Do Hallucination Detectors Understand Factuality? A Robustness Study with Paraphrased Summaries

Scott Stossel, Jing Tan
University of California, Berkeley, School of Information
scott_stossel@berkeley.edu, jt3124@berkeley.edu

Abstract

Hallucinations in Large Language Models (LLMs) pose significant risks when these systems are used in domains where factual accuracy is essential. Although hallucination-detection models often report strong performance, it remains unclear whether they genuinely assess factual consistency or simply exploit superficial linguistic cues present in standard benchmarks. To investigate this question, we fine-tune five widely used architectures—BERT, RoBERTa, FLAN-T5, ELECTRA, and DeBERTa—on a dataset of article-summary pairs labeled for factual consistency. Across models, we found that strong in-distribution performance often fails to generalize, with substantial degradation under paraphrasing. Notably, DeBERTa shows comparatively stronger robustness, suggesting that certain architectures capture factual consistency more reliably than others. Our findings show that many hallucination detectors rely on shallow linguistic cues rather than true semantic grounding, highlighting the need for robustness-focused evaluation frameworks.

1 Introduction

Large Language Models (LLMs) have become central to modern information systems, supporting tasks such as summarization, question answering, and decision-support across many industries. Despite their growing capabilities, a persistent limitation undermines their reliability: LLMs frequently generate hallucinations—statements that are factually incorrect, unsupported by the source text, or logically inconsistent. As a result, detecting hallucinations has emerged as a critical area of research, leading to a variety of detection models built on architectures such as BERT, RoBERTa, FLAN-T5, ELECTRA, and DeBERTa. Although these detectors often report high accuracy, a fundamental question remains: Do they genuinely understand factual consistency, or do they perform well only because existing evaluations reward superficial pattern matching?

Many hallucination detectors are trained and tested on curated datasets where hallucinated and non-hallucinated summaries follow predictable linguistic patterns. Surveys of hallucination in natural language generation and LLMs consistently highlight this behavior as a core obstacle to trustworthy deployment in real-world settings. (Ji et al., 2022) Under these conditions, models may succeed not by reasoning about truthfulness but by relying on shallow cues such as lexical signatures, stylistic differences, or dataset-specific heuristics. If so, their performance may underperform in real-world settings where summaries are reworded or paraphrased. Understanding whether detectors generalize beyond such surface patterns is especially important as they are increasingly used in safety-critical workflows, including clinical summarization, legal analysis, financial research, and public policy decision-making. (Guo et al., 2024)

To probe this issue, we conduct a comparative evaluation of five widely used architectures—BERT, RoBERTa, FLAN-T5, ELECTRA, and DeBERTa—using a two-stage robustness framework. Many of these detectors are implemented using encoder-based architectures (e.g., BERT, RoBERTa, DeBERTa) or encoder–decoder

models (e.g., T5 variants), and are frequently reported to achieve high accuracy on standard test sets. (Liu et al., 2025) First, each model is fine-tuned on hallucination-labeled summaries in their original form. Second, we evaluate all models on a paraphrased version of the same dataset, where summaries are rewritten while preserving their factual validity. This setup allows us to isolate whether detectors rely on genuine semantic understanding or on form-based regularities. A model that truly recognizes factual consistency should maintain stable performance across paraphrastic variation; substantial degradation would indicate dependence on superficial cues.

Through this approach, our study assesses the extent to which current hallucination detectors understand factual grounding versus stylistic structure. Our findings contribute to the broader question of whether these systems can be trusted to generalize beyond the benchmarks they are trained on, or whether their apparent success reflects hidden shortcut learning.

2 Background

2.1 Overview of Models

BERT is a bidirectional transformer encoder trained with masked-language modeling and next-sentence prediction (Devlin et al., 2019). It has been widely used in hallucination-detection and NLI-based factuality tasks, though prior work shows it often relies on dataset-specific cues rather than true factual reasoning.

