SummaCoz: A Dataset for Improving the Interpretability of Factual Consistency Detection for Summarization

Ge Luo, Weisi Fan, Miaoran Li, Guoruizhe Sun, Runlong Zhang, Chenyu Xu
and **Forrest Sheng Bao***
Iowa State University, Ames, IA, USA
{gluo,weisifan,limr,gsun1,runlongz,chenyux}@iastate.edu
{forrest.bao}@gmail.com

Abstract

Summarization is an important application of Large Language Models (LLMs). When judging the quality of a summary, factual consistency holds a significant weight. Despite numerous efforts dedicated to building factual inconsistency detectors, the exploration of explanability remains limited among existing effort. In this study, we incorporate both human-annotated and model-generated natural language explanations elucidating how a summary deviates and thus becomes inconsistent with its source article. We build our explanation-augmented dataset on top of the widely used SummaC summarization consistency benchmark. Additionally, we develop an inconsistency detector that is jointly trained with the collected explanations. Our findings demonstrate that integrating explanations during training not only enables the model to provide rationales for its judgments but also enhances its accuracy significantly.

1 Introduction

Factual consistency checking in summarization assesses whether the information presented in a machine-generated summary aligns, and thus is consistent, with its source document(s). This task has gained prominence in recent years due to concerns about abstractive summarization systems generating erroneous or "hallucinated" content and thus compromising their reliability (Kryscinski et al., 2020). Traditionally, the task of factual consistency checking has been formulated as a binary classification problem, where the output label indicates whether the summary is consistent or not. However, a binary label alone offers limited insights into the nature of inconsistencies. When an inconsistency is identified, it would be better to pinpoint which part of the summary is inconsistent, cite corresponding information from the source document, and explain the differences between the summary and the source. This explanatory information serves as valuable guidance for manually or automatically post-editing the summary to rewrite and rectify any inconsistencies (Dong et al., 2020; Mishra et al., 2024).

Current research focuses on detecting inconsistencies in summaries without delving into explanations . Approaches such as MFMA (Lee et al., 2022), FalseSumm (Utama et al., 2022), and NonFactS (Soleimani et al., 2023) employ entailment classification methods that only yield binary classification outputs. There have been a few datasets (Maynez et al., 2020; Wu et al., 2023) that annotate the spans in the summaries that are inconsistent to their respective sources.But the information or text spans in the source documents that correspond to and falsify such inconsistent spans are missing in these datasets.

In this paper, to facilitate the research in summarization consistency, we curate a dataset that includes not only the binary consistency labels but also natural language explanations as to why a summary is inconsistent to its source. To do so, we extend the SummaC (Laban et al., 2022) summarization consistency benchmark by augmenting its binary labels with both human-annotated and LLM-generated explanations. The resulting dataset, called SummaCoz, not only adds interpretablity into factual consistency evaluation but also, as to be shown later in this paper, sheds lights on the challenges of detecting inconsistent summaries.

With SummaCoz[1], we then train a text generation model that serves as not only a classifier but also a reasoner that justifies its classification judgement. Empirical evaluation demonstrates that leveraging explanations during training results in

* Forrest Bao is now with Funix.io

[1] Our code and data is publicly available at https://github.com/NKWBTB/SummaCoz.

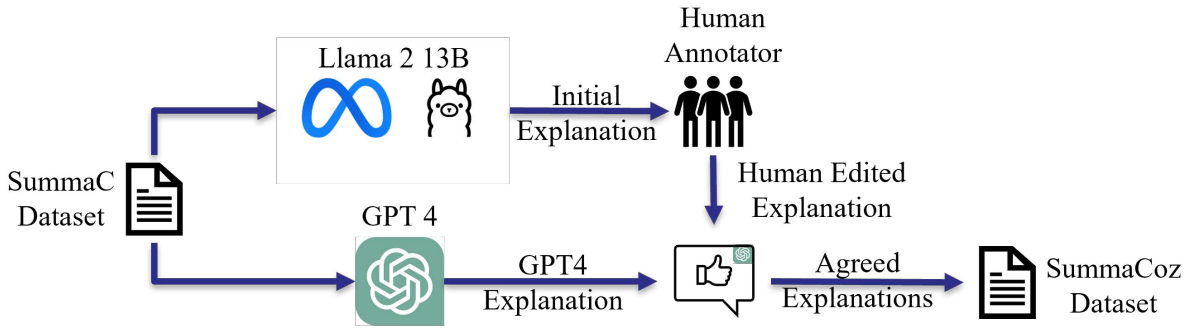

Figure 1: The collection process for the SummaCoz dataset. Two LLM-based explanations for each sample are generated by the Llama-2 and GPT model. Human annotators post edit on the Llama-2 explanations. Finally, the agreed explanations produced by human and GPT-4 are included in the dataset. The disagreed explanations are discarded.

a factual consistency detector not only adds interpretability to it but also makes it more accurate.

