SIFiD: Reassess Summary Factual Inconsistency Detection with LLM

Jiuding Yang [*1] **Hui Liu** [*2] **Weidong Guo**[†2] **Zhuwei Rao** [2] **Yu Xu** [2] **Di Niu** [1]
[1]University of Alberta
[2]Platform and Content Group, Tencent
[1]{jiuding,dniu}@ualberta.ca
[2]{pvopliu,weidongguo,evanyiu,henrysxu}@tencent.com

Abstract

Ensuring factual consistency between the summary and the original document is paramount in summarization tasks. Consequently, considerable effort has been dedicated to detecting inconsistencies. With the advent of Large Language Models (LLMs), recent studies have begun to leverage their advanced language understanding capabilities for inconsistency detection. However, early attempts have shown that LLMs underperform traditional models due to their limited ability to follow instructions and the absence of an effective detection methodology. In this study, we reassess summary inconsistency detection with LLMs, comparing the performances of GPT-3.5 and GPT-4. To advance research in LLM-based inconsistency detection, we propose SIFiD (**S**ummary **I**nconsistency Detection with **Fi**ltered **D**ocument) that identify key sentences within documents by either employing natural language inference or measuring semantic similarity between summaries and documents.

1 Introduction

Document summarization, the process of distilling key information from extensive texts, has become indispensable across various real-world applications, propelled by advancements in Natural Language Generation (NLG) (Pilault et al., 2020; Ma et al., 2022). The advent of Large Language Models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023) has notably enhanced models' capabilities to generate natural and factually consistent summaries (Chang et al., 2023). However, the rapid evolution in summarization techniques may lead to factually inconsistent summaries which are very close to facts (Zhang et al., 2023). Such inconsistencies could pose significant challenges, resulting in hallucinations that traditional detection models struggle to identify. As LLMs evolve, there is a critical demand for more robust methods to detect factual inconsistencies, leveraging the advanced capabilities of LLMs themselves.

Luo et al. (2023) were among the first to utilize LLMs for the detection of factual inconsistencies, employing a universal zero-shot prompt across various benchmarks in SUMMAC (Laban et al., 2022) and inputting the full document along with its summary into GPT-3.5 for evaluation. Despite these innovations, their approach was limited by the plain application, early GPT-3.5 model's constraints and a lack of adaptation to the specific requirements of different benchmarks. Consequently, their method did not achieve superior performance compared to existing models, such as those detailed in the SUMMAC paper.

This paper revisits the challenge of inconsistency detection in document summarization through zero-shot inference with LLMs, specifically examining the latest versions of GPT-3.5 and GPT-4 on the SUMMAC dataset. We aim to set up new LLM-based baselines for research in this domain. Moreover, we introduce a novel methodology, SIFiD (**S**ummary **I**nconsistency Detection with **Fi**ltered **D**ocument), designed to significantly enhance the efficiency and effectiveness of factual inconsistency detection. SIFiD focuses on identifying crucial sentences within documents by evaluating their entailment scores or semantic similarity with summary sentences, subsequently retaining only the most relevant sentences for further analysis. This approach not only refines the assessment of factual consistency but also reduces the computational resources required for evaluation by decreasing the number of input tokens.

Our comprehensive evaluation on the SUMMAC dataset reveals that, while the updated GPT-3.5 model still falls short of outperforming traditional

[*] These authors contributed equally to this work.
[†] Corresponding author.

Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), pages 27–31
August 16, 2024 ©2024 Association for Computational Linguistics

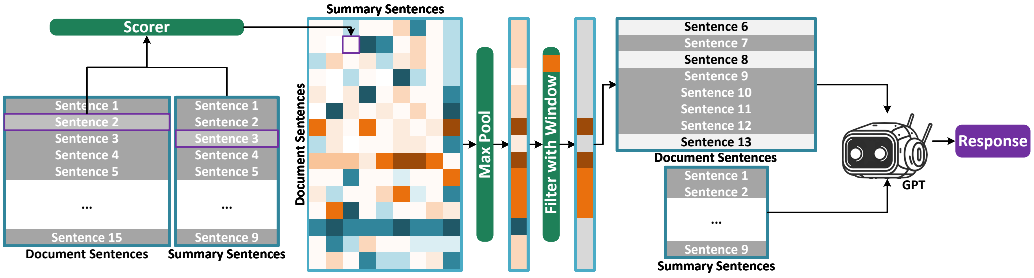

Figure 1: An illustration of SIFiD. The Score could either be entailment score or semantic cosine similarity.

baseline methods, GPT-4 significantly excels in detecting factual inconsistencies. The integration of SIFiD further amplifies GPT-4's detection capabilities, highlighting the potency of our proposed method. To support continued research and collaboration in this field, we make our code available open source at `https://github.com/XpastaX/SIFiD`, fostering advancements and exploration in factual inconsistency detection.

2 Related Work

The evaluation of summary factual consistency has traditionally relied on methods such as Question Answering and Question Generation (QAG) (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021), synthetic classifiers (Kryściński et al., 2020), and pairing-based approaches (Goodrich et al., 2019; Goyal and Durrett, 2020). These methodologies focus on identifying discrepancies between documents and their summaries. Laban et al. (2022) later demonstrated that Natural Language Inference (NLI) could be effectively employed for inconsistency detection at appropriate levels of text granularity, thereby advancing the field of summary inconsistency detection.

The emergence of Large Language Models (LLMs) has recently shifted the focus towards integrating these models into the assessment of summary factual consistency. Luo et al. (2023) pioneered the application of GPT-3.5 for this purpose, tailoring prompts to various evaluation tasks including summary factual inconsistency detection, summary ranking, and consistency evaluation. Despite this innovative approach, the early iteration of GPT-3.5, coupled with an insufficient detection methodology, did not yield improvements over conventional techniques in identifying factual inconsistencies.

In our research, we revisit the approach proposed by Luo et al. (2023), employing the most recent versions of GPT-3.5 and GPT-4. We integrate these advanced LLMs with our newly developed Summary Inconsistency Detection with Filtered Document (SIFiD) method. This combination aims to enhance the accuracy and efficiency of factual inconsistency detection, leveraging the state-of-the-art capabilities of LLMs to set new benchmarks in the field.

3 Approach

In this section, we detail our approach to reevaluating summary factual consistency using the latest GPT models and introduce the novel SIFiD method.

3.1 Summary Factual Inconsistency Detection with Large Language Models

As underscored in the Introduction, leveraging Large Language Models (LLMs) for detecting summary factual inconsistencies is crucial to addressing the challenges posed by rapidly improving document summarization capabilities. While Luo et al. (2023) were pioneers in utilizing LLMs for this task, their methodology was constrained by the plain application, the limitations of early GPT models and a lack of differentiation in benchmark requirements. Our objective is to reevaluate this detection process using the most recent GPT models and a refined prompt template for the Polytope benchmark.

Initially, we applied the prompt template used by Luo et al. (2023) to assess the performance of GPT-3.5 Turbo and GPT-4 Turbo on SUMMAC. Recognizing the distinct requirements of Polytope benchmark in SUMMAC, we crafted a tailored prompt template to better suit Polytope and reevaluated the models' performance. The revised prompt template is detailed below:

Decide if the following summary have any of the specified problems in relation to the corresponding article.

The problems are categorized as omission, addition, or inaccuracy. Omission means Key point is missing from the summary. Addition means Unnec-

essary and irrelevant snippets from the Article are included in the summary. Inaccuracy means some information in the summary is not supported by the article.
Article:
{{ Article }}
Summary:
{{ Summary }}
If the summary has any of the above problems, answer 'No'. Otherwise, answer 'Yes'. Answer (Yes or No):

Comparing with the original prompt, we let the model detect omission, addition, and inaccuracy summary to fit the annotation of Polytope. With the experiments above, we set a new baseline for summary factual inconsistency detection with LLMs.

3.2 SIFiD

Building on prior research in Summary Inconsistency Detection, we propose SIFiD (**S**ummary **I**nconsistency Detection with **Fi**ltered **D**ocument), a method designed to enhance detection capabilities by filtering irrelevant content from documents. Inspired by the SUMMAC methodology, which calculates sentence-level entailment scores to identify factual inconsistencies, SIFiD constructs a relevance matrix to filter out irrelevant sentences, focusing the inconsistency check solely on the filtered document and its summary. An illustrative depiction of this process is presented in Figure 1.

Given a document $D = \{d_k\}_{0 \leq k \leq M}$ and its summary $S = \{s_k\}_{0 \leq k \leq N}$, where d_k and s_k represent the k^{th} sentence in D and S, respectively, and M, N are the total number of sentences in each, we first calculate a relevance matrix R:

$$\begin{aligned} R &= \{\texttt{Scorer}(d_i, s_j)\}_{0 \leq i \leq M, 0 \leq j \leq N} \\ &= \{r_{i,j}\}_{0 \leq i \leq M, 0 \leq j \leq N}. \end{aligned} \tag{1}$$

Here, $r_{i,j}$ denotes the relevance score between the document-summary sentence pair (d_i, s_j), computed using either entailment scores as per the SUMMAC method or semantic cosine similarity via the sentence-transformers library[1].

Subsequently, we apply max pooling across matrix rows to extract the highest relevance score $R^p = \{d_i^p\}_{0 \leq i \leq M}$ for each document sentence. We then establish a threshold β to filter sentences, employing a window method to ensure contextual continuity:

$$D^{\texttt{filtered}} = \{d_{x-1}, d_x, d_{x+1}\}_{d_x > \beta, 0 \leq x \leq M}. \tag{2}$$

This approach retains a sentence d_x (and its immediate neighbors) if $d_x > \beta$, as demonstrated in Figure 1, where Sentence 6 is included within the window of Sentence 7.

The filtered document $D^{\texttt{filtered}}$ and the summary S are then integrated into the prompt template for evaluation by an LLM. Following Luo et al. (2023), we simply determine factual consistency by identifying whether the LLM's response contains "Yes" (indicating consistency) or "No".

3.3 Scorer

We use one of the two distinct scoring mechanisms to evaluate the relevance between document sentences and summary sentences.

Entailment Scorer: We adopt the entailment scoring approach as proposed by Laban et al. (2022), which utilizes a Natural Language Inference (NLI) model (Schuster et al., 2021). The net entailment score is calculated by $\texttt{score}_{i,j}^{\texttt{ent}} = e_{i,j}^0 - c_{i,j}$, where $e_{i,j}^0$ and $c_{i,j}$ are the initial entailment score and contradiction score directly calculated by the NLI model on (d_i, s_j). The net entailment score reflects the degree to which the summary sentence is supported by the document sentence without contradiction.

Semantic Similarity Scorer: For assessing semantic similarity, we leverage the sentence-transformers library to generate embeddings for both document and summary sentences, denoted as h_i^d and h_j^s, respectively. The cosine similarity between these embeddings serves as the measure of semantic similarity, which is $\texttt{score}^{\texttt{sim}} = \cos(h_i^d, h_j^s)$, where $\texttt{score}^{\texttt{sim}}$ quantifies the semantic closeness between the document and summary sentences. This metric enables us to identify and assess the degree of semantic overlap.

4 Experiments

In this section, we detail the experiments conducted with GPT models and the SIFiD method on SUMMAC (Laban et al., 2022). We evaluated the performance of GPT-3.5, GPT-4, and SIFiD against a range of state-of-the-art approaches, including traditional methods such as DAE (Goyal and Durrett, 2020), FEQA (Durmus et al., 2020), QuestEval (Scialom et al., 2021), SummaC-ZS, SummaC-Conv (Laban et al., 2022), and an LLM-based method proposed by Luo et al. (2023).

Following previous research (Luo et al., 2023; Laban et al., 2022), we report the balanced ac-

[1] https://huggingface.co/sentence-transformers

Table 1: Experiment results on SUMMAC. Values in brackets represent balanced accuracy without redesigned prompt template. "+CoT" means using chain-of-thought method.

Method	**CoGenSum**	**XsumFaith**	**Polytope**	**FactCC**	**SummEval**	**FRANK**	**Avg.**
DAE	63.4	50.8	62.8	75.9	70.3	61.7	64.2
FEQA	61.0	56.0	57.8	53.6	53.8	69.9	58.7
QuestEval	62.6	62.1	70.3	66.6	72.5	82.1	69.4
SUMMAC-ZS	70.4	58.4	62.0	83.8	78.7	79.0	72.1
SUMMAC-Conv	64.7	**66.4**	62.7	89.5	81.7	81.6	74.43
Luo et al. (2023)	63.3	64.7	56.9	74.7	76.5	80.9	69.5
+CoT	74.3	63.1	61.4	79.5	83.3	82.6	74.0
GPT-3.5 Turbo	59.9	67.6	41.0(57.9)	71.3	81.4	80.2	66.9(69.7)
+CoT	65.2	62.3	49.5(59.1)	79.1	77.4	81.4	69.2(70.8)
SIFiD-Entailment	65.5	63.9	37.5	81.0	79.0	81.6	68.1
+CoT	65.7	60.3	52.7	82.3	79.3	81.6	70.3
SIFiD-Similarity	65.4	64.7	35.3	76.0	74.5	80.1	66.0
+CoT	64.3	59.7	52.8	81.7	76.6	80.4	69.2
GPT-4 Turbo	80.9	61.0	66.0(60.9)	89.6	**88.0**	87.4	78.8(78.0)
+CoT	80.2	**66.4**	62.1(61.4)	87.8	86.2	85.6	78.1(78.0)
SIFiD-Entailment	82.8	58.9	**74.4**	89.4	87.5	86.1	**79.9**
+CoT	**83.2**	60.6	61.7	89.4	87.1	85.8	78.0
SIFiD-Similarity	83.1	60.2	71.0	**90.6**	86.8	**87.7**	**79.9**
+CoT	82.9	65.0	69.3	91.3	84.6	86.0	79.8

curacy for SUMMAC. The experimental results were obtained from Luo et al. (2023). Our experiments utilized `gpt-3.5-turbo-1106` and `gpt-4-1106-preview`[2]. For the SIFiD configuration, we applied $\beta = 0.0$ for entailment-based filtering and $\beta = 0.5$ for semantic similarity-based filtering, observing a 61.3% and 67% sentence removal rate on average across benchmarks, respectively. We use `all-mpnet-base-v2` for sentence-transformers.

4.1 Results and Analysis

The experimental outcomes are summarized in Table 1, leading to several insights on LLM-based summary factual inconsistency detection:

Prefer GPT-4 Over GPT-3.5. Analysis indicates that previous LLM-based methods, though superior to many traditional techniques, underperform compared to SUMMAC-Conv. This discrepancy is attributed to the limited capabilities of the GPT-3.5 model. Our reevaluation with the GPT-3.5 Turbo model yielded results similar to those of Luo et al. (2023). However, substituting GPT-3.5 with GPT-4 Turbo significantly enhanced performance, from 69.7 to 78.0, underscoring GPT-4's advanced language comprehension.

Adopt Benchmark-Specific Prompt Templates. The effectiveness of a single prompt template across different benchmarks is limited due to the unique requirements of each benchmark. Traditional methods typically incorporate benchmark-specific training, which mitigates task variance. In contrast, LLMs rely on the provided instructions, necessitating tailored prompt templates. Adjusting the prompt template for Polytope increased GPT-4's performance from 60.9 to 66.0, elevating the overall average to 78.8. However, this adjustment resulted in a performance decline for GPT-3.5 on Polytope, from 57.9 to 41.0, highlighting GPT-3.5's inferior prompt comprehension.

Enhanced Performance with SIFiD on GPT-4. Integrating SIFiD with GPT-4 further improved its performance to 79.9. SIFiD's selective filtering of sentences enhances document relevance to the summary, simplifying factual inconsistency detection. This approach did not yield similar benefits for GPT-3.5, possibly due to its reduced efficacy in processing less fluent filtered documents.

Mixed Results with Chain-of-Thought (CoT). Applying CoT techniques did not uniformly benefit all methods. While GPT-3.5 saw improvements, GPT-4's performance declined, suggesting GPT-4's innate proficiency in inconsistency detection without CoT. Additionally, CoT might introduce biases that could negatively influence outcomes.

5 Conclusion

In this study, we advance the field of LLM-based summary factual inconsistency detection by evaluating the performance of the latest GPT models, thereby establishing new benchmarks for future research. We introduce SIFiD, a novel, efficient, and effective approach that computes a relevance matrix at the sentence level between the document and its summary. This method filters out irrelevant sentences from the document before employing LLMs for inconsistency detection. Our experimental findings on the SUMMAC dataset demonstrate that SIFiD significantly enhances the performance of advanced GPT models in detecting factual inconsistencies, highlighting its potential to facilitate

[2] https://platform.openai.com/docs/models

more accurate and resource-efficient research in this domain.

Limitations

The principal constraint of employing LLMs for summary factual inconsistency detection lies in the costs associated with using such powerful models. As elaborated in Section 4, this task necessitates LLMs with substantial capabilities, where only models at or beyond the level of GPT-4 are deemed sufficient. Despite our SIFid method's ability to eliminate over 60% of document sentences, thereby reducing the input size, the financial implications of utilizing GPT-4 for inconsistency detection remain considerable. Nonetheless, given the swift advancements in LLM technology, we anticipate a substantial reduction in these costs. This progression is expected to make the application of such models more feasible and economically viable for widespread real-world applications.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

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.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 166–175.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603.

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

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

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

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Christopher Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.