RoBERTa improves on BERT through larger training data, dynamic masking, and removal of next-sentence prediction (Liu et al., 2019). Its stronger representations make it a common baseline for factual-consistency classifiers, but studies find it still fails under paraphrasing, suggesting reliance on surface-level patterns.

DeBERTa improves transformer attention by disentangling content and positional embeddings, yielding stronger semantic representations (He et al., 2021). Its architecture suggests potential advantages for factuality detection, though its robustness under paraphrastic variation has been less explored.

FLAN-T5 is an instruction-tuned encoder–decoder model that generalizes well across tasks and input styles (Chung et al., 2022). While its generative architecture supports summarization and consistency evaluation, it is unclear whether its performance reflects genuine semantic understanding or sensitivity to prompt phrasing.

ELECTRA uses a discriminative pretraining objective—replaced-token detection—which yields strong performance with less training data (Clark et al., 2020). Although effective for classification tasks, its token-level training signal may limit its ability to capture global factual consistency when summaries are paraphrased.

We selected these models due to their diverse pretraining objectives. This selection compares various masked-language modeling approaches (BERT/RoBERTa/DeBERTa), replaced-token detection (ELECTRA), and instruction-tuned encoder-decoder modeling (FLAN-T5). This provides a clear comparison into whether robustness varies by architecture.

2.2 Present Research

Recent research has documented that hallucination-detection systems often perform well only within the narrow

distributions of their training benchmarks. Surveys by (Ji et al. 2023) and (Anh-Hoang et al. 2025) emphasize that many detectors—especially those built on BERT- or RoBERTa-style encoders—tend to overfit surface patterns rather than capture true factual consistency. Work such as FactCC (Kryściński et al., 2020) and SummaC (Laban et al., 2022) shows that NLI-based and classifier-based detectors can achieve strong performance on standard datasets, yet follow-up studies consistently find notable drops when summaries are rephrased or generated in more abstractive styles. More recent perturbation-based approaches (e.g., Wang et al., 2024; Zhang et al., 2024) highlight this brittleness by demonstrating that even small controlled rewrites can cause detectors to misclassify factually equivalent content. Building on these insights, our study directly examines this issue by comparing multiple model families (BERT, RoBERTa, DeBERTa, ELECTRA, FLAN-T5) under an original-versus-paraphrased evaluation, providing a clearer test of whether current detectors truly capture semantic factuality or rely on dataset-specific cues.

3 Methods

3.1 Dataset The dataset used in this study was retrieved from the **HaluEval** benchmark (pminervini/HaluEval), which provides paired source documents, system-generated summaries, and human-annotated labels indicating the presence of hallucinated content, making it an appropriate choice for supervised factuality evaluation. After loading the dataset, we reviewed its structure, field completeness, and annotation validity to ensure the document–summary pairs were suitable for modeling. Hallucination descriptors (e.g., "yes," "no") were standardized into a binary label to support classification tasks, and basic text-quality diagnostics—including character-level length distributions for both documents and summaries—were performed to assess variability in textual complexity. These descriptive analyses (Appendix A1) confirmed that the dataset did not exhibit systematic length-based artifacts or structural irregularities, providing a clean and reliable foundation for training transformer-based hallucination detection models.

3.2 Baseline Model

In our study, BERT-base serves as the baseline model because it represents one of the earliest and most widely adopted transformer encoder architectures for semantic classification tasks. Using BERT as a baseline allows us to anchor our comparisons to a well-understood, historically influential model whose strengths and limitations are extensively documented. By evaluating more recent architectures against BERT, we can clearly observe whether newer models provide meaningful gains in robustness—particularly when summaries are paraphrased and superficial linguistic cues are removed.