2 Curating the SummaCoz Dataset

The SummaCoz dataset is curated (Figure 1) in a semi-automatic manner followed by rigorous quality check. The curation began with inconsistent summaries and their corresponding sources from the validation split of the SummaC summarization benchmark (Laban et al., 2022). Next, we employed a Large Language Model (LLM) to generate an initial explanation for each of such inconsistent summaries as to why it deviates from its source. The LLM-generated explanation is subsequently refined by human annotators. Finally, to ensure the quality of the explanations, we cross-check the human-edited annotations with a more advanced LLM and exclude samples where the human annotators and the advanced LLM disagree.

We focus on inconsistent summaries only in SummaCoz for several reasons. First, recognizing consistent summaries is relatively easy. ChatGPT can accurately recognize over 95% of consistent summaries (Luo et al., 2023). Second, it's challenging to pinpoint why a summary is supported by its corresponding article. Hence, this study skips consistent summaries. Our effort is focused on understanding and addressing the nuances and challenges associated with identifying inconsistent summaries.

2.1 Label-elicited Initial Explanation Generation

Given the labor-intensive nature of requesting humans to generate explanations from scratch, we adopt a semi-automatic approach to produce the explanations. In this approach, human annotators post-edit explanations generated by an LLM. We devise a label-first prompt to elicit post-hoc reasoning. Specifically, the prompt explicitly informs the LLM that the summary is inconsistent with the source (highlighted in red below) and instructs the LLM to justify why. The prompt template is shown below:

> *Note that consistency means all information in the summary is supported by the article. It's known that the following summary is not consistent with the article. Find out why.*
> *<Article>{Article}</Article>*
> *<Summary>{Summary}</Summary>*
> *Explain your reasoning step by step:*

Using the prompt template above, initial explanations can be generated from the samples and labels in an summarization factual consistency dataset by an LLM. In this study, we use inconsistent summary and document pairs from the SummaC benchmark's validation split and employe `Llama-2-13b-chat-hf` (Touvron et al., 2023) as the LLM.

2.2 Explanation Post-editing by Humans

Although elicited by the ground truth label, the explanations generated by an LLM above may be wrong. Therefore, we incorporate human annotators to examine the explanations and post-edit them if needed.

The human annotators are given the typology of factual errors from the FRANK (Pagnoni et al., 2021) dataset to understand what kinds of issues in a summary are considered as consistency errors. Frank's typology provides 8 categories of factual

consistency errors: Predicate Error, Entity Error, Circumstance Error, Coreference Error, Discourse Link Error, Out of Article Error, Grammatical Error and Other Error.

The annotators are then furnished with examples and guidelines illustrating how to edit inaccurate explanations. They are instructed to organize each explanation in a numbered list. The annotators are tasked with incorporating specific aspects into the explanation process. These include identifying the inconsistent text spans in the summary, citing the corresponding information in the source, and optionally, specifying the differences. The annotations are conducted by 6 authors of this study, who are undergraduate and graduate students possessing backgrounds in computer science and natural language processing. Each sample has one human annotation. Details of the annotation guidelines and examples are in Appendix A.6.

In order to diversify the explanations collected in SummaCoz and validate the quality of human annotations, we employee a more powerful LLM, OpenAI's GPT-4 (OpenAI, 2023) [2], to generate rationales from the same data that is fed into Llama-2 to generate initial explanations. This addition allows for a broader spectrum of explanations.

Finally, we filter out samples where the explanation from GPT-4 contradicts the human post-edited explanation, ensuring consistency and accuracy in the dataset. Details can be found in Appendix A.2. As a result, there are 755 distinctive pairs of inconsistent summaries and source documents in SummaCoz. Each pair has two explanations, one human post-edited from Llama-2's initial explanation and the other from GPT4 without human intervention. The statistics of SummaCoz are shown in Appendix Table 5.

