SummEQuAL: Summarization Evaluation via Question Answering using Large Language Models

Junyuan Liu and **Zhengxiang Shi** and **Aldo Lipani**
University College London
{junyuan.liu.22,zhengxiang.shi.19,aldo.lipani}@ucl.ac.uk

Abstract

Summarization is hard to evaluate due to its diverse and abstract nature. Although N-gram-based metrics like BLEU and ROUGE are prevalent, they often do not align well with human evaluations. While model-based alternatives such as BERTScore improve, they typically require extensive labelled data. The advent of Large Language Models (LLMs) presents a promising avenue for evaluation. To this end, we introduce SummEQuAL, a novel content-based framework using LLMs for unified, reproducible summarization evaluation. SummEQuAL evaluates summaries by comparing their content with the source document, employing a question-answering approach to gauge both recall and precision. To validate SummEQuAL's effectiveness, we develop a dataset based on MultiWOZ. We conduct experiments on SummEval and our MultiWOZ-based dataset, showing that SummEQuAL largely improves the quality of summarization evaluation. Notably, SummEQuAL demonstrates a 19.7% improvement over QuestEval in terms of sample-level Pearson correlation with human assessments of consistency on the SummEval dataset. Furthermore, it exceeds the performance of the BERTScore baseline by achieving a 17.3% increase in Spearman correlation on our MultiWOZ-based dataset. Our study illuminates the potential of LLMs for a unified evaluation framework, setting a new paradigm for future summarization evaluation.

1 Introduction

Summary evaluation remains a complex task, and to this day, it cannot be adequately accomplished by automatic metrics (Chen et al., 2022; Goyal et al., 2022). While N-gram-based metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are widely used, they often show a poor correlation with human judgment, particularly in content assessment (Kasai et al., 2022a; Reiter and Belz, 2009). In addition, these methods rely on references as a "gold standard", which diminishes their effectiveness especially when assessing varied and abstract summaries due to the limited availability of reference texts. Summary evaluation has higher requirements for diversity and accuracy.

The rise of Large Language Models (LLMs) offers a promising way to evaluate generative texts more effectively, due to their understanding and reasoning abilities. While previous studies have been conducted in various contexts, such as machine translation Kocmi and Federmann (2023) and summarization (Chen et al., 2023; Fu et al., 2023; Liu et al., 2023) (without relying on reference summaries), the use of LLMs for evaluation has its limitations. Specifically, there are large variation across different prompts and LLMs, which complicates the use of a unified evaluation framework. While the question-answering (QA) approach offers a structured method for evaluation to mitigate this issue, existing QA approaches still have several limitations (Durmus et al., 2020; Manakul et al., 2023; Scialom et al., 2021; Wang et al., 2020): (1) previous works are more confined to direct answers, and answers based on reasoning or hidden information are often tricky; (2) models in generating questions lack focus so they may introduce irrelevant information; and (3) models need expensive pre-training and the performance will be influenced by the coverage and quality of training data.

To address these issues, we propose a novel unified framework for summarization evaluation *SummEQuAL*, which can effectively identify abstract information across a wide range of topics and perform complex inference. Specifically, SummEQuAL streamlines the evaluation process by breaking it down into separate tasks and conducts QA to approximate human-like evaluation. It guides LLMs with a structured schema to identify key information and employs a QA mechanism to complete the content evaluation, producing more re-

Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024), pages 46–55

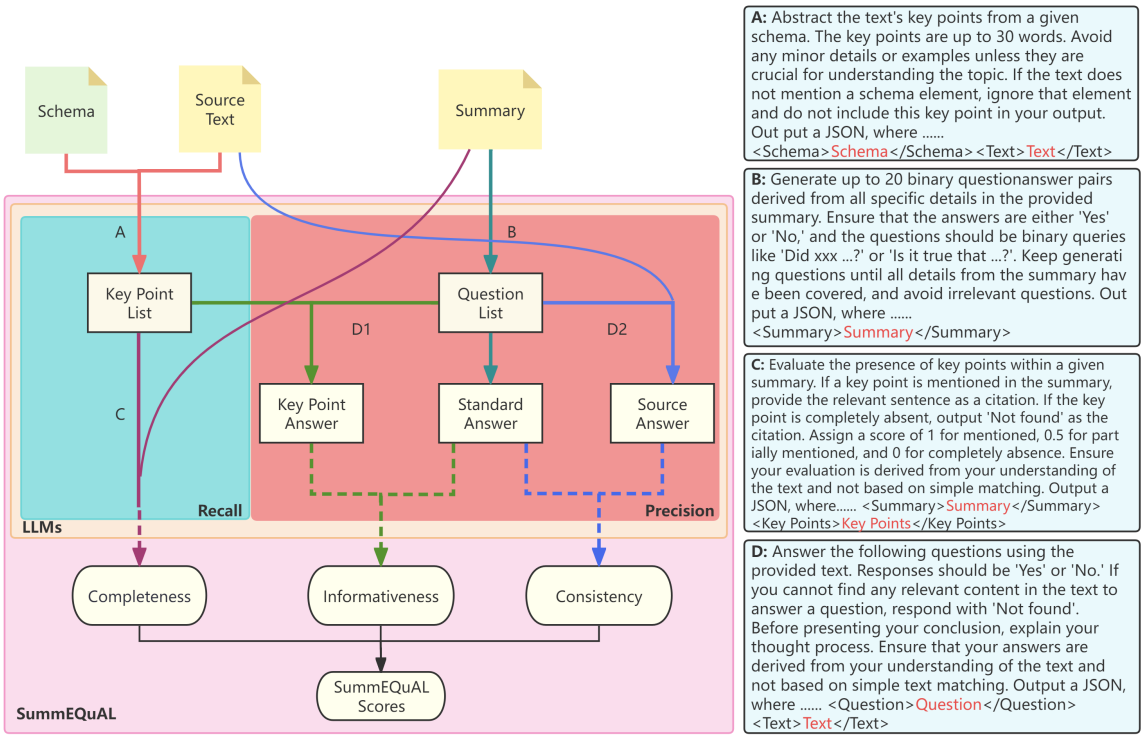

Figure 1: Illustration of the SummEQuAL framework. The blue and red areas represent the recall-oriented and the precision-oriented framework respectively. Solid colored lines and letters represent subtasks of LLMs. Other lines are simple comparison and calculation. The right side is an example of prompts for the subtasks.

liable results. Our experimental results on the SummEval dataset and MultiWOZ-based dataset demonstrate that the QA process benefits from LLMs. Particularly remarkable is that SummEQuAL improves the Spearman correlation with human evaluation by 19.7% on SummEval consistency, compared with our baseline QuestEval.

In summary, the main contributions of this paper are as follows:

1. We propose a unified summarization evaluation framework, SummEQuAL, for content correctness via QA using LLMs, minimising human works (§3).

2. We introduce a new dataset based on MultiWOZ to assess the quality of summarization content evaluation (§4.2).

3. As demonstrated in our experiments, our proposed framework, SummEQuAL, improves the baseline by 19.7% on the SummEval benchmark (§5.1.2) and 17.3% on the MultiWOZ-based dataset (§5.2.2).

4. Our further analysis reveals that LLMs can evaluate uniformly and robustly by designing a workflow with objective sub-tasks (§6).

2 Related Works

2.1 N-gram-based Evaluation Metrics

N-gram-based metrics measure the overlap between the generated summary and a reference summary. BLEU (Papineni et al., 2002) quantifies the concurrence of n-grams in a precision-oriented manner. ROUGE (Lin, 2004) assesses summarization from recall orientation. Since the above methods mainly focus on n-gram matching, they may fail to capture higher-level issues such as sentence structure, syntax, and semantics. Also, it can be affected by the repeated use of common phrases. As a result, they cannot accurately measure the content quality of a generated text (Reiter and Belz, 2009). Although many new metrics have appeared, nearly 70% of research works are still based on the old BLUE and ROUGE metrics (Kasai et al., 2022b).