3.3 Experimental Design

Hallucination detectors are often evaluated on datasets where hallucinated and non-hallucinated summaries follow highly predictable linguistic patterns. As a result, models may achieve strong performances without truly understanding factual consistency. They may instead memorize correlations tied to the dataset's writing style or structural cues. Our proposed approach addresses this weakness by examining how well different model families generalize when the wording of summaries changes but the underlying truth value does not. By placing models under controlled

paraphrase-based distribution shifts, we directly test whether they rely on meaningful semantic reasoning or brittle surface-level heuristics. Specific examples are located in Appendix A2.

A model that depends on superficial cues may misclassify the paraphrased version, revealing that it never learned factual grounding in the first place. This example highlights why robustness testing is critical: real-world summaries are rarely phrased identically to training examples.

Our method tackles this by using a two-stage evaluation pipeline. First, each model is fine-tuned to predict using AdamW with learning rate 5e-5, linear decay, no warmup, batch size 16, two epochs, mixed precision off, and default optimizer/scheduler settings from Hugging Face. Models are evaluated with both the standard and paraphrase summaries, and changes in accuracy, precision, recall, and calibration reveal whether they are recognizing deep semantic meaning or simply exploiting dataset regularities.

3.4 Evaluation Methodology

These experiments directly address our central question—*Do hallucination detectors truly understand factual consistency?*—by isolating linguistic form while without altering semantic content. If a model's performance remains stable across paraphrased and original data, this suggests genuine semantic understanding of factual grounding. If performance collapses under paraphrasing, the model likely depends on shallow heuristics rather than understanding factual relationships. This distinction is critical for real-world settings such as medical summarization, legal analysis, and financial reporting, where confidence in a model's predictions must remain stable even when the surface form of input text varies.

We evaluate both the in-distribution (ID) test set and the paraphrased out-of-distribution (OOD) test set using accuracy, F1-score to measure classification performance, and Expected Calibration Error (ECE) to measure probability reliability. ECE matters in hallucination detection because the predicted probability of "hallucinated" is often interpreted as a confidence score. A well-calibrated model should assign high probability only when hallucinations are truly likely

ECE measures how well a model's predicted probabilities align with empirical frequencies. We compute the positive class ECE by binning predicted probabilities of hallucination into M equal-width bins and then comparing average predicted probabilities with the empirical fraction of hallucinations within each bin:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| acc(B_m) - conf(B_m) \right|.$$

4 Results and Discussion

Across all models, we observe strong performance on the original hallucination-labeled summaries, but noticeable differences emerge once the summaries are paraphrased. Compared to the BERT-baseline, DeBERTa-base stands out as the most robust model, maintaining high accuracy (97.8% → 92.9%) and strong F1 scores (0.9777 → 0.9294). Its low Expected Calibration Error (ECE) further indicates stable and reliable confidence estimates under distribution shift. By contrast, RoBERTa-base—despite nearly perfect accuracy on the original dataset (99.8%)—fails dramatically on paraphrased summaries, dropping to 47% accuracy and showing a significant rise in ECE from 0.0019 to 0.5329. This collapse highlights its strong dependence on surface-level

	Original Summary					Paraphrased Summary				
	ECE	Accuracy	F-1 Score	Precision	Recall	ECE	Accuracy	F-1 Score	Precision	Recall
BERT	0.054	0.9288	0.932	0.892	0.9756	0.180	0.799	0.984	0.718	0.830
RoBERTa	**0.002**	**0.998**	**0.9980**	**1.000**	**0.996**	0.533	0.470	0.6300	0.480	0.900
DeBERTa	0.019	0.9776	0.978	0.976	0.980	**0.061**	**0.929**	**0.92940**	**0.922**	0.937
Electra	0.084	0.9053	0.9103	0.8645	0.961	0.241	0.744	0.791	0.669	**0.967**
T5	0.0396	0.549	0.433	0.583	0.344	0.091	0.4959	0.4156	0.494	0.359

Table 1: Detailed model specific results for performance on the original summary dataset and the paraphrased summary dataset.

patterns in the original text.

