G-E VAL ： G PT -4を用いたNLG評価とより良い人間アライメント

楊 劉 伊 伊 陽 徐 周 周 周 周 忠

マイクロソフト認知サービス研究{yaliu10、イテダン、イククス、スフオウ、ルオックス、チェズフ}@microsoft.com

Abstract

自然言語生成(NLG)システムによって生成されたテキストの品質を自動的に測定することは困難である。BLEU や ROUGE などの従来の参照ベースの指標は、特に創造性と多様性を必要とするタスクにおいて、人間の判断との相関が比較的低いことが示されている。最近の研究では、大規模言語モデル(LLM)をNLG評価のための参照不要なメトリクスとして使用することが提案されており、これは人間の参照を欠いた新しいタスクに適用できるという利点がある。しかし、これらのLLMベースの評価者は、中規模のニューラル評価者よりもまだ人間の対応関係が低い。本研究では、大規模言語モデルと思考連鎖(CoT)、およびフォームフィリングパラダイムを用いて、NLG出力の品質を評価するフレームワークであるG-E VALを紹介する。テキスト要約と対話生成の2つの生成タスクで実験を行う。GPT-4をバックボーンモデルとしたG-E VALは、要約タスクにおいて、人間と0.514のスピアマン相関を達成し、全ての先行手法を大きく上回る性能を示すことを示す。また、LLMベースの評価者の動作に関する分析を提案し、LLMが生成するテキストに偏りを持つLLMベースの評価者の潜在的な懸念を強調する。

1 Introduction

自然言語生成システムの品質を評価することは、大規模言語モデルが高品質で多様なテキストを生成でき、人間が書いたテキストと区別がつかないことが多い場合でも、困難な問題である(Ouyang et al.、2022)。BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) などの従来の自動評価指標は、NLG評価に広く用いられているが、以下のことが示されている。は、特にオープンエンドの生成タスクにおいて、人間の判断との相関が比較的低いことがわかる。さらに、これらのメトリクスは関連する参照出力を必要とし、これは新しいタスクのために収集するためにコストがかかる。最近の研究では、LLMを参照不要のNLG評価器として直接使用することが提案されている(Fu et al., 2023; Wang et al., 2023)。LLMは高品質で流暢なテキストに高い確率を割り当てることを学習したと仮定し、参照対象がない場合の生成確率に基づいて候補出力をスコアリングするアイデアである。しかし、LLMをNLG評価器として使用することの妥当性・信頼性については、体系的に検討されていない。また、メタ評価では、これらのLLMベースの評価者は、中規模のニューラル評価者よりも依然として人間の対応関係が低いことが示されている(Zhong et al.、2022)。したがって、LLMをNLG評価に利用するための、より効果的で信頼性の高いフレームワークが必要である。本論文では、LLM with chain-of-thoughts (CoT) (Wei et al., 2022) を用いて生成テキストの品質を形埋めパラダイムで評価するフレームワーク、G-E VAL を提案する。タスクの紹介と評価基準をプロンプトとしてのみ与えることで、LLMに詳細な評価ステップのCoTを生成するよう依頼する。次に、生成されたCoTとともにプロンプトを使用して、NLGの出力を評価する。評価者の出力は、フォームとしてフォーマットされています。さらに、出力された評価トークンの確率を利用して、最終的な指標を洗練させることができる。我々は、2つのNLGタスク(テキスト要約と対話生成)の3つのメタ評価ベンチマークについて広範な実験を行った。この結果から、G-E VAL は既存の NLG 評価器と人間の評価との相関において、大きなマージンをもって上回ることができることがわかる。最後に、LLMベースの評価者の動作に関する分析を行い、LLMベースの評価者がLLM生成テキストに偏りを持つという潜在的な問題を浮き彫りにする。

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

G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment

Yang Liu Dan Iter Yichong Xu
Shuohang Wang Ruochen Xu Chenguang Zhu

Microsoft Cognitive Services Research
{***yaliu10***, *iterdan, yicxu, shuowa, ruox, chezhu*}*@microsoft.com*

Abstract

The quality of texts generated by natural language generation (NLG) systems is hard to measure automatically. Conventional reference-based metrics, such as BLEU and ROUGE, have been shown to have relatively low correlation with human judgments, especially for tasks that require creativity and diversity. Recent studies suggest using large language models (LLMs) as reference-free metrics for NLG evaluation, which have the benefit of being applicable to new tasks that lack human references. However, these LLM-based evaluators still have lower human correspondence than medium-size neural evaluators. In this work, we present G-EVAL, a framework of using large language models with chain-of-thoughts (CoT) and a form-filling paradigm, to assess the quality of NLG outputs. We experiment with two generation tasks, text summarization and dialogue generation. We show that G-EVAL with GPT-4 as the backbone model achieves a Spearman correlation of 0.514 with human on summarization task, outperforming all previous methods by a large margin. We also propose analysis on the behavior of LLM-based evaluators, and highlight the potential concern of LLM-based evaluators having a bias towards the LLM-generated texts. [1]

1 Introduction

Evaluating the quality of natural language generation systems is a challenging problem even when large language models can generate high-quality and diverse texts that are often indistinguishable from human-written texts (Ouyang et al., 2022). Traditional automatic metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), are widely used for NLG evaluation, but they have been shown to have relatively low correlation with human judgments, especially for open-ended generation tasks. Moreover, these metrics require associated reference output, which is costly to collect for new tasks.

Recent studies propose directly using LLMs as reference-free NLG evaluators (Fu et al., 2023; Wang et al., 2023). The idea is to use the LLMs to score the candidate output based on its generation probability without any reference target, under the assumption that the LLMs have learned to assign higher probabilities to high-quality and fluent texts. However, the validity and reliability of using LLMs as NLG evaluators have not been systematically investigated. In addition, meta-evaluations show that these LLM-based evaluators still have lower human correspondence than medium-size neural evaluators (Zhong et al., 2022). Thus, there is a need for a more effective and reliable framework for using LLMs for NLG evaluation.

In this paper, we propose G-EVAL, a framework of using LLMs with chain-of-thoughts (CoT) (Wei et al., 2022) to evaluate the quality of generated texts in a form-filling paradigm. By only feeding the Task Introduction and the Evaluation Criteria as a prompt, we ask LLMs to generate a CoT of detailed Evaluation Steps. Then we use the prompt along with the generated CoT to evaluate the NLG outputs. The evaluator output is formatted as a form. Moreover, the probabilities of the output rating tokens can be used to refine the final metric. We conduct extensive experiments on three meta-evaluation benchmarks of two NLG tasks: text summarization and dialogue generation. The results show that G-EVAL can outperform existing NLG evaluators by a large margin in terms of correlation with human evaluations. Finally, we conduct analysis on the behavior of LLM-based evaluators, and highlight the potential issue of LLM-based evaluator having a bias towards the LLM-generated texts.

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

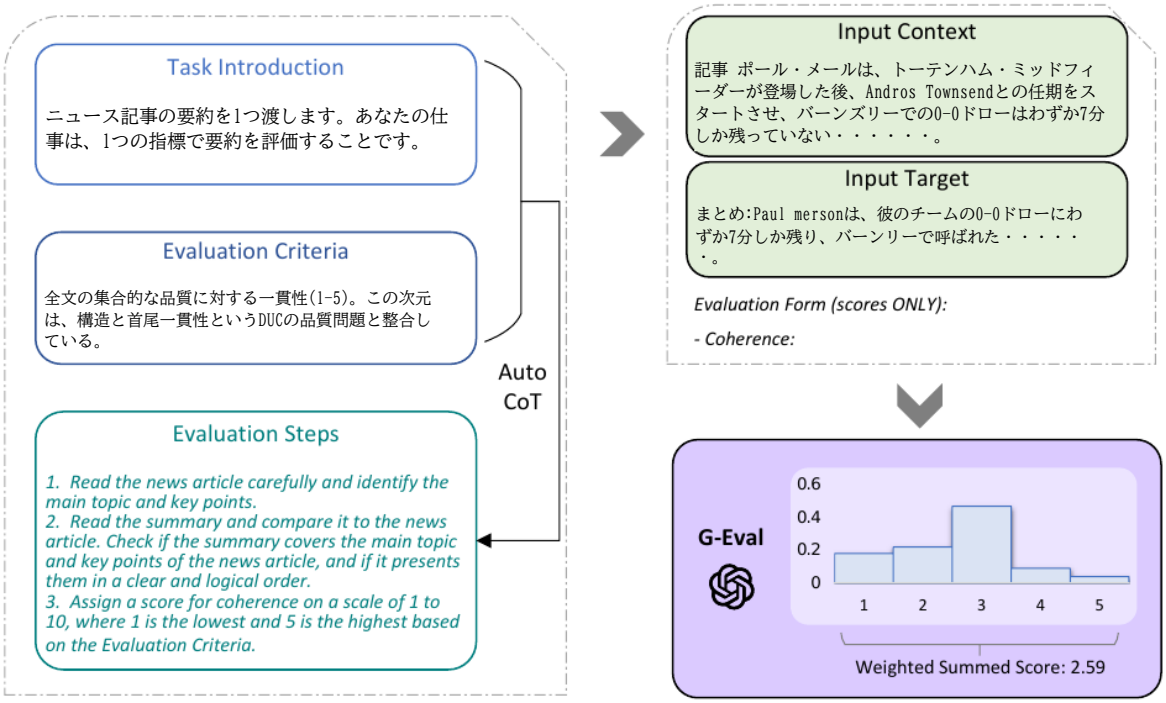

図1:G-E VALの全体的なフレームワーク.まず、タスクの紹介と評価基準をLLMに入力し、詳細な評価ステップのCoTを生成するよう依頼する。次に、生成されたCoTとともにプロンプトを使用し、フォームフィリングパラダイムでNLG出力を評価する。最後に、出力スコアの確率加重和を最終的なスコアとする。