3 Experiments

We conduct a series of experiments to assess the effectiveness of utilizing natural language explanations in building a factual consistency detector. Through these experiments, we aim to evaluate the performance our approach in providing interpretable explanations for summarization factual inconsistencies.

3.1 Settings

While existing methods formulate the problem as a text classification problem, such as FactFT(Luo et al., 2024) which demonstrated the effectiveness of transferring knowledge from NLI models to build a robust summarization consistency classifier, we adopt a different framework that is purely text-to-text. Specifically, we employ a text generation model capable of jointly outputting both the label text and the accompanying explanation. This approach enables us to leverage the flexibility and expressiveness of text generation techniques in providing more comprehensive and nuanced explanations for summarization factual inconsistencies.

Foundation Models: Several foundation models of varying sizes, ranging from 0.8B to 11B parameters, are employed:

- **Flan-T5-0.8B/3B/11B** (Chung et al., 2022), is an Encoder-Decoder transformer trained on the Flan collection (Wei et al.), which comprises seven NLI datasets.
- **Mistral-7B-Instruct-v0.2** (Jiang et al., 2023), is a Decoder-only transformer with 7B parameters.

Prompt: We employ the following NLI prompt to elicit the NLI knowledge that is embedded in the model:

> *Is the hypothesis true based on the premise?*
> *Premise: {article}*
> *Hypothesis: {summary}*

Target Output:

> *Yes, the hypothesis is true.*
>
> *OR*
>
> *No, the hypothesis is not true. {explanation}*

Training Settings: Given that the SummaCoz dataset exclusively contains inconsistent summaries, we randomly sample an equal number of consistent summaries from the SummaC's validation set to construct a balanced training set. The remaining samples from the SummaC's validation set are reserved for validation. Following the methodology established in FactFT (Luo et al., 2024), we utilize LoRA (Hu et al., 2021) as the parameter-efficient fine-tuning technique to train the model with only a subset of parameters compared to full fine-tuning. The hyperparameters used in this process are provided in Appendix Table 6. The training is conducted on a single NVIDIA A100 GPU with 80GB of VRAM.

[2] Specifically `gpt-4-1106-preview`

SummaC Test Set	Flan-T5-0.8B		Flan-T5-3B		Mistral-7B		Flan-T5-11B	
	LabelOnly	w/Explain	LabelOnly	w/Explain	LabelOnly	w/Explain	LabelOnly	w/Explain
CoGenSum	79.8	80.3	73.2	80.3	77.7	77.3	86.0	86.2
FactCC	87.4	86.4	87.0	87.6	82.4	88.3	89.1	88.8
Frank	86.5	85.7	83.9	87.3	86.2	87.7	87.6	87.7
SummEval	84.5	82.2	85.0	85.4	84.2	87.8	81.6	86.4
XSumFaith	58.6	63.6	64.0	63.4	65.6	64.3	61.3	60.8
Average	79.4	**79.6**	78.6	**80.8**	79.2	**81.1**	81.1	**82.0**

Table 1: Balanced accuracy (BA) on SummaC test set

		P	R	F1	BA
Flan-T5-0.8B	LabelOnly	61.2	25.5	36.0	60.4
	w/Explain	60.0	7.4	13.1	53.0
Flan-T5-3B	LabelOnly	79.2	18.6	30.2	58.6
	w/Explain	72.1	24.0	36.0	60.6
Mistral-7B	LabelOnly	**92.6**	12.3	21.7	56.0
	w/Explain	69.1	27.5	39.3	61.9
Flan-T5-11B	LabelOnly	68.8	35.3	46.6	65.3
	w/Explain	68.6	**48.5**	**56.9**	**71.0**

Table 2: Results for the RAGTruth test set: (P)recision, (R)ecall, F1 score for inconsistent sample as a hit. BA stands for Balanced Accuracy.

