EVA-Score: Evaluating Abstractive Long-form Summarization on Informativeness through Extraction and Validation

Anonymous ACL submission

Abstract

Since LLMs emerged, more attention has been paid to abstractive long-form summarization, where longer input sequences indicate more information contained. Nevertheless, the automatic evaluation of such summaries remains underexplored. Current evaluation metrics either use traditional metrics like ROUGE and BERTScore, which rely on surface-level similarity and fail to consider informativeness, or simple LLM-based metrics, which are not robust and easily overwhelmed by the long contexts. In this paper, we propose a new evaluation metric called EVA-SCORE to extract all information from the given summaries, identify overlapped information based on reference, and calculate the information score. We test EVA-SCORE on several datasets and the experimental results reveal that EVA-SCORE shows the highest correlation with humans. We also re-evaluate the performance of LLMs on long-form summarization from the information perspective. The results indicate that responses of LLMs still have a gap with the human-written answers. Moreover, we provide a detailed analysis of the effectiveness of EVA-SCORE, forecasting future ways to automatically evaluate abstractive long-form summarization.

1 Introduction

Long-form summarization is challenging due to its extensive knowledge and lengthy context. It can be categorized into extractive (Zhang et al., 2023) and abstractive (Wang et al., 2023b), where the former retains exact content from the document, while the latter produces a paraphrased and coherent output. With advancements in LLMs (OpenAI et al., 2024; Touvron et al., 2023), significant progress has been made (Chang et al., 2024; Chhikara et al., 2024; Kim et al., 2024). However, the automatic evaluation of abstractive long-form summarization remains underexplored and is more challenging than evaluating extractive ones due to the implicit equivalence of information involved. Therefore, an information-sensitive metric is needed to effectively capture the depth and breadth of summaries.

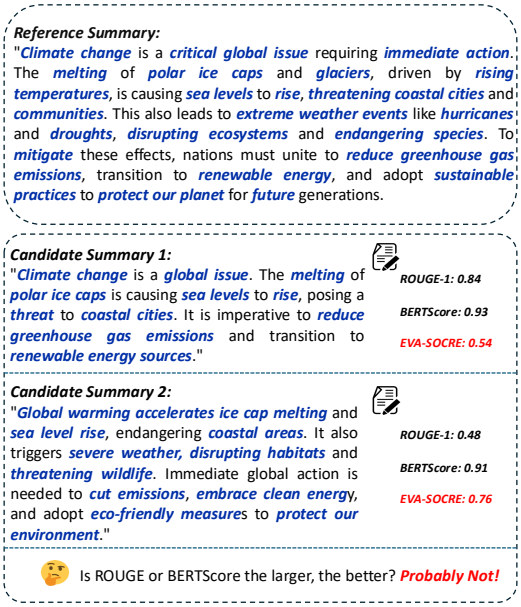

Figure 1: The example shows the dilemma of traditional evaluation metrics, including ROUGE and BERTScore. *Candidate Summary 1* and *Candidate Summary 2* are summaries to be evaluated. The information within the summary is denoted using blue. Even though *Candidate Summary 1* is rewarded with a higher ROUGE and BERTScore, it contains less information and is given a lower EVA-SCORE which meets human preference. It inspires us that traditional metrics are not reasonable for evaluating long-form and knowledge-intensive summaries due to their insensitivity to information.

Past evaluation metrics for abstractive summarization include some static evaluation methods, like ROUGE-k and ROUGE-L (Lin, 2004). With the development of pre-trained language models, BERTScore (Zhang et al., 2020) is widely used as an evaluation metric based on word embeddings to measure the similarity between the reference and the candidate. Also, some works use LLMs to eval-

uate long-form summarization, defining criteria, and prompting LLMs to assign scores for different aspects like fluency and consistency. For clarity, we categorize metrics that utilize prompts and obtain responses from LLMs as LLM-based evaluation metrics (Kim et al., 2024; Chang et al., 2024), while other metrics like ROUGE and BERTScore are referred to as traditional metrics. Each category, however, has its limitations and challenges, especially when applied to complex, meticulous content in abstractive long-form summarization.

Traditional metrics, such as ROUGE and BERTScore, are primarily based on surface-level similarity—either word-level or embedding-level similarity to the reference (Krishna et al., 2021). Therefore, they often fail to capture information-level similarity, which involves understanding the actual meaning or key ideas conveyed, making them less suitable in the era of LLMs, where responses are highly flexible and often use synonyms or different syntactic structures. We show an example in Figure 1, indicating the inherent drawback of ROUGE and BERTScore. On the other hand, LLM-based Metrics struggle with capturing minor information differences between candidates and references. Also, their evaluations can be unreliable, even with detailed evaluation criteria in such a complex scenario, where LLMs must first identify key information from the candidate and then assess it against the reference (Shen et al., 2023).

To fill the gap, we propose a new evaluation metric called **E**xtraction-and-**Va**lidation Score (EVA-SCORE), designed to provide an objective and quantitative analysis while capturing the full range of information in a summary. Our approach begins by extracting atomic facts from reference and candidate. Unlike FACTSCORE (Min et al., 2023), we recognize that certain pieces of information are context-dependent, resulting in some extracted facts not being atomic but cascaded between them instead. To address this, we reformat these atomic facts into logic chains that preserve contextual integrity. Additionally, since atomic fact extraction operates on a sentence-level granularity, it often misses document-level relationships. For complement, we incorporate Document-Level Relation Extraction (Xue et al., 2024; Xu et al., 2022) to capture the hierarchical relations and provide a more complete understanding of the content. To verify the accuracy of each atomic fact, we identify the most relevant context from the reference and prompt LLMs to determine whether the candidate information is entailed by the reference. In our implementation, we ensure that only one piece of information is evaluated at a time, which simplifies the validation and enhances reliability. The overall evaluation pipeline is illustrated in Figure 2.

To the best of our knowledge, we are the first to evaluate long-form summarization from the informativeness perspective. We conduct comprehensive experiments to compare EVA-SCORE with existing evaluation metrics, including both traditional and LLM-based metrics. The results reveal that EVA-SCORE achieves the highest alignment with human evaluations, significantly outperforming past metrics. At the system level, EVA-SCORE demonstrates a strong correlation with human judgments, achieving a Pearson correlation of **0.975** and perfect scores of **1.0** in both Spearman and Kendall correlations. Additionally, we re-evaluate long-form summarization performance based on EVA-SCORE. The results highlight the impressive capability of GPT-4, especially as the context length increases, where it ranks first on two test datasets and second on the remaining two. We also analyze why certain LLMs fail on specific datasets and provide insights into scenarios where EVA-SCORE aligns with human annotations versus cases where discrepancies occur. These analyses shed light on the future development of automatic evaluation metrics for abstractive long-form summarization.

2 Related Work

2.1 Abstractive Long-form Summarization

With the advent of models like ChatGPT and GPT-4 (OpenAI et al., 2024), LLMs have demonstrated considerable potential in traditional summarization tasks (Lu et al., 2023; Zhang et al., 2023; Wang et al., 2023a). However, as context length grows, LLMs still face significant issues (Bai et al., 2023). Recent research has explored methods to extend the capability of LLMs to handle longer contexts, which meanwhile facilitates long-form summarization (Su et al., 2023; Li* et al., 2023). Long-form summarization requires language models to understand the purpose of the context and paraphrases core information from extensive text which can be categorized into extractive and abstractive summarization, with the latter providing a more fluent and flexible summary that better aligns with human preferences compared to extractive methods. Therefore, many studies focus on abstractive

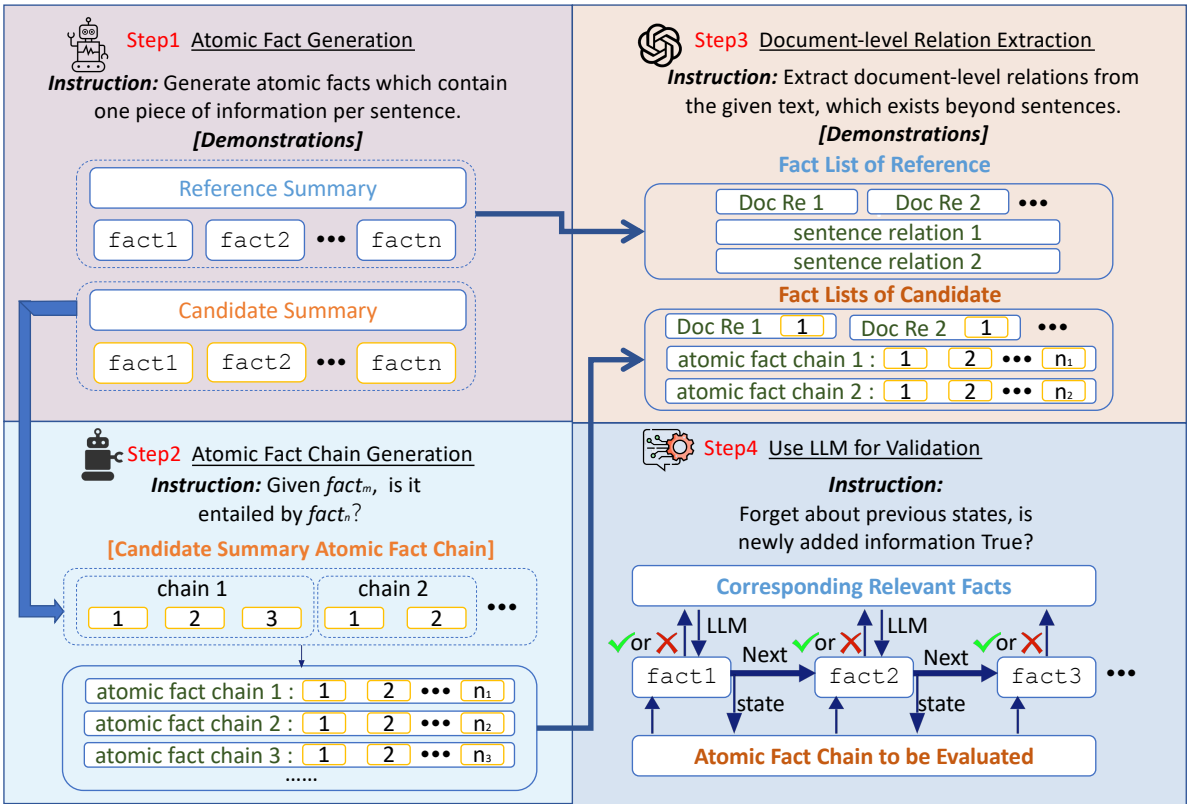

Figure 2: Pipeline of EVA-SCORE. We begin with **Atomic Fact Generation**. For extracted facts, we use LLMs to check whether they can be treated as a logic chain and use **Document-level Relation Extraction** as a complement. "Doc Re" in the figure is actually "Document-level Relation" for abstraction. When analyzing the truthfulness of a given fact, we retrieve and find several reference facts and provide the evaluation results of previous facts within a logic chain for help. Thus, we ensure that one piece of information is analyzed once.

long-form summarization in different domains, including news reports (See et al., 2017), government reports (Cohan et al., 2018), and books (Kryściński et al., 2022). Despite these efforts, abstractive summarization for book-level and report-level texts still remains a challenging task (Chang et al., 2024).

2.2 Evaluation of Summarization

Tradition evaluation metrics for summarization include ROUGE and its variants, and BERTScore (Zhang et al., 2020). However, (Krishna et al., 2021) highlights that ROUGE is not such an informative metric, implying its drawback for abstractive long-form summarization. Similar challenges apply to BERTScore. Currently, several LLM-based evaluation metrics (Zheng et al., 2024; Ke et al., 2023) have been proposed to aligning model evaluation with human evaluation. Some works apply LLMs to the evaluation of long-form summarization, for example, Luo et al. (2023) uses ChatGPT to evaluate factual inconsistency for text summarization and Song et al. (2024) leverages LLMs to evaluate summaries based on faithfulness, completeness, and conciseness. However, we argue that only using the inherent ability of LLMs is not enough to measure the quality of a summary due to the inferior ability of LLMs to differentiate between summaries with nuanced information. We are seeking to find a reasonable yet explainable method for the evaluation of long-form summarization.

3 Preliminaries

In this section, we provide definitions for terms used in our evaluation pipeline.

Our work begins with extracting atomic facts from the reference and the candidate which we denote as Sum and as $\hat{Sum}$ respectively. The extracted atomic fact lists are S and $\hat{S}$ with s_i and $\hat{s}_i$ being each atomic fact in the list. After **Atomic Fact Chain Generation**, we obtain fact chain lists $\hat{C}$ from $\hat{S}$, while each element in it is defined $\hat{C}_i$

and each atomic fact in logic chains is defined as $\hat{C_{i_j}}$. Finally, we conduct **Document-level Relation Extraction** to complement the original atomic fact list with hierarchical information. The final lists are $DocS$ and $Doc\hat{S}$ separately and the elements in them are denoted as D_i and $\hat{D_{i_j}}$.

4 Methods

In this section, we introduce the detailed implementation of EVA-SCORE to evaluate informativeness in long-form summarization.

4.1 Atomic Fact Generation

In any given sentence within a model prediction, multiple pieces of information may coexist. Consequently, evaluating sentences as a single unit renders the true-or-false outcomes, which are inadequate for assessing the quality. Min et al. (2023) finds that splitting the prediction into atomic facts, defined as concise sentences consisting of a single piece of information, enables a more fine-grained evaluation and improves the evaluation by a large margin. Following their work, we use ChatGPT with the same prompt to generate atomic facts from both Sum and $\hat{Sum}$ and obtain S and $\hat{S}$. We share an example in Appendix A.1. The prompt used for **Atomic Fact Generation** is displayed in Appendix C.1. The temperature is set to 0.0 to ensure the effectiveness of demonstrations.

4.2 Atomic Fact Chain Generation (AFCG)

Although claimed atomic, certain information, like time and location, cannot be independently extracted due to its reliance on contextual semantics. Consequently, parts of $\hat{S}$ may appear as cascaded atomic fact lists, where subsequent facts incorporate or entirely encompass information from preceding facts. In this case, if any of the preceding facts are incorrect, all subsequent facts are deemed wrong, even if the newly added information is accurate. Since we try to measure the quality of a summary from the information perspective, the former incorrect information should not affect the latter. To mitigate the impact, we reform the atomic facts into several logic chains so that in one logic chain, when analyzing a certain atomic fact, we can explicitly focus on newly added information. We show an example of *AFCG* in Appendix A.2. In particular, *AFCG* applies only to $\hat{S}$ since all the information in Sum is considered ground truth. Therefore, the cascade effect does not exist.

We use Mistral-7B-instruct (Jiang et al., 2023a) in our work to find logical relationships between atomic facts, treating it as an NLI task between one fact and the exact next fact, due to the special structure of $\hat{S}$ (Min et al., 2023), and reforming atomic facts as a list of logic chains. The prompt used for *AFCG* is listed in Appendix C.2. The temperature is set at 1.0 to encourage more divergent responses.

4.3 Document-level Relation Extraction

Extracting atomic facts from text mainly captures sentence-level relations—those between entities within the same sentence. However, document-level relations, involving entities across an entire document or beyond sentence boundaries, are crucial for understanding contextual logic. To address this, we implement **Document-level Relation Extraction** (DocRE) using GPT-4, starting with **Named Entity Recognition** (NER) followed by in-context learning to extract document-level relations from Sum and $\hat{Sum}$. The outputs are tuples $[(h_1, r_1, t_1), (h_2, r_2, t_2), \dots]$ and $[(\hat{h_1}, \hat{r_1}, \hat{t_1}), (\hat{h_2}, \hat{r_2}, \hat{t_2}), \dots]$, representing head entities, relations, and tail entities respectively. The temperature for *NER* and *DocRE* is set at 1.0 to ensure diverse outputs. Appendix A.3 provides an example illustrating differences between sentence-level and document-level relations, and Appendices C.3 and C.4 show the prompts for *NER* and *DocRE*, respectively. Finally, the triples are paraphrased into natural language for post-processing.

To prevent the inclusion of sentence-level relations, we use cosine similarity based on BERT (Devlin et al., 2019) to filter out overlapping information between the atomic fact list and the document-level relation list. To determine the threshold, we randomly sample a subset from the datasets listed in Section 5.1, test various thresholds, and select the optimal one, as detailed in Appendix B.2. We ultimately choose a threshold of 0.65. The filtered relations are then incorporated into C and $\hat{C}$ to form $DocS$ and $D\hat{o}cS$ where $DocS$ includes atomic facts from Section 4.1 and document-level relations, whereas $D\hat{o}cS$ comprises atomic fact chains from Section 4.2 and document-level relations with each relation as a single-element list.

4.4 LLM for Validation

For each atomic fact $\hat{D_{i_j}}$ in $D\hat{o}cS$, we retrieve k most relevant atomic facts according to BERTScore (Zhang et al., 2020) from $DocS$, using $\hat{D_{i_j}}$ as hypothesis and each D_i as the reference. k is a hy-

perparameter which is a trade-off between effectiveness and efficiency, and we use $k = 3$ in our experiment settings. We use Mistral-7B-instruct as our validation model with a temperature equal to 1.0. The matching process between two statements is represented using $f(\cdot)$. By providing the correctness of previous statements and prompting it not to pay attention to them but to focus on newly added information, we ensure that one piece of information is checked each time. We provide an example in Appendix A.4 and the prompt used for LLMs is listed in Appendix C.5.

We calculate our Precision and Recall as follows:

$$\text{Precision} = \frac{1}{|Doc\hat{S}|} \sum_{i=0}^{n} \mathbb{1}\left(f(\hat{D_{i_j}}, D_i)\right) \quad (1)$$

$$\text{Recall} = \frac{1}{|DocS|} \sum_{i=0}^{n} \mathbb{1}\left(f^{-1}(D_i, \hat{D_{i_j}})\right) \quad (2)$$

where n is the number of atomic facts in $Doc\hat{S}$. We define EVA-SCORE as the harmonic mean of Precision and Recall.

5 Experiment

In this section, we detail our implementations and the results of the experiments.

5.1 Dataset

To evaluate the applicability of EVA-SCORE, we gather long-form summarization across varying context lengths and domains. We adopt **CNN Dailymail** (See et al., 2017), **PubMed**, **GovReport** (Cohan et al., 2018), **arXiv**, and **BookSum** (Kryściński et al., 2022). In particular, we select the chapter-level summarization version of **BookSum** for less difficulty. For all datasets, we use validation datasets for evaluation. We list a detailed description of the dataset in Appendix B.1.

5.2 Baseline

We compare the results of EVA-SCORE with traditional evaluation metrics, i.e., ROUGE-1, ROUGE-2 (Lin, 2004) and BERTScore (Zhang et al., 2020) which are based on text overlap or word embeddings. For LLM-based evaluation metrics, we include ChatGPT and GPT-4-0125-preview as our evaluators since they are always considered one of the strongest LLMs, and we denote these methods as prompt-only metrics. Moreover, we validate on fine-tuned LLMs including AUTO-J-13B (Li et al., 2023) and CRITIQUE-LLM-6B (Ke et al., 2024).We denote these two metrics as fine-tuned metrics. For ChatGPT and GPT-4, we instruct them using the prompt from Table 17. For all the LLMs, we set temperatures at 1.0.

5.3 Human Annotation

We aim to develop an evaluation metric that resonates with human judgment. To achieve this, we randomly select 50 records from the datasets mentioned in Section 5.1 and assess the informativeness of each record. We employ five annotators, including some of the authors, all of whom are either M.S. or Ph.D. students. Initially, the annotators identify atomic facts at the sentence and document level, denoted as $DocS$ and $\hat{DocS}$, from Sum and $\hat{Sum}$. They then determine the overlap between $DocS$ and $\hat{DocS}$. Redundant information, which appears multiple times, is recorded only once. Subsequently, they calculate Precision and Recall for these overlaps and derive the F1 Score as their harmonic mean. Furthermore, we ensure that each record is evaluated by at least two annotators to maintain the fairness of the assessment. Each annotation costs about 15 minutes. The high time cost implies the complexity and high quality of our annotation. The detailed annotation recipe is listed in Appendix B.3. After the annotation process, we assessed the quality by measuring the agreement between the annotators. We achieve an Intraclass Correlation Coefficient (ICC) of **0.999** using a two-way random-effects model, and a Pearson correlation of **0.814**. These high correlations demonstrate the robustness and reliability of our annotations.

5.4 Main Results

In this section, we report the correlation between each evaluation metric and human annotations. We adopt text-level and system-level Pearson (σ), Spearman (γ), and Kendall (τ) correlation. Text-level correlation is calculated by computing the correlation coefficients between human evaluations and automatic evaluation metrics and then averaging these coefficients for all generated texts. System-level correlation, on the other hand, is determined by computing the correlation coefficients between human evaluations and automatic evaluation metrics for the average score of each dataset. The results are listed in Table 1. From the results, we observe that our EVA-SCORE outperforms all other traditional and LLM-based evaluation metrics in achieving the highest correlation with human judgments across all indicators at both levels. At

Level	Text-Level			System-Level		
Metric	σ	γ	τ	σ	γ	τ
Traditional Evaluation Metric						
ROUGE-1	0.305	0.160	0.114	0.455	0.800	0.667
ROUGE-2	0.413	0.203	0.127	0.014	0.399	0.333
BERT-Score	0.534	0.533	0.413	0.623	0.600	0.333
Trained Evaluation Metric						
Auto-J-13B	0.285	0.233	0.167	0.885	0.947	0.913
CritiqueLLM-6B	0.510	0.490	0.353	0.865	0.800	0.667
Prompt-only Evaluation Metric						
ChatGPT	0.135	0.210	0.162	0.430	0.600	0.333
GPT-4	0.308	0.329	0.229	0.505	0.600	0.333
EVA-SCORE	**0.710**	**0.673**	**0.503**	**0.975**	**1.000**	**1.000**

Table 1: Text-level and system-level Pearson(σ), Spearman(γ), and Kendall(τ) correlations between different evaluation metrics and human annotations. The highest correlation among the methods is denoted in **bold**. Our EVA-SCORE achieves the best performance with a perfect correlation at the System-Level.

the System Level, EVA-SCORE achieves perfect Spearman and Kendall correlations of **1.00**, demonstrating complete alignment with human assessments. In contrast, Traditional Evaluation Methods, while somewhat correlated with human evaluations, generally lack strong correlations, highlighting their limitations. Models such as ChatGPT and GPT-4 exhibit a relatively low correlation with human annotations, which we attribute to the complexity involved in assessing the informativeness of a given summary against a reference, requiring them to decompose the task and finish the task step by step. Interestingly, evaluation models fine-tuned specifically for NLG achieve a high correlation with human annotations at the System Level. This improved performance is likely due to their fine-tuning process, which explicitly includes informativeness as an evaluation criterion, suggesting a promising direction for the future development of larger evaluation models.

5.5 Error Analysis

To have a thorough understanding of EVA-SCORE, we provide error analysis from the pipeline of EVA-SCORE in Appendix B.5 showing several situations where EVA-SCORE fails. To step further into why ROUGE and BERTScore are more suitable in relatively shorter settings than in longer settings, we provide our explanation in Appendix B.6.

5.6 Re-evaluation on Abstractive Long-form Summarization

To have a thorough understanding of the performance of LLMs on abstractive long-form summarization, we test on a wide range of models. For open source models, we evaluate on Mistral-7B-instruct (Jiang et al., 2023b), Llama-2-chat-7B, Llama-2-chat-13B (Touvron et al., 2023), Vicuna-7B, Vicuna-13B (Zheng et al., 2023), and Gemma-7B-it (Team et al., 2024). For close-source models, we experiment with ChatGPT and GPT-4-0125-preview. For all the LLMs, we set the temperatures as 1.0. We use three different evaluation metrics to indicate their performance, i.e. EVA-SCORE, ROUGE-1, and BERTScore. Since this work focuses on the evaluation of long-form summarization, we do not pursue feeding long context into LLMs without information loss. Instead, we believe that useful information usually appears at the beginning or end of a text as Bai et al. (2023), and we simply truncate the context from the beginning and end to meet the maximum input sequence length of the models. The length of the context window can be viewed as part of the models' inherent abilities, so it will not introduce unfairness in our evaluation. For all GLMs, our prompt is *"Summary the following text. You can use your own words instead of extracting from the original document."*, ensuring an abstractive setting.

Referring to Table 2, as context length increases,

Dataset	CNN Dailymail			Pubmed			arXiv			Gov Report			BookSum		
Metric	EVA	R-1	BS	EVA	R-1	BS	EVA	R-1	BS	EVA	R-1	BS	EVA	R-1	BS
Vicuna-7B	**0.501**	0.441	0.873	0.327	0.398	0.832	0.327	0.344	0.838	0.239	0.308	0.846	0.145	0.212	0.825
Vicuna-13B	0.462	0.432	0.872	0.431	0.388	0.837	0.346	0.348	0.832	0.154	0.298	0.849	0.144	0.188	0.827
Mistral-7B-Instruct	0.499	0.422	0.872	0.447	0.447	0.847	0.444	0.448	0.845	0.313	0.399	0.854	**0.259**	0.377	0.839
Llama-2-7B-chat	0.399	0.386	0.868	0.358	0.421	0.845	0.411	0.435	0.851	0.193	0.198	0.798	0.186	0.268	0.826
Llama-2-13B-chat	0.453	0.423	0.868	0.416	0.405	0.841	0.444	0.332	0.830p	0.296	0.378	0.857	0.154	0.280	0.827
Gemma-7B-it	0.431	0.383	0.858	**0.492**	0.417	0.842	0.385	0.359	0.831	0.250	0.291	0.843	0.154	0.256	0.818
ChatGPT	0.472	0.373	0.864	0.444	0.359	0.838	0.443	0.368	0.838	0.302	0.261	0.854	0.182	0.202	0.826
GPT-4	0.466	0.376	0.863	0.454	0.407	0.838	**0.453**	0.361	0.835	**0.325**	0.332	0.854	0.257	0.261	0.824

Table 2: The results of several LLMs on different long-form summarization datasets. We denote EVA-SCORE as EVA, R-1 as ROUGE-1, and BS as BERTScore. All the values are normalized between 0 and 1. We indicate the highest score of EVA-SCORE using **bold**.

all models experience varying degrees of performance decline, particularly in the **BookSum** dataset, where results are notably poorer. However, models capable of handling longer input sequences (Su et al., 2023) demonstrate superior performance in abstractive long-form summarization compared to those with limited context windows when dealing with larger contexts, largely due to their ability to comprehend extensive content. Conversely, when the context length is shorter, these models perform comparably or even worse than models with normal sequence lengths, such as Vicuna-7B. GPT-4 demonstrates robust capabilities compared to others, particularly in situations of longer input sequences under EVA-SCORE. However, it fails in relatively shorter settings, probably resulting from the tendency to respond in more words than others. Surprisingly, Mistral-7B-instruct, compared with other models with the same amount of parameters or even more parameters, shows great potential to fulfill abstractive long-form summarization and even beats GPT-4 by a small margin in **BookSum**, which meets the findings of Chang et al. (2024).

Moreover, compared to other evaluation metrics, EVA-SCORE shows a more pronounced and consistent decrease in performance as the context length grows, with fewer fluctuations. This pattern mirrors human cognitive processes, where increasing information load makes maintaining summarization quality more challenging. The consistent decline highlights the effectiveness of EVA-SCORE from another perspective.

5.7 Analysis of Summarization Length

To elucidate why GPT-4 underperforms under EVA-SCORE in the situation of a shorter context, we analyze the summary lengths produced by different models, as detailed in Table 3.

We observe that shorter contexts typically result in shorter ground truth. Consequently, if a model generates a lengthy summary, this verbosity is penalized with a reduced EVA-SCORE due to decreased precision. As the context lengthens, the references become richer in information. If these models tend to maintain a similar response length, lacking significant information addition, they fail to capture enough relevant points. GPT-4, which is refined through RLHF (Ouyang et al., 2022), typically caps its responses to summarizations to a narrow range. This behavior contributes to GPT-4's underachievement in **Booksum**, where more extensive and detailed responses are crucial to capture the full scope of the content. Additionally, we provide a meta-evaluation of human annotations in Appendix B.4, highlighting the underlying approach adopted by humans and models for summarization.

5.8 Ablation Study

We report the effect of *AFCG* and *DocRE* on the number of atomic facts to be verified. From Table 4, datasets with longer context and richer information tend to result in more facts, and by *AFCG*, we largely reduce the number of cascaded facts. Also, *DocRE* complements *AFG* with document-level relations, which is helpful for the evaluation of the hierarchical information in the summary. The two processes work together to form a concise yet comprehensive list of facts to be verified.

5.9 Analysis of Different LLM Verifier

Since our LLM Verifier should consider previous states in a logic chain and verify the accuracy of current atomic fact, it is a rather hard problem compared to simple verification of an NLI problem. Intuitively, with stronger models, we will get a more human-aligned evaluation metric. Therefore,

Dataset	**CNN Dailymail**	**Pubmed**	**arXiv**	**Gov Report**	**Booksum**
Metric	*Avg Length*	*Avg Length*	*Avg Length*	*Avg Length*	*Avg Length*
Vicuna-7B	108	214	133	258	124
Vicuna-13B	106	146	205	188	99
Mistral-7B-Instruct	109	207	212	**353**	**251**
Llama-2-7B-chat	118	143	108	140	185
Llama-2-13B-chat	124	184	221	236	179
Gemma-7B-it	98	139	140	177	151
ChatGPT	102	118	124	137	107
GPT-4	**153**	**227**	**272**	242	212

Table 3: The average length of summary counted in words generated by different models denoted by *Avg Length*. The largest count is marked using **bold**. It can be seen that GPT-4 prefers to summarize with the length of responses in a narrow range.

Settings	*AFG only*	*AFCG*	*AFCG + DocRE*
CNN Dailymail	25	19	24
Pubmed	38	25	30
arXiv	33	22	26
Gov Report	44	26	30
BookSum	36	27	31

Table 4: Number of facts (fact chains) to be evaluated under three settings, averaging 9 models we used for experiments, i.e., **Atomic Fact Generation** (*AFG*), **Atomic Fact Chain Generation** (*AFCG*) and **Atomic Fact Chain Generation** Complemented with **Document-level Relation Extraction** (*AFCG + DocRE*).

we report the results using ChatGPT for a simple proof which are listed in Table 5.

Leveraging ChatGPT results in a higher human correlation in Pearson and Kendall correlation and comparable Spearman correlation, which is consistent with our recognition. However, using ChatGPT results in much more costs than open-source LLMs because of the great number of validations.

Metric	σ	γ	τ
EVA-SCORE-Mixtral-7B-Instruct	0.710	**0.673**	0.503
EVA-SCORE-ChatGPT	**0.737**	0.672	**0.543**

Table 5: Human correlation using ChatGPT as LLM verifier of EVA-SCORE compared with EVA-SCORE with Mistral-7B-Instruct. σ, γ, τ are Pearson, Spearman, and Kendall correlation respectively. The larger one is denoted in **bold**.

5.10 Case Study

To provide a more intuitive and straightforward insight into the pipeline of EVA-SCORE, we provide a detailed example in Appendix A.5. From the case, we can see that EVA-SCORE provides not only a quantitative and objective information score but also provides a human-like explanation.

6 Conclusion and Future Work

In this work, we introduce a novel evaluation metric, termed EVA-SCORE, to evaluate the informativeness in abstract long-form summarization. EVA-SCORE begins by extracting atomic facts from given summaries, which are then restructured into logical chains. We also extract document-level relations to complete the information. Subsequently, LLMs are employed for further validation of the information. Experimental results demonstrate that, compared to existing automatic evaluation metrics, EVA-SCORE achieves the highest correlation with human judgments, providing an objective and explainable assessment. Furthermore, we use EVA-SCORE to re-evaluate abstractive long-form summarization models, highlighting the challenges LLMs face in handling extended contexts. We also conduct a detailed analysis of cases where EVA-SCORE and other evaluation metrics fall short in evaluating abstractive long-form summaries. We believe that EVA-SCORE, designed to quantify information overlap between reference and candidate, has potential applications beyond long-form summarization and can be extended to broader evaluation scenarios.

Limitations

EVA-SCORE is time-consuming to run

EVA-Score includes several pre-processing methods. We should first extract atomic facts and document-level relations from two summaries. We also use LLMs for paraphrasing to make the relations more natural. The preprocessing takes around two minutes per record. Then for each atomic fact in the candidate, we select k most relevant from the reference and use LLMs for validation. Since the number of atomic facts is very large, it takes quite a lot of time. We report the time consumed correlated to the sum of the length of the candidate and reference counted in tokens.

Length Interval	0-300	300-600	600-900	900-1200
Time Consuming/s	195	402	578	629

Table 6: Time consumed measured in seconds for different summary lengths measured in tokens. The time consumption shows an approximately linear correlation with text length.

From Table 6, for a summary of relatively short length, it takes about three minutes to finish the evaluation, while for a longer summary, it may take about ten minutes which is a long time compared to traditional methods.

EVA-SCORE only considers Informativeness

EVA-SCORE is designed to evaluate informativeness. Some existing evaluation metrics consider many other factors, like faithfulness, and consistency. EVA-SCORE may be combined with other evaluation metrics to provide a more comprehensive and fair evaluation of abstractive long-form summarization.

EVA-SCORE Leverages GPT-4 for Doc-RE

EVA-SCORE incorporates *DocRE* to address the lack of hierarchical information. However, *DocRE* by GPT-4 may be incomplete, as GPT-4 may struggle to fully capture the core meaning at the document level due to the inherent complexity of the task. Leveraging more advanced models specifically fine-tuned for *DocRE* could potentially improve results. Furthermore, the filter used in the current approach is simplistic, and experimental findings suggest that a more effective filter is needed to better distinguish document-level relations from sentence-level ones.

Acknowledgments

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. *Preprint*, arXiv:2310.00785.

Garima Chhikara, Anurag Sharma, V. Gurucharan, Kripabandhu Ghosh, and Abhijnan Chakraborty. 2024. Lamsum: Creating extractive summaries of user generated content using llms.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *Preprint*, arXiv:1804.05685.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023b. Mistral 7b. *Preprint*, arXiv:2310.06825.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2023. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation. *Preprint*, arXiv:2311.18702.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. *Preprint*, arXiv:2311.18702.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo,

Tanya Goyal, and Mohit Iyyer. 2024. Fables: Evaluating faithfulness and content selection in book-length summarization. *Preprint*, arXiv:2404.01261.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. *Preprint*, arXiv:2103.06332.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization. *Preprint*, arXiv:2105.08209.

Dacheng Li*, Rulin Shao*, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can open-source llms truly promise on context length?

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *Preprint*, arXiv:2310.05470.

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.

Guang Lu, Sylvia B. Larcher, and Tu Tran. 2023. Hybrid long document summarization using c2f-far and chatgpt: A practical study. *Preprint*, arXiv:2306.01169.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *Preprint*, arXiv:2303.15621.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Bar-

ret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Preprint*, arXiv:1704.04368.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. *Preprint*, arXiv:2305.13091.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. Finesure: Fine-grained summarization evaluation using llms. *Preprint*, arXiv:2407.00908.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Zero-shot cross-lingual summarization via large language models. *Preprint*, arXiv:2302.14229.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *Preprint*, arXiv:2305.13412.

Wang Xu, Kehai Chen, Lili Mou, and Tiejun Zhao. 2022. Document-level relation extraction with sentences importance estimation and focusing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2920–2929, Seattle, United States. Association for Computational Linguistics.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autore: Document-level relation extraction with large language models. *Preprint*, arXiv:2403.14888.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via chatgpt for faithful summary generation. *Preprint*, arXiv:2304.04193.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Danna Zheng, Danyang Liu, Mirella Lapata, and Jeff Z. Pan. 2024. Trustscore: Reference-free evaluation of llm response trustworthiness. *Preprint*, arXiv:2402.12545.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Case Study

In this section, we list all all the cases for display and comprehension.

A.1 Atomic Fact Generation (AFG)

We display an example of our *AFG* process in Figure 3. It decomposes a text into a combination of multiple pieces of information.

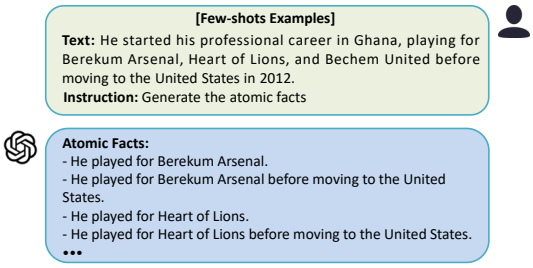

Figure 3: Examples show the generation of atomic facts. We implement it using In-Context-Learning for better extraction.

A.2 Atomic Fact Chain Generation (AFCG)

We display an example of our *AFCG* process in Figure 4. We convert the original 8 atomic facts into 2 fact chains to ease validation because each fact in a certain fact chain shares some information overlap. Also, by *AFCG*, we ensure checking pieces of information one by one is feasible.

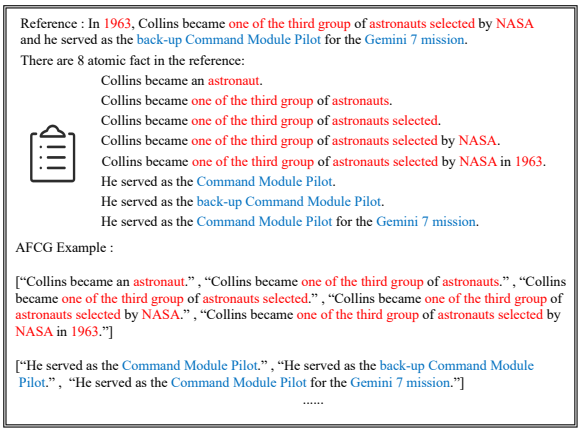

Figure 4: An example displaying the process of *AFCG*. We first extract atomic facts from a given summary using *AFG*. Then based on the logic overlap, we reformulate them into several logic chains. The texts denoted in red and blue are information cascaded which need to be verified later.

A.3 Document-level Relation Extraction (DocRE)

We display an example illustrating the difference between document-level relations and sentence-level relations in Table 7. To make it clear, sentence-level relations are relations that exist within a sentence while document-level relations are relations beyond a single sentence that can be found via implicit or explicit reasoning between sentence-level relations.

Text:
TechCorp, headquartered in Silicon Valley, is known for its innovative products in the tech industry. John recently joined TechCorp as a software engineer. He is excited to work on cutting-edge AI technologies.
Sentence-level Relation:
(John, employed by, TechCorp)
Document-level Relations:
(John, lived in, Silicon Valley)

Table 7: An example indicating the difference between sentence-level relations and document-level relations.

A.4 LLM Validation

We provide an example revealing the process of using LLMs for final validation in Figure 5. This validation process verifies newly added information only rather than treating the whole sentence as a basic unit. Therefore, we can measure the informativeness from a fine-grained dimension.

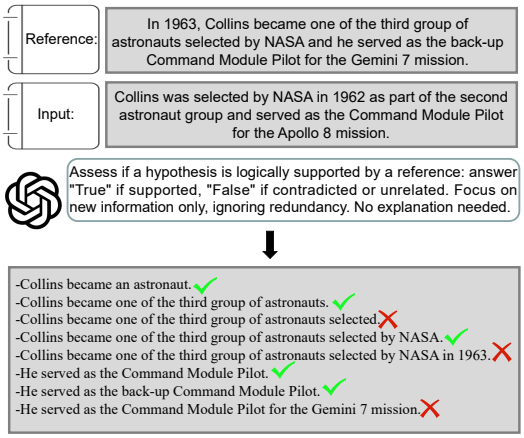

Figure 5: An example displaying the process of LLM Validation. For each fact chain illustrated in Figure 4, we verify the atomic fact within it one by one by explicitly prompting LLMs to focus on newly added information rather than redundant information.

Dataset	CNN Dailymail	Pubmed	arXiv	Gov Report	BookSum
Average Text Length	781	3111	5894	7797	**14337**
Average Summary Length	56	216	172	**573**	570

Table 8: The *Average Text Length* and the *Average Summary Length* counted in words in our experimented dataset. The largest values are denoted using **bold**.

A.5 Overall Process

To provide a more intuitive and straightforward insight into the pipeline of EVA-SCORE, we provide a detailed case in Table 9. From the case, we can see that EVA-SCORE provides not only a quantitative and objective information score but also provides a human-like explanation.

B Detailed Experiments

In this section, we provide details that are not listed in the main papers due to the page limitation.

B.1 Dataset Details

We provide the average text length and the average gold summary length, both measured in words, in Table 8. Among the datasets, **BookSum** contains the longest average text length, while **Gov Report** has the longest average summary length.

B.2 Threshold for Filtering

In this section, we explain our approach for selecting the threshold used to filter document-level relations from sentence-level relations. Our principle is to filter out as many similar statements as possible while retaining those that are distinct. We utilize cosine similarity based on BERT as the filtering criterion. Specifically, for each element in the set of extracted document-level relations, we calculate the cosine similarity between it and each fact in the list of extracted atomic facts and then determine if the highest cosine similarity exceeds the threshold.

To guide our threshold selection, we consider three indicators: (1) *Accuracy Rate*, which measures the proportion of filtered document-level relations that genuinely required filtering; (2) *Remaining Rate*, which represents the ratio of document-level relations that are retained after filtering; and (3) *Remaining Accuracy Rate*, which indicates the proportion of retained document-level relations that genuinely should not have been filtered. The results are listed in Table 10. From the experimental results, a larger threshold results in fewer statements being filtered out, as only those highly similar to one of the extracted atomic facts are removed. This leads to a higher accuracy rate and remaining rate but a lower remaining accuracy rate. Based on this trade-off, we set the threshold at 0.65 to minimize redundancy while ensuring sufficient remaining accuracy. The *Remaining Accuracy Rate* is not higher primarily due to the straightforward use of GPT-4 for *DocRE*. However, since the *Remaining Rate* is lower and there is a general lack of document-level relations within the task, the impact of this limitation is not particularly pronounced due to their little contribution to the final results.

B.3 Annotation Recipe

Identify Atomic Information We identify information contained in the candidate summary and the reference summary. Specifically, for a sentence, there may be several pieces of information so that we can treat it as a combination of information. For duplicated information, we only record once to ensure fairness and accuracy.

Compare Candidate Based on Reference To check whether the information contained in the reference also exists in the candidate, we verify information piece by piece. A piece of information will only be recorded as overlapping if and only if it exists both in the reference and the candidate. After that, we can get Precision and Recall for each record and we use the F1 Score as our final results.

After Annotation Validation To ensure reliability, we ask at least two annotators for annotation for each record. After annotation, we calculate the correlation between the annotators. If there is an apparent low correlation between annotators, we ask a third annotator for re-annotation.

B.4 Meta-evaluation of Human Annotation

By observing the results of human annotations, we find that the ground truth prefers to summarize more elegantly, using core arguments from the original context in addition to examples supported with other a little bit trivial arguments. However, LLMs try to list all of the arguments, ignoring the exam-

REFERENCE SUMMARY:
The mpp decreased the number of dopaminergic neurons by 40% , and increased the release of lactate dehydrogenase (ldh) into the culture medium . The tq significantly rescued dopaminergic neurons and decreased the release of ldh at the concentrations of 0.1 and 1 m

ATOMIC FACT CHAINS
[- The mpp decreased dopaminergic neurons, - The mpp decreased the number of dopaminergic neurons, - The mpp decreased the number of dopaminergic neurons by 40%]
[- The mpp increased the release of lactate dehydrogenase (ldh), - The mpp increased the release of lactate dehydrogenase (ldh) into the culture medium]
[- The tq significantly rescued dopaminergic neurons]
[- The tq decreased the release of ldh, - The tq decreased the release of ldh at the concentrations, - The tq decreased the release of ldh at the concentrations of 0.1 and 1 m]
DOCUMENT-LEVEL RELATIONS:
tq causes Thymoquinone
mpp leads to 1-methyl-4-phenylpyridinium

CANDIDATE SUMMARY:
The MPP significantly decreased the survival of dopaminergic neurons and increased the release of lysosomal hydrolase (LDH) into the culture medium. The neurotoxicity of MPP was attributed to its selective uptake by dopaminergic neurons through the dopamine transporter.

ATOMIC FACT CHAINS
[- The MPP significantly decreased dopaminergic neurons, -The MPP significantly decreased the survival of dopaminergic neurons]
[- The MPP increased the release of lysosomal hydrolase (LDH), - The MPP increased the release of lysosomal hydrolase (LDH) into the culture medium]
[- The neurotoxicity was attributed to dopaminergic neurons, - The neurotoxicity of MPP was attributed to dopaminergic neurons, - The neurotoxicity of MPP was attributed to its selective uptake by dopaminergic neurons, - The neurotoxicity of MPP was attributed to dopaminergic neurons through the dopamine transporter, - The neurotoxicity of MPP was attributed to its selective uptake by dopaminergic neurons through the dopamine transporter.]
DOCUMENT-LEVEL RELATIONS:
mpp causes 1-methyl-4-phenylpyridinium
neurotoxicity infects neuroprotective effect

MATCHING PROCESS
- “The mpp decreased dopaminergic neurons” match “The MPP significantly decreased dopaminergic neurons”
- “The mpp increased the release of lactate dehydrogenase (ldh)” match “The mpp increased the release of lactate dehydrogenase (ldh)”
- “The mpp increased the release of lactate dehydrogenase (ldh) into the culture medium” match “The MPP increased the release of lysosomal hydrolase (LDH) into the culture medium”
- “The mpp decreased dopaminergic neurons and increased the release of lactate dehydrogenase (ldh)” match ‘The MPP significantly decreased dopaminergic neurons and increased the release of lysosomal hydrolase (LDH)”
- “The mpp decreased dopaminergic neurons and increased the release of lactate dehydrogenase (ldh) into the culture medium” match “The MPP significantly decreased dopaminergic neurons and increased the release of lysosomal hydrolase (LDH) into the culture medium”
- “mpp causes 1-methyl-4-phenylpyridinium” match “mpp leads to 1-methyl-4-phenylpyridinium”

Table 9: Case Study For EVA-SCORE from Pubmed Dataset.

Threshold	0.60	0.65	0.70	0.75
Accuracy Rate	0.823	0.861	0.886	**0.920**
Remaining Rate	0.075	0.142	0.258	**0.368**
Remaining Accuracy Rate	**0.512**	0.476	0.437	0.413

Table 10: Results of the performance of the filter under different thresholds. We use three metrics, *Accuracy Rate*, *Remaining Rate*, and *Remaining Accuracy Rate*, for help. The largest value is denoted using **bold**.

ples or data in the context, leading to differences with ground truth. However, there does exist some verbosity in the ground truth, indicating the natural drawback of the summarization dataset.

B.5 Error Analysis

We provide an error analysis of why EVA-SCORE underperforms compared to human annotations. Since EVA-SCORE is built on *Atomic Fact Generation*, *Atomic Fact Chain Generation*, *Document-level Relation Extraction (DocRE)*, and *LLM Validation*, we randomly select 30 records to identify the sources of errors. Upon manual inspection, we find that most errors occurred during *DocRE* and *LLM Validation*. This is because *AFG* is fundamental for LLMs, and *AFCG*, which involves NLI, is generally handled well by LLMs. However, for *DocRE*, GPT-4 struggles to differentiate between sentence-level and document-level relations, leading to inaccuracies. Additionally, the static threshold used for filtering document-level relations may be either too strict or too lenient, further contributing to errors. In *LLM Validation*, we prompt the model to focus on newly added information, which requires a high level of reasoning capability of LLMs. We also display examples in Table 11. To make EVA-SCORE more aligned with human judgments, we should either use more powerful LLMs for *DocRE* or develop novel methods for relation extraction. Furthermore, the LLM validation process can be improved through techniques like majority voting or self-consistency, which could enhance the reliability of the evaluations.

B.6 Why Traditional Metrics Fail

The challenge of using traditional metrics in long contexts stems from the fact that surface-level similarity does not always indicate informativeness. Metrics like BERTScore and ROUGE measure the similarity between generated text and a reference in terms of content and structure, but they often fall short in assessing the quality and relevance of the conveyed information. In long contexts, a text may be highly informative and similar in meaning to the reference, yet exhibit low similarity in terms of word choice or embeddings.

BERTScore, for instance, faces limitations in longer texts due to semantic dilution, where embeddings capture a wide range of content, including irrelevant details, diluting the importance of key informative elements. Its reliance on contextual embeddings also means that the meaning of words can shift with varying contexts. Furthermore, as the number of tokens increases, the complexity of token matching may not align with the most informative parts of the text, leading to lower similarity scores despite the presence of valuable information. ROUGE, on the other hand, struggles with lexical overlap insufficiency, as it evaluates texts based on overlapping n-grams, failing to account for paraphrases or semantic equivalents. This can penalize informative texts that use different wording to convey the same ideas. In longer context, the number of possible n-gram overlaps grows, but the proportion of meaningful overlaps decreases, causing the metric to become less sensitive to informativeness.

C Instructions for EVA-Score

C.1 Instruction for Atomic Fact Generation

We use the prompt in Table 12 to generate atomic facts.

C.2 Instruction for Atomic Fact Chain Generation

As discussed before, we use the prompt in Table 13 to decide whether one fact is entailed by others.

C.3 Instruction for Named Entity Recognition

We use the prompt in Table 14 to get named entities from the given text.

C.4 Instruction for Document-level Relation Extraction

We use the prompt in Table 15 with GPT-4 to conduct Document-level Relation Extraction.

C.5 Instruction for LLM validation

We use the prompt in Table 16 to instruct LLMs to focus on one piece of information each time.

C.6 Instruction for LLM-based Evaluation of Informativeness

We use the following prompt in Table 17 to instruct LLMs to generate the informative score.

Type: Mixture of Doument-level Relations with Sentence-level Relations
Text:
The paper presents a theoretical study on the precession of vortex lines in a superfluid in the presence of an external potential. The author uses the Gross-Pitaevskii equation to study the stationary states of the system and finds that there are two types of vortex lines that can be described by the lyapunov functional. The author also finds that the angular momentum of the vortex line as a function of the precession frequency. The paper also compares the results with previous works and finds that the difference in the density of the condensates affects the precession frequency. The author concludes that the present results and the density of the condensates.
Document-level Relation:
"The theoretical study is on the precession of vortex lines." ✓ "The Gross-Pitaevskii equation is used to study the stationary states of the system."✓ "The author finds that there are two types of vortex lines." ✗ → It is a sentence-level relation "The author finds the angular momentum of the vortex line."✗ → It is a sentence-level relation "The angular momentum of the vortex line is a function of the precession frequency."✓ "Stationary states are a feature of Vortex lines"✓ "Angular momentum is a property of vortex lines"✓
Type: Miss of Document-level Relations
Text:
in that light, it seems surprising that they can say much about inflation at all , but it turns out that the flow equations can be viewed as a (rather complicated) algorithm for generating functions which have a suitable form to be interpreted as inflationary model implied that the flow equation predictions ought to be little changed by moving to the braneworld , although this modifies the friedmann equation .
Doument-level Relations Caught
"The Friedmann equation was modified in the braneworld inflation scenario"
Document-level Relations Missed
"The flow equation predictions little changed in the braneworld inflation scenario"
Type: LLM Validation Error
Reference:
"The scheme uses the two-channel Raman interaction." "The interaction is between two atoms." "The scheme involves two atoms."
Hypothesis:
"The scheme is for implementing an unconventional geometric two-qubit phase gate." "The scheme involves using the two-channel Raman interaction." "The two atoms are in a cavity."
Matching Pool:
"The scheme uses the two-channel Raman interaction." → "The scheme involves using the two-channel Raman interaction." ✓ "The scheme involves two atoms." → "The scheme uses the two-channel Raman interaction." ✗

Table 11: The error analysis of why EVA-SCORE fails. We display the example of ***Mixture of Document-level Relations with Sentence-level Relations***, ***Miss of Document-level Relations***, and ***LLM Validation Error***.

INSTRUCTION: Please breakdown the following sentence into independent facts:

DOMONSTRATIONS

Example 1

He made his acting debut in the film The Moon is the Sun's Dream (1992), and continued to appear in small and supporting roles throughout the 1990s.

- He made his acting debut in the film.
- He made his acting debut in The Moon is the Sun's Dream.
- The Moon is the Sun's Dream is a film.
- The Moon is the Sun's Dream was released in 1992.
- After his acting debut, he appeared in small and supporting roles.
- After his acting debut, he appeared in small and supporting roles throughout the 1990s.

Example 2 In 1963, Collins became one of the third group of astronauts selected by NASA and he served as the back-up Command Module Pilot for the Gemini 7 mission.

- Collins became an astronaut.
- Collins became one of the third group of astronauts.
- Collins became one of the third group of astronauts selected.
- Collins became one of the third group of astronauts selected by NASA.
- Collins became one of the third group of astronauts selected by NASA in 1963.
- He served as the Command Module Pilot.
- He served as the back-up Command Module Pilot.
- He served as the Command Module Pilot for the Gemini 7 mission.

Table 12: The instruction for *AFG*.

INSTRUCTION: Given a hypothesis and a reference, determine if the hypothesis is entailed by the reference.

Entailed means that if we use the information from reference, we can infer the hypothesis, e.g. the hypothesis is in a logical relationship with the reference.

If the hypothesis is entailed by the reference, answer True.

If the hypothesis contradicts the reference or the hypothesis is neural,answer False.

If the answer is ambiguous, answer False.

The output should be True and False only. You should not explain why you made this choice.

Now let's begin the question:

DOMONSTRATIONS

Example 1

Hypothesis: The cat is on the mat.

Reference: The cat is sleeping on the mat.

Answer: True

Example 2

Hypothesis: The cat is on the mat.

Reference: The dog and the rat are on the mat.

Answer: False

Table 13: The instruction for *AFCG*.

INSTRUCTION: Given a text, your task is to extract all the entities as comprehensive as possible from the text. Below is the list of entities that you need to extract from the text.
Types	Content
PER	People, including fictional
ORG	Companies, universities, institutions, political or religious groups, etc.
LOC	Geographically defined locations, including mountains, waters, etc. Politically defined locations, including countries, cities, states, streets, etc. Facilities, including buildings, museums, stadiums, hospitals, factories, airports, etc.
TIME	Absolute or relative dates or periods.
NUM	Percents, money, quantities Products, including vehicles, weapons, etc.
MISC	Events, including elections, battles, sporting events, etc. Laws, cases, languages, etc
Respond with a list of dictionaries, where each dictionary contains the entity type and the entity text. The list should be sorted by the entity type, and the entities should be sorted by their start position.
Note: You should respond in a list of dictionaries, where each dictionary contains the entity type and the entity text. To make it specific, the format is like:
[
"type": "", "text": "",
"type": "", "text": "",
"type": "", "text": ""
]
I will provide you with some examples to help you understand the task.

DOMONSTRATIONS
Example 1
"Apple is looking at buying U.K. startup, U.S. company and China startup for 1 billion. Mr. John Doe is the CEO of the company."
[
"type": "ORG", "text": "U.K. startup",
"type": "ORG", "text": "U.S. company",
"type": "ORG", "text": "China startup",
"type": "ORG", "text": "Apple",
"type": "PER", "text": "John Doe",
"type": "MISC", "text": "CEO",
"type": "NUM", "text": "1 billion"
]
Example 2
"Skai TV is a Greek free-to-air television network based in Piraeus. It is part of the Skai Group, one of the largest media groups in the country. It was relaunched in its present form on 1st of April 2006 in the Athens metropolitan area, and gradually spread its coverage nationwide. Besides digital terrestrial transmission, it is available on the subscription-based encrypted services of Nova and Cosmote TV. Skai TV is also a member of Digea, a consortium of private television networks introducing digital terrestrial transmission in Greece. At launch, Skai TV opted for dubbing all foreign language content into Greek, instead of using subtitles. This is very uncommon in Greece for anything except documentaries (using voiceover dubbing) and children's programmes (using lip-synced dubbing), so after intense criticism the station switched to using subtitles for almost all foreign shows."
[
"type": "ORG", "text": "Skai TV",
"type": "LOC", "text": "Greek",
"type": "MISC", "text": "television network",
"type": "LOC", "text": "Piraeus",
....]
Respond with a json format strictly.

Table 14: The instruction for using LLMs for NER.

INSTRUCTION: Given a text and entities in the given text, your task is to extract document-level relation triples from the text.
A relation triple is a 3-tuple (subject, relation, object) where subject and object are named entities and relation is the relation between them. For example, given the text "Bill Gates is the founder of Microsoft", the extracted relation triple would be (Bill Gates, founder, Microsoft).
In document-level RE, subject and object entities in a given triple might be dispersed across distinct sentences, and certain entities may have aliases in the form of distinct entity mentions. Therefore, you should pay attention to relations across sentences and entity aliases rather than sentence-level relations. i.e. You should not respond with sentence-level relations, but document-level relations.
The answer format is listed below, you should obey it for any situation. Any additional information should not be added. Neither head entity nor tail entity should be empty strings.
The relation triples should be in the following format: ('head entity', 'relation', 'tail entity') ('head entity', 'relation', 'tail entity') ('head entity', 'relation', 'tail entity')

DEMONSTRATIONS
Text
Alexeni is a commune in Ialomița County, Romania, some 65 km north-east of Bucharest, near the town of Urziceni. It is composed of a single village, Alexeni. Until 2001 a Romanian Air Force military helicopters unit was located at the nearby airfield. In 2007, as the airfield was not used by the Romanian Air Force any longer, the former Minister of Transport Radu Berceanu suggested to use the location for Bucharest's new low-cost flights airport(as the operational tariffs for Bucharest's previous low-cost hub, Aurel Vlaicu Airport, were set to grow). However, some analysts considered the project unrealistic and doomed to fail due to the poor conditions of the infrastructure in the area. Eventually, those plans were abandoned and all low-cost flights were moved in March 2012 at Bucharest main airport Henri Coandă International Airport.
Entities
['Urziceni', 'Aurel Vlaicu Airport', '2001', '2007', 'Radu Berceanu', 'Romania', 'Romanian Air Force', 'Alexeni', 'Ialomița County', 'March 2012', 'Henri Coandă International Airport', '65 km', 'Bucharest']
Response
('Ialomița County', 'located in the administrative territorial entity', 'Romania')
('Ialomița County', 'country', 'Romania')
('Ialomița County', 'contains administrative territorial entity', 'Alexeni')
('Romania', 'contains administrative territorial entity', 'Ialomița County')
('Bucharest', 'country', 'Romania')
('Bucharest', 'located in the administrative territorial entity', 'Romania')
('Romanian Air Force', 'country', 'Romania')
('Aurel Vlaicu Airport', 'country', 'Romania')
('Aurel Vlaicu Airport', 'located in the administrative territorial entity', 'Bucharest')
('Henri Coandă International Airport', 'country', 'Romania')
('Urziceni', 'country', 'Romania')
('Radu Berceanu', 'country of citizenship', 'Romania')
('Alexeni', 'located in the administrative territorial entity', 'Ialomița County')
('Alexeni', 'country', 'Romania')

Table 15: The instruction for *DocRE*.

INSTRUCTION: Given a hypothesis and a reference, determine if the hypothesis is entailed by the reference.
Entailed means that if we use the information from reference, we can infer the hypothesis, e.g. the hypothesis is in a logical relationship with the reference.
The hypothesis may be part of a logical chain, so we may provide you with the correctness of previous states.
If we provide you with the correctness of previous states, you should not focus on the information that is redundant, that is, the information that exists both in previous facts and the current fact. Whether they are correct or not, you should not care them. You should only concentrate on new information.
If the hypothesis is entailed by the reference, answer True.
If the hypothesis contradicts the reference or the hypothesis is neural,answer False.
If the answer is ambiguous, answer False.
The output should be True and False only. You should not explain why you make this choice.

Table 16: The instruction for LLM Validation

INSTRUCTION:

Given a Reference Answer and a Candidate Answer, your task is to first extract all information important to the Reference Answer and the Candidate Answer, and then calculate the Precision, Recall, and F1-score of the Candidate Answer Information pool concerning the Reference Answer Information pool.

Information is defined as words, phrases or sentences that are important to the Reference Answer and the Candidate Answer. The subjective, objective, adverbial and attribute may all be included in the information. The information pool is the collection of all information extracted from the Reference Answer and the Candidate Answer.

You can use the following function to calculate the Precision, Recall, and F1-score of the Candidate Answer Information pool with respect to the Reference Answer Information pool.

```
def calculate_metrics(gold_key_information_pool, model_key_information_pool, matching_pool):
    precision = len(matching_pool) / len(gold_key_information_pool)
    recall = len(matching_pool) / len(model_key_information_pool)
    f1_score = 2 * precision * recall / (precision + recall)
    return f1_score
```

Finally, only return with a floating point number rounded to 3 decimal places.

DEMONSTRATIONS

Reference Summary

background: coronary computed tomography angiography (ccta) is a frequently performed examination for coronary artery disease . when performed with retrospective gating , there is an opportunity to derive functional parameters of left ventricle utilizing automated software . complementary information , if validated with established standards , will enhance the total value of study.objective:study evaluates the usefulness of fully automated software for the assessment of left ventricular ejection fraction (lvef) using 64-slice ccta data and to correlate ct results with echocardiography (echo) . role of ct derived lv function is reviewed in the light of emerging technologies and recent developments in multidetector ct (mdct).materials and methods : a total of 113 patients referred for mdct ccta for evaluation of coronary artery disease . all patients were scanned on 64 slice ge - helical ct scanner and had an echo done within 1 week of the ct scan . retrospectively electrocardiogram (ecg)-correlated image reconstruction was performed with the reconstruction at 10% r - r interval increment . axial image sets were analyzed with advanced workstation using a program - auto ejection fraction , circulation : ge medical solutions.results:the mean lvef calculated by clinical echo was 58.6 4.5% and by fully automated software based on cta data was 58.9 5.4% . the pearson 's regression analysis showed a large correlation , with a correlation coefficient of 0.503 (p < 0.001) . bland - altman analysis showed a trend towards mdct resulting in slightly higher values for lvef when compared with echo.conclusion:the fully automated software is simple , reliable , and user - friendly , and can provide rapid assessment of lv functional parameters with good reproducibility . despite of good correlation , fewer patients are likely to benefit , in future , from this function due to smaller number of patients undergoing ccta with retrospective gating

Candidate Summary

The text summarizes the results of a study on the use of multidetector computed tomography (MDCT) for assessing left ventricular (LV) function in patients undergoing coronary angiography. The study found that MDCT can provide useful functional information in addition to coronary angiography, and that fully automated analysis software can be used to measure LV ejection fraction (EF) with good reproducibility. However, the study also noted that the radiation dose involved in MDCT is a limitation, and that additional clinically indicated coronary CT examination with a lowest possible radiation dose is invaluable in patients known or suspected of ischemic heart disease. The study also found that validating consistency of results with magnetic resonance imaging (MRI) will further lend ‘ to the use of MDCT derived results.

Thought

Gold key information pool = [coronary computed tomography angiography, examination, coronary artery disease, retrospective gating,derive functional parameters,utilizing automated software, validated with established standards,enhance, total value, evaluates the usefulness, fully automated software, left ventricular ejection fraction, 64-slice ccta data, ct results, echocardiography (echo), reviewed, emerging technologies,multidetector ct (mdct), 113 patients, evaluation, scanned,helical ct scanner, 1 week of,electrocardiogram, performed with the reconstruction, 10%r - r interval increment, axial image,analyzed with advanced workstation,auto ejection fraction,the mean, 58.6 4.5%, cta data was 58.9 5.4%, pearson 's regression analysis,large correlation,coefficient of 0.503 (p < 0.001), altman analysis,slightly higher values for lvef, compared, friendly,rapid assessment, good reproducibility, fewer patients, reliable, lvef]

Model Key information pool = [multidetector computed tomography (MDCT), assessing left ventricular (LV), patients, coronary angiography, provide,useful functional information, automated analysis software, Measure, ejection fraction (EF), good reproducibility, radiation dose, limitation, clinically indicated, coronary CT examination, lowest possible, radiation dose, invaluable, patients known,suspected, ischemic heart disease, validating consistency, magnetic resonance imaging (MRI), further lend support, derived]

Matching pool: [multidetector computed tomography (MDCT), assessing left ventricular (LV) match assessing left ventricular (LV), patients match patients, coronary angiography match coronary angiography, coronary angiography match coronary angiography, ejection fraction match lvef, good reproducibility match good reproducibility, slightly higher values match invaluable, derive functional parameters match derived]

```
F1 = calculate_metrics(gold_key_information_pool, model_key_information_pool, matching_pool)
print(F1)
```

Output:

0.228

Table 17: The instruction for prompting ChatGPT and GPT-4 for evaluation.