To summarize, our main contributions in this paper are:

1. LLMベースのメトリクスは、特に対話応答生成のようなオープンエンドで創造的なNLGタスクにおいて、人間の品質判断との相関の点で、一般的にリファレンスベースやリファレンスフリーのベースラインメトリクスを凌駕する。

2. LLMベースのメトリクスは指示やプロンプトに敏感であり、思考の連鎖はより多くの文脈とガイダンスを提供することによってLLMベースの評価者のパフォーマンスを向上させることができる。

3. LLMベースのメトリクスは、離散スコアをそれぞれのトークン確率で再重み付けすることで、よりきめ細かい連続スコアを提供することができる。

4. LLMベースのメトリクスは、人間が書いたテキストよりもLLMが生成したテキストを好むという潜在的な問題があり、LLMベースのメトリクスを自己改善のための報酬信号として使用した場合、LLMの自己強化につながる可能性がある。

2 Method

G-E VAL はプロンプトベースの評価器であり、主に 3 つの要素からなる。1）評価タスクの定義と望ましい評価基準を含むプロンプト、2）詳細な評価ステップを記述するLLMによって生成された中間命令のセットである思考の連鎖(CoT)、3）LLMを呼び出し、リターントークンの確率に基づいてスコアを計算するスコアリング関数、である。

自然言語処理による評価のプロンプト 評価タスクと希望する評価基準を定義する自然言語による指示である。例えば、テキスト要約の場合、プロンプトは次のようになる。

> ニュース記事の要約を1つ渡します。あなたの仕事は、1つの指標で要約を評価することです。
>
> これらの説明をよく読んで、理解した上でください。この文書は、レビュー中は自由に保管し、必要に応じて参照してください。

'プロンプトにはカスタマイズされた評価も含まれていなければなりません。

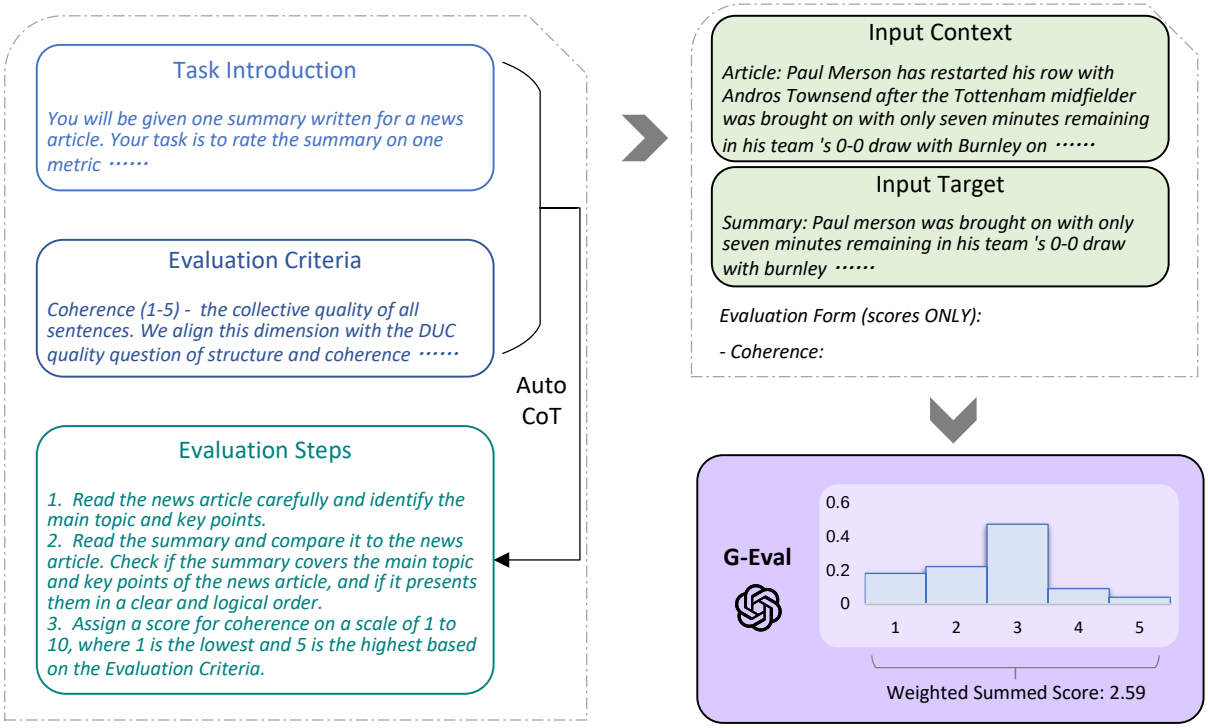

Figure 1: The overall framework of G-EVAL. We first input Task Introduction and Evaluation Criteria to the LLM, and ask it to generate a CoT of detailed Evaluation Steps. Then we use the prompt along with the generated CoT to evaluate the NLG outputs in a form-filling paradigm. Finally, we use the probability-weighted summation of the output scores as the final score.

To summarize, our main contributions in this paper are:

1. LLM-based metrics generally outperform reference-based and reference-free baseline metrics in terms of correlation with human quality judgments, especially for open-ended and creative NLG tasks, such as dialogue response generation.

2. LLM-based metrics are sensitive to the instructions and prompts, and chain-of-thought can improve the performance of LLM-based evaluators by providing more context and guidance.

3. LLM-based metrics can provide a more fine-grained continuous score by re-weighting the discrete scores by their respective token probabilities.

4. LLM-based metrics have a potential issue of preferring LLM-generated texts over human-written texts, which may lead to the self-reinforcement of LLMs if LLM-based metrics are used as the reward signal for improving themselves.

2 Method

G-EVAL is a prompt-based evaluator with three main components: 1) a prompt that contains the definition of the evaluation task and the desired evaluation criteria, 2) a chain-of-thoughts (CoT) that is a set of intermediate instructions generated by the LLM describing the detailed evaluation steps, and 3) a scoring function that calls LLM and calculates the score based on the probabilities of the return tokens.

Prompt for NLG Evaluation The prompt is a natural language instruction that defines the evaluation task and the desired evaluation criteria. For example, for text summarization, the prompt can be:

> *You will be given one summary written for a news article. Your task is to rate the summary on one metric.*
>
> *Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.*

The prompt should also contain customized eval-

また、首尾一貫性、簡潔性、文法など、さまざまなNLGタスクの評価基準も満たしています。例えば、テキスト要約の一貫性を評価するために、プロンプトに以下の内容を追加する。

Evaluation Criteria:

全文の集合的な品質に対する一貫性(1-5)。この次元は、構造と一貫性に関するDUCの品質問題と整合しており、「要約はよく構成され、よく組織されるべきである」と述べている。要約は単に関連情報の山であるだけでなく、文から文、トピックに関する首尾一貫した情報の体系へと構築されるべきである。”

Auto Chain-of-Thoughts for NLG Evaluation CoT(思考の連鎖)は、テキスト生成プロセス中にLLMによって生成される中間表現のシーケンスである。評価タスクの場合、いくつかの基準では単純な定義を超えたより詳細な評価指示が必要であり、各タスクのためにそのような評価ステップを手動で設計するのは時間がかかる。LLMはこのような評価ステップを単独で生成できることがわかった。CoTは、生成されたテキストを評価するためのLLMのためのより多くのコンテキストとガイダンスを提供することができ、また、評価プロセスと結果を説明するのに役立つことができる。例えば、テキスト要約の一貫性を評価するために、プロンプトに「評価ステップ:」の行を追加し、LLMに以下のCoTを自動生成させる。

1. ニュース記事を注意深く読み、主なトピックとキーポイントを特定する。

2. まとめを読み、ニュース記事と比較する。概要がニュース記事のメイントピックとキーポイントをカバーしているかどうか、そして、それらを明確かつ論理的な順序で提示しているかどうかを確認する。