Evaluation Settings: We conduct evaluations on the trained model using the test set of the SummaC benchmark, excluding Polytope (Huang et al., 2020) due to reported consistency labeling issues in previous works (Fabbri et al., 2022; Tang et al., 2023). Additionally, we incorporate the test split from the summarization task of the RAGTruth dataset (Wu et al., 2023), which poses a greater challenge as it contains longer inconsistent summaries generated by more recent LLMs. For all trained models, we use greedy decoding during the generation process, and limit the number of generated tokens to 512 using the NLI prompt.

3.2 Results

We compare models trained under two different settings to examine the impact of introducing explanations in the text generation process:

- **LabelOnly**: the model is trained to output the label only, without incorporating any explanation during training.
- **w/Explain**: the model is trained to first output the label and then provide an explanation if the predicted label is "inconsistent".

We present the balanced accuracy results on the SummaC benchmark in Table 1. While the performances of models depend on their sizes, the average balanced accuracy for models trained under the w/Explain setting is superior to that under the LabelOnly setting.

For the more challenging RAGTruth benchmark, we provide additional metrics such as precision, recall, and F1 score, alongside balanced accuracy, for a more comprehensive understanding of the classification results in Table 2. The w/Explain setting exhibits a higher recall, F1 score, and balanced accuracy than the LabelOnly setting for model sizes ranging from 3B to 11B. The higher recall suggests that training with explanations enhances the model's ability to recognize more inconsistent summaries. We posit that the additional signal provided by explanations during training contributes to the improvement in model performance. However, the `Flan-T5-0.3B` model does not benefit from training with explanations. We hypothesize that the limited model size may constrain its reasoning ability to make use of the explanations effectively.

In addition to the classification results, we also incorporate quantitative study on the generated explanation in Appendix A.1.

3.3 Error Analysis

By leveraging the explanations generated by the inconsistency detector, we can delve into why the model believes certain parts of the summary are inconsistent, including cases where the model mistakenly classifies a consistent summary as inconsistent. We showcase some of these cases, where the model makes false inferences, in Table 3.

In example 1, the article discusses a woman whose identity remained unknown for decades until recent advancements in genealogy technology successfully identified her. The summary describes her as having remained a "Jane Doe" for decades. The explanation attributes the inconsistency to the usage of the phrase "Jane Doe." However, "Jane Doe" is a commonly used placeholder name in the United States and the United Kingdom when the

ID	Article sentence(s)	Summary sentences(s)	Inaccurate model explanation
1	*In a significant development in a 1988 cold case, the Georgia Bureau of Investigation (GBI) announced they have successfully identified the human remains discovered in a Jenkins County dumpster on February 14, 1988 ... According to officials, in 2023, through advanced DNA analysis, a genealogical search generated investigative leads, eventually leading to the identification of Chong Un Kim.*	*In 1988, a woman's body was found in a dumpster in Georgia. Despite efforts to identify her, she remained a Jane Doe for decades. Recently, the Georgia Bureau of Investigation used advanced DNA testing and genealogy technology to identify her as Chong Un Kim.*	The hypothesis states that the victim remained a **Jane Doe** for decades, but the premise states that the victim was identified as Chong Un Kim in 2023. The hypothesis incorrectly states that the victim remained a Jane Doe for decades.
2	*... Dickens, 33, appeared in court Wednesday on a charge of disseminating information related to terrorist acts. A judge set a $10,000 bond for her and banned her from social media. She was then transferred to Fulton County Jail, where she **remained until bonding out** at 6:19 a.m. Thursday, according to county sheriff's office spokeswoman Tracy Flanagan.*	*... She was charged with disseminating information related to terrorist acts and was released on a $10,000 bond with a condition to stay off social media.*	The hypothesis states that Ebony Dickens was charged with disseminating information related to terrorist acts and released on a $10,000 bond with a condition to stay off social media. However, the premise does not mention anything about Ebony Dickens being **released** on a $10,000 bond.
3	*Alaska Airlines Flight 448 was just barely on its way to Los Angeles from Seattle-Tacoma International Airport on Monday afternoon when the pilot reported hearing unusual banging from the cargo hold. "There could be a person in there so **we're going to come back around**," he told air traffic control.*	*An Alaska Airlines flight returned to Seattle-Tacoma International Airport shortly after takeoff due to a person trapped in the cargo hold.*	The hypothesis states that the Alaska Airlines flight returned to Seattle-Tacoma International Airport shortly after takeoff due to a person trapped in the cargo hold. However, the premise does not mention anything about the plane **returning** to the airport.

