推論過程の忠実性を担保した 大規模言語モデルによる自動評価手法開発への取り組み

田中杏 [1]　小林一郎 [1]
[1] お茶の水女子大学
{g2120526,koba}@is.ocha.ac.jp

概要

近年，大規模言語モデル（LLM）を用いた生成文の自動評価は，人手コスト削減と従来の統計的な指標よりも内容に即した評価が可能な点で注目されている．本研究は，評価理由と結果の整合性（推論忠実性）を崩さずに評価性能を確保する自動評価手法構築を目的に，二段階学習を提案する．学習段階 I で理由生成と理由に基づく推論形式を学習し，学習段階 II で人評価に近い推論過程を選抜して理由生成を重点的に学習する．実験より，推論忠実性向上の可能性を示す一方，評価精度との両立が課題であることが分かった．

1　はじめに

生成 AI の普及に伴い生成文が増加・多様化する中，生成文評価の重要性が高まっている．従来は BLEU [1] や ROUGE [2] など参照文との一致度に基づく指標が用いられてきたが，多様な内容を十分に捉えられない場合がある．そこで LLM に評価を行わせる自動評価が注目されている [3, 4, 5]．

LLM による自動評価は，人手コストを要さず，従来指標より人手評価に近いと報告されている [3]．

一方で Hu ら [6] は，LLM が観点別評価において他観点を誤って考慮し，これが評価精度に悪影響を及ぼしている可能性を指摘している．この問題を解決し，人間が納得できる信頼性を確保して評価精度を改善するためには，評価理由と最終評価結果が一貫していることが前提となる．そのため，評価理由に基づく一貫した評価，すなわち推論過程と評価結果の忠実性の担保は重要課題である．

また，生成文の種類が多様化する中で，適切な観点選択と観点ごとの理由・採点の整合的生成を学習することは，特定タスクに特化した評価器ではなく，未知タスクへの汎用化を見据えた基盤能力として重要である．

そこで本研究では，観点選択を含む文章評価において，評価理由と結果の忠実性を重視した推論過程を学習する枠組みを提案する．要約および対話生成の評価データで実験し，推論忠実性を維持しつつ評価精度を確保するための設計上の知見を示す．

2　提案手法

2.1　研究概要

図 1 に提案手法を示す．学習段階 I では，生成文を評価するように指示する入力を受け取ったとき，文章の評価理由を生成し，その理由を踏まえた推論により，最終的に評価観点とスコアの組である出力を生成する形式を学習する．学習段階 II では，人評価に近い結果を出す理由を選抜し，その理由生成を重点的に学習することで推論性能の向上を図る．

2.2　データ作成

モデル学習に必要な，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 [3] に従い，テキスト要約タスクには SummEval [7]，対話生成タスクには USR [8] で提供されている Topical-Chat [9] のデータを用いる．いずれのデータセットも，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によるスコアを含む．SummEval では，Coherence, Consistency, Fluency, Relevance の 4 観点について 1 から 5 の整数スコアが付与され，Topical-Chat では，Naturalness, Coherence, Engagingness の 3 観点に 1 から 3 の整数スコア，Groundedness に 0 または 1 のスコアが付与されている．

質問部分：　質問部分は G-EVAL [3] のプロンプトを参考にタスク説明，評価対象となる生成文章などから構成する．指示として，SummEval [7],

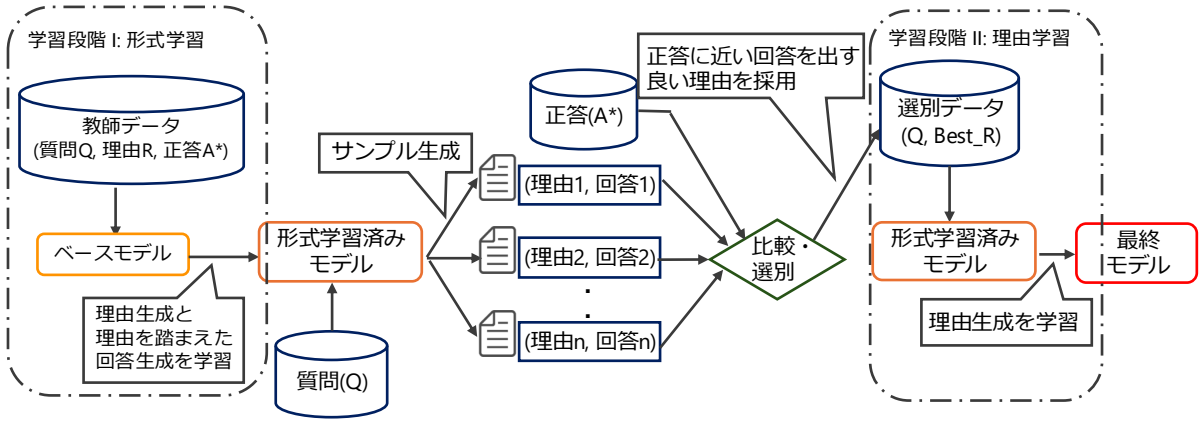

図 1　選抜した推論過程の学習

Topical-Chat [9] で使用されている評価観点に GPT-4o に生成させた観点を加えたリストから適切な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使用する．評価観点ごとの差をなくすため，すべてのスコアを 1 から 5 の整数に変換した上で平均を算出する．

評価理由部分：　Few-Shot の形式のプロンプトを GPT-4o に与えタスク別に作成する．Few-Shot データとしては評価観点の選択方法やスコア付与理由の説明を指示したプロンプトに対する GPT-4o の回答を用い，テキスト要約と対話生成タスクそれぞれについて 5 つずつ使用する．Few-Shot のデータは，正答を導く理由を生成させる指示と評価観点リスト内の観点説明の追加により正答を導きやすく観点の意味を考慮した理由になるようにしたものである．理由生成時の 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．

2.3　学習段階 I

準備段階として，評価理由生成と評価理由を踏まえた推論の両方を行えるようにモデルに形式を学習させる．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．

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

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

2.4　学習段階 II

学習段階 I 終了時点でのモデルを用いて，2step で質問から理由と最終回答を複数 (今回は 8) サンプル生成させる．その後，最終回答を人の評価と比較して以下の式で定義する Alignment and Coverage Score (ACS) スコアを算出し，各質問ごとに top-2 あるいは閾値 0.8 以上のものを学習データとして採用する．出力した予測評価を $pred$，正答を $truth$ とし，それぞれの観点を $pred\_keys$，$truth\_keys$，

$$cmn\_keys = truth\_keys \cap pred\_keys \tag{3}$$

$$\begin{aligned} penalty = &|pred\_keys \setminus truth\_keys| \\ &+ |truth\_keys \setminus pred\_keys| \end{aligned} \tag{4}$$

とすると，

$$ACS = \max\left(0, \frac{\sum_{k \in cmn\_keys}\left(1 - \frac{|pred[k]-truth[k]|}{4}\right)}{|cmn\_keys|} - 0.2 \times penalty\right). \tag{5}$$

採用したデータを用いて学習させる際には，損失関数として，式 (1) を採用する．ここで，質問から理由生成のみを学習させ回答生成を学習させないのは，評価結果の直接最適化による推論忠実性の低下を防ぎ，理由生成能力を高めるためである．

3　実験

3.1　実験設定

データセットには 2.2 節で述べたものを用いる．テキスト要約では学習 1,196 件・検証 239 件，対話生成では学習 266 件・検証 53 件を使用する．各質問回答に対し 20 件の評価理由を生成するため，学習段階 I で用いる学習データ数は計 29,240 件となる．学習段階 II ではデータ選抜後，top-2 で 2,864 件，閾値方式で 8,062 件を用いる．

基盤モデルには Llama-3.1-8B-Instruct[1] を使用す

1)　https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

る．学習段階 I では評価理由生成と推論学習を各 50 ステップ交互に計 200 ステップ行う．学習段階 II では 8,000 ステップ学習し 100 ステップごとに評価を行い，結果は ACS が最大となるステップを報告する．評価時の生成は greedy 設定とした．

3.2 比較手法

以下の手法について，同一の検証データを用いて評価を行う．

- **directQA**: 評価理由を用いず，質問から直接正答を導くように学習させたモデル
- **G-EVAL**: GPT-4 を用いた G-EVAL [3][2)]
- **StageI-only**: 学習段階 I のみを適用したモデル
- **Long-StageI**: 学習段階 I を 50,000 ステップ適用[3)]したモデル
- **GPT-4.1**: GPT-4.1 に質問部分のみを入力した結果

directQA では以下の損失関数を用いる．

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

3.3 評価指標

本研究では，評価精度および推論忠実性の観点から，以下の指標を用いて評価を行う．

3.3.1 評価精度

人手評価との一致度を測る指標として，相関係数，ACS，および完全一致率（EMR）を用いる．

相関係数 予測結果と人手評価結果との観点ごとの関係を見るため，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算出する．提案手法では評価観点の選択も行うため，人手評価と対応する観点が存在する場合のみ算出対象とする．

独自指標 ACS 評価観点ごとのスコア誤差に加え，観点の過不足も考慮する指標として，式 (5) で定義される独自指標 ACS を用いる．ACS は 0 以上 1 以下の小数値をとり，1 に近いほど正解に近いことを示す．

完全一致率（EMR） 予測評価結果が，人手評価結果と完全に一致した割合 Exact Match Rate（EMR）を算出する．

3.3.2 推論忠実性

評価理由と最終評価結果との整合性を測る指標として，RAEMR および RAS を用いる．

RAEMR Rationale-Answer Exact Match Rate（RAEMR）を独自に定義して用いる．評価理由内に記述された各評価観点のスコアと，最終評価スコアの完全一致率として測定する．A を評価理由内のスコア辞書，B を最終評価のスコア辞書とし，Keys$(\cdot)$ を辞書に含まれる評価観点（キー）の集合とする．$C = \text{Keys}(A) \cap \text{Keys}(B)$ とすると，RAEMR は以下の式で定義される．

$$\text{RAEMR}(A, B) = \frac{1}{|C|} \sum_{k \in C} \mathbf{1}[A_k = B_k],$$

ただし

$$\mathbf{1}[A_k = B_k] = \begin{cases} 1 & \text{if } A_k = B_k \\ 0 & \text{otherwise} \end{cases}$$

$|C| = 0$ の場合は算出対象外とする．

RAS RAEMR はスコアの完全一致のみに基づく指標であるため，スコア差の大きさを考慮した評価として，本研究では独自に Rationale-Answer Similarity（RAS）を定義する．

$$\text{RAS}(A, B) = \frac{1}{|C|} \sum_{k \in C} \left(1 - \frac{|A_k - B_k|}{4}\right),$$

ただし $|C| = 0$ の場合は算出対象外とする．

3.4 結果

表 1 にテキスト要約データ，表 2 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は，相関係数（γ, ρ, τ）は Fluency を除く全観点で directQA が最も高い値を示した．一方，評価精度指標である ACS および EMR は，提案手法の top-2 バージョンが最良であった．推論忠実性指標（RAEMR, RAS）では StageI-only が最も高く，提案手法も近い水準を維持したが，Long-StageI は他手法と比べて低下した．

対話生成タスクでは，相関係数は Coherence を除く全観点で GPT-4.1 が最も高かった．一方，ACS および EMR は Long-StageI が最良であり，提案手法も ACS が 0.5 以上となるなど一定の評価精度を示した．特に提案手法は，推論忠実性を StageI-only に近い水準で維持しつつ，評価精度では StageI-only を上

2) プロンプトは GitHub [https://github.com/nlpyang/geval] に載っているものを参考に作成

3) 既存設定に従い，評価理由生成用の学習と推論用の学習は 500 ステップずつ交互に適用．本研究の目的はステップ数を揃えた性能比較ではなく，学習目標の違いによる忠実性の変化の分析である．

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

手法	Coherence			Consistency			Fluency			Relevance			ACS	EMR	RAEMR	RAS
	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ				
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	0.821	—	—	—
directQA	**0.627**	**0.646**	**0.561**	**0.677**	**0.53**	**0.517**	0.497	0.502	0.482	**0.643**	**0.614**	**0.557**	0.748	0.159	—	—
StageI-only	0.414	0.390	0.340	0.598	0.454	0.436	0.410	0.397	0.383	0.499	0.479	0.434	0.882	0.201	**0.927**	**0.981**
Long-StageI	0.610	0.614	0.536	0.546	0.495	0.472	**0.568**	**0.566**	**0.538**	0.459	0.485	0.439	0.885	0.222	0.618	0.893
GPT-4.1	0.601	0.576	0.503	0.605	0.505	0.470	0.445	0.455	0.414	0.598	0.554	0.501	0.343	0.000	—	—
ours(top-2)	0.490	0.488	0.434	0.502	0.410	0.393	0.258	0.210	0.203	0.494	0.465	0.424	**0.886**	**0.247**	0.838	0.957
ours(閾値)	0.405	0.395	0.352	0.501	0.464	0.450	0.309	0.288	0.278	0.373	0.315	0.288	0.879	0.222	0.922	0.980

表 2　対話生成データでの実験結果 (γ: ピアソンの積率相関係数，ρ: スピアマンの順位相関係数，τ: ケンドールの順位相関係数)

手法	Naturalness			Coherence			Engagingness			Groundedness			ACS	EMR	RAEMR	RAS
	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ				
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	0.684	—	—	—
directQA	0.445	0.363	0.312	**0.581**	**0.598**	**0.508**	0.734	0.737	0.643	0.623	0.613	0.574	0.757	0.038	—	—
StageI-only	0.023	-0.022	-0.022	0.350	0.347	0.272	0.610	0.57	0.480	0.697	0.693	0.667	0.141	0.000	**0.860**	**0.965**
Long-StageI	0.496	0.503	0.433	0.498	0.525	0.447	0.661	0.677	0.594	0.738	0.726	0.666	**0.779**	**0.132**	0.388	0.815
GPT-4.1	**0.807**	**0.794**	**0.758**	0.517	0.555	0.484	**0.876**	**0.875**	**0.806**	**0.865**	**0.862**	**0.785**	0.022	0.000	—	—
ours(top-2)	0.379	0.353	0.303	0.439	0.480	0.393	0.608	0.619	0.504	0.536	0.576	0.522	0.547	0.057	0.729	0.931
ours(閾値)	0.331	0.341	0.296	0.516	0.533	0.463	0.549	0.574	0.479	0.181	0.195	0.175	0.544	0.019	0.810	0.951

回る場合が多かったが，相関係数は全体として比較手法より低い傾向が見られた．

3.5　考察

本研究では，評価理由を介さずに質問と評価スコアの対応関係のみを学習する振る舞いを，推論ショートカットと定義する．テキスト要約タスクにおいて directQA が高い相関係数を示したことは，評価理由を明示的に生成せず，質問と人手評価スコアの対応関係を直接学習することで，数値的一致度を高めていると解釈できる．このような手法は相関係数の観点では有効である一方，評価理由と評価結果の整合性は保証されない．

実際に，両タスクにおいて Long-StageI は，相関係数や ACS が一定程度高いにもかかわらず，RAEMR や RAS が大きく低下しており，評価結果を直接最適化するショートカット的学習が強まった可能性が示唆される．これに対し提案手法は，推論忠実性を高く保ちつつ StageI-only を上回る評価精度を示す場合が多く，理由生成のみを重点的に学習する二段階学習により評価理由と結果の整合性を保った推論過程を維持できていると考えられる．

一方，相関係数の観点では比較手法に及ばない場合も多く，推論忠実性を維持しながら評価精度をさらに向上させることは今後の課題である．StageI-only と比べて提案手法で相関係数が低下する場合があるが，これは，学習段階 II において理由生成のみを学習し，さらに ACS に基づくデータ選抜を行うため，観点選択や観点別相関を直接最適化していないことが影響している可能性がある．今後，観点ごとの対象数やスコア分布の分析により要因を切り分ける．

4　おわりに

本研究では，評価理由と評価結果の整合性を維持しつつ評価性能を確保する自動評価手法の構築を目的として，データ選抜を組み込んだ二段階学習手法を提案した．要約および対話生成データでの実験により，提案手法は評価精度指標（ACS，EMR）を一定程度確保しつつ，推論忠実性指標（RAEMR，RAS）を高く維持できることが確認された．一方で，相関係数の観点では既存手法に及ばない場合もあり，推論忠実性と評価精度の両立には依然として課題が残ることが明らかになった．

今後は，推論忠実性を保ちながら相関係数を含む評価精度を向上させるため，データ選抜基準や学習方法の改良を検討する．また，本研究では推論忠実性の指標として観点ごとのスコア整合性に着目したが，評価理由文自体の意味的妥当性や説明性を評価する指標の導入も重要である．さらに，より汎用的な評価モデルの実現に向けて，対象とする生成タスクやデータセットの拡張を進めていきたい．

参考文献

[1] Kishore Papineni, et al. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[2] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[3] Yang Liu, et al. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[4] Peiwen Yuan, et al. BatchEval: Towards human-like text evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15940–15958, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[5] Yuxuan Liu, et al. HD-eval: Aligning large language model evaluators through hierarchical criteria decomposition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7641–7660, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] Xinyu Hu, et al. Are LLM-based evaluators confusing NLG quality criteria? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9530–9570, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

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

[8] Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In Dan Jurafsky, et al., editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 681–707, Online, July 2020. Association for Computational Linguistics.

[9] Karthik Gopalakrishnan, et al. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In **Proc. Interspeech 2019**, pp. 1891–1895, 2019.

A　生成された評価理由の例

表 3　実際に生成・選抜されて学習段階 II に用いた評価理由の例

- **Coherence (4)**:
The summary effectively presents a main storyline about the topic of UFOs in Japan.
However, it could improve by including more context about the political discussion
or the involvement of politician Antonio Inoki to provide a clearer narrative flow.
- **Consistency (5)**:
The summary accurately reflects the key facts from the source text,
such as the game S̈pace Invadersänd the lack of UFO encounters by the Japanese military,
ensuring factual alignment with the original information.
- **Fluency (5)**:
The text is grammatically correct and easy to read, with well-structured sentences
that facilitate smooth reading and understanding.
- **Relevance (4)**:
While it captures the essential information about UFOs in Japan, it omits details about the political context
and the specific involvement of Antonio Inoki, which are important aspects of the full story.
Aspects like creativity, empathy, and humor are not relevant to the task of summarizing factual information.
Depth and inclusivity were not selected as they are more applicable to analysis rather than summarization.
Cultural sensitivity and humor were also not chosen as they are not critical to the factual summary provided.

B　データセットごとの正答スコア分布

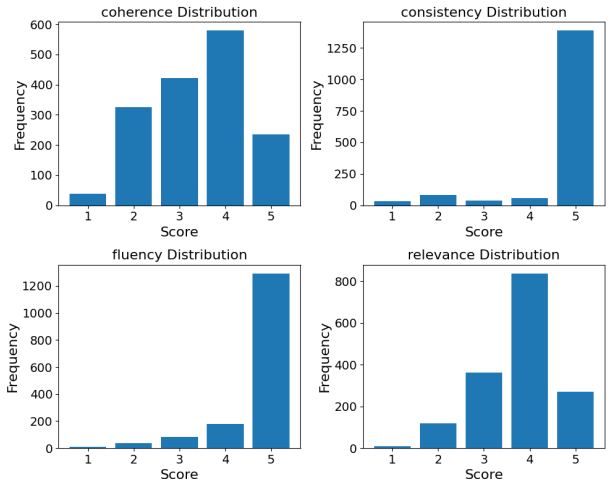

図 2　テキスト要約データの正答スコア分布

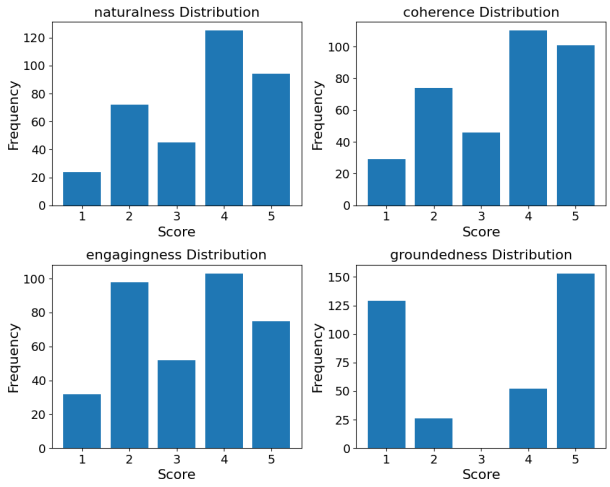

図 3　対話生成データの正答スコア分布

C　データセット

表 4　評価指示プロンプト（テキスト要約の場合）

For the following task, if the following output is obtained, choose appropriate evaluation aspects from the following aspects list to assess the quality of the output and assign scores(1-5) to those aspects as well.
aspects list: ["coherence", "consistency", "fluency", "relevance", "naturalness", "engagingness", "groundedness", "clarity", "creativity", "empathy", "adaptability", "depth", "accuracy", "inclusivity", "persuasiveness", "formatting", "cultural sensitivity", "humor or emotional appeal", "interactivity", "robustness"]
task:
　　You will be given one summary written for a news article(the following Source Text).
　　Your task is to rate the summary on appropriate metrics.
　　Source Text: {要約原文}
　　output: {評価対象であるモデルの出力}