2.2 Pre-trained Model-based Evaluation

Pre-trained model metrics are a category of methods that utilize pre-trained language models to evaluate. The language models can better capture semantic information and consequently assess the quality of generated text more accurately. BERTScore (Zhang et al., 2020) compares text em-

beddings, calculating similarity scores through the alignment of generated and reference summaries at a token level. MoverScore (Zhao et al., 2019) based on BERT considers the movement between words to evaluate the similarity between the generated text and the reference text. BARTScore (Yuan et al., 2021) regards evaluation as a text generation problem, which calculates the probability of a text generating or generated from other texts to evaluate.

2.3 QA-based Evaluation

A direct approach to compare the information of summaries and source documents is through the QA process. Previous research has explored this in fine-tuned language models. SummaQA (Scialom et al., 2019) generates questions from the source document and compares the confidence of the QA model in answering based on the summary, but fake answers are not considered. FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020), from another angle, generate questions from summarizations, using the original text for answering, to evaluate factual consistency. QuestEval (Scialom et al., 2021) subsequently applied QA in both directions, and MQAG (Manakul et al., 2023) introduced a multiple-choice QA model to deal with highly abstractive summarizations or multiple-answer spans.

2.4 LLM-based Evaluation

LLMs have been pre-trained on vast text data, enabling them to address a variety of domains without specific training. They can identify complex logical content. GPTscore (Fu et al., 2023) utilizes LLMs for zero-shot instruction and in-context learning, suggesting that higher-quality texts are generated with a higher likelihood. Incorporating CoT (Wei et al., 2022), G-EVAL (Liu et al., 2023) yields superior outcomes, allowing LLMs to autonomously generate evaluation methods. There's a bias observed favouring LLM-created content. Additionally, Chen et al. (Chen et al., 2023) determined that explicit LLM evaluations are more effective than implicit ones. In conclusion, while LLMs offer promising capabilities in text evaluation, challenges such as sensitivity, model bias, and score distribution bias remain.

3 SummEQuAL Framework

We propose a QA-based framework utilizing LLMs to evaluate summarization systems without needing human references for each summary. SummEQuAL incorporates both recall-oriented and precision-oriented approaches, featuring three submetrics, as illustrated in Figure 1. Below, we provide detailed explanations for each metric in our framework.

3.1 Completeness

Summarization aims to extract core information from texts to present the main content of the original text in a shorter form. A good summary should include as much important information from the original text as possible. Based on this concept, we first consider how much important information from the original text is included in the summary.

A *schema* module is introduced to facilitate question generation from the source texts. Rather than generating questions from the entire source texts, we utilize LLMs to generate a concise list of key points with the schema. The schema is a set of predefined important information for certain summarization tasks, which includes important information such as the main character of a story, the purpose of a product, or the time required to reserve a restaurant table, depending on the purpose and topic of the summarization. Experts can define schema according to their professional knowledge about the purpose of a summarization task. In this way, we do not need to create references for all summaries but only design one schema for one summarization task. Also, it is possible to generate the schema by LLMs with the description of the task topic. An example of schema for news is:

```
{
  "Time": "When does the event happen?",
  "Location": "Where does the event happen?",
  "Figure": "Who are the participants involved?",
  "Description": "What is the main event?",
  "Cause": "Why does the main event happen?",
  "Result": "What are the results of the event?"
}
```

LLMs generate the list of key points by the schema, and then compare whether these key points are included in the summary. The comparison is presented by a match function M, of which the value is between 0 and 1. A simple way is to answer boolean questions by LLMs like "Is it included in the summary that ...", and then transform it into numbers. Given a key point list K including n key points k_1 ... k_n, and a summary S, we define the *completeness* (Cons.) of a summary as:

$$\texttt{Comp.} = \frac{1}{n}\sum_{i=1}^{n} M(S, k_i), \tag{1}$$

3.2 Informativeness

A good summary should also avoid incorporating irrelevant information, and ideally, all the information within summaries should be of significance. In evaluation, we employ LLMs to generate question-answer pairs pertaining to the details of summaries. Subsequently, we prompt LLMs to answer questions using a key point list and then assess whether the answers to the two sets of questions match. To streamline the comparison process, we also generate boolean questions in the experiments of this paper. Denote the question list generated from summary S by LLMs as $Q(S)$, and q are questions in it. Given a text T and a question q, the answer generated by LLMs is $A(T, q)$. We define *informativeness* (Inf.) as:

$$\texttt{Inf.} = \frac{1}{|Q(S)|} \sum_{q \in Q(S)} \delta(A(S, q), A(K, q)), \quad (2)$$

where $|\cdot|$ represents the size of the set and $\delta(a, b)$ is equal to 1 when $a = b$ otherwise 0.