Table 3: Examples from the RAGTruth test set that are wrongfully explained by the trained `FLAN-T5-11B` model.

true identity of a person is unknown. This case highlights a scenario where knowledge solely from the article may not suffice to determine the consistency of a summary; additional commonsense knowledge is also required.

Example 2 illustrates a scenario where subtle nuances in expression influence the judgment of the trained model. While the summary states that a woman was "released," the article does not explicitly mention her release. Instead, it states that the woman remained in jail until bonding out. This subtle connection may not be recognized by the trained model, leading to an inaccurate explanation.

Example 3 presents a challenging case where subtle reasoning is required to understand the information accurately. The summary mentions that an airplane "returned shortly after takeoff." However, the article does not explicitly state the airplane's return; rather, the information is inferred from the pilot's words indicating a "come back around." The trained model fails to capture this nuanced reasoning, resulting in an inaccurate explanation.

These examples highlight some of the new challenges in recognizing factual inconsistency in summarization, particularly with modern LLMs. Subtle expressions, nuanced reasoning, and the need for commonsense knowledge pose hurdles for automated systems in accurately assessing consistency between summaries and source articles. Addressing these challenges requires advancements in natural language understanding, including the ability to infer context, detect subtle cues, and incorporate external knowledge sources. As summarization models continue to evolve, it becomes increasingly important to develop robust methods for detecting and explaining factual inconsistencies to ensure the reliability and trustworthiness of generated summaries.

4 Conclusion

In this study, we explore the integration of explanations into the inconsistency detector, with the goal of offering insights into the underlying processes driving evaluation outcomes. To achieve this, we compile a dataset named SummaCoz, comprising both human-written and LLM-generated natural language explanations for inconsistent summary and article pairs. Leveraging this dataset, we showcase the effectiveness of training a text-generation model to output both the consistency judgment and explanation simultaneously. This approach empowers users to understand and interpret evaluation results effectively, thereby enhancing transparency and trust in the assessment process.

Acknowledgment

This work is partially supported by NSF grants CNS-2141153.

Limitation

For evaluating natural generation tasks with references, the quality of the reference texts can affect the evaluation quality. The SummaCoz dataset is created with human annotators doing post-editing on LLM's reasoning. In our task, there may be multiple reasons why a summary is not consistent, it is possible our referential explanation does not cover all the reasons. The problem can be mitigated by creating multiple references with different annotators writing reason for the same sample. We would consider the option in the next version of the dataset if time and budget allow.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rockt"aschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Hou Pong Chan, Qi Zeng, and Heng Ji. 2023. Interpretable automatic fine-grained inconsistency detection in text summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6433–6444, Toronto, Canada. Association for Computational Linguistics.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

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. 2023. Mistral 7b.

Pride Kavumba, Ana Brassard, Benjamin Heinzerling, and Kentaro Inui. 2023. Prompting for explanations improves adversarial NLI. is this true? Yes it is true because it weakens superficial cues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2165–2180, Dubrovnik, Croatia. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Ffci: A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, 73:1553–1607.

Wojciech Kryscinski, 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, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. 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.

Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee, and Kyomin Jung. 2022. Masked summarization to generate factually inconsistent summaries for improved factual consistency checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1019–1030, Seattle, United States. Association for Computational Linguistics.

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.

Ge Luo, Weisi Fan, Miaoran Li, Youbiao He, Yinfei Yang, and Forrest Bao. 2024. On the intractability to synthesize factual inconsistencies in summarization. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1026–1037, St. Julian's, Malta. Association for Computational Linguistics.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models.

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

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2023. NonFactS: NonFactual summary generation for factuality evaluation in document summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6405–6419, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Appendix

A.1 Explanation Quality

In addition to evaluating the classification results, we are also keen on assessing the quality of explanations generated by the best performing model, Flan-T5-11B, on the 99 samples classified as true negatives from the RAGTruth test set. To facilitate this assessment, we utilize the span annotations provided by the RAGTruth dataset as a reference. Human annotators are tasked with rating the model-generated explanations on three scales:

- **1** - the explanation is fully correct: the generation mentions at least one of the error span, and no incorrect information is given in the explanation.
- **0.5** - the explanation is partially correct: the explanation mentions at least one of the error span, but some details are not accurate. Or the explanation contains reasons that unrelated to inconsistency.
- **0** - the explanation is not correct, the generation fails to capture the inconsistency.

For the inter-annotator agreement, two annotators reported the interval Krippendorff's alpha of 0.68. The final rating is determined by averaging the scores provided by the annotators, resulting in a final rating on a scale of 0, 0.25, 0.5, 0.75, or 1.

The average rating distribution is illustrated in Figure 2. Approximately 78% of the generated explanations accurately or partially accurately represent the inconsistency in the true negative samples. This finding underscores that while a model may correctly predict the consistency label, the explanation of the decision process may not always be correct.

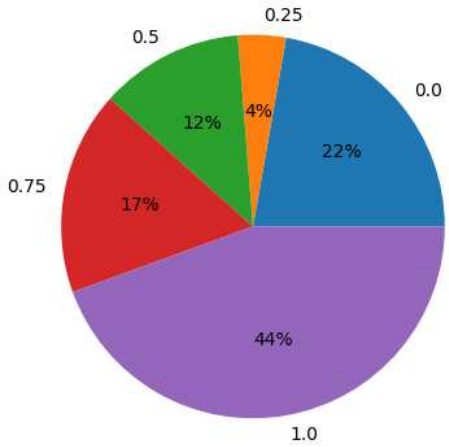

Figure 2: Average human rating to the explanations generated by `Flan-T5-11B` for the true negative samples from RAGTruth test set.

A.2 Explanation filtering

This section demonstrate the steps to filter out the disagreed samples between human explanation and GPT-4 generated explanation. Collecting human feedbacks for the alignment between the two explanations can be laborious, we aim to evaluate the agreement of explanations automatically. Similarity measures like ROUGE (Lin, 2004) and BertScore (Zhang et al., 2019) rely on lexical or embedding similarity. They may not be suitable to evaluate the logical reasoning between texts because the reasoning trace could contain many citation texts from the summary and article, making two logically different reasoning traces lexically or semantically similar.

We prompt an LLM[3] to rating the reasoning steps. With the human explanation segmented in to bullet points, we set the rating policy as such: each bullet point in the reference is worth 1 point and the GPT-4 reasoning gets 1 point if it covers or mentions one bullet point. The prompt used for the LLM judge is:

You are an assessor to give judgment on

[3] `gpt-3.5-turbo-0301` model is used in this work.

a reasoning problem. Here is the text to be assessed:
<text></text>
Does the above text mention or contain the following reference reasoning step:
<reference></reference>
Answer (yes or no):

To achieve the rating process mentioned above, we iteratively feed each bullet point of the human explanation as reference to the LLM to obtain the agreement score between human explanation and GPT-4 explanation. We only include samples with full-point of agreement in the SummaCoz dataset. There were 1323 human annotated samples in total, 755 samples of the human edited explanations agree with GPT-4 explanations. The disagreed 568 samples are discarded.

A.3 Zero-shot Results

Setting	Model	Prec	Recall	F1	BA
Zero-shot	Mistral-7B	34.2	6.4	10.7	51.4
	Flan-T5-11B	70.0	6.9	12.5	53.0
	GPT-4	83.8	30.4	44.6	64.3
	GPT-3.5-turbo	82.2	18.1	29.7	58.5
Finetune-LabelOnly	Mistral-7B	**92.6**	12.3	21.7	56.0
	Flan-T5-11B	68.8	35.3	46.6	65.3
Finetune w/Explain	Mistral-7B	69.1	27.5	39.3	61.9
	Flan-T5-11B	68.6	**48.5**	**56.9**	**71.0**

Table 4: Results for the RAGTruth test set: (P)recision, (R)ecall, F1 score for inconsistent sample as a hit. BA stands for Balanced Accuracy. GPT-4 as `GPT-4-0613` and GPT-3.5-turbo as `GPT-3.5-Turbo-0125`.

In addition to the finetuning results, we present the zero-shot results in Table 4. For both Mistral-7B and Flan-T5-11B, an improvement in metric scores is observed when comparing the zero-shot versions with the finetuned models. Notably, the Flan-T5-11B model finetuned with explanations outperforms GPT-4 (zero-shot) in terms of F1 score and balanced accuracy.

A.4 Related Work

Factual consistency checking is a crucial aspect in summarization evaluation (Koto et al., 2022). The task shares strong similarities with Natural Language Inference (NLI) (Bowman et al., 2015), which involves determining the truthfulness of a "hypothesis" given a "premise." In NLI, models classify whether a hypothesis is true (entailment), false (contradiction), or undetermined (neutral) based on a given premise. NLI datasets, like SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), and ANLI (Nie et al., 2020), typically consist of short hypotheses and premises, with the hypothesis often being a single sentence containing only a few atomic facts. A key distinction between NLI and summarization factual consistency checking is that while a summary may be logically entailed by the article, it may still lack consistency. For instance, the statement that *"John died before 2022"* can be logically entailed from the other statement *"John died before 1945"* but the former statement is not consistent to the latter. This illustrates the nuanced nature of consistency checking in summarization. Recognizing the overlap between the two tasks, previous work such as FactFT (Luo et al., 2024) has successfully leveraged knowledge transfer from NLI models to develop efficient summarization consistency classifiers. We extend upon the FactFT framework by incorporating natural language explanations into a consistency detector, thereby enhancing the interpretability and utility of the model.

In terms of providing explanation in consistency evaluation, FINEGRAINFACT (Chan et al., 2023) incorporates highlighted semantic frames and a classification of factual error types in its approach; however, the results still require further inspection for clear interpretation. Therefore, we leverage the free-form natural language, the easiest way for humans to comprehend, as the explanation in our method. Efforts to integrate natural language explanations into the NLI task have been undertaken in various studies. For instance, E-SNLI (Camburu et al., 2018) expands upon the SNLI dataset by including human-annotated explanations. Additionally, segments of the ANLI dataset feature human-written explanations regarding the entailment relationships. Previous research has demonstrated the benefits of prompting for explanations in improving Adversarial NLI (Kavumba et al., 2023). This approach enhances model robustness by mitigating reliance on superficial cues through training with explanations. In this study, we shift our focus towards incorporating human explanations into the more challenging task of detecting summarization factual inconsistencies.

In comparison with some recent works using LLM-as-judge for consistency checking, FactScore (Min et al., 2023) and FacTool (Chern et al., 2023) break the evaluation into atomic facts. Evaluating multiple atomic facts needs multiple

LLM calls and introduces extra costs. The decoupled atomic fact may also be hallucinated and affect the evaluation result. Our method performs the end-to-end evaluation in a single LLM call. In the domain of study, FactScore focuses evaluations on Biography generation, Long-form response. FacTool targets at QA, Code, Math evaluations. Our SummaCoz dataset serves as an explanation extended summarization consistency evaluation dataset. The dataset also allows fine-tuning open-source LLMs for a more transparent evaluation rather than relying on closed-source GPT models.

A.5 Dataset Statistics & Training Hyperparameters

# of Sample	Total	755×2
	-XSUM	362×2
	-CNNDM	393×2
# of Source Tokens	Min	66
	Max	1978
	Avg.	710
# of Summary Tokens	Min	8
	Max	189
	Avg.	43
# of Explanation Tokens	Min	24
	Max	584
	Avg.	227

Table 5: Dataset statistics of the SummaCoz. Article sources are from the CNNDM (See et al., 2017) and XSum (Narayan et al., 2018) datasets.

Hyperparameter	**Value**
epochs	10
learning rate	1e-4
warmup ratio	0.1
batch size	1
gradient accumulation	8
optimizer	paged_adamw_8bit
LoRA	
lora r	16
lora alpha	32
lora dropout	0.05
bias	lora_only

Table 6: Hyperparameter setting for training

See Table 5 and 6.

A.6 Guidelines of annotations

We gave the following guidelines to the annotators on how to edit the unfaithful explanations:

1. Put your annotation after "Corrected:". Even the reasoning is correct, still copy the steps under "Corrected:".
2. Delete incorrect reasoning and keep the **correct** reasoning.
3. Delete irrelevant sentences and **only** keep the reasoning, e.g. "Sure! Here's why the summary is not consistent with the article:" should be deleted.
4. Unify the format as an ordered list, following the format provided as Figure 3.
5. Add your own reasoning if necessary. Keep the reasoning **short** and **simple**.

In addition to the guide, we also provide example annotations to the annotators as in Figure 4 5 6 and 7.

