Explainable Hallucination through Natural Language Inference Mapping

Wei-Fan Chen[1,3] **Zhixue Zhao**[2] **Akbar Karimi**[1,3] **Lucie Flek**[1,3]
[1]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany
[2]Computer Science School, University of Sheffield, United Kingdom
[3]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
chen@bit.uni-bonn.de

Abstract

Large language models (LLMs) often generate hallucinated content, making it crucial to identify and quantify inconsistencies in their outputs. We introduce HaluMap, a post-hoc framework that detects hallucinations by mapping entailment and contradiction relations between source inputs and generated outputs using a natural language inference (NLI) model. To improve reliability, we propose a calibration step leveraging intra-text relations to refine predictions. HaluMap outperforms state-of-the-art NLI-based methods by five percentage points compared to other training-free approaches, while providing clear, interpretable explanations. As a training-free and model-agnostic approach, HaluMap offers a practical solution for verifying LLM outputs across diverse NLP tasks. The resources of this paper are available at https://github.com/caisa-lab/acl25-halumap.

1 Introduction

Despite their advanced capabilities in generating fluent and human-like texts, Large Language Models (LLMs) are still limited by their *hallucination* problem (Augenstein et al., 2024; Huang et al., 2025), where LLMs generate text containing fabricated information, *factuality hallucinations*, or inconsistent with the input, *faithfulness hallucinations* (Maynez et al., 2020). Hallucination in LLMs occurs in many NLP applications, particularly evident in summarization, where models may generate summaries that are partially or entirely unsupported by the source document (Maynez et al., 2020). Accurately identifying and quantifying these hallucinated segments within the summary is crucial not only to improve the reliability of LLM outputs but also to enable targeted revisions for enhanced accuracy and trustworthiness in practical deployments.

In particular, we are interested in the faithfulness hallucination, where the generated output misses *logical consistency* compared to the input (Huang et al., 2025). To find such (in)consistency, many existing methods use NLI to approach the faithfulness hallucination (Falke et al., 2019; Goyal and Durrett, 2020; Laban et al., 2022), where they mainly use the entailment or contradiction probabilities from an NLI model to capture the consistency or the inconsistency between a document and its summary. While such NLI approaches successfully capture the overall hallucination in general, it is still unclear which parts of the summary are hallucinated as shown in Figure 1(a).

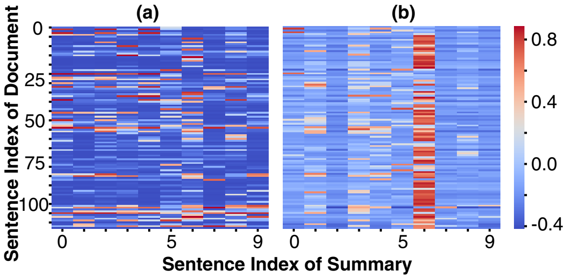

Figure 1: (a) Vanilla NLI visualization doesn't reveal where the hallucination is. (b) Calibrated HaluMap shows that the 6th sentence in the summary is hallucinated in context of the given document. Higher values indicate hallucinations.

To close the gap, we propose **HaluMap**, a framework designed to build heatmap-like relations between the source input and the generated output for faithful hallucination detection. As shown in Figure 1(b), HaluMap highlights the hallucinated sections of the summary, thereby enhancing the explainability of the hallucination detection tasks. HaluMap operates by first decomposing the source and the output into segments. It then employs an NLI model to examine the entailment and contradiction relations between any two segments. The two relations enable the identification of sections in the texts that do not align, indicating potential hallucinations. To enhance the precision of entailment

Findings of the Association for Computational Linguistics: ACL 2025, pages 1888–1896
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

and contradiction relations as mentioned, we introduce **SelfHaluMap**, which computes the HaluMap within the same text, i.e. within the source document. This self-referential analysis serves as a calibration mechanism, effectively filtering out background noise and ensuring that highlighted discrepancies accurately reflect genuine inconsistencies.