The baseline BERT model and ELECTRA exhibit moderate declines when exposed to paraphrasing. BERT's accuracy drops from 92.9% to 79.9%, with precision falling from 0.8922 to 0.7181, indicating that paraphrasing causes the model to incorrectly flag more factually correct summaries as hallucinations. ELECTRA declines similarly, from 90.5% to 74.4% accuracy, and shows weakened precision (0.8645 → 0.669). These results suggest that while both models capture some semantic signals, they still rely partially on lexical cues and structural patterns present in the unaltered summaries. FLAN-T5-base performs the weakest overall, with accuracy remaining near chance level on both original (54.9%) and paraphrased (49.6%) datasets, implying that encoder–decoder architectures may require more specialized methods for hallucination detection beyond standard classification.

To further interpret model behavior, additional metrics such as precision, recall, and ECE provide insight into different types of failure. Precision drops sharply in RoBERTa and ELECTRA under paraphrasing—indicating an overprediction of hallucinations when surface markers are removed—while DeBERTa's precision and recall remain stable. ECE highlights how well-calibrated a model's confidence is: DeBERTa's minimal increase suggests consistent reliability, whereas RoBERTa's more than 25× increase demonstrates severe overconfidence on incorrect predictions. These trends reinforce that paraphrasing acts as an effective stress test, revealing hidden weaknesses in models that otherwise appear strong under in-distribution evaluation.

The error patterns observed can be traced back to architectural differences. RoBERTa and BERT, which rely heavily on masked-language modeling and contextual token patterns, appear to latch onto superficial lexical cues that do not generalize to paraphrased text. ELECTRA's discriminator-style training encourages sensitivity to token-level anomalies, leaving it less equipped for global factual reasoning when sentence structure changes. FLAN-T5, optimized for generative tasks rather than binary classification, struggles to form stable decision boundaries in this setting. In contrast, DeBERTa's disentangled attention and enhanced positional encoding enable it to track meaning across different phrasings, demonstrating better semantic generalization. Together, these results support the conclusion that high accuracy on standard datasets does not imply genuine factual-consistency understanding, underscoring the value of paraphrase-based robustness evaluations in hallucination detection research.

5 Conclusion

This study examined whether common hallucination-detection models truly understand factual consistency or simply rely on superficial linguistic patterns. By

evaluating five architectures on both original and paraphrased summaries, we tested their ability to generalize beyond the phrasing seen during training. Our results show that while models like BERT, RoBERTa, and ELECTRA perform well in-distribution, their accuracy and calibration degrade sharply under paraphrasing, indicating reliance on shallow cues rather than semantic understanding. In contrast, DeBERTa demonstrates strong robustness across all metrics, suggesting it learns deeper factual relationships and is less sensitive to surface-level variation. These findings confirm that high benchmark accuracy does not necessarily reflect reliable hallucination detection. Future work could explore adversarial paraphrasing, long-context evaluation settings, hybrid generative–discriminative detection approaches, or training methods explicitly designed to improve robustness. Together, these directions can help develop hallucination detectors that better reflect true semantic understanding and real-world reliability.

References

Li, J., Chen, J., Ren, R., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2024). *The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models.* In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Long Papers)* (pp. 10879–10899). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.586 ACL Anthology+1

Liu, S., Halder, K., Qi, Z., Xiao, W., Pappas, N., Htut, P. M., John, N. A., Benajiba, Y., & Roth, D. (2025). *Towards Long Context Hallucination Detection.* In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 7827–7835). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.436 ACL Anthology+1

Anh-Hoang, D., et al. (2025). *Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. Frontiers in Artificial Intelligence.* https://pmc.ncbi.nlm.nih.gov/articles/PMC12518350/

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., … Wei, J. (2022). *Scaling instruction-finetuned language models.* https://arxiv.org/abs/2210.11416

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training text encoders as discriminators rather than generators.* Proceedings of ICLR. https://arxiv.org/abs/2003.10555

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* Proceedings of NAACL-HLT. https://arxiv.org/abs/1810.04805