3. 3. コヒーレンスのスコアを1～5で設定する(1が最低、5が最高)。

スコアリング関数 スコアリング関数は、設計されたプロンプト、自動CoT、入力コンテキスト、評価する必要のあるターゲットテキストでLLMを呼び出します。G-E VALは、ターゲットテキストを生成する条件付き確率を評価指標とするGPTScore (Fu et al., 2023)とは異なり、フォームフィリングパラダイムで直接評価タスクを実行する。例えば、テキスト要約の一貫性を評価するために、プロンプト、CoT、ニュース記事、要約を連結し、定義された基準に基づいて、各評価側面について1から5までのスコアを出力するLLMを呼び出します。しかし、この直接採点関数には2つの問題があることに気がつく。

1. ある評価課題では、通常1桁の数字が点数の分布を支配し、1 5スケールの場合は3桁の数字が中心となります。これは、スコアの分散が小さく、人間の判断との相関が低いことにつながる可能性がある。

2. LLMは通常、プロンプトが明示的に小数点以下を要求しても、整数スコアを出力するだけである。これは、生成されたテキスト間の微妙な違いを捉えていない評価スコアに多くの関連性をもたらす。

これらの問題を解決するために、我々はLLMからの出力トークンの確率を用いてスコアを正規化し、それらの重み付き和を最終結果として取ることを提案する。形式的には、プロンプトS = {s_1, s_2, --, s_n }であらかじめ定義されたスコア(1から5まで)の集合が与えられたとき、各スコアp(s_i)の確率はLLMによって計算され、最終スコアは以下の通りである。

$$score = \sum_{i=1}^{n} p(s_i) \times s_i \tag{1}$$

この方法は、生成されたテキストの品質と多様性をよりよく反映した、よりきめ細かい連続的なスコアを得ることができる。

3 Experiments

Zhongら(2022)に従い、要約と対話応答生成の2つのNLGタスクのSummEval、Topical-Chat、QAGSの3つのベンチマークで評価者をメタ評価した。

3.1 Implementation Details

LLMにはOpenAIのGPTファミリーをGPT-3.5(text-davinci-003)、GPT-4などで使用しています。GPT-3.5では、モデルの決定性を高めるために復号温度を0に設定しました。GPT-4では、トークン確率の出力をサポートしていないため、'n = 20, temperature = 1, top p = 1'を設定し、20回サンプリングしてトークン確率を推定しています。

uation criteria for different NLG tasks and, such as coherence, conciseness, or grammar. For example, for evaluating coherence in text summarization, we add the following content to the prompt:

> *Evaluation Criteria:*
>
> *Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic."*

Auto Chain-of-Thoughts for NLG Evaluation The chain-of-thoughts (CoT) is a sequence of intermediate representations that are generated by the LLM during the text generation process. For evaluation tasks, some criteria need a more detailed evaluation instruction beyond the simple definition, and it is time-consuming to manually design such evaluation steps for each task. We find that LLM can generate such evaluation steps by itself. The CoT can provide more context and guidance for the LLM to evaluate the generated text, and can also help to explain the evaluation process and results. For example, for evaluating coherence in text summarization, we add a line of *"Evaluation Steps:"* to the prompt and let LLM to generate the following CoT automatically:

> *1. Read the news article carefully and identify the main topic and key points.*
>
> *2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.*
>
> *3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.*

Scoring Function The scoring function calls the LLM with the designed prompt, auto CoT, the input context and the target text that needs to be evaluated. Unlike GPTScore (Fu et al., 2023) which uses the conditional probability of generating the target text as an evaluation metric, G-EVAL directly performs the evaluation task with a form-filling paradigm. For example, for evaluating coherence in text summarization, we concatenate the prompt, the CoT, the news article, and the summary, and then call the LLM to output a score from 1 to 5 for each evaluation aspect, based on the defined criteria.

However, we notice this direct scoring function has two issues:

1. For some evaluation tasks, one digit usually dominates the distribution of the scores, such as 3 for a 1 - 5 scale. This may lead to the low variance of the scores and the low correlation with human judgments.

2. LLMs usually only output integer scores, even when the prompt explicitly requests decimal values. This leads to many ties in evaluation scores which do not capture the subtle difference between generated texts.

To address these issues, we propose using the probabilities of output tokens from LLMs to normalize the scores and take their weighted summation as the final results. Formally, given a set of scores (like from 1 to 5) predefined in the prompt $S = \{s_1, s_2, ..., s_n\}$, the probability of each score $p(s_i)$ is calculated by the LLM, and the final score is:

$$score = \sum_{i=1}^{n} p(s_i) \times s_i \tag{1}$$

This method obtains more fine-grained, continuous scores that better reflect the quality and diversity of the generated texts.

3 Experiments

Following Zhong et al. (2022), we meta-evaluate our evaluator on three benchmarks, SummEval, Topical-Chat and QAGS, of two NLG tasks, summarization and dialogue response generation.

3.1 Implementation Details

We use OpenAI's GPT family as our LLMs, including GPT-3.5 (text-davinci-003) and GPT-4. For GPT-3.5, we set decoding temperature to 0 to increase the model's determinism. For GPT-4, as it does not support the output of token probabilities, we set '$n = 20, temperature = 1, top\_p = 1$' to sample 20 times to estimate the token probabilities. We use G-EVAL-4 to indicate G-EVAL with GPT-4

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BARTScore	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377
GPTScore	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-EVAL-3.5	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
- Probs	0.359	*0.313*	0.361	*0.344*	0.339	*0.323*	0.327	*0.288*	0.346	*0.317*
G-EVAL-4	**0.582**	**0.457**	**0.507**	**0.425**	**0.455**	**0.378**	**0.547**	**0.433**	**0.514**	**0.418**
- Probs	0.560	*0.472*	0.501	*0.459*	0.438	*0.408*	0.511	*0.444*	0.502	*0.446*
- CoT	0.564	0.454	0.493	0.413	0.403	0.334	0.538	0.427	0.500	0.407

表 1: SummEval ベンチマークにおける各メトリクスのサマリーレベルのスピアマン(ρ)およびケンドール-タウ(τ)相関。確率のないG-E VAL(イタリック体)は、τに関する他の指標との公正な比較として考えるべきでない、なぜなら、それはスコアに多くの関連性をもたらすからである。この結果、Kendall-Tau相関は高くなりますが、真の評価能力を公平に反映するものではありません。詳細はセクション 4 を参照されたい。

GPT-4をバックボーンモデルとしたG-E VALをG-E VAL -4、GPT-3.5をバックボーンモデルとしたG-E VALをG-E VAL -3.5と表記する。各タスクのプロンプトの例は付録のとおりである。

3.2 Benchmarks

G-E VAL と人間の判断の相関を測定するために、3 つのメタ評価ベンチマークを採用した。

SummEval (Fabbri et al., 2021) は、要約のための異なる評価方法を比較するベンチマークである。各要約の4つの側面(流暢さ、一貫性、一貫性、関連性)について、人間の評価を与える。CNN/DailyMailデータセット(Hermann et al.,2015)を基に構築されている。

Topical-Chat (Mehri and Eskenazi, 2020) は、知識を用いた対話応答生成システムにおいて、異なる評価者をメタ評価するためのテストベッドである。我々は(Zhong et al., 2022)に従って、自然さ、首尾一貫性、魅力、接地性の4つの側面で人間の評価を使用する。

QAGS (Wang et al., 2020) は、要約タスクにおける幻覚を評価するためのベンチマークである。2つの異なる要約データセットにおける要約の一貫性の次元を測定することを目的とする。

3.3 Baselines

G-E VAL を、最先端の性能を達成した様々な評価者と比較評価した。

BERTScore (Zhang et al., 2019) measures the BERT (Devlin et al., 2019) の文脈埋め込みに基づく2つのテキスト間の類似性。

MoverScore (Zhao et al., 2019) improves BERTScoreは、より頑健な類似性指標を得るために、ソフトアライメントと新しい集計方法を追加することで、BERTScoreを拡張した。

BARTScore (Yuan et al., 2021) is a unified evaluatorは、事前に学習したエンコーダ・デコーダモデルBARTの平均尤度で評価する(Lewis et al., 2020)。ソースとターゲットのフォーマットによって、異なるスコアを予測することができます。

FactCC and **QAGS** (Kryściński et al., 2020; Wang et al., 2020)は、生成された要約の事実上の一貫性を測定する2つの評価者である。FactCC は、要約がソース文書と一致するかどうかを予測する BERT ベースの分類器です。QAGSは質問応答ベースの評価者であり、要約から質問を生成し、その回答が原文に記載されているかどうかを チェックします。

USR (Mehri and Eskenazi, 2020) is evaluator は、対話応答生成を様々な角度から評価するものである。ターゲットの回答ごとに異なるスコアを割り当てるバージョンがいくつかあります。

UniEval (Zhong et al., 2022) is a unified evalua-

テキスト生成の様々な側面をQAタスクとして評価することができる。

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BARTScore	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377
GPTScore	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-EVAL-3.5	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
- Probs	0.359	*0.313*	0.361	*0.344*	0.339	*0.323*	0.327	*0.288*	0.346	*0.317*
G-EVAL-4	**0.582**	**0.457**	**0.507**	**0.425**	**0.455**	**0.378**	**0.547**	**0.433**	**0.514**	**0.418**
- Probs	0.560	*0.472*	0.501	*0.459*	0.438	*0.408*	0.511	*0.444*	0.502	*0.446*
- CoT	0.564	0.454	0.493	0.413	0.403	0.334	0.538	0.427	0.500	0.407

Table 1: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on SummEval benchmark. G-EVAL without probabilities (*italicized*) should not be considered as a fair comparison to other metrics on τ, as it leads to many ties in the scores. This results in a higher Kendall-Tau correlation, but it does not fairly reflect the true evaluation ability. More details are in Section 4.

as the backbone model, and G-EVAL-3.5 to indicate G-EVAL with GPT-3.5 as the backbone model. Example prompts for each task are provided in the Appendix.

