タスクで生成される文章の評価については，Coherence，Engagingness の 2 つの観点での結果を除いて提案手法の **fsv2-4o** の結果が最も高くなった．

### 3.5　考察

テキスト要約タスクで生成される文章の評価においては Fluency を除いた指標で，対話生成タスクで生成される文章の評価においても Coherence，Engagingness の 2 つの観点での結果を除いて提案手法の結果が最も良くなったことから，評価理由に基づいて評価を行うように学習させたモデルの有効性が示唆される．

形式を揃える目的のみで評価理由生成の際の Few-Shot データを作成した **fsv1-mini**，正答を導くような理由を生成させる指示と評価観点の説明を追加することで Few-Shot データを改善した **fsv2-4o**，**fsv2-mini** に対して，精度の良い順を見ると，理由生成の際の Few-Shot データの改善や評価理由を生成するモデルの性能を高くすることがより人と近い評価の実現につながるという示唆が得られる．

Fluency での精度が高くならない結果については，Fluency とスコアが 5 に偏っている分布になっている点で似ている Consistency では精度が良い結果になっていることから，正答のスコア分布のパターンによるものとは考えにくく，Fluency の観点の評価理由説明に改善の余地があるのではないかと考えられる．

## 4　まとめ

本研究では，人と近い評価を出せる自動評価手法の構築を目的として，評価理由の生成と生成した評価理由を踏まえた推論を行うようにモデルの学習を行った．GPT-4o mini や GPT-4o を用いて事前に生成した評価理由を学習させたモデルに与えると，人と類似の評価が可能になり，理由生成の際の Few-Shot データの改善や評価理由を生成するモデルの性能を高くすることがより人と近い評価の実現につながるという示唆が得られた．

今後は，評価精度向上に寄与すると期待される更なる評価理由の改良や，データセットを増やして学習することなどに挑戦したい．

## 参考文献


[1] Yang Liu, et al. G-eval: NLG evaluation using gpt-4 with better human alignment. pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[2] Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. Are LLM-based evaluators confusing NLG quality criteria? pp. 9530–9570, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Alexander R. Fabbri, et al. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 391–409, 2021.

[4] Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. pp. 681–707, Online, July 2020. Association for Computational Linguistics.

[5] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pp. 1891–1895, 2019.


[2]https://github.com/nlpyang/geval

## 付録 A　使用したプロンプト (テキスト要約の場合)

| | プロンプト |
|---|---|
| 評価指示用（質問） | For the following task, if the following output is obtained,<br>choose appropriate evaluation aspects from the following aspects list<br>to assess the quality of the output<br>and assign scores(1-5) to those aspects as well.<br>aspects list: [”coherence”, ”consistency”, ”fluency”, ”relevance”,<br>”naturalness”, ”engagingness”, ”groundedness”,<br>”clarity”, ”creativity”, ”empathy”, ”adaptability”, ”depth”, ”accuracy”,<br>”inclusivity”, ”persuasiveness”, ”formatting”, ”cultural sensitivity”,<br>”humor or emotional appeal”, ”interactivity”, ”robustness”]<br>task:<br>You will be given one summary<br>written for a news article(the following Source Text).<br>Your task is to rate the summary on appropriate metrics.<br>Source Text: { 要約原文 }<br>output: { 評価対象であるモデルの出力 } |
| 評価理由生成用 | Please explain the reasons why you choose the aspects from the aspects list<br>and assign each scores to evaluate output described in the input.<br>Ensure that your response is coherent and does not end abruptly,<br>and output complete sentences within 150 tokens.<br>(input, knowledge が 5 組 (Few-Shot))<br>input: { 評価指示 (質問)}<br>knowledge: |

## 付録 B　対話生成データでの実験結果

| 手法 | Naturalness | | | Coherence | | | Engagingness | | | Groundedness | | | AVG | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\gamma$ | $\rho$ | $\tau$ | $\gamma$ | $\rho$ | $\tau$ | $\gamma$ | $\rho$ | $\tau$ | $\gamma$ | $\rho$ | $\tau$ | $\gamma$ | $\rho$ | $\tau$ |
| G-EVAL | 0.495 | 0.481 | 0.376 | 0.490 | 0.534 | 0.417 | 0.166 | 0.652 | 0.510 | 0.544 | 0.551 | 0.469 | 0.424 | 0.555 | 0.443 |
| directQA | 0.482 | 0.530 | 0.468 | **0.619** | **0.557** | **0.493** | **0.654** | **0.662** | **0.584** | 0.577 | 0.552 | 0.508 | 0.583 | 0.575 | 0.513 |
| fsv1-mini | 0.421 | 0.393 | 0.347 | 0.369 | 0.386 | 0.337 | 0.566 | 0.534 | 0.454 | 0.355 | 0.388 | 0.366 | 0.428 | 0.425 | 0.376 |
| fsv2-mini | 0.348 | 0.362 | 0.314 | 0.369 | 0.423 | 0.363 | 0.613 | 0.607 | 0.523 | 0.781 | 0.750 | 0.673 | 0.528 | 0.536 | 0.468 |
| fsv2-4o | **0.576** | **0.594** | **0.529** | 0.454 | 0.509 | 0.442 | 0.619 | 0.617 | 0.534 | **0.844** | **0.815** | **0.746** | **0.623** | **0.634** | **0.563** |

($\gamma$: ピアソンの積率相関係数，$\rho$: スピアマンの順位相関係数，$\tau$: ケンドールの順位相関係数)

## 付録 C　生成された評価理由の例

### C.1　1 種類目の Few-Shot を使用した場合

| | |
|---|---|
| 1 種類目の Few-Shot に対し GPT-4o mini で生成された評価理由の例 (fsv1-mini) | To evaluate the summary, I selected the following aspects:<br>1. **Relevance (Score: 5)**:<br>The summary effectively captures the main event of the guard slipping during the Changing of the Guard, which is the core focus of the source text.<br>2. **Clarity (Score: 4)**:<br>The summary is mostly clear, but could benefit from slightly more straightforward language to enhance understanding.<br>3. **Consistency (Score: 4)**:<br>The summary is consistent with the source, accurately reflecting the incident and the guard's reaction, though it could include more details about the context.<br>4. **Fluency (Score: 3)**:<br>While the summary flows reasonably well, some phrasing could be improved for better readability.<br>Aspects like "creativity" and "humor" are not applicable, as the focus is on factual reporting.<br>Overall, the summary effectively conveys the incident while leaving room for minor improvements. |

### C.2　2 種類目の Few-Shot を使用した場合

| | |
|---|---|
| 2 種類目の Few-Shot に対し GPT-4o mini で 生成された 評価理由の例 (fsv2-mini) | - **Coherence (4)**:<br>The summary effectively captures the main event<br>but could improve the flow by better connecting the sentences.<br>The narrative is mostly logical, but transitions could enhance clarity.<br>- **Consistency (5)**:<br>The summary accurately reflects the facts from the source text without introducing<br>any discrepancies, maintaining alignment with the original information.<br>- **Fluency (5)**:<br>The text is grammatically correct and well-structured,<br>allowing for smooth reading and comprehension.<br>- **Relevance (4)**:<br>It includes essential details about the incident and the guard's reaction<br>but could benefit from additional context, such as the significance<br>of the Changing of the Guard.<br>Aspects like creativity and humor were not selected<br>as they do not pertain to the factual nature of the summary.<br>Overall, the summary effectively conveys the incident<br>while leaving room for slight improvements in coherence and relevance. |
| 2 種類目の Few-Shot に対し GPT-4o で 生成された 評価理由の例 (fsv2-4o) | - **Coherence (4)**:<br>The summary provides a logical sequence of events, capturing the main incident<br>of the guard slipping and the subsequent reactions. However, it could benefit from<br>a more structured narrative that ties the details together more seamlessly.<br>- **Consistency (5)**:<br>The summary accurately reflects the key facts from the source text,<br>including the incident and the guard's reaction, ensuring full factual alignment<br>without introducing discrepancies.<br>- **Fluency (5)**:<br>The text is grammatically correct and easy to read, with well-constructed sentences<br>that facilitate smooth comprehension.<br>- **Relevance (4)**:<br>The summary includes essential details about the incident and the guard's reaction<br>but omits some context, such as the setting and the role of the tourists,<br>which could provide a fuller picture.<br>Aspects like creativity, empathy, and humor were not selected<br>as they are not relevant to the task of summarizing factual content.<br>Depth and inclusivity were also not chosen, as they pertain more to |