3.3 Consistency

Generative models could mistakenly create information that seems plausible but is not right, known as hallucinations. In the summarization task, this property introduces information outside the source document, thus damaging the consistency of the summary with the source document. To measure consistency, we use the summary to generate question-answer pairs, then answer the question by the source text with LLMs. After these, we compare the two sets of answers. For a given source document D, *consistency* (Cons.) is defined as:

$$\texttt{Cons.} = \frac{1}{|Q(S)|} \sum_{q \in Q(S)} \delta(A(S, q), A(D, q)) \quad (3)$$

3.4 SummEQuAL Score

SummEQuAL score is a comprehensive index combining completeness, informativeness, and consistency. The equation is structured to reflect the correct proportion of important and effective information. Completeness and informativeness are combined using the harmonic mean to balance the quantity and relevance of information. Consistency is then used as a multiplier to calculate the correction rate, reflecting the amount of accurate information in the summary. SummEQuAL score is computed as follows:

$$\texttt{SummEQuAL} = 2 \cdot \frac{\texttt{Comp} \cdot \texttt{Inf}}{\texttt{Comp} + \texttt{Inf}} \cdot \texttt{Cons} \quad (4)$$

4 Datasets

4.1 SummEval

SummEval (Fabbri et al., 2021) is one of the largest human annotated datasets for summarization evaluation tasks, built on the foundation of the CNN/-Daily Mail (Hermann et al., 2015) dataset. SummEval collects and releases both expert and crowdsourced human evaluation for 16 model outputs on 100 articles across 4 dimensions to advance research into human-correlated evaluation metrics. For each summary, there are 3 expert and 5 crowdsourced evaluations, totalling 12,800 human annotations. We will compare the correlation between LLMs' evaluations within the SummEQuAL framework and human evaluations in this dataset.

4.2 MultiWOZ

MultiWOZ (Budzianowski et al., 2018) is an open dataset released by the University of Cambridge, serving as a widely-used multi-domain task-oriented dialogue dataset with detailed human annotations for tracking dialogue information. As a task-oriented dataset, MultiWOZ offers objective schemas for specific tasks and annotated slot values for all dialogues. The dataset presents complex logic and multiple topics within individual dialogue texts, making it suitable for evaluating the framework's proficiency in working with a designed schema. We choose the newest version MultiWOZ 2.4 (Ye et al., 2022) for the experiment.

Dataset Construction Task-oriented summarization focuses on specific results. For booking tickets, we expect the summaries to fully contain information on user demands and the booking result. With this in mind, we created summaries on the MultiWOZ dataset and manually evaluated the completeness of the summaries. We randomly sample 50 dialogues from the test set of the MultiWOZ 2.4 dataset, and generate both short summaries and detailed summaries for each dialogue using the text-curie-001 model, GPT-3.5-turbo-0613 model, and GPT-4 model (OpenAI, 2023), totalling 300 summaries. The statistics of the dialogues and the summaries are shown in Table 1. Then, we annotated the summaries' completeness by checking how much proportion of the schema information was contained to provide a score. We used three annotators for marking. If the annotations by the first

Table 1: Statistics for dialogue and summary. The lengths are described by the number of words.

Dialogue		Average Summary Lengths	
Metric	**Value**	**Summarization**	**Length**
Dialogue Turn	6.98	gpt-3.5 simple	109.7
Mean Length	205.8	gpt-3.5 detailed	126.5
Min. Length	70	gpt-4 simple	69.0
Max. Length	470	gpt-4 detailed	89.9
Median Length	210	text-curie-001 simple	68.7
Std. Deviation	81.2	text-curie-001 detailed	72.7

two annotators were inconsistent, the third annotator adjudicated to determine the final annotation.