3.2 Benchmarks

We adopt three meta-evaluation benchmarks to measure the correlation between G-EVAL and human judgments.

SummEval (Fabbri et al., 2021) is a benchmark that compares different evaluation methods for summarization. It gives human ratings for four aspects of each summary: `fluency`, `coherence`, `consistency` and `relevance`. It is built on the CNN/DailyMail dataset (Hermann et al., 2015)

Topical-Chat (Mehri and Eskenazi, 2020) is a testbed for meta-evaluating different evaluators on dialogue response generation systems that use knowledge. We follow (Zhong et al., 2022) to use its human ratings on four aspects: `naturalness`, `coherence`, `engagingness` and `groundedness`.

QAGS (Wang et al., 2020) is a benchmark for evaluating hallucinations in the summarization task. It aims to measure the `consistency` dimension of summaries on two different summarization datasets.

3.3 Baselines

We evaluate G-EVAL against various evaluators that achieved state-of-the-art performance.

BERTScore (Zhang et al., 2019) measures the similarity between two texts based on the contextualized embedding from BERT (Devlin et al., 2019).

MoverScore (Zhao et al., 2019) improves BERTScore by adding soft alignments and new aggregation methods to obtain a more robust similarity measure.

BARTScore (Yuan et al., 2021) is a unified evaluator which evaluate with the average likelihood of the pretrained encoder-decoder model, BART (Lewis et al., 2020). It can predict different scores depending on the formats of source and target.

FactCC and **QAGS** (Kryściński et al., 2020; Wang et al., 2020) are two evaluators that measure the factual consistency of generated summaries. FactCC is a BERT-based classifier that predicts whether a summary is consistent with the source document. QAGS is a question-answering based evaluator that generates questions from the summary and checks if the answers can be found in the source document.

USR (Mehri and Eskenazi, 2020) is evaluator that assess dialogue response generation from different perspectives. It has several versions that assign different scores to each target response.

UniEval (Zhong et al., 2022) is a unified evaluator that can evaluate different aspects of text gen-

これは、事前に学習されたT5モデル(Raffel et al., 2020)を用いて、評価タスク、ソース、ターゲットテキストを質問と回答として符号化し、評価スコアとしてQAスコアを計算するものである。また、質問形式を変更することで、異なる評価タスクを扱うことができます。GPTScore (Fu et al., 2023) は、GPT-3 のような生成的な事前学習モデルでテキストを評価する新しいフレームワークである。これは、生成的な事前学習モデルが、与えられた指示と文脈に従って高品質の生成テキストの確率を高く割り当てることを想定している。G-E VALとは異なり、GPTScoreは評価タスクをフォームフィリング問題ではなく、条件付き生成問題として定式化する。

3.4 Results for Summarization

要約レベルのスピアマン相関とケンドール・タウ相関を用いて、Zhongら(2022)と同じアプローチで、異なる要約指標を評価する。表 1 の前半は、モデル出力と参照テキストの意味的類似性を比較したメトリクスの結果である。これらのメトリクスは、ほとんどの次元で低いパフォーマンスを示す。第二部では、人間の要約品質評価からニューラルネットワークを用いて学習するメトリクスの結果を示す。これらのメトリクスは、類似性ベースのメトリクスよりも相関が高く、要約の評価においてより信頼性が高いことが示唆される。表1の最後の部分はGPTベースの評価者に対応し、GPTスコアも要約テキストの評価にGPTを使用するが、与えられたターゲットの条件付き確率に依存する。G-E VAL は、SummEvalベンチマークにおいて、これまでの最先端評価者を大幅に上回る性能を示した。G-E VAL -4 は G-E VAL -3.5 と比較して、スピアマン相関、ケンドール-タウ相関ともに非常に高い人 間対応性を達成しており、GPT-4 のモデルサイズが大きいことが要約評価に有効であることが示された。G-E VALもGPTScoreをいくつかの次元で上回り、単純なフォームフィリングパラダイムの有効性を実証している。

3.5 Results for Dialogue Generation

我々は、Mehri and Eskenazi (2020)のTopical-chatベンチマークを使用して、対話応答の品質に関する人間の評価と異なる評価者がどの程度一致しているかを測定する。ダイアログの各ターンについて、ピアソン相関とスピアマン相関を計算します。表2より、類似度ベースのメトリクスは、回答がどの程度魅力的で根拠があるかについては人間と良い一致を示すが、他の側面については一致しないことがわかる。

学習型評価者に関しては、G-E VAL の前に、UniEval はあらゆる側面で人間の判断と最も一致するスコアを予測する。前節で示したように、G-E VAL も Topical-Chat ベンチマークにおいて、これまでの最先端評価器を大幅に上回っている。注目すべきは、G-E VAL -3.5がG-E VAL -4でも同様の結果を達成できることである。これは、このベンチマークがG-E VALモデルに対して比較的容易であることを示している。

3.6 Results on Hallucinations

高度なNLGモデルはしばしば文脈入力と一致しないテキストを生成し(Cao et al., 2018)、最近の研究では、強力なLLMでさえも幻覚の問題に悩まされていることが判明している。このことは、要約における一貫性アスペクトを測定するための評価器を設計するための最近の研究の動機付けとなる(Kryściński et al., 2020; Wang et al., 2020; Cao et al., 2020; Durmus et al., 2020)。QAGSメタ評価ベンチマークをテストした。このベンチマークには2つの異なる要約データセットが含まれている。CNN/DailyMailとXSum (Narayan et al., 2018) 表3は、BARTScoreがより抽出的なサブセット(QAGS-CNN)では良好なパフォーマンスを示すが、より抽象的なサブセット(QAGS-Xsum)では低い相関を持つことを示している。UniEvalは、データの両サブセットで良好な相関がある。G-E VAL -4 は、QAGS において、QAGS-Xsum において、すべての最先端評価者を平均して上回り、大きなマージンを獲得している。一方、G-E VAL -3.5 はこのベンチマークで良好な結果を得ることができず、整合性が LLM の容量に敏感であることがわかる。この結果は、表1と一致する。

4 Analysis

G-E VALはLLMベースの出力を好むか?LLMを評価者として使用する際の懸念点として、人間が書いた高品質のテキストよりも、LLM自身が生成した出力を好む可能性があることが挙げられます。この問題を調べるために、要約タスクの実験を行い、LLM生成要約と人間が書いた要約の評価スコアを比較する。Zhang et al. (2023)で収集されたデータセットを使用し、まずフリーランスの書き手にニュース記事の高品質な要約を書くよう依頼し、次にアノテーターに人間が書いた要約とLLMが生成した要約を比較するよう依頼する(GPT-3.5、text-davinci003を使用)。

eration as QA tasks. It uses a pretrained T5 model (Raffel et al., 2020) to encode the evaluation task, source and target texts as questions and answers, and then computes the QA score as the evaluation score. It can also handle different evaluation tasks by changing the question format.

GPTScore (Fu et al., 2023) is a new framework that evaluates texts with generative pre-training models like GPT-3. It assumes that a generative pre-training model will assign a higher probability of high-quality generated text following a given instruction and context. Unlike G-EVAL, GPTScore formulates the evaluation task as a conditional generation problem instead of a form-filling problem.

3.4 Results for Summarization

We adopt the same approach as Zhong et al. (2022) to evaluate different summarization metrics using summary-level Spearman and Kendall-Tau correlation. The first part of Table 1 shows the results of metrics that compare the semantic similarity between the model output and the reference text. These metrics perform poorly on most dimensions. The second part shows the results of metrics that use neural networks to learn from human ratings of summary quality. These metrics have much higher correlations than the similarity-based metrics, suggesting that they are more reliable for summarization evaluation.