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with disentangled attention.* Proceedings of ICLR. https://arxiv.org/abs/2006.03654

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). *Survey of hallucination in natural language generation. ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2020). *Evaluating the factual consistency of abstractive text summarization.* Proceedings of EMNLP 2020. https://arxiv.org/abs/1910.12840

Laban, P., Schnabel, T., Bennett, P. N., & Hearst, M. (2022). *SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Transactions of the ACL, 10*, 163–177. https://arxiv.org/abs/2111.09525

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach.* https://arxiv.org/abs/1907.11692

Wang, T., Kulkarni, N., & Qi, Y. (2024). *Less is more for improving automatic evaluation of factual consistency.* https://arxiv.org/abs/2404.06579

Zhang, D., et al. (2024). *Enhancing hallucination detection through perturbation-based data generation. Findings of ACL 2024.* https://arxiv.org/abs/2402.09234

Appendix

A1: Dataset Characteristics

In the summary sample HaluEval dataset that we used, there were 10,000 total samples. The distribution of hallucinated vs not hallucinated samples are as follows:

Hallucination Distribution in Dataset	
Hallucination	**Number of Samples**
Yes	5010
No	4990

The descriptive statistics for the dataset are as follows:

Summary Statistics of Dataset		
	Document Length	**Summary Length**
Min	294	51
Max	11585	2880
Mean	3878.88	367.22
Standard Deviation	2011.92	157.20

We then assigned an 80/10/10 training/validation/test split prior to conducting our experiments.

A2: Examples of Summaries and Paraphrased Summaries

Table A2.1 contains various samples of the summaries and their paraphrased versions contained in the test set. The T5-paraphrase-paws model performed the paraphrasing portions.

Summary	Paraphrased Summary
Dove's latest ad campaign shows actresses bullying women in a café while reading their personal notebooks. Women react to the harsh comments made by the actresses. The campaign aims to reduce bullying by showing the negative effects of harsh words.	Dove's latest ad campaign shows actresses bullying women in a café while reading their personal notebooks . The campaign aims to reduce bullying by showing the negative effects of harsh words .
His Blue Origin company completed a successful test in West Texas. The New Shepard vehicle rose to a height of 58 miles before landing. It was unmanned but will ultimately take six people into space. The launch was conducted in secrecy before being released to the public.	His Blue Origin company completed a successful test in West Texas . The New Shepard vehicle rose to a height of 58 miles before landing but will ultimately take six people into space . The launch was conducted in secrecy before being released to the public .
Nicholas Soukeras, 37, of Queens, New York wants his future son to be called Spyridon, after his father. His wife, Kseniya, 33, doesn't like the 'archaic' name and prefers to call their child Michael. Nicholas needs his	Nicholas Soukeras, 37, of Queens, New York wants his future son to be called Spyridon, after his father , his wife Kseniya, 33, doesn't like the 'archaic' name and prefers to call their child Michael Nicholas needs his

online petition to earn 100,000 signatures for his wife to relent. Ironically, the couple don't actually know for certain that they are having a boy.	online petition to earn 100,000 signatures for his wife to relent , the couple don't actually know for sure that they are having a boy .
Manchester City have been linked with summer move for Raheem Sterling. Sterling has two years left on his contract and is stalling on a new deal. Brendan Rodgers says a move to City would not be step up for Sterling. Indicating it will take the Manchester club 20 years to eclipse Liverpool.	Manchester City have been linked with a summer move for Raheem Sterling , who has two years left on his contract and is stalling on a new deal . Brendan Rodgers says a move to City would not be for Sterling , indicating it will take the Manchester club 20 years to eclipse Liverpool .

A2: Paraphrase Verification

In order to verify that the paraphrasing model did not accidentally change the semantic meaning of the text itself, we performed two verification steps. First, using the roberta-large-mnli model, we scored the entailment and contradiction of the summary/paraphrased-summary pairs. We then flagged the ones that contained a contradiction score of 0.3 or higher in either direction (summary → paraphrased or paraphrased → summary) as suspicious. From here, we employed an LLM-as-a-judge to label the pairs as the following: "no_drift", "added_facts", "removed_facts", "contradiction". Table A2.1 shows how the 20 suspicious pairs were categorized.

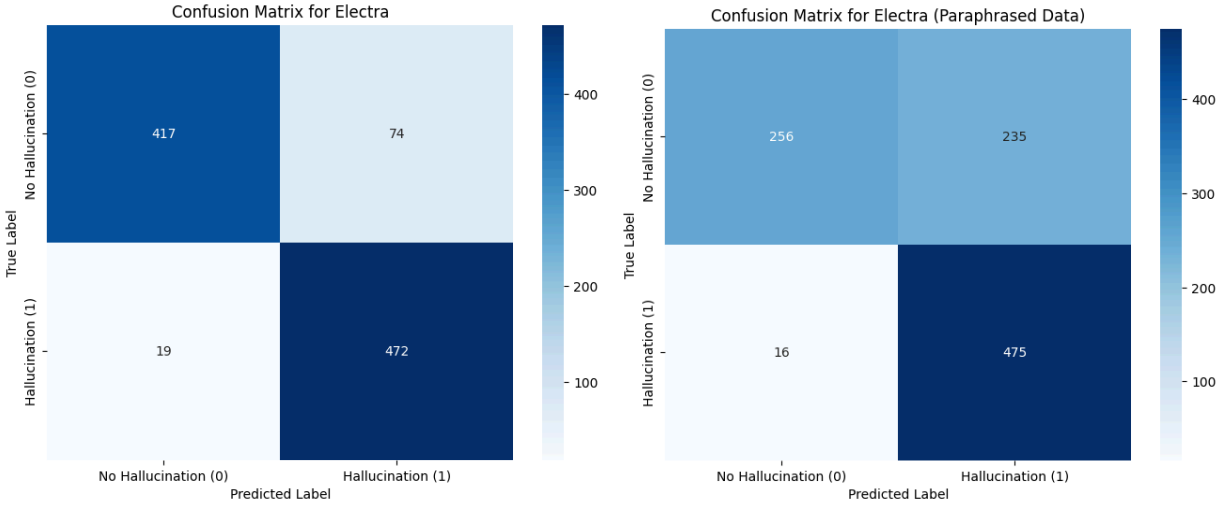
LLM Judge Category	**Number of Instances**
removed_facts	16
no_drift	3
contradiction	1
added_facts	0

We ultimately removed the rows that contained a contradiction between summaries or removed facts. While paraphrasing may occasionally shorten a passage, this helps remove the risk that the potential hallucination in the original summary was removed during the paraphrasing portion and best ensures that the paraphrasing process was faithful to the original.

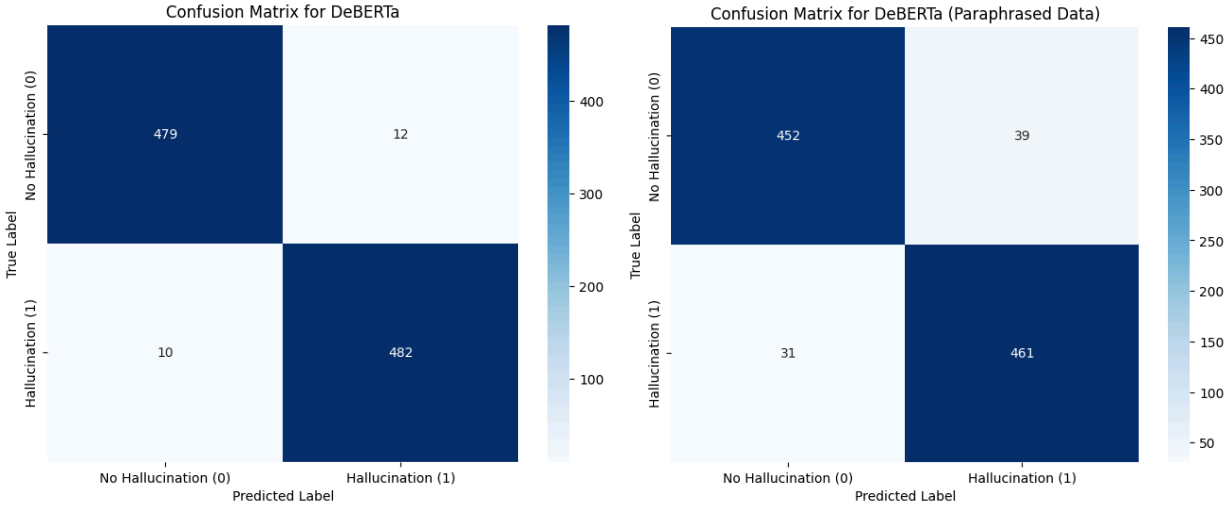
A3: Confusion Matrices

Images of the confusion matrices for all models for both the original and paraphrased summaries.

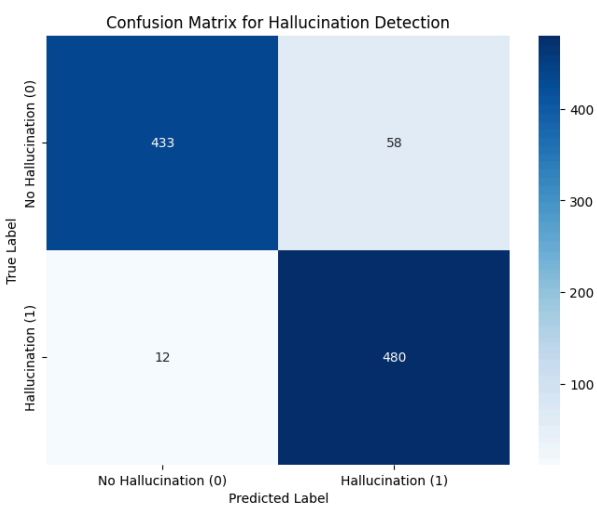
ELECTRA:

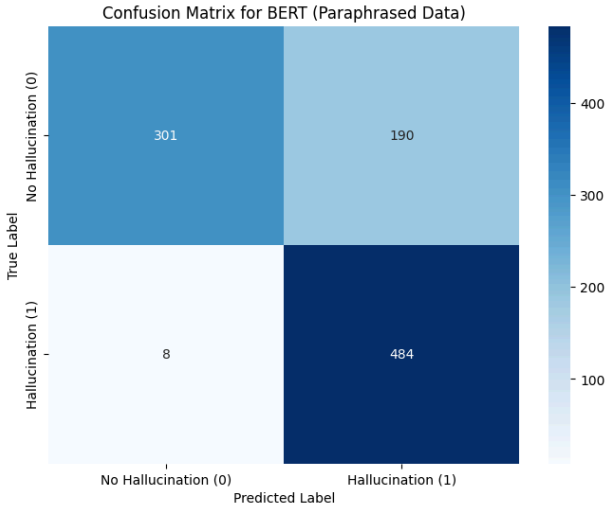


DeBERTa:

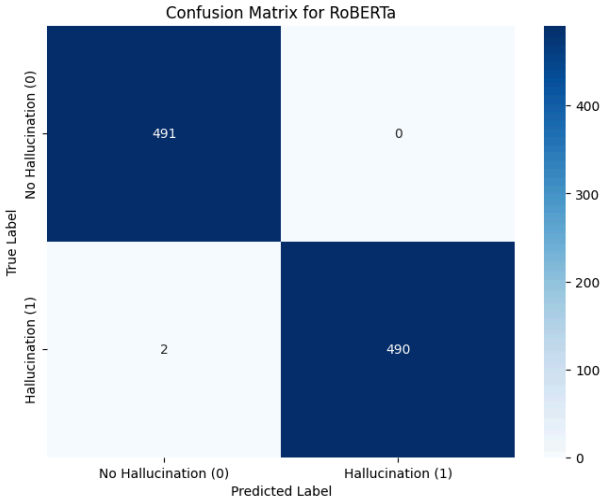


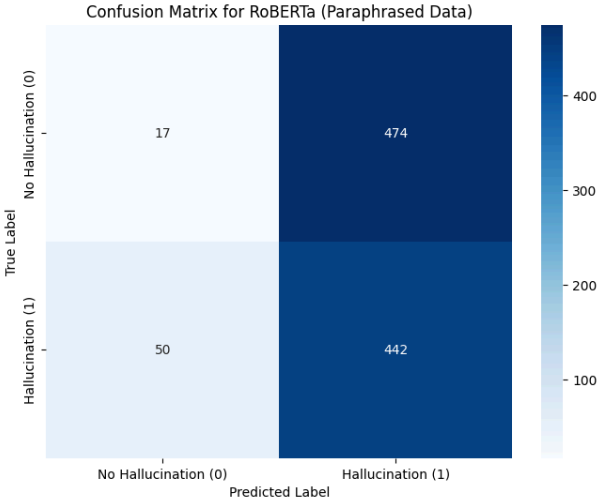
BERT:



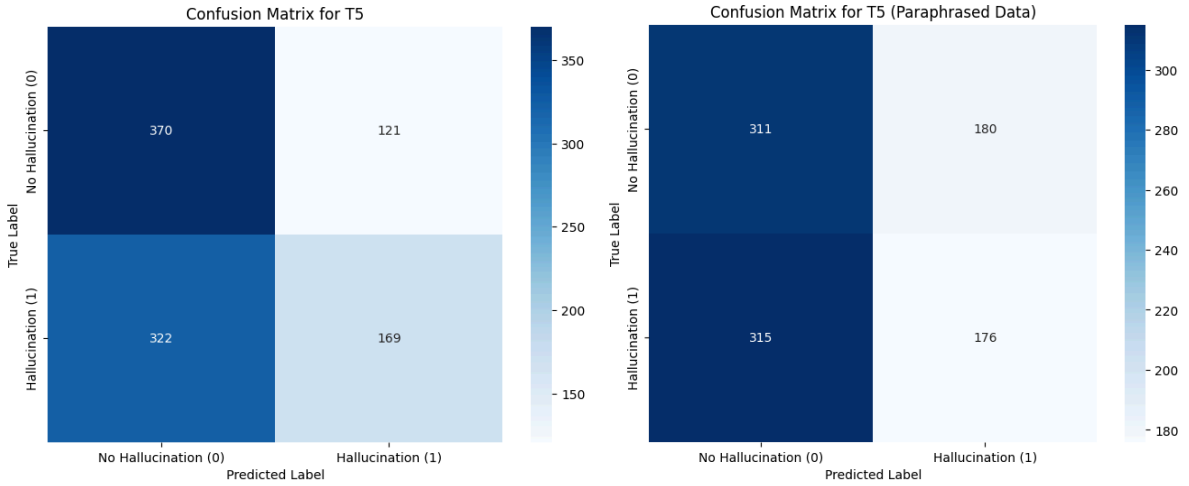


RoBERTa:





T5-FLAN:



Contributions

Jing and Scott were involved in the conceptualization of the project. Jing and Scott developed the experimental design. Scott did the initial EDA and data engineering. Scott built the baseline model and fine-tuned the BERT model, RoBERTa model and DeBERTa model. Jing fine-tuned the ELECTRA model, and the T5-FLAN model. Jing and Scott performed the analysis and Jing prepared the draft manuscript. Jing and Scott reviewed and edited the manuscript. All authors read and approved the final manuscript.