Our approach achieves a five-percentage-point improvement over state-of-the-art NLI-based, training-free methods in detecting hallucinations across various tasks, including summarization, dialogue generation, and question answering. Moreover, HaluMap, as a training-free method, outperforms several supervised training methods while providing interpretable results through the entailment and contradiction scores for segments.

2 Related Work

Several works have used NLI as a tool to detect hallucinations (Goyal and Durrett, 2020; Laban et al., 2022; Chrysostomou et al., 2024). For instance, Laban et al. (2022) divide both the document and summary into sentences and using an NLI matrix, they find that using both entailment and contraction probabilities for each pair leads to boosts in performance. In addition, they introduce a benchmark for inconsistency detection in summarization and carry out the NLI score calculation on the sentence level. Similarly, Goyal and Durrett (2020) break down the hypotheses into dependency arcs and measure the entailment of each resulting relation with the source document.

Migrating from pre-trained encoder models to large language models, Yang et al. (2024) utilize ChatGPT variants as the scoring tool. Dhuliawala et al. (2023) propose the chain of verification technique which makes the LLM correct a hallucinated answer by asking simpler additional questions that are easier for the model to answer. Min et al. (2023) break down generated biographies by LLMs into atomic facts and propose a score for evaluating the factuality of an LLM generation based on the number of correct atomic facts that the language model has generated. Similarly, Mündler et al. (2023) use an LLM with the Chain-of-Thought (CoT) prompting (Wei et al., 2022) to detect self-contradictory sentences. In addition, prompting ChatGPT (OpenAI, 2022), Luo et al. (2023) show that it can detect hallucinations in summaries in a zero-shot setting and CoT prompting further improves its performance. Farquhar et al. (2024) cluster factoids in the LLM responses using semantic entropy and bidirectional entailment, identifying the clusters with lower semantic entropy as likely confabulations. While previous work has explored various ways to detect inconsistencies, an explainable component in these modes is lacking. Our approach consists of an explainable map that highlights the detected inconsistencies.

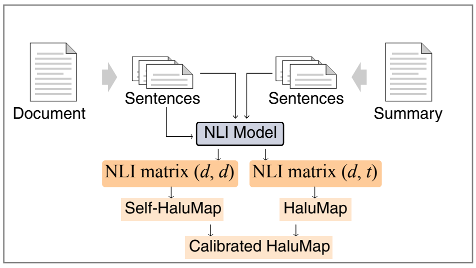

Figure 2: Overview of the proposed HaluMap and calibrated HaluMap.

Besides detecting hallucination, previous work aimed to reduce the hallucination in the generated output. Approaches include modifying the LLM distribution of output to improve the context consistency (Shi et al., 2024) or leveraging knowledge bases to identify hallucination and further guide the generation process (Choi et al., 2023). For example, Chen et al. (2024) use references to generate more faithful outputs in an autoencoder-based model, where they also map the NLI relations to hallucination.

3 Method

Our method consists of two stages: 1) building an NLI map of the two inputs, e.g. a document and its summary; and 2) calibrating the NLI map. Using the NLI map, we can then compute the overall hallucination strength as well as visualize the hallucinated parts of the text (see Figure 2).

3.1 Building NLI Matrix

Given a document $d = [d_1, d_2, ..., d_M]$ and an associated text $t = [t_1, t_2, ..., t_N]$ (e.g. a summary of d), let M and N be the numbers of textual segments in d and t, respectively. We apply an NLI model[1] predicting the relationship (i.e. entailment, neutral, or contradiction) between each pair of segments (d_m, t_n), where m and n are the index of d and t. A text segment can be a paragraph, sentence, or clause, depending on the task and the application.

[1] microsoft/deberta-large-mnli

The three scores, $\text{ent}_{m,n}$, $\text{neu}_{m,n}$, and $\text{con}_{m,n}$ sum to one. After computing these scores for all segment pairs, we construct a three-dimensional NLI matrix $\mathbb{N} \in \mathbb{R}^{M \times N \times 3}$, where the last dimension corresponds to the three scores.