In the last part of Table 1 which corresponds to GPT-based evaluators, GPTScore also uses GPTs for evaluating summarization texts, but relies on GPT's conditional probabilities of the given target. G-EVAL substantially surpasses all previous state-of-the-art evaluators on the SummEval benchmark. G-EVAL-4 achieved much higher human correspondence compared with G-EVAL-3.5 on both Spearman and Kendall-Tau correlation, which indicates that the larger model size of GPT-4 is beneficial for summarization evaluation. G-EVAL also outperforms GPTScore on several dimension, demonstrating the effectiveness of the simple form-filling paradigm.

3.5 Results for Dialogue Generation

We use the Topical-chat benchmark from Mehri and Eskenazi (2020) to measure how well different evaluators agree with human ratings on the quality of dialogue responses. We calculate the Pearson and Spearman correlation for each turn of the dialogue. Table 2 shows that similarity-based metrics have good agreement with humans on how `engaging` and `grounded` the responses are, but not on the other aspects. With respect to the learning-based evaluators, before G-EVAL, UniEval predicts scores that are most consistent with human judgments across all aspects.

As shown in the last part, G-EVAL also substantially surpasses all previous state-of-the-art evaluator on the Topical-Chat benchmark. Notably, the G-EVAL-3.5 can achieve similar results with G-EVAL-4. This indicates that this benchmark is relatively easy for the G-EVAL model.

3.6 Results on Hallucinations

Advanced NLG models often produce text that does not match the context input (Cao et al., 2018), and recent studies find even powerful LLMs also suffer from the problem of hallucination. This motivates recent research to design evaluators for measuring the `consistency` aspect in summarization (Kryściński et al., 2020; Wang et al., 2020; Cao et al., 2020; Durmus et al., 2020). We test the QAGS meta-evaluation benchmark, which includes two different summarization datasets: CNN/DailyMail and XSum (Narayan et al., 2018) Table 3 shows that BARTScore performs well on the more extractive subset (QAGS-CNN), but has low correlation on the more abstractive subset (QAGS-Xsum). UniEval has good correlation on both subsets of the data.

On average, G-EVAL-4 outperforms all state-of-the-art evaluators on QAGS, with a large margin on QAGS-Xsum. G-EVAL-3.5, on the other hand, failed to perform well on this benchmark, which indicates that the consistency aspect is sensitive to the LLM's capacity. This result is consistent with Table 1.

4 Analysis

Will G-EVAL prefer LLM-based outputs? One concern about using LLM as an evaluator is that it may prefer the outputs generated by the LLM itself, rather than the high-quality human-written texts. To investigate this issue, we conduct an experiment on the summarization task, where we compare the evaluation scores of the LLM-generated and the human-written summaries. We use the dataset collected in Zhang et al. (2023), where they first ask freelance writers to write high-quality summaries for news articles, and then ask annotators to compare human-written summaries and LLM-generated summaries (using GPT-3.5, text-davinci-

Metrics	Naturalness		Coherence		Engagingness		Groundedness		AVG	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
ROUGE-L	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327	0.243	0.244
BLEU-4	0.180	0.175	0.131	0.235	0.232	0.316	0.213	0.310	0.189	0.259
METEOR	0.212	0.191	0.250	0.302	0.367	0.439	0.333	0.391	0.290	0.331
BERTScore	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317	0.262	0.273
USR	0.337	0.325	0.416	0.377	0.456	0.465	0.222	0.447	0.358	0.403
UniEval	0.455	0.330	0.602	0.455	0.573	0.430	0.577	0.453	0.552	0.417
G-EVAL-3.5	0.532	0.539	0.519	0.544	**0.660**	**0.691**	**0.586**	0.567	0.574	0.585
G-EVAL-4	**0.549**	**0.565**	**0.594**	**0.605**	0.627	0.631	0.531	0.551	**0.575**	**0.588**

表 2: Topical-Chat ベンチマークにおける各メトリクスのターンレベルのスピアマン(ρ)およびケンドール-タウ(τ)相関。

データセットは3つのカテゴリに分けられる。1）人間による審査でGPT-3.5要約より高い評価を得た人間による審査、2）人間による審査でGPT-3.5要約より低い評価を得た人間による審査、3）人間による審査でGPT-3.5要約と同等の評価を受けた人間による審査、である。G-E VAL -4 を用いて各カテゴリの要約を評価し、平均したスコアを比較する。その結果を図2に示す。G-E VAL -4は、人間の審査員が人間が書いた要約も好む場合に、人間が書いた要約に高いスコアを割り当て、人間の審査員がGPT-3.5の要約を好む場合に低いスコアを割り当てていることが分かる。しかし、G-E VAL -4は、人間が書いた要約を好む場合でも、常にGPT-3.5要約の方が人間が書いた要約よりも高いスコアを与えている。この現象の理由として、2つの可能性が考えられる。

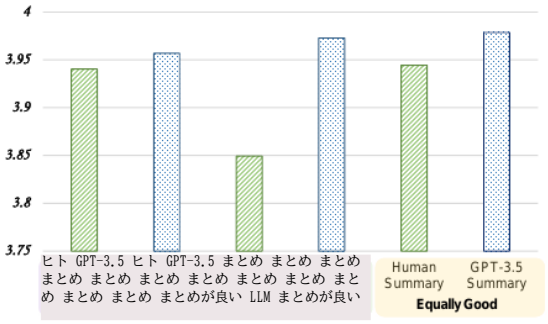

図2:人間が書いた要約とGPT-3.5の要約のG-E VAL -4の平均スコアを人間の好みで割ったもの。

1. 高品質なシステムからのNLG出力は、当然ながら評価が困難である。原著論文の著者らは、人間が書いた要約とLLMが生成した要約の判断に関するアノテーター間の一致度は非常に低く、Krippendorffのαは0.07であったことを明らかにした。

2. G-E VALは、生成時と評価時に同じ評価基準の概念を共有する可能性があるため、LLMで生成された要約に偏りがある可能性がある。

我々の研究は、この問題に関する予備的な研究として考慮されるべきであり、LLMベースの評価者は、LLM生成テキストに対する固有のバイアスを軽減するために、LLM生成テキストに対する固有のバイアスを軽減する。我々は、評価スコアを更なるチューニングのための報酬信号として使用した場合、LLMベースの評価者がLLMの自己強化につながる可能性があるという文脈で、この懸念を強調する。また、これはNLGタスクの真の評価基準ではなく、LLMを独自の評価基準にオーバーフィットさせる結果になる可能性がある。

思考の連鎖の効果 SummEvalベンチマークにおいて、思考の連鎖(CoT)がある場合とない場合のG-E VALの性能を比較する。表 1 より、CoT を用いた G-E VAL -4 は、CoT を用いない G-E VAL -4 よりも、すべての次元で、特に流暢さにおいて高い相関があることがわかる。このことから、CoTは生成されたテキストを評価するためのLLMにより多くの文脈とガイダンスを提供し、評価プロセスと結果を説明するのに役立つことが示唆された。

.

確率正規化の効果 SummEvalベンチマークにおいて、確率正規化を行った場合と行わなかった場合のG-E VALの性能を比較した。表1より、KendallTau相関では、SummEvalでは確率のあるG-E VAL -4が確率のないG-E VAL -4より劣っていることがわかる。

[2] この実験では、要約タスクの評価におけるG-E VAL -4が優れているため、G-E VAL -4を使用する。GPT-3.5では分布が異なるが、2つのLLMはテキスト生成の点で類似した挙動を示すはずである。

Metrics	Naturalness		Coherence		Engagingness		Groundedness		AVG	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
ROUGE-L	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327	0.243	0.244
BLEU-4	0.180	0.175	0.131	0.235	0.232	0.316	0.213	0.310	0.189	0.259
METEOR	0.212	0.191	0.250	0.302	0.367	0.439	0.333	0.391	0.290	0.331
BERTScore	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317	0.262	0.273
USR	0.337	0.325	0.416	0.377	0.456	0.465	0.222	0.447	0.358	0.403
UniEval	0.455	0.330	0.602	0.455	0.573	0.430	0.577	0.453	0.552	0.417
G-EVAL-3.5	0.532	0.539	0.519	0.544	**0.660**	**0.691**	**0.586**	0.567	0.574	0.585
G-EVAL-4	**0.549**	**0.565**	**0.594**	**0.605**	0.627	0.631	0.531	0.551	**0.575**	**0.588**

Table 2: Turn-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on Topical-Chat benchmark.

003).

The dataset can be divided in three categories: 1) human-written summaries that are rated *higher* than GPT-3.5 summaries by human judges, 2) human-written summaries that are rated *lower* than GPT-3.5 summaries by human judges, and 3) human-written summaries and GPT-3.5 summaries are rated *equally* good by human judges. We use G-EVAL-4 to evaluate the summaries in each category, and compare the averaged scores. [2]

The results are shown in Figure 2. We can see that, G-EVAL-4 assigns higher scores to human-written summaries when human judges also prefer human-written summaries, and assigns lower scores when human judges prefer GPT-3.5 summaries. However, G-EVAL-4 always gives higher scores to GPT-3.5 summaries than human-written summaries, even when human judges prefer human-written summaries. We propose two potential reasons for this phenomenon:

1. NLG outputs from high-quality systems are in natural difficult to evaluate. The authors of the original paper found that inter-annotator agreement on judging human-written and LLM-generated summaries is very low, with Krippendorff's alpha at 0.07.