5 Experiments and Results

Evaluation Strategy. We evaluate automatic metrics by comparing their alignment with reference human evaluations. Three prevalent correlations, the Spearman correlation, the Person correlation and the Kendall's Tau correlation are employed. Given n source texts and m summary models, the i-th text's summary generated by the j-th model is denoted as $s_{i,j}$. The formula of correlation at a sample level is as follows:

$$\text{Corr} = \frac{1}{n}\sum_{i=1}^{n}\rho\left([e_{\text{auto}}(s_{i,1}),\ldots,e_{\text{auto}}(s_{i,M})],\right. \\ \left.[e_{\text{ref}}(s_{i,1}),\ldots,e_{\text{ref}}(s_{i,M})]\right) \quad (5)$$

where ρ denotes the function of correlation metrics. e_{auto} and e_{ref} denote the automatic evaluation and reference evaluation functions, respectively.

5.1 Evaluation on SummEval

5.1.1 Experimental Setup

We evaluate the summaries in SummEval dataset with our SummEQuAL framework and compare the evaluation results with human evaluations. In this paper, if not specified, we use the GPT-3.5-turbo-0613 model with a temperature of 0 as the LLM of our framework, since other LLMs do not always follow the instructions strictly to work with the SummEQuAL framework. To avoid confusion across texts, each step of our framework separately inputs data into the LLM and is linked by processing output JSON. The human evaluation is set using the average scores from three experts. The determined input schema is the same as the example schema shown in Section 3.1. To make the evaluation for our framework convincing, we also use a rephrased version of the prompts besides the example prompts in Figure 1 and calculate the average correlation.

5.1.2 Results

LLMs have demonstrated competitive results within the SummEQuAL framework in Table 2. On the aspect of consistency, SummEQuAL has shown a clear superiority over traditional automatic metrics, improving by 19.7% compared with the best-performing QA models and more than 10% compared with G-EVAL. SummEQuAl benefits both from LLMs and QA process, effectively verifying whether the information in the summary is consistent with the original text.

In terms of relevance, the correlations are high on G-EVAL and QuestEval. The primary reason for not outstanding results lies in the differences in the definitions of relevance. SummEQuAL scores does not precisely reflect the definition of relevance in SummEval, which emphasizes selecting important content from the source without stressing correctness. In contrast, G-EVAL directly prompt the definition of relevance. QuestEval, when utilizing summaries to answer questions generated from the original text, relies on the answerability confidence instead of an answer, thus aligning more closely with the definition of relevance in SummEval.

5.2 Task-oriented Summarization Evaluation

Task-oriented summarization requires a focus on specific information within the text. For instance, in ticket booking dialogues, details about the ticket such as time and location are crucial, while in a doctor's diagnostic interview notes, the patient's symptoms and feelings are of importance. To accurately evaluate the effectiveness of task-oriented summarization, a well-defined schema that lists all the relevant information is helpful. SummEQuAL can then be used to provide a desirable evaluation for the tasks, ensuring that the summarization meets the specific requirements. To test the ability of the SummEQuAL framework on task-oriented summarization evaluation, we build a MultiWOZ-based dataset, which contains dialogues between users and conversational systems, involving a range of tasks, such as restaurant reservations, travel bookings, information queries, and so forth, and then we conduct experiments with the SummEQuAL framework.

5.2.1 Experimental Setup

A dictionary with the 35 types of slots tracked in the dialogue dataset is set as the schema for generating key points. Within the SummEQuAL framework, we input schemas into both the GPT-3.5 and

Table 2: Sample-level Spearman correlation (Spear.), Pearson correlation (Pear.), and Kendall's Tau correlation (Kend.) of relevance and consistency on SummEval. #Ref. is the number of reference summaries in evaluation. Results of QA metrics are from Scialom's work (Scialom et al., 2021).

Metrics	#Ref.	Relevance			Consistency		
		Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	1	0.199	0.220	0.152	0.157	0.200	0.132
ROUGE-2	1	0.145	0.174	0.105	0.137	0.162	0.116
ROUGE-L	1	0.203	0.221	0.156	0.149	0.198	0.125
ROUGE-1	11	0.311	0.347	0.237	0.153	0.216	0.125
ROUGE-2	11	0.248	0.298	0.189	0.122	0.181	0.101
ROUGE-L	11	0.293	0.329	0.224	0.103	0.180	0.082
BERTScore	11	0.269	0.304	0.203	0.168	0.242	0.140
BARTScore	11	0.264	0.290	0.197	0.311	0.321	0.256
MoverScore	11	0.282	0.313	0.215	0.166	0.221	0.137
SummaQA	0	–	0.262	–	–	0.083	–
QAGS	0	–	0.204	–	–	0.091	–
QuestEval	0	–	**0.392**	–	–	0.420	–
G-EVAL-3.5	0	**0.385**	–	**0.293**	0.386	–	0.318
SummEQuAL	0	0.311	0.337	0.241	0.274	0.324	0.227
-Completeness	0	0.252	0.274	0.204	0.151	0.182	0.131
-Informativeness	0	0.181	0.187	0.143	0.161	0.188	0.136
-Consistency	0	0.228	0.265	0.193	**0.432**	**0.503**	**0.403**

GPT-4 models, generating two lists of key points. To evaluate the LLMs' capability to generate key points based on the schema, we manually compared these generated lists against the slot values annotated in MultiWOZ. Considering the target of this ticket-booking summarization, only slot values in the final turn of user-system dialogues are extracted as the ground truth. After this, we use a simple match, ROUGE-L, BERTscore and LLMs to compare the key point list with the source document (C in Figure 1) and generate a completeness score. We compared these completeness scores with human scores to demonstrate whether this framework benefits from LLMs. Last but not least, we compare the performance of different evaluation metrics like ROUGE, BERTScore and BARTScore on the 300 summaries consisting of simple and detailed summaries generated by the text-curie-001, the GPT-3.5-turbo-0613 and GPT-4 models. Based on our observation, the GPT-4 model is better than the GPT-3.5 model, and both outperform the text-curie-001 model; the detailed version is better than the simple version of the same model. So we assigned them scores from high to low as ground truth. Finally, we calculated the correlation coefficients of metrics and assigned scores accordingly.

5.2.2 Results

As shown in Table 3, both models can extract key points according to the schema with relatively good results, highlighting the flexibility and adaptability of the SummEQuAL framework. If the schema is given, GPT-3.5 with relatively weak reasoning ability can obtain high recall. Compared with GPT-4, the precision of GPT-3.5 is lower. We checked the output and found this is because some unimportant information outside the schema is introduced, which is often repeated information or redundant content as a result of not correctly following the schema.

Table 3: Key Point Generation Abilities: Scores are derived by comparing model predictions with the last dialogue turn's state value manually.

Model	Recall	Precision	F1
GPT-3.5	0.919	0.758	0.830
GPT-4	0.958	0.875	0.915

Table 4: Key Point Comparison Abilities: Model correlation with human on comparison of key point and summary.

Model	Spearman	Pearson	Kendall Tau
Simple match	0.298	0.229	0.211
ROUGE-L	0.328	0.282	0.232
BERTScore	0.644	0.634	0.476
GPT-3.5	0.742	0.756	0.599
GPT-4	**0.929**	**0.828**	**0.946**

Table 4 shows that the evaluation of the SummEQuAL framework benefits from the reasoning ability of LLMs. The comparison based on LLMs is the most consistent with human evaluations, surpassing

Table 5: Sample-level Correlations of Summarization Evaluation on Our MultiWOZ-based Dataset.

Metrics	Spear.	Pear.	Kend.
ROUGE-1	-0.010	0.077	-0.007
ROUGE-L	0.003	0.079	0.001
BERTScore	0.589	0.657	0.489
BARTScore	0.423	0.519	0.317
MoverScore	-0.137	-0.360	-0.109
QuestEval	0.511	0.568	0.429
SummEQuAL	0.240	0.283	0.192
-Completeness	**0.691**	**0.764**	**0.573**
-Informativeness	0.212	0.257	0.161
-Consistency	-0.028	-0.058	-0.031

other methods significantly. Upon careful observation of the results, we discovered that GPT-3.5 tends to make errors and give lower scores, which suggests that in complex scenarios, GPT-3.5's reasoning ability may not be sufficient to capture all the details of the summaries. We will discuss this in the analysis section (§6) in detail.