3.2 Calibrating NLI Matrix

While the initial NLI matrix can be used directly to identify which segment pair of the document and the text are not aligning with each other (i.e. which segment within t is hallucinated), we found that using the matrix directly does not effectively indicate hallucinated segments. As illustrated in Figure 1(a), it is not visible which segments are likely to be hallucinated.

Therefore, we propose calibrating the NLI matrix using a *self-NLI matrix*. The intuition behind this is that, within one document, there are unlikely contradicted parts, and any two text segments in one document should have an NLI relation of entailment or neutral. With this in mind, we build a self-NLI matrix within one document. We construct a self-NLI matrix $\mathbb{N}_{\text{self}} \in \mathbb{R}^{M \times M \times 3}$ via computing any two text segments of . Figure 4 in Appendix A.2 shows the contradiction matrix of a document, where the red spots demonstrate the contradiction. Such a phenomenon could be the result of an unreliable NLI model, or some inherent feature of the text. To reduce the impact of such an issue, we first compute an average vector of $\mathbb{N}_{self}$ along one of its second axes. The averaged vector $\bar{\mathbb{N}}_{self} \in R^{M,3}$ presents an inherent contradiction value of each text segment. Then, we compute the calibrated NLI matrix $\tilde{\mathbb{N}}$ where each element of it for one NLI score s is: $\tilde{\mathbb{N}}_{m,n,s} = \mathbb{N}_{m,n,s} - \bar{\mathbb{N}}_{self,m,s}$, meaning that each element of $\tilde{\mathbb{N}}$ is subtracted by its background noise. After the subtraction, the calibrated NLI matrix keeps its shape as $[M, N, 3]$. As shown in Figure 1(b), the calibrated NLI matrix can now be used in the later process.

3.3 Overall Hallucination Strength

To estimate the overall hallucination, Laban et al. (2022) proposed to rely on the entailment values of the NLI matrix . In practice, they first find out the maximum entailment values among all sentences of d in t_n , and then apply a mean operator on the resulting vector. Following the notation used in Laban et al. (2022), the final entailment strength is $mean(max(\mathbb{N}_e, axis = 0), axis = 1)$, where $\mathbb{N}_e$ is the entailment NLI matrix. Conceptually, this equation can be seen as the average pooling of the maximum pooling of $\mathbb{N}_e$. Laban et al. (2022) only tested on the entailment values $\mathbb{N}_e$. We suggest the final contradiction strength to be $max(max(\mathbb{N}_c, axis = 0), axis = 1)$, where $\mathbb{N}_c$ is the contradiction NLI matrix.

The intuition for choosing the two operations for the entailment value is as follows. From t's perspective, as long as one part of d can entail t_n, t_n can be seen as entail (or consistent). Therefore, the first operation should be maximum to find out the most likely entailed part and its entailment value. Given that all segments are considered to be equally important, the second operation should be mean.

For the contradiction values, a segment t_n is hallucinated (contradicted with d), as long as any segments in d contradict t_n. Therefore, the first operation is taking the maximum. Second, we consider t to be hallucinated even if just part of it is hallucinated. Hence, the second operation should be maximum as well[2].

4 Experiments and Results

4.1 Baselines

We compare our method with state-of-the-art related approaches that utilize NLI. We briefly described their approach in Section 2: **MNLI-doc** uses RoBERTa (Liu et al., 2019) fine-tuned on MNLI (Williams et al., 2018). **DAE** (Goyal and Durrett, 2020) uses ELECTRA (Clark et al., 2020) to obtain NLI scores (Laban et al., 2022) while **SIFiD** (Yang et al., 2024) uses LLMs (ChatGPT 3.5/4 Turbo). Given the size of our utilized DeBERTa model (He et al., 2020), and for a fair comparison, we compare our results with their results from ChatGPT 3.5 Turbo. **SummaC** (Laban et al., 2022) is a benchmark for inconsistency detention, using BERT-base and BERT-large models (Devlin et al., 2019). In addition to these models, we compare our results with ChatGPT in zero-shot and zero-shot with CoT prompting.