2. G-EVAL may have a bias towards the LLM-generated summaries because the model could share the same concept of evaluation criteria during generation and evaluation.

Our work should be considered as a preliminary study on this issue, and more research is needed to fully understand the behavior of LLM-based evaluators to reduce its inherent bias towards LLM-generated text. We highlight this concern in the context that LLM-based evaluators may lead to self-reinforcement of LLMs if the evaluation score is used as a reward signal for further tuning. And this could result in the over-fitting of the LLMs to their own evaluation criteria, rather than the true evaluation criteria of the NLG tasks.

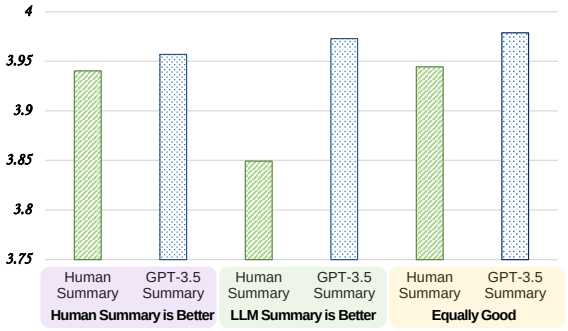

Figure 2: Averaged G-EVAL-4's scores for human-written summaries and GPT-3.5 summaries, divided by human judges' preference.

The Effect of Chain-of-Thoughts We compare the performance of G-EVAL with and without chain-of-thoughts (CoT) on the SummEval benchmark. Table 1 shows that G-EVAL-4 with CoT has higher correlation than G-EVAL-4 without CoT on all dimensions, especially for `fluency`. This suggests that CoT can provide more context and guidance for the LLM to evaluate the generated text, and can also help to explain the evaluation process and results.

The Effect of Probability Normalization We compare the performance of G-EVAL with and without probability normalization on the SummEval benchmark. Table 1 shows that, on Kendall-Tau correlation, G-EVAL-4 with probabilities is

[2]We use G-EVAL-4 in this experiment, because its superiority in evaluating summarization tasks. Although it has different distribution with with GPT-3.5, the two LLMs should share similar behaviors in terms of text generation.

Metrics	QAGS-CNN			QAGS-XSUM			Average		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-2	0.459	0.418	0.333	0.097	0.083	0.068	0.278	0.250	0.200
ROUGE-L	0.357	0.324	0.254	0.024	-0.011	-0.009	0.190	0.156	0.122
BERTScore	0.576	0.505	0.399	0.024	0.008	0.006	0.300	0.256	0.202
MoverScore	0.414	0.347	0.271	0.054	0.044	0.036	0.234	0.195	0.153
FactCC	0.416	0.484	0.376	0.297	0.259	0.212	0.356	0.371	0.294
QAGS	0.545	-	-	0.175	-	-	0.375	-	-
BARTScore	**0.735**	0.680	0.557	0.184	0.159	0.130	0.459	0.420	0.343
CTC	0.619	0.564	0.450	0.309	0.295	0.242	0.464	0.430	0.346
UniEval	0.682	0.662	0.532	0.461	0.488	0.399	0.571	0.575	0.465
G-EVAL-3.5	0.477	0.516	0.410	0.211	0.406	0.343	0.344	0.461	0.377
G-EVAL-4	0.631	**0.685**	**0.591**	**0.558**	**0.537**	**0.472**	**0.599**	**0.611**	**0.525**

表 3: QAGS ベンチマークにおける各メトリクスのピアソン(r)、スピアマン(ρ)、ケンドール・タウ(τ)相関。

これは、Kendall-Tau相関の計算に関連していると考えており、一致するペアと不一致のペアの数に基づいています。確率を除いた直接のスコアリングは、多くの同点になり、それらは一致するものと不一致のいずれともカウントされない。これはKendall-Tauの相関を高くするかもしれませんが、生成されたテキストを評価するモデルの真の能力を反映したものではありません。一方、確率正規化により、生成されたテキスト間の微妙な違いをよりよくとらえた、よりきめ細かい連続的なスコアを得ることができる。これは、G-E VAL -4 の高いスピアマン相関と、スコアの順位に基づく確率に反映されている。

モデルサイズの影響 SummEvalとQAGSのベンチマークにおいて、モデルサイズを変えた場合のG-EVALの性能を比較した。表1および表3より、Topical-Chatベンチマークでは、G-E VAL -4はG-E VAL -3.5よりも、魅力と接地性を除くほとんどの次元およびデータセットで高い相関があることがわかる。これは、モデルサイズを大きくすることで、特に一貫性や関連性といった、より困難で複雑な評価タスクにおいて、G-E VALの性能を向上させることができることを示している。

5 Related Work

ngram-based metrics ngram-based metricsとは、生成されたテキストと参照テキストの語彙的重複を測定することで、NLGモデルを評価するためのスコアのことである。BLEU (Papineni et al., 2002) は機械翻訳評価に最も広く用いられている指標で、修正N-gram精度の幾何平均と簡潔性ペナルティを計算するものである。

ROUGE (Lin, 2004) は要約評価のための想起指向の指標であり、生成された要約と参照要約の集合の間の n-gram オーバーラップを測定するものである。NLGに関する最近の論文の60%以上は、ROUGEやBLEUにのみ依存してシステムを評価していることが示されている(Kasai et al., 2021)。しかし、これらのメトリクスは、コンテンツ品質の測定(Reiter and Belz, 2009)や構文エラーの捕捉(Stent et al., 2005)に失敗するため、NLGシステムの信頼性を正確に反映することができない。

埋め込みベースメトリクス 埋め込みベースメトリクスとは、単語や文の埋め込みに基づいて生成されたテキストと参照テキストとの間の意味的類似性を測定することによって、NLGモデルを評価するためのスコアを指す。WMD (Kusner et al., 2015) は、単語埋め込みに基づいて2つのテキスト間の距離を測定するメトリックである。BERTScore (Zhang et al., 2019) は、BERT (Devlin et al., 2019) からの文脈埋め込みに基づいて、2つのテキスト間の類似性を測定する。MoverScore (Zhao et al., 2019) は、より堅牢な類似性指標を得るために、ソフトアライメントと新しい集約方法を追加することによってBERTScoreを改善する。(Clark et al., 2019) は、文の埋め込みに基づいて生成されたテキストと参照テキストの間の類似度を計算することによって、複数文のテキストを評価するメトリックを提案している。

タスク固有の評価者 タスク固有の評価指標とは、特定のタスク要件に基づいて生成されたテキストの品質を測定することによって、NLGモデルを評価するためのスコアを指す。例えば、要約タスクは生成された要約の一貫性を評価する必要があり(Kryściński et al., 2020

Metrics	QAGS-CNN			QAGS-XSUM			Average		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-2	0.459	0.418	0.333	0.097	0.083	0.068	0.278	0.250	0.200
ROUGE-L	0.357	0.324	0.254	0.024	-0.011	-0.009	0.190	0.156	0.122
BERTScore	0.576	0.505	0.399	0.024	0.008	0.006	0.300	0.256	0.202
MoverScore	0.414	0.347	0.271	0.054	0.044	0.036	0.234	0.195	0.153
FactCC	0.416	0.484	0.376	0.297	0.259	0.212	0.356	0.371	0.294
QAGS	0.545	-	-	0.175	-	-	0.375	-	-
BARTScore	**0.735**	0.680	0.557	0.184	0.159	0.130	0.459	0.420	0.343
CTC	0.619	0.564	0.450	0.309	0.295	0.242	0.464	0.430	0.346
UniEval	0.682	0.662	0.532	0.461	0.488	0.399	0.571	0.575	0.465
G-EVAL-3.5	0.477	0.516	0.410	0.211	0.406	0.343	0.344	0.461	0.377
G-EVAL-4	0.631	**0.685**	**0.591**	**0.558**	**0.537**	**0.472**	**0.599**	**0.611**	**0.525**

Table 3: Pearson (r), Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on QAGS benchmark.

inferior to G-EVAL-4 without probabilities on SummEval. We believe this is related to the calculation of Kendall-Tau correlation, which is based on the number of concordant and discordant pairs. Direct scoring without probabilities can lead to many ties, which are not counted as either concordant or discordant. This may result in a higher Kendall-Tau correlation, but it does not reflect the model's true capacity of evaluating the generated texts. On the other hand, probability normalization can obtain more fine-grained, continuous scores that better capture the subtle difference between generated texts. This is reflected by the higher Spearman correlation of G-EVAL-4 with probabilities, which is based on the rank order of the scores.