By the comparison in Table 5, the completeness of SummEQuAL has achieved the best correlation, but the correlations of other parts are not high. Metrics based on pre-trained language models perform well, but ROUGE performs poorly. This result is from the generation of summarization. The GPT models tend to cover more content when summarizing on MultiWOZ, so the summarization contains comprehensive important information but is not concise. Moreover, since both GPT-3.5 and GPT-4 have good summary capabilities for the MultiWOZ data set, the actual summarization quality of GPT-3.5 is not lower than GPT-4 in some cases, especially considering the consistency. As a result, the assigned scores are not completely accurate, which affects the overall correlation coefficient.

6 Further Analysis

6.1 Discrepancy of Summarization Tasks

The evaluation result is affected by the dataset's features, the evaluation criteria and the evaluation target. Articles in SummEval are longer and contain more detailed descriptions, making it easier for summarizations to introduce non-essential information and produce inconsistent information, while the dialogues are shorter. Moreover, we need to consider the practical meaning of the comparison scores. The dialogues in MultiWOZ are based on specific tasks, and their themes and key information are more clearly defined. In the case of providing a schema reflecting the content of the task as a reference, the completeness score is more consistent with the target of evaluation. When using SummEQuAL for evaluation tasks, we can clarify the purpose of the task and use a more appropriate schema and metric or combination of metrics for a better evaluation.

6.2 Error Analysis

Inferencing capability. As illustrated in Table 6, the dialogue contains implicit shifts of needed information. GPT-3.5 only capture the user's initial request for a restaurant serving canapes and did not correctly comprehend the system's rejection of the user's proposed dining time of 16:15. This comparison underscores the importance of inferencing abilities of the model's reasoning capabilities. Similar errors could also occur in the evaluation of informativeness and consistency.

Subjectivity. Evaluation on the SummEval dataset is influenced significantly by subjectivity. The evaluation results can vary depending on the chosen schema. Our schema may focus on different information from the human annotation in SummEval. Consequently, the generated key points do not align with information in the reference summaries used by human experts during evaluation, resulting in a potential distortion of the summarization evaluation.

Stability. When generating key points, GPT-3.5 could generate duplicate information or possible schema values that do not appear in the summary, causing the SummEQuAL framework to fail. During the entire evaluation process, GPT-3.5 is occasionally affected by the text content, so that it is unable to correctly execute the framework tasks according to the instructions, and outputs wrong or incomplete content.

6.3 Reproducibility

An inherent characteristic of LLMs, especially when interfaced through APIs, is the potential variability in their outputs upon repeated experiments. We evaluate the potential difference level of this issue by repeating the experiment. Here we use the GPT-3.5-turbo as the base model of the SummEQuAL framework, and we experimented on the SummEval benchmark. Table 7 shows that the majority of the results across various metrics fell within the minimal absolute difference percentage

Table 6: Case Study: Comparison of GPT-3.5 and GPT-4 for Key Points Generation. Table (top) provides a brief overview of the dialogue and Table (bottom) summarizes both the actual key points and the predictions. Key points highlighted in green indicate correct summarization, while those in red represent incorrect predictions by the model.

Role	Dialogue
User:	I am looking for a restaurant that serves canapes in the east.
Sys:	Unfortunately there are no restaurants serving canapes in the east.
User:	Ah, well, too bad. In that case, I think that'll be everything that I needed. Thanks and have a good day!
Sys:	Are you sure? I can find other options in other parts of town?
User:	How about Italian food?
Sys:	There is the Pizza Hut Fen Ditton in the east serving Italian food.
User:	Great! Please book a table for 6 at 16:15 on Saturday.
Sys:	Sorry, but no tables are available for that time slot. Would you like to change the time?
User:	How about 15:15 then?
Sys:	Your table is booked. Your reference number is qw8jzwzk. Can I help you with anything else?
User:	Great. Thank you for your help today. That is all.

Key Points (Name)	Slot Values (Truth)	GPT-3.5 (Prediction)	GPT-4 (Prediction)
restaurant-food	italian	canapes	italian
restaurant-area	east	east	east
restaurant-book day	saturday	saturday	saturday
restaurant-book people	6	6	6
restaurant-book time	15:15	16:15	15:15
restaurant-name	pizza hut fenditton	pizza hut fen ditton	pizza hut fen ditton

Table 7: Comparison of Repeated Experiment Results

Metrics	Abs Mean	Variance	Abs Difference Percentage				
			0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	>0.4
S.E.	0.050	0.010	81.31%	11.75%	4.31%	1.69%	0.94%
Comp.	0.052	0.012	78.94%	11.69%	5.69%	2.38%	1.31%
Inf.	0.056	0.016	78.75%	10.19%	5.62%	3.81%	1.62%
Cons.	0.018	0.004	92.31%	4.56%	2.06%	0.75%	0.31%

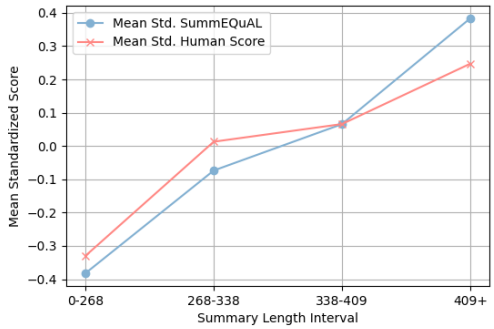

Figure 2: Standardized SummEQuAL and Human Scores by Summary Length Interval

range (0-0.1), indicating a high degree of reproducibility. Among all the metrics, the consistency score shows the smallest discrepancy.

6.4 Comparison of Text Length

We conduct a comparative analysis with human scores to evaluate the possible bias of our SummEQuAL score across various text lengths. First, we scale the SummEQuAL and human scores using z-score normalization to ensure comparability. We then grouped the summaries into four equal-sized intervals based on length and computed the mean standardized score for both SummEQuAL and human evaluations within each group. Figure 2 indicates a general trend where both SummEQuAL and human scores increased with the length of the text. The apparent preference in the SummEQuAL model for longer texts could result from the fact that longer summaries are better in the SummEval benchmark. The overall parallel trends in SummEQuAL and human scoring across different text lengths demonstrate a degree of consistency.

7 Conclusion

This work proposes a novel summarization evaluation framework, SummEQuAL, based on LLMs. The SummEQuAL framework provides a reliable and effective approach for the evaluation of summarization, opening up a new direction for future work on unified and reproducible summarization evaluation using LLMs. SummEQuAL sheds new light on developing reliable and consistent summarization evaluation methods, expected to help researchers more precisely understand and evaluate the performance of summarization models, thereby improving the quality of summarization content.

Limitations

SummEQuAL's performance depends on the capabilities of LLMs. Any limitations these LLMs have, especially in parsing complex logic or discerning implicit information, will directly influence SummEQuAL's evaluation. The framework's reliance on multi-step reasoning, involving several interactions, means it can be time-intensive and resource-heavy compared to simpler one-step evaluations. Moreover, although SummEQuAL has shown promise in initial tests, its effectiveness across different models, languages, and text domains still needs further evaluation.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *ArXiv preprint*, abs/2304.00723.

Yulong Chen, Naihao Deng, Yang Liu, and Yue Zhang. 2022. DialogSum challenge: Results of the dialogue summarization shared task. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 94–103, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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*, abs/2302.04166.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv preprint*, abs/2209.12356.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022a. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022b. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *ArXiv preprint*, abs/2302.14520.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Mqag: Multiple-choice question answering and generation for assessing information consistency in summarization. *ArXiv preprint*, abs/2301.12307.

OpenAI. 2023. Openai platform models documentation. Accessed: 2023-10-02.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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, Hong Kong, China. Association for Computational Linguistics.