4.2 Datasets

To fully test HaluMap, we consider SummaC (Laban et al., 2022), HaluEval (Li et al., 2023), and LLM-AggreFact datasets (Tang et al., 2024), the experiment settings are as follows.

SummaC consistent label means not hallucination and the inconsistent label means hallucination.

[2] Other combinations of operations can be found in A.1.

Model Name	CGS	XSF	Poly	FactCC	Frank	Avg.
MNLI-doc	57.6	57.5	61.0	61.3	63.6	61.3
DAE	63.4	50.8	62.8	75.9	61.7	64.2
SIFiD-Ent	65.5	63.9	37.5	81.0	81.6	68.1
SIFiD-Ent$_{CoT}$	65.7	60.3	52.7	82.3	81.6	70.3
SummaC$_{ZS}$	70.4	58.4	62.0	83.8	79.0	70.7
ChatGPT$_{ZS}$	63.3	64.7	56.9	74.7	80.9	69.5
ChatGPT$_{ZS-CoT}$	74.3	63.1	61.4	79.5	**82.6**	74.0
HaluMap$_e$	**74.9**	**66.8**	61.8	86.3	80.7	74.4
HaluMap$_e$ (Cal.)	76.5	66.3	**66.2**	**89.5**	80.6	**75.8**

Table 1: Balanced accuracies of the baselines, SummaC, and our approach (HaluMap) with the entailment values.

Benchmark	Dataset	Size	% Hallu.
SummaC	CGS	400	50.2
	XSF	1250	89.8
	Poly	634	93.4
	FactCC	503	15.0
	FRANK	1575	66.8
HaluEval	Summarization	20,000	50.0
	Dialogue w/ history	20,000	50.0
	QA w/ Knowledge	20,000	50.0
LLM-AggreFact	AggreFact-CNN	558	10.2
	AggreFact-XSum	558	48.9
	TofuEval-Media	726	23.7
	TofuEval-MeetB	772	19.4
	Wice	358	68.9
	Reveal	1710	76.7
	ClaimVerify	1088	27.5
	FactCheck	1566	75.0
	ExpertQA	3702	19.7
	LFQA	1911	41.3

Table 2: Number of instances and percentage of hallucination in SummaC, HaluEval, and the test set of LLM-AggreFact. Note that Summeval from SummaC was discard in the experiments.

Model Name	Sum.	Dialogue	QA	Avg.
ChatGPT (gpt-3.5-turbo)	58.5	**72.4**	62.6	64.5
Llama3-8b	56.9	62.8	49.8	56.5
SummaC$_{ZS}$	62.1	56.0	65.5	61.2
HaluMap$_e$	72.7	64.7	69.3	68.9
HaluMap$_e$ (Cal.)	**73.3**	65.2	**69.7**	**69.4**

Table 3: Accuracies of the baselines, Llama3-8b, SummaC by running their codes ourselves, and our approach (HaluMap) on the three HaluEval subsets (summarization, dialogue generation and QA).

In our experiments, we follow the setting in SummaC. The statistics for this dataset can be found in Table 2. The utilized datasets from SummaC are as follows: **CGS** contains articles from the CNN/DM dataset (Nallapati et al., 2016) and model-generated summaries (Falke et al., 2019). **XSF** (Maynez et al., 2020) is a collection of human-annotated abstractive summaries from XSum (Narayan et al., 2018). Both extrinsic and intrinsic hallucination are considered. **Poly** includes summaries from the CNN/DM dataset with eight different errors (Huang et al., 2020). We consider accuracy errors to be the inconsistent label. **FactCC** (Kryściński et al., 2020) also comprises summaries from the CNN/DM dataset. It comes with a test split where the consistent and inconsistent labels are annotated by the authors. **FRANK** (Pagnoni et al., 2021) contains summaries from both CNN/DM and XSum articles. Summaries are considered to be consistent if a majority of the annotations of the summary contain no error. We consider accuracy errors to be the inconsistent label.

HaluEval contains three specific tasks, namely summarization, question answering, and dialogue generation (Li et al., 2023). The authors used ChatGPT to label the hallucination answers. There are 10,000 instances per task to form 20,000 article-summary pairs pairs (10,000 with valid summaries and 10,000 containing hallucinations.

LLM-AggreFact is one of the largest collections of studying hallucination in LLMs (Tang et al., 2024). It contains summaries (AggreFact and TofuEval), responses to search queries (ClaimVerify, LFQA, ExpertQA, Reveal, and FactCheck-GPT), as well as Wikipedia claims (Wice). In our experiments, we treat their *negative* label as hallucination, and we use only the test set for a fair comparison.

4.3 Experiments on SummaC Benchmarks

Table 1 shows the balanced accuracies of HaluMap as well as baselines on the SummaC benchmarks. Overall, calibrated HaluMap (75.8%) has a five-percentage-point improvement against the SummaC-zero-shot approach (70.7%), and it even achieves better performance compared to ChatGPT with CoT prompting. Comparing HaluMap with calibrated HaluMap, they achieve very similar performance and the main difference is in the Polytope and FactCC datasets, where calibrated HaluMap has around three percentage points improvement on both datasets. Given that these two datasets are the most skewed ones (93.4% hallucination and 15% hallucination), it suggests that the calibrated HaluMap is more robust on the imbalanced distributed datasets.

Model Name	AggreFact		TofuEval		Wice	Reveal	Claim Verify	Fact Check	Expert QA	LFQA	Avg.
	CNN	XSum	MediaS	MeetB							
SummaC$_{ZS}$	51.1	61.5	69.5	71.0	62.8	85.3	69.7	75.2	55.2	77.6	67.9
MiniCheck-FT5	**69.9**	**74.3**	**73.6**	**77.3**	**72.2**	**86.2**	**74.6**	**74.7**	**59.0**	**85.2**	**74.7**
HaluMap$_e$	65.1	60.5	66.2	65.8	63.2	81.9	73.3	68.9	56.9	76.0	67.8
HaluMap$_e$ (Cal.)	67.7	65.2	65.8	67.4	67.1	85.4	73.2	71.4	56.3	76.1	69.6

Table 4: Balanced accuracies of the SummaC, MiniCheck-FT5, and our approach (HaluMap) with the entailment values on the test set of LLM-AggreFact. Except the HaluMap results, other results are directly from the original LLM-AggreFact paper.

4.4 Experiments on HaluEval

In Table 3, we compare HaluMap with SummaC using their codebase[3], ChatGPT `gpt-3.5-turbo` from Li et al. (2023), and Llama3-8b from Chekalina et al. (2024)—the most recent, best approach on HaluEval. Overall, HaluMap outperforms other baselines by at least 5 percentage points. However, ChatGPT outperforms HaluMap in dialogue generation. Given that ChatGPT was developed as an AI agent, it is not surprising that the developers may further focus on its abilities in dialogue generation. Comparing HaluMap with the calibrated HaluMap, they have similar performance on the three tasks, and the latter is slightly better in all three cases.

4.5 Experiments on LLM-AggreFact

In Table 4, we compare HaluMap with the best variation of MiniCheck (MiniCheck-FT5) on the LLM-AggreFact dataset. We observe a similar trend in other datasets: HaluMap performs close to SummaC$_{ZS}$ while the calibrated HaluMap outperforms its non-calibrated variation. Given that MiniCheck requires training, it achieves the best performance in all subsets, as expected. Still, we argue that the calibrated HaluMap is the best-performing among all the training-free models.

5 Qualitative Analysis

We conducted a qualitative analysis to gain a better insight into the calibrated HaluMap. In Figure 3, we show a vanilla HaluMap of contradiction, an average vector of self-HaluMap $\bar{\mathbb{N}}_{cself}$, and the calibrated HaluMap where the summary is labeled as non-hallucination. We found that the sentence *'Psychology can explain why somebody would turn rage inward on themselves about the fact'* contradicts many other sentences in the document (Appendix A.3). These high contradiction values can be seen as the by-product of NLI models. Therefore, calibrating these values can further improve the pooling process when computing the overall hallucination strength. It can also benefit in spotting the possible hallucination sentences in the document. In this example, using vanilla HaluMap would wrongly get a high contradiction due to some high values in the map. On the other hand, calibrated HaluMap has a lower overall contradiction value after pooling.

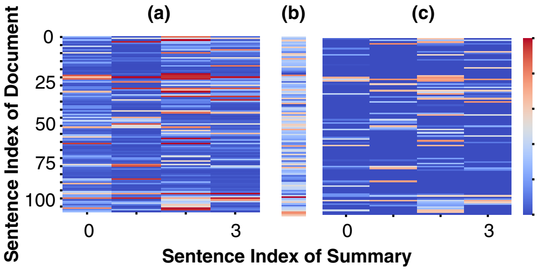

Figure 3: (a) The vanilla contradiction HaluMap of a non-hallucination document-summary pair. (b) The average self-HaluMap of contradiction shows that a few sentences (especially the second-to-last sentence) have a very high average contradiction against other sentences. (c) The calibrated HaluMap shows there is almost no contradiction in any place. More red indicates hallucinations.

6 Conclusion

We introduced HaluMap, a post-hoc framework that enhances explainability in hallucination detection. HaluMap is a metric based on an informed calibration of NLI outcomes and it outperforms existing NLI-based methods with improved hallucination detection performance as well as a clearer visual interpretation. By providing interpretable explanations without requiring model access, HaluMap serves as a practical tool for verifying model outputs across various NLP tasks.

[3] https://github.com/tingofurro/summac/

7 Limitations

This paper only studies one certain type of hallucination: faithfulness. There exists also factuality (Augenstein et al., 2024), and intrinsic or extrinsic hallucinations (Maynez et al., 2020). We expect the proposed method to also apply to these definitions of hallucinations as long as we can define the entailment and/or the contradiction relation in these definitions. The scope of this paper covers a limited number of tasks, namely summarization, question answering, and dialogue generation. Hallucination also exists in other text generation tasks, for which this study can be extended if suitable datasets become available.

8 Ethical Concerns

The hallucination in LLMs itself draws the attention of ethical concerns in using LLMs. However, this paper focuses on how to better combat and explain hallucinations with interpretable methods. With the proposed method, we hope to better reveal the hallucinations in specific text segments, contribute to a better understanding of hallucinations, and therefore, help to mitigate the ethical concerns of using LLMs.

Acknowledgments

This work was conducted during Zhixue Zhao's research visit to the Lamarr Institute for Machine Learning and Artificial Intelligence. We would like to especially thank our colleagues Ipek Baris Schlicht, Shaina Ashraf, and Tianyi Zhang, for their discussions in shaping the idea of this paper. This work was partially supported by the AISafety Project, funded by the BMBF under the grant proposal 05D23PD1, and by the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Viktoriia Chekalina, Anton Razzigaev, Elizaveta Goncharova, and Andrey Kuznetsov. 2024. Addressing hallucinations in language models with knowledge graph embeddings as an additional modality. *arXiv preprint arXiv:2411.11531*.

Wei-Fan Chen, Milad Alshomary, Maja Stahl, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2024. Reference-guided style-consistent content transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13754–13768.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14053.

George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12:1163–1181.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

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

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2214–2220.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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 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.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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

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*, page 1906. 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.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

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*. Association for Computational Linguistics.

OpenAI. 2022. ChatGPT. https://openai.com/blog/chatgpt.

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.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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.

Jiuding Yang, Hui Liu, Weidong Guo, Zhuwei Rao, Yu Xu, and Di Niu. 2024. Reassess summary factual inconsistency detection with large language model. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 27–31.

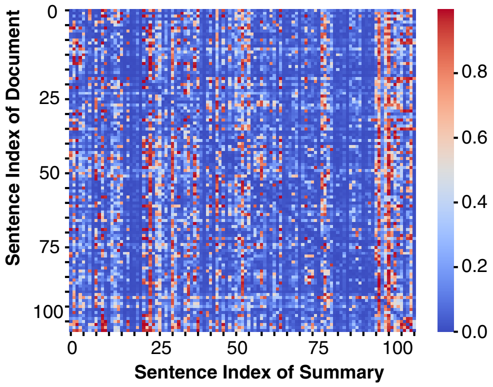

Figure 4: Self-HaluMap of contradiction showing many contradictions (red) in one document.

A Appendix

A.1 Different Pooling Operations

The performance of using different combinations of operations can be in Table 5 and Table 6.

	1st Operation		
2nd Operation	**Max**	**Mean**	**Min**
Max	57.6	**75.8**	62.9
Mean	58.9	61.0	60.0
Min	61.0	61.3	58.6

Table 5: The performance of different pooling operations by using entailment value in SummaC benchmarks.

	1st Operation		
2nd Operation	**Max**	**Mean**	**Min**
Max	**74.9**	63.8	50.4
Mean	74.1	65.3	50.3
Min	73.6	66.8	50.3

Table 6: The performance of different pooling operations by using contradiction value in SummaC benchmarks.

A.2 Self-HaluMap

The contradiction values of a self-HaluMap can be found in Figure 4.

A.3 Full Text in the Qualitative Analysis

Table 7 shows the full text in Sec5, including the document, right summary and the hallucinated summary.

Document:
Marseille, France (CNN)The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that "so far no videos were used in the crash investigation." He added, "A person who has such a video needs to immediately give it to the investigators." Robin's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps. All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. The two publications described the supposed video, but did not post it on their websites. The publications said that they watched the video, which was found by a source close to the investigation. "One can hear cries of 'My God' in several languages," Paris Match reported. "Metallic banging can also be heard more than three times, perhaps of the pilot trying to open the cockpit door with a heavy object. Towards the end, after a heavy shake, stronger than the others, the screaming intensifies. Then nothing." "It is a very disturbing scene," said Julian Reichelt, editor-in-chief of Bild online. An official with France's accident investigation agency, the BEA, said the agency is not aware of any such video. Lt. Col. Jean-Marc Menichini, a French Gendarmerie spokesman in charge of communications on rescue efforts around the Germanwings crash site, told CNN that the reports were "completely wrong" and "unwarranted." Cell phones have been collected at the site, he said, but that they "hadn't been exploited yet." Menichini said he believed the cell phones would need to be sent to the Criminal Research Institute in Rosny sous-Bois, near Paris, in order to be analyzed by specialized technicians working hand-in-hand with investigators. But none of the cell phones found so far have been sent to the institute, Menichini said. Asked whether staff involved in the search could have leaked a memory card to the media, Menichini answered with a categorical "no." Reichelt told "Erin Burnett: Outfront" that he had watched the video and stood by the report, saying Bild and Paris Match are "very confident" that the clip is real. He noted that investigators only revealed they'd recovered cell phones from the crash site after Bild and Paris Match published their reports. "That is something we did not know before. ... Overall we can say many things of the investigation weren't revealed by the investigation at the beginning," he said. What was mental state of Germanwings co-pilot? German airline Lufthansa confirmed Tuesday that co-pilot Andreas Lubitz had battled depression years before he took the controls of Germanwings Flight 9525, which he's accused of deliberately crashing last week in the French Alps. Lubitz told his Lufthansa flight training school in 2009 that he had a "previous episode of severe depression," the airline said Tuesday. Email correspondence between Lubitz and the school discovered in an internal investigation, Lufthansa said, included medical documents he submitted in connection with resuming his flight training. The announcement indicates that Lufthansa, the parent company of Germanwings, knew of Lubitz's battle with depression, allowed him to continue training and ultimately put him in the cockpit. Lufthansa, whose CEO Carsten Spohr previously said Lubitz was 100% fit to fly, described its statement Tuesday as a "swift and seamless clarification" and said it was sharing the information and documents – including training and medical records – with public prosecutors. Spohr traveled to the crash site Wednesday, where recovery teams have been working for the past week to recover human remains and plane debris scattered across a steep mountainside. He saw the crisis center set up in Seyne-les-Alpes, laid a wreath in the village of Le Vernet, closer to the crash site, where grieving families have left flowers at a simple stone memorial. Menichini told CNN late Tuesday that no visible human remains were left at the site but recovery teams would keep searching. French President Francois Hollande, speaking Tuesday, said that it should be possible to identify all the victims using DNA analysis by the end of the week, sooner than authorities had previously suggested. In the meantime, the recovery of the victims' personal belongings will start Wednesday, Menichini said. Among those personal belongings could be more cell phones belonging to the 144 passengers and six crew on board. Check out the latest from our correspondents. The details about Lubitz's correspondence with the flight school during his training were among several developments as investigators continued to delve into what caused the crash and Lubitz's possible motive for downing the jet. A Lufthansa spokesperson told CNN on Tuesday that Lubitz had a valid medical certificate, had passed all his examinations and "held all the licenses required." Earlier, a spokesman for the prosecutor's office in Dusseldorf, Christoph Kumpa, said medical records reveal Lubitz suffered from suicidal tendencies at some point before his aviation career and underwent psychotherapy before he got his pilot's license. Kumpa emphasized there's no evidence suggesting Lubitz was suicidal or acting aggressively before the crash. Investigators are looking into whether Lubitz feared his medical condition would cause him to lose his pilot's license, a European government official briefed on the investigation told CNN on Tuesday. While flying was "a big part of his life," the source said, it's only one theory being considered. Another source, a law enforcement official briefed on the investigation, also told CNN that authorities believe the primary motive for Lubitz to bring down the plane was that he feared he would not be allowed to fly because of his medical problems. Lubitz's girlfriend told investigators he had seen an eye doctor and a neuropsychologist, both of whom deemed him unfit to work recently and concluded he had psychological issues, the European government official said. But no matter what details emerge about his previous mental health struggles, there's more to the story, said Brian Russell, a forensic psychologist. "Psychology can explain why somebody would turn rage inward on themselves about the fact that maybe they weren't going to keep doing their job and they're upset about that and so they're suicidal," he said. "But there is no mental illness that explains why somebody then feels entitled to also take that rage and turn it outward on 149 other people who had nothing to do with the person's problems." Germanwings crash compensation: What we know. Who was the captain of Germanwings Flight 9525? CNN's Margot Haddad reported from Marseille and Pamela Brown from Dusseldorf, while Laura Smith-Spark wrote from London. CNN's Frederik Pleitgen, Pamela Boykoff, Antonia Mortensen, Sandrine Amiel and Anna-Maja Rappard contributed to this report.

Right Summary:
Marseille prosecutor says "so far no videos were used in the crash investigation" despite media reports. Journalists at Bild and Paris Match are "very confident" the video clip is real, an editor says. Andreas Lubitz had informed his Lufthansa training school of an episode of severe depression, airline says.

Hallucination Summary:
A video showing the final moments of Germanwings Flight 9525 has been recovered by investigators from the wreckage site. Marseille prosecutor Brice Robin urged anyone who might have more footage to turn it over immediately. Andreas Lubitz, the co-pilot accused of deliberately crashing the plane, had a history of severe depression and suicidal tendencies.

Table 7: The full document used in Sec 5.