The Effect of Model Size We compare the performance of G-EVAL with different model sizes on the SummEval and QAGS benchmarks. Table 1 and Table 3 show that G-EVAL-4 has higher correlation than G-EVAL-3.5 on most dimensions and datasets, except for `engagingness` and `groundedness` on the Topical-Chat benchmark. This demonstrates that larger model size can improve the performance of G-EVAL, especially for more challenging and complex evaluation tasks, such as `consistency` and `relevance`.

5 Related Work

Ngram-based Metrics Ngram-based metrics refer to the scores for evaluating the NLG models by measuring the lexical overlap between a generated text and a reference text. BLEU (Papineni et al., 2002) is the most widely used metric for machine translation evaluation, which calculates the geometric mean of modified n-gram precision and a brevity penalty. ROUGE (Lin, 2004) is a recall-oriented metric for summarization evaluation, which measures the n-gram overlap between a generated summary and a set of reference summaries. It has been shown that more than 60% of recent papers on NLG only rely on ROUGE or BLEU to evaluate their systems (Kasai et al., 2021). However, these metrics fail to measure content quality (Reiter and Belz, 2009) or capture syntactic errors (Stent et al., 2005), and therefore do not reflect the reliability of NLG systems accurately.

Embedding-based Metrics Embedding-based metrics refer to the scores for evaluating the NLG models by measuring the semantic similarity between a generated text and a reference text based on the word or sentence embeddings. WMD (Kusner et al., 2015) is a metric that measures the distance between two texts based on the word embeddings. BERTScore (Zhang et al., 2019) measures the similarity between two texts based on the contextualized embedding from BERT (Devlin et al., 2019). MoverScore (Zhao et al., 2019) improves BERTScore by adding soft alignments and new aggregation methods to obtain a more robust similarity measure. (Clark et al., 2019) propose a metric that evaluates multi-sentence texts by computing the similarity between the generated text and the reference text based on the sentence embeddings.

Task-specific Evaluators Task-specific metrics refer to the scores for evaluating the NLG models by measuring the quality of the generated texts based on the specific task requirements. For example, summarization tasks need to assess the `consistency` of the generated sum-

; Wang et al., 2020; Cao et al., 2020; Durmus et al., 2020)、対話応答生成タスクは生成応答の一貫性を評価する必要がある(Dziri et al., 2019; Ye et al., 2021)。しかし、これらのメトリクスは他のNLGタスクに一般化できず、生成されたテキストの全体的な品質を測定することができない。

統一評価者 近年、入力と出力の内容(Yuan et al., 2021)や使用するモデルの変種(Mehri and Eskenazi, 2020)を変化させることで、多次元からテキスト品質を評価する評価者も開発されている。UniEval (Zhong et al., 2022) は、QAタスクとしてテキスト生成の様々な側面を評価することができる統一的な評価者である。質問形式を変更することで、様々な評価タスクを処理することができます。

LLMベースの評価者 Fuら(2023)は、GPT-3のような生成的な事前学習モデルでテキストを評価する新しいフレームワークであるGPTScoreを提案している。これは、生成的な事前学習モデルが、与えられた指示と文脈に従って高品質の生成テキストのより高い確率を割り当てることを想定している。Wangら(2023)は、NLG評価器としてChatGPTを使用する際の事前調査を実施している。Kocmi and Federmann (2023)は、機械翻訳タスクの評価にGPTモデルを使用することを提案した。

6 Conclusion

本論文では、生成されたテキストの品質を評価するために、思考の連鎖を伴うLLM(CoT)を用いるフレームワークであるG-E VALを提案する。我々は、テキスト要約と対話生成という2つの自然言語処理タスクについて広範な実験を行い、G-E VALが最新の評価器を凌駕し、より高い人間の対応を達成できることを示す。また、LLMベースの評価者の動作に関する予備的な分析を提案し、LLMベースの評価者がLLM生成テキストに偏りを持つという潜在的な問題を浮き彫りにする。我々は、我々の研究が、LLMをNLG評価に用いる研究のきっかけとなり、また、LLMを評価者として用いることの潜在的なリスクと課題に対する認識を高めることを期待している。

References

Gill・ト・サン・エヴ・バネルジー、アロン・ラヴィー2005. Meteor: 人間の判断との相関を改善した mt 評価のための自動的な指標。機械翻訳および/または要約のための内在的および外在的評価尺度に関する cl ワークショップの議事録、65-72 ページ。

孟曹、郭優、呉佳峰、キット・ジェイクチ。2020. 抽象化要約モデルに対する事実上の誤り訂正。自然言語処理における経験的手法(EMNLP)に関する2020年会議論文集，ページ 6251-6258.

曹志強、ウェイ古、李文傑、李杉建。2018. オリジナルに忠実である。事実を意識した神経抽象的要約。第30回AAAI人工知能会議において。

エリザベス・クラーク、アスリ・チェリキルマズ、ノア・A・スミス。2019. 文移動者の類似性。複数文のテキストに対する自動評価。第57回計算言語学会年次大会予稿集，2748-2760 ページ.

ジェイコブ・デブリン、ミン・ウェイ・チャン、ケントン・リー、クリスティナ・トゥタノワ。2019. Bert: 言語理解のための深い双方向変換器の事前学習。計算言語学会の北米支部の2019年大会の議事録にて。Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171- 4186.

エジン・デュルマス、へ、モナ・ディアブ2020. Feqa: 抽象化要約における忠実度評価のための質問応答評価フレームワーク.第58回計算言語学会年次大会予稿集，ページ 5055- 5070.

ノウハ・ディジリ、エハンサン・カマルー、コリ・マシューソン、オスマル・R・ザイアン。2019. entailmentを用いた対話システムにおける一貫性の評価。計算言語学会の北米支部の2019年大会の議事録にて。Human Language Technologies, Volume 1 (Long and Short Papers), pages 3806-3812.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev.2021. 要約 要約評価。要約評価の再評価。計算言語学会論文誌 9:391-409.

周志蘭、Ng 聡孔、江正宝、劉鵬飛。2023. Gptscore。好きなように評価する。arXiv preprint arXiv:2302.04166.

カール・モリッツ・ハーマン、トマス・コシスキー、エドワード・グレフェンストット、ラッセ・エスペホルト、ウィル・ケイ、ムスタファ・スレーマン、フィル・ブランソム。2015. 読み取りと理解のための機械の教育神経情報処理システムの進歩,28.

maries (Kryściński et al., 2020; Wang et al., 2020; Cao et al., 2020; Durmus et al., 2020), and dialogue response generation tasks need to assess the `coherence` of the generated responses (Dziri et al., 2019; Ye et al., 2021). However, these metrics are not generalizable to other NLG tasks, and they are not able to measure the overall quality of the generated texts.

Unified Evaluators Recently, some evaluators have been developed to assess text quality from multiple dimensions by varying the input and output contents (Yuan et al., 2021) or the model variants (Mehri and Eskenazi, 2020) they use. UniEval (Zhong et al., 2022) is a unified evaluator that can evaluate different aspects of text generation as QA tasks. By changing the question format, it can handle different evaluation tasks.

LLM-based Evaluators Fu et al. (2023) propose GPTScore, a new framework that evaluated texts with generative pre-training models like GPT-3. It assumes that a generative pre-training model will assign a higher probability of high-quality generated text following a given instruction and context. Wang et al. (2023) conduct a preliminary survey of using ChatGPT as a NLG evaluator. Kocmi and Federmann (2023) proposed to use GPT models for evaluating machine translation tasks.

6 Conclusion

In this paper, we propose G-EVAL, a framework of using LLM with chain-of-thoughts (CoT) to evaluate the quality of generated texts. We conduct extensive experiments on two NLG tasks, text summarization and dialogue generation, and show that G-EVAL can outperform state-of-the-art evaluators and achieve higher human correspondence. We also propose preliminary analysis on the behavior of LLM-based evaluators, and highlight the potential issue of LLM-based evaluator having a bias towards the LLM-generated texts. We hope our work can inspire more research on using LLMs for NLG evaluation, and also raise awareness of the potential risks and challenges of using LLMs as evaluators.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

笠井淳悟、坂口圭介、Ronan Le Bras、Lavinia Dunagan、Jacob Morrison、Alexander R Fabbri、Yejin Choi、Noah A Smith.2021. 二次元のリーダーボード。手元にある言語を生成し、評価する。arXiv preprint arXiv:2112.04139.

トム・コクミ、クリスチャン・フェダーマン2023. 大規模言語モデルは翻訳品質の最先端評価者である。arXiv preprint arXiv:2302.14520.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher.2020. 抽象的テキスト要約の事実上の一貫性を評価する。自然言語処理における経験的手法(EMNLP)に関する2020年会議論文集, ページ9332-9346.

マット・クズナー、ユ・サン、ニコラス・コルキン、キリアン・ワインバーガー。2015. 単語埋め込みから文書距離へ機械学習に関する国際会議、ページ957-966で。PMLR.