1.<what the summary states>
2.<what the article actually states>
3.<where the inconsistency is >
4.<optionally more steps>
5.<.....>

Figure 3: Annotation Format

Note that consistency means all information in the summary is supported by the article.
It's known that the following summary is not consistent with the article.
Find out why.

<Article>
Mr Katter said the Government believes Mr Gordon would quit after he was recently accused of domestic violence.
</Article>

<Summary>
Mr Katter said he would quit after he was accused of domestic violence.
</Summary>

Explain your reasoning step by step:
1. The summary states that Mr Katter would quit after being accused of domestic violence.
2. However, the article states that the Government believes Mr Gordon would quit after being recently accused of domestic violence.
3. The inconsistency is due to the use of the word'recently'. The article implies that the Government believes Mr Gordon would quit after being recently accused, while the summary states that Mr Katter would quit after being accused.
4. To maintain consistency, the summary should use the same word as the article,'recently'.

Corrected:
1. The summary states that Mr Katter would quit after being accused of domestic violence.
2. However, the article states that the Government believes Mr Gordon would quit after being recently accused of domestic violence.
3. The inconsistency is due to the error that the article implies that the Government believes Mr Gordon would quit, while the summary states that Mr Katter would quit.

Figure 4: Annotation Example 1

Note that consistency means all information in the summary is supported by the article.
It's known that the following summary is not consistent with the article.
Find out why.

<Article>
Barcelona club president Josep Maria Bartomeu has insisted that the La Liga leaders have no plans to replace Luis Enrique and they're 'very happy' with him.
</Article>

<Summary>
Josep Maria Bartomeu says the La Liga leaders are very happy with him.
</Summary>

Explain your reasoning step by step:
1. The summary claims that Bartomeu is happy with himself, which contradicts the article's claim that the La Liga leaders are very happy with him.
2. The article provides evidence for the claim that the La Liga leaders are very happy with him, while the summary does not.
3. Therefore, the summary is not consistent with the article.

Corrected:
1. The summary claims that Bartomeu says the La Liga leaders are very happy with him.
2. The article provides evidence for the claim that the La Liga leaders have no plans to replace Luis Enrique and they're 'very happy' with him.
3. Therefore, the summary is not consistent with article on that the summary implies the La Liga leaders are very happy with Bartomeu, but the article implies the La Liga leaders are very happy with **Luis Enrique**.

Figure 5: Annotation Example 2

Note that consistency means all information in the summary is supported by the article.
It's known that the following summary is not consistent with the article.
Find out why.

<Article>
Goldfish are being caught weighing up to 2kg and koi carp up to 8kg and one metre in length.
</Article>

<Summary>
Goldfish are being caught weighing up to 8kg and one metre in length.
</Summary>

Explain your reasoning step by step:
1. The article states that goldfish can weigh up to 8kg, while the summary states that koi carp can weigh up to 8kg.
2. The article also states that koi carp can reach a length of 1 metre, while the summary states that goldfish can reach a length of up to 8cm.
3. Therefore, the summary is not consistent with the article.

Corrected:
1. The summary states that **goldfish** can weigh up to 8kg, while the article states that **koi carp** can weigh up to 8kg.

Figure 6: Annotation Example 3

Note that consistency means all information in the summary is supported by the article.
It's known that the following summary is not consistent with the article.
Find out why.

<Article>
Paul Merson had another dig at Andros Townsend after his appearance for Tottenham against Burnley ...Townsend hit back at Merson on Twitter after scoring for England against Italy.
</Article>

<Summary>
Paul Merson had another dig at andros townsend after scoring for England against Italy.
</Summary>

Explain your reasoning step by step:
1. The summary states that Paul Merson had another dig at Andros Townsend after scoring for England against Italy.
2. However, the article mentions that Andros Townsend was not even in the match against Italy.
3. Therefore, the summary is not consistent with the article.

Corrected:
1. The summary states that Paul Merson had another dig at Andros Townsend after scoring for England against Italy.
2. However, the article mentions that Paul Merson had another dig at Andros Townsend after his appearance for Tottenham against Burnley.
3. Therefore, the summary is not consistent with the article on the time of event where the summary states **after scoring for England against Italy**, while the article states **after his appearance for Tottenham against Burnley**.

Figure 7: Annotation Example 4