マイク・ルイス、イインハン・リュー、ナマン・ゴヤル、マルジャン・ガズヴィニンジャッド、アブダラマン・モハメド、オメル・レヴィ、ヴェセリン・ストイノフ、ルーク・ゼトルモイヤー。2020. BART: 自然言語生成、翻訳、理解のためのsequence-to-sequence事前学習のノイズ除去。計算言語学会第58回年次大会予稿集, ACL 2020, Online, July 5-10, 2020, pages 7871-7880. Association for Computational Linguistics(計算言語学会).

林チン・ユー. 2004. Rouge: 要約の自動評価のためのパッケージ.テキスト要約の分岐点、74-81ページ。

Shikib MehriとMaxine Eskenazi。2020. Usr: 対話生成のための教師なしかつ参照不要な評価指標。arXiv preprint arXiv:2005.00456.

Shashi Narayan, Shay B Cohen, and Mirella Lapata.2018. 詳細は説明せず、ただ要約してくださいトピックを考慮した畳み込みニューラルネットワークの開発により、極端な要約を行う。自然言語処理における経験的方法に関する2018年会議論文集, ページ1797-1807.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022年。人間のフィードバックで指示に従うための言語モデルの学習。神経情報処理システムの進歩, 35:27730-27744.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu.2002. Bleu:機械翻訳の自動評価法。計算言語学会第40回年次大会予稿集, page 311-318.

コリン・ラフェル、ノーム・シャゼール、アダム・ロバーツ、キャサリン・リー、シャラン・ナラン、マイケル・マテナ、ヤンチー・ズー、ウェイ・リー、ピーター・J・リュー。2020. 統一的なテキストからテキストへの変換器を用いた転移学習の限界を探る。機械学習研究』21:1- 67.

エフード・レイター、アンジャ・ベルツ。2009. 自然言語生成システムを自動的に評価するためのいくつかのメトリクスの妥当性に関する調査。計算言語学, 35(4):529-558.

Amanda Stent, Matthew Marge, and Mohit Singhai.2005. バリエーションが存在する場合の生成のための評価方法の評価。第6回国際会議「計算言語学と知的テキスト処理」予稿集, 341-351 ページ.

Alex Wang, Kyunghyun Cho, and Mike Lewis.2020. 要約の事実上の一貫性を評価するための質問と回答。第58回計算言語学会年次大会予稿集, 5008-5020 ページ.

王建安、梁雲龍、孟範東、石浩祥、李志xu、徐晋、Qu建峰、周傑。2023. チャットグラプトは良い評価者なのか?arXiv preprint arXiv:2303.04048.

ウェイジェイソン、王雪司、ダル・シュールマンズ、マーティン・ボスマ、エド・チ、クオック・ル、デニー・ズー。2022. 思考連鎖プロンプトは大規模言語モデルにおいて推論を誘発する。arXiv preprint arXiv:2201.11903.

鄭麗、呂劉俊、黄石山、林亮、梁暁ダン。2021. 定量化可能な対話の一貫性評価に向けて第59回計算言語学会年次大会および第11回自然言語処理国際合同会議(第1巻:Long Papers)予稿集、2718-2729ページ。

袁偉哲、グラハム・ノイビッヒ、劉鵬飛。2021. バートスコア 生成されたテキストをテキスト生成として評価する。神経情報処理システムの進歩, 34.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi.2019. Bertscore。bertでテキスト生成を評価する。arXiv preprint arXiv:1904.09675.

張天義、ファサル・ラダック、エジン・デュルムス、パーシー・リャン、キャサリン・マッケウン、タツノリ・B.橋本. 2023. ニュース要約のための大規模言語モデルのベンチマーク.

趙偉、ピイヤード最上級、劉飛、高陽、ミヒャス・ミヒャル、エガー・ステッフェン。2019. ムーバースコア。文脈に応じた埋め込みと土手移動距離で評価したテキスト生成。自然言語処理における経験的方法に関する2019年会議および第9回自然言語処理に関する国際合同会議(EMNLP-IJCNLP)の議事録、ページ563-578にて。

丁 玲、丁 芳、毛 雲、佳 静、劉 鵬飛、朱 成陽、江 恒、韓 嘉偉。2022. テキスト生成のための統一された多次元評価器に向けて。arXiv preprint arXiv:2210.07197.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R Fabbri, Yejin Choi, and Noah A Smith. 2021. Bidimensional leaderboards: Generate and evaluate language hand in hand. *arXiv preprint arXiv:2112.04139*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, pages 341–351.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2718–2729.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

A Example Prompts

要約タスクにおける一貫性の評価

ニュース記事の要約を1つ渡します。

あなたの仕事は、1つの指標で要約を評価することです。

これらの説明をよく読んで、理解した上でください。この文書は、レビュー中は自由に保管し、必要に応じて参照してください。

Evaluation Criteria:

全文の集合的な品質に対する一貫性(1-5)。この次元は、構造と一貫性に関するDUCの品質問題と整合しており、「要約はよく構成され、よく組織されるべきである」と述べている。要約は単に関連情報の山であるだけでなく、文から文、トピックに関する首尾一貫した情報の体系へと構築されるべきである。”

Evaluation Steps:

1. ニュース記事を注意深く読み、主なトピックとキーポイントを特定する。

2. まとめを読み、ニュース記事と比較する。概要がニュース記事のメイントピックとキーポイントをカバーしているかどうか、そして、それらを明確かつ論理的な順序で提示しているかどうかを確認する。

3. 3. コヒーレンスのスコアを1～5で設定する(1が最低、5が最高)。

Example:

Source Text:

{{*Document*}}

Summary:

{{*Summary*}}

Evaluation Form (scores ONLY):

- *Coherence:*

対話生成タスクにおけるエンゲージメントの評価

あなたは二人の個人間の会話をされます。そして、次のターンの会話で、一つの可能な返答が与えられることになります。この回答は、興味深い事実に関するものであり、これについても提供される予定である。

あなたのタスクは、1つの指標で回答を評価することです。

これらの説明をよく読んで、理解した上でください。この文書は、レビュー中は自由に保管し、必要に応じて参照してください。

Evaluation Crieteria:

エンゲージメント（1-3）鈍い/面白い反応か?

スコアが1(dull)であれば、一般的な回答で鈍い回答であることを意味する。

2(やや面白い)のスコアは、その回答がやや面白いという意味で、会話に参加させることができる(例:意見、思考)。

3(興味深い)のスコアは、その回答が非常に興味深い、

Evaluation Steps:

A Example Prompts

Evaluate Coherence in the Summarization Task

You will be given one summary written for a news article.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic."

Evaluation Steps:

1. Read the news article carefully and identify the main topic and key points.

2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.

3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

Example:

Source Text:

{{*Document*}}

Summary:

{{*Summary*}}

Evaluation Form (scores ONLY):

- Coherence:

Evaluate Engagingness in the Dialogue Generation Task

You will be given a conversation between two individuals. You will then be given one potential response for the next turn in the conversation. The response concerns an interesting fact, which will be provided as well.

Your task is to rate the responses on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Crieteria:

Engagingness (1-3) Is the response dull/interesting?

- A score of 1 (dull) means that the response is generic and dull.

- A score of 2 (somewhat interesting) means the response is somewhat interesting and could engage you in the conversation (e.g., an opinion, thought)

- A score of 3 (interesting) means the response is very interesting or presents an interesting fact

Evaluation Steps:

1．会話、対応する事実、応答をよく読んでください。

2．上記の基準に従って、エンゲージメントの度合いを1～3で評価する。

または興味深い事実を示すことを意味します。回答と会話の特定の側面について、あなたの評価について簡単な説明をしてください。

Example:

Conversation History:

{{*Document*}}

Corresponding Fact:

{{*Fact*}}

Response:

{{*Response*}}

Evaluation Form (scores ONLY):

- Engagingness:

Evaluate Hallucinations

テキスト要約システムの人間による評価。

事実の一致。事実の一致:原文に裏付けられていない事実の要約は、真実でないか、誤解を招くか?

Source Text:

{{*Document*}}

Summary:

{{*Summary*}}

要約は事実の矛盾を含んでいるか?

Answer:

1. Read the conversation, the corresponding fact and the response carefully.

2. Rate the response on a scale of 1-3 for engagingness, according to the criteria above.

3. Provide a brief explanation for your rating, referring to specific aspects of the response and the conversation.

Example:

Conversation History:

{{*Document*}}

Corresponding Fact:

{{*Fact*}}

Response:

{{*Response*}}

Evaluation Form (scores ONLY):

- Engagingness:

Evaluate Hallucinations

Human Evaluation of Text Summarization Systems:

Factual Consistency: Does the summary untruthful or misleading facts that are not supported by the source text?

Source Text:

{{*Document*}}

Summary:

{{*Summary*}}

Does the summary contain factual inconsistency?

Answer: