

RadFlag: A Black-Box Hallucination Detection Method for Medical Vision Language Models

Serena Zhang*

*Department of Computer Science, Stanford University
Department of Biomedical Informatics, Harvard Medical School*

SERENA2Z@STANFORD.EDU

Sraavya Sambara*

Department of Biomedical Informatics, Harvard Medical School

SRAAVYA_SAMBARA@COLLEGE.HARVARD.EDU

Oishi Banerjee

Department of Biomedical Informatics, Harvard Medical School

OISHI.BANERJEE@G.HARVARD.EDU

Julian N. Acosta

Department of Biomedical Informatics, Harvard Medical School

JULIAN_ACOSTA@HMS.HARVARD.EDU

L. John Fahrner

Midwest Radiology

FAHRNER@GMAIL.COM

Pranav Rajpurkar

Department of Biomedical Informatics, Harvard Medical School

PRANAV_RAJPURKAR@HMS.HARVARD.EDU

Abstract

Generating accurate radiology reports from medical images is a clinically important but challenging task. While current vision language models show promise, they are prone to generating hallucinations, potentially compromising patient care. We introduce RadFlag, a black-box method to enhance the accuracy of radiology report generation. Our method uses a sampling-based flagging technique to find hallucinatory generations that should be removed. We first sample multiple reports at varying temperatures and then use a large language model to identify claims that are not consistently supported across samples, indicating that the model has low confidence in those claims. Using a calibrated threshold, we flag a fraction of these claims as likely hallucinations, which should undergo extra review or be automatically rejected. Our method achieves high precision when identifying both individual hallucinatory sentences and reports that contain hallucinations. As an easy-to-use, black-box system that only requires access to a model’s temperature parameter, RadFlag is compatible with a wide range of radiology report generation models and has the potential to broadly improve the quality of automated radiology reporting.

Keywords: Hallucination Detection, Radiology Report Generation, Selective Prediction,

Chest X-Rays, Large Language Models, Vision Language Models, Uncertainty Estimation

Data and Code Availability The primary dataset used in this research is MIMIC-CXR, an open-source chest X-ray report generation dataset publicly available through PhysioNet (Johnson et al., 2019). Code will be made publicly available at <https://github.com/rajpurkarlab/RadFlag>.

1. Introduction

Generating accurate radiology reports from medical images is a challenging task that requires a deep understanding of medical imaging and the ability to precisely interpret and communicate complex findings. Medical Vision Language Models (VLMs) that automatically generate radiology reports from chest X-ray images have the potential to mitigate shortages of radiological expertise and improve clinical efficiency and accuracy (Bannur et al., 2024; Tanida et al., 2023; Chen et al., 2024). However, such models are prone to hallucinations, where the system generates incorrect claims that are not supported by the input image (Liu et al., 2024a; Bai et al., 2024). For example, even high-performing radiology report generation models can hallucinate in approximately 40% of generated sentences (see Table 2 in Appendix D). Such errors can mislead clinicians, with potentially severe consequences for patient care.

* These authors contributed equally.

To reduce the adverse impact of hallucinations, we propose RadFlag, a new, easy-to-use method for detecting hallucinations in AI-generated radiology reports without needing access to the AI model’s inner workings. RadFlag is the first black-box hallucination detection method specifically designed for VLMS, and flags both sentences and reports as containing hallucinated findings. We provide two key contributions:

- We develop a novel entailment scoring method specifically designed for radiology, which accurately identifies medical findings that a VLM has low confidence in. We combine this scoring method with conformal risk control thresholding to detect hallucinated sentences with high precision.
- We further aggregate the sentence-level flagging results to identify reports with a high rate of hallucinations. This analysis enables selective prediction at the report level, making it possible to abstain from generating reports when there is a high chance of hallucinating.

Our empirical results show that RadFlag can accurately flag 28% of hallucinatory sentences while maintaining a flagging precision of 73% on Medversa, a recent high-performing report generation model [Zhou et al. \(2024\)](#). At the report level, our method analyzed 208 reports generated by MedVersa and divided them into two sets: a flagged set with 4.2 hallucinations per report ($n = 57$) and an accepted set with only 1.9 hallucinations per report ($n = 151$). This approach can unlock selective prediction, where models refrain from generating reports in cases of heightened uncertainty.

2. Related Work

2.1. Uncertainty Modeling and Hallucination Mitigation

Techniques for uncertainty modeling in LLMs fall into three categories based on how much access to the model is required: white-box (using internal components such as intermediate activations), gray-box (using output probabilities), and black-box (using only generated outputs) ([Liu et al., 2024b](#); [Huang et al., 2023](#)). Recently, black-box methods measuring self-consistency have emerged as accurate, easy-to-use approaches to assess model output confidence

([Zhang et al., 2024](#); [Mündler et al., 2024](#)). Self-Check GPT, a black-box hallucination detection method, compares multiple sampled responses, measures their consistency using tools like BERTscore and LLM prompting, and labels claims that appear inconsistently across samples as hallucinations. Despite using only model outputs, this black-box method demonstrates superior performance compared to gray-box approaches ([Manakul et al., 2023](#)). Discrete semantic entropy offers another black-box approach that clusters multiple generated outputs based on semantic similarity and computes entropy over these clusters to measure uncertainty ([Farquhar et al., 2024](#)). While these methods have proven effective for text-only models, to the best of our knowledge, RadFlag is the first work to extend this concept to VLMS.

In the domain of radiology report generation, various techniques have been explored to mitigate hallucinations and improve accuracy. Existing methods are generally not black-box and often involve fine-tuning a model with new losses or datasets ([Banerjee et al., 2024](#); [Anonymous, 2024](#); [Ramesh et al., 2022](#)). Additionally, the intersection of uncertainty modeling and hallucination detection remains relatively unexplored for radiology report generation. To the best of our knowledge, only one study has applied uncertainty modeling to remove hallucinations in radiology report generation by using Monte Carlo dropout to estimate uncertainty at both the image and text levels, integrating these estimates into a novel loss function ([Wang et al., 2024](#)). While effective, this method relies on access to model weights. In contrast, our sampling-based method provides a versatile, easy-to-use solution that is model-agnostic and compatible with both proprietary and closed-source models.

2.2. Selective Prediction for Abstention

The concept of selective prediction, where models abstain from generating predictions when there is a high chance of error, has also been explored in Natural Language Processing (NLP) ([Geifman and El-Yaniv, 2017](#); [Xin et al., 2021](#); [Vazhentsev et al., 2023](#)). Selective prediction relies on uncertainty estimation to decide whether to withhold a response, using the assumption that the model is likely to be incorrect when uncertain. A study by [Yadkori et al. \(2024\)](#) reduces hallucinations in LLM question-answering by combining selective prediction with self-consistency sampling. They evaluate response similarity across samples and use a calibration dataset to set a selective

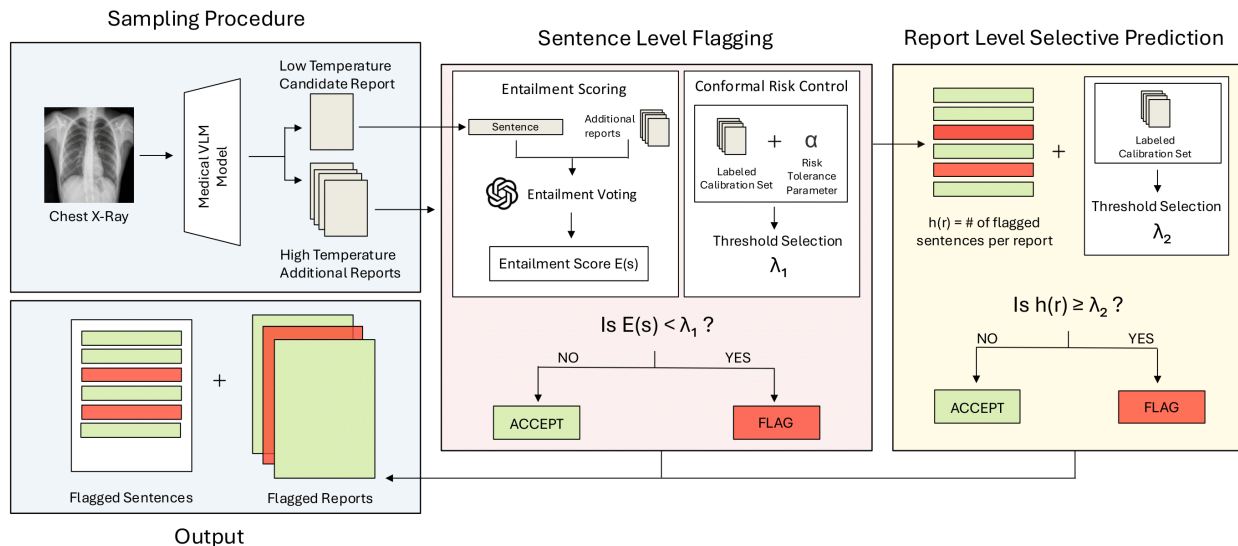


Figure 1: RadFlag flags hallucinatory text in a candidate report by comparing it against a corpus of additional reports sampled at a higher temperature. GPT-4 is used to count how many of the high-temperature reports support each sentence in the candidate report. We flag sentences with little support, as well as reports that contain multiple flagged sentences.

prediction threshold, offering strong theoretical guarantees on error rates. Selective prediction has also been applied to medical imaging. In chest X-ray image classification, selective prediction has shown to improve overall accuracy when models abstain when uncertain about chest X-ray findings (Ghesu et al., 2021). We are the first to develop domain-specific uncertainty estimation techniques for selective prediction in radiology report generation.

3. Methods

The RadFlag method is designed to flag hallucinatory content in generated radiology reports as depicted in Figure 1. The method consists of the following key steps:

1. Given an input chest X-ray image and a report generation model, we generate one candidate report at a low temperature setting. We also generate a corpus of reports sampled at a higher temperature setting.
2. We tokenize the candidate report into sentences. For each sentence, we compute an entailment score using GPT-4 (Achiam et al., 2023), where this score represents the number of sampled re-

ports in the high-temperature corpus that support this particular sentence.

3. The computed entailment scores for each sentence are compared against a sentence-level threshold, which is chosen using a calibration dataset. Sentences with scores below this threshold are flagged as likely hallucinations.
4. We tally the number of flagged sentences in each candidate report. The total number of flagged sentences in each candidate report is compared against a report-level threshold, also determined using the calibration dataset. If the count exceeds this report-level threshold, the candidate report is flagged as being especially likely to contain hallucinations.

3.1. Radiology-Specific Entailment Function

To determine whether a sentence s is entailed by a report r , we define a few-shot entailment function tailored for GPT-4 using radiology-specific knowledge. The detailed entailment prompt can be found in Appendix A. We categorize entailment into three key groups:

- **Completely Entailed (CE):** This category includes sentences that are fully confirmed by the

report. For example, the sentence “atelectasis in the left lung” would be confirmed by a report mentioning “bilateral atelectasis.” Notably, negative findings, like “the lungs are clear” or “there are no bony abnormalities,” are also classified as CE even when not explicitly confirmed by the report, as it is common clinical practice to leave negative findings implied rather than exhaustively list them.

- **Partially Entailed (PE):** This category captures sentences that are mostly confirmed by the report but have slightly different details, such as severity, size, or anatomical location. For example, the sentence “there is mild cardiomegaly” would be partially entailed by a report that mentions “moderate cardiomegaly.” Additionally, this category is used when entailment is unclear due to references to prior exams, such as when a sentence states “heart size is normal” and a report states “heart size is unchanged.”
- **Not Entailed (NE):** This category includes findings that are either directly contradicted by the report or, in the case of positive findings, omitted entirely. For example, if the sentence claims “there is no pneumothorax” but the report states “there is a pneumothorax,” the sentence is NE. Furthermore, if the sentence claims “there is consolidation” but the report does not mention consolidation, the sentence is NE, as reports typically list all abnormalities that are present.

3.2. Entailment Scoring for Candidate Sentences using High-Temperature Reports

We use an entailment score to represent how much support each sentence in the candidate report, which is sampled at a low temperature, has from the corpus of additional reports sampled at a fixed higher temperature. We define this score as $E(s, \{r_1, \dots, r_n\})$ where s is a sentence from the candidate report, and r_i is the i th report in the corpus of n high-temperature samples. To compute this score for a sentence s , we use the entailment function described above to compare s against every high-temperature report, and we then sum the number of reports categorized as “Completely Entailed” or “Partially Entailed.” The final score can range from 0, indicating that no high-temperature report samples support the

sentence, to n , indicating that all sampled reports support the sentence.

In this step, we choose to count “Partially Entailed” sentences as entailed because models often make small mistakes at higher temperatures despite being correct at lower temperatures. For example, we observe a model correctly noting that an endotracheal tube is 2.5 inches from the carina in a candidate report but claiming that it is 3.5 inches away in a high-temperature sample. To maintain precision and avoid incorrectly flagging factual sentences as hallucinations, we combine “partial entailment” with “complete entailment” when comparing sentences against high-temperature samples; we only categorize a sentence as “not entailed” when there is a substantive mismatch in content. The final score can range from 0, indicating that no high-temperature report samples support the sentence, to n , indicating that all samples support the sentence.

3.3. Sentence-Level Flagging Method

$$f_{\text{sent}}(s; r_1, \dots, r_n; \lambda_1) = \begin{cases} 1 & \text{if } E(s, \{r_1, \dots, r_n\}) < \lambda_1, \\ 0 & \text{if } E(s, \{r_1, \dots, r_n\}) \geq \lambda_1. \end{cases} \quad (1)$$

When flagging sentences, the entailment score for each sentence is compared against a calibrated threshold λ_1 . If the entailment score falls below this threshold, the sentence is flagged as a likely hallucination, as it is poorly supported by the high-temperature reports. Otherwise, the sentence is accepted, not flagged.

3.4. Calibrating Sentence-Level Thresholds Using Conformal Risk Control

We employ a Conformal Risk Control (CRC) framework to choose λ_1 , our sentence-level hallucination threshold, using an automatically labeled calibration dataset. CRC provides rigorous theoretical guarantees on performance by controlling how often factual sentences are incorrectly flagged as possible hallucinations. Informally speaking, it aims to keep the proportion of factual statements that are flagged by our method from exceeding α , where α is a parameter that users set based on their risk tolerance. Using a low value for α allows us to achieve high precision when flagging.

To label our calibration set, we define a binary entailment function $E_{\text{cal}}(s, r_{\text{gt}})$. s is a sentence from the candidate report, which was generated by a VLM at a

low temperature, and r_{gt} is the corresponding expert-written ground-truth report. As above, we apply our GPT-4 entailment function and treat a sentence as entailed if GPT-4 chooses CE and as not entailed if GPT-4 chooses PE or NE, indicating any mismatch in content. If a sentence is entailed, $E_{cal}(s, r_{gt}) = 1$; otherwise, $E_{cal}(s, r_{gt}) = 0$.

We then define a loss function $l(s; r_{gt}; \lambda)$ to calibrate the sentence-level threshold λ_1 . This loss function is designed to penalize the incorrect flagging of non-hallucinated sentences. The loss function is expressed as:

$$l(s; r_{gt}; \lambda) = f_{\text{sent}}(s; r_1, \dots, r_n; \lambda_1) \cdot E_{cal}(s, r_{gt}) \quad (2)$$

This loss function equals 1 when a factual sentence entailed by the ground truth is incorrectly flagged, and it equals 0 otherwise. We next follow the CRC framework to choose a threshold λ_1 that will control how often factual sentences are incorrectly flagged. To do so, we set λ_1 as follows:

$$\lambda_1 = \inf \left\{ \lambda : \frac{c+1}{c} L_c(\lambda) + \frac{1}{c+1} \leq \alpha \right\} \quad (3)$$

where for simplicity, $L_c(\lambda)$ represents the average loss on the calibration set and c is the size of the calibration set. For more theoretical details, please see the original paper on CRC (Angelopoulos et al., 2023).

3.5. Report-Level Selective Prediction

To flag problematic reports, we start by counting the number of flagged sentences in each candidate report. This count is denoted as $h(r)$ for candidate report r . We compare each report’s count against a report-level threshold λ_2 to decide if the report should be flagged:

$$f_{rep}(r, \lambda_2) = \begin{cases} 1, & \text{if } h(r) \geq \lambda_2 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Reports with more than λ_2 hallucinations are flagged as being especially likely to contain hallucinations, while the rest are accepted. The value of λ_2 can be chosen based on its behavior on the calibration set; we explore values of λ_2 that flag 5-25% of candidate reports.

3.6. Models and Sampling

To evaluate the performance of RadFlag, we apply it to Medversa and RaDialog, two of the highest-performing radiology report generation models on ReXrank, a leaderboard developed for evaluating radiology report generation models from medical images (Lab, 2024; Zhou et al., 2024; Pellegrini et al., 2023). Both Medversa and RaDialog leverage LLMs for report generation, allowing for temperature-based sampling as required by our method. Based on model specifications for ideal hyperparameters, we chose temperature = 0.1 for the low-temperature generation. For the high temperature generations, prior literature referenced a range of 0.9-1.0 as a reasonable temperature setting (Yadkori et al., 2024; Manakul et al., 2023). After qualitatively examining generations from both MedVersa and RaDialog at various temperatures, we chose temperature 0.5 and 5.0, respectively, since they resulted in diverse generations which were still stylistically coherent.

3.7. Datasets

MIMIC-CXR is a publicly available dataset of chest X-ray images and associated free-text radiology reports (Johnson et al., 2019). We use the MIMIC-CXR test set to evaluate our pipeline for models trained on MIMIC-CXR. We curate a subset of 508 image-report pairs (comprising approximately 3,500 sentences), enriched for reports that have few to no references to prior exams (eg. “the heart is stable”) that would make judging entailment especially difficult. We randomly sampled 300 out of 508 reports to create the calibration set. The remaining 208 reports formed our test set.

4. Experiments

4.1. Choosing λ_1 as a Threshold for Flagging Sentences

Applying conformal prediction to determine the desired threshold λ_1 requires setting α , a risk tolerance parameter that controls how often factual sentences are flagged as hallucinations. Previous research has used values as high as 0.1 (Yadkori et al., 2024). For our purposes, we decide on a more conservative alpha value of 0.05 to prioritize precision when flagging hallucinations. λ_1 is then calculated through the CRC procedure described in Methods. Using this value,

we arrive at 6 for λ_1 for MedVersa and 4 for RaDialog, and we use these values throughout our Results section. We also give results from setting α to 0.02 in Appendix D.

4.2. Choosing λ_2 as a Threshold for Flagging Reports

λ_2 represents the number of flagged sentences in a report as detected by RadFlag. Reports with fewer flagged sentences than λ_2 are kept, while those with λ_2 errors or more are filtered out. This approach results in varying percentages of reports being filtered from the dataset. We explore three thresholds for λ_2 . For both models, $\lambda_2 = 4$ provides a conservative threshold, flagging only 5% of all reports in the test set. $\lambda_2 = 3$ flags around 10% of all reports in the test set, while $\lambda_2 = 2$ flags 25%.

4.3. Metrics

We assess report quality using four report-level metrics: RadCliQ-v1, RadGraph entity precision, RadGraph entity recall, RadGraph relation precision, and RadGraph relation recall (Yu et al., 2023; Jain et al., 2021). Lower RadCliQ-v1 scores indicate better performance, while higher scores indicate better performance on all RadGraph metrics. Of these metrics, RadGraph entity precision serves as the most direct proxy for hallucinations in radiology reports as it measures how many medical entities in the generated report appear in the ground truth. A lower RadGraph entity precision value suggests that hallucinations have introduced unwanted topics into the report.

4.4. Defining a True Hallucinations Metric

We introduce a custom metric termed “true hallucinations per report” to quantify the occurrence of true hallucinations. We use the same entailment function used to label our calibration set, using GPT-4 to assess whether sentences are supported by their corresponding ground-truth reports or not. To check GPT-4’s assessments, one radiologist and one senior internal medicine resident, who are listed in the acknowledgements, and one clinician, who is a contributor to the paper, labeled 547 sentences as “entailed” or “not entailed” by a ground-truth report. We then compare GPT-4’s results to the clinician labels and find that GPT-4 achieves 84% accuracy. A table with a detailed breakdown is shown in Appendix B.

5. Results

We demonstrate that RadFlag effectively flags hallucinations with high precision in both Medversa and RaDialog. Furthermore, the report-level selective prediction method reliably identifies reports with a high number of hallucinations.

5.1. RadFlag Flags Hallucinatory Sentences with High Precision

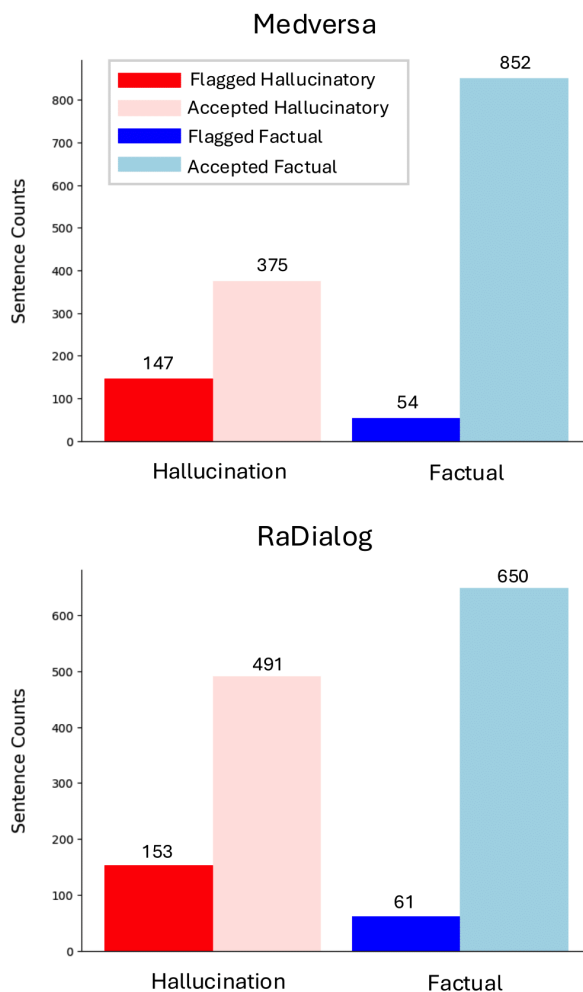


Figure 2: RadFlag flags hallucinatory sentences with high precision for Medversa (top) and RaDialog (bottom).

Figure 2 shows the number of factual and hallucinatory sentences flagged by RadFlag with respect to the remaining factual and hallucinatory sentences. When evaluating Medversa, RadFlag achieves 73%

precision while flagging 28% of all hallucinations. When evaluating RaDialog, RadFlag achieves 71% precision while flagging 24% of all hallucinations.

For a more detailed analysis, we categorized each sentence into one of six medical findings: Lungs, Pleural, Cardiomedastinal, Musculoskeletal, Devices, and Other. These categories were developed with input from radiologists (further details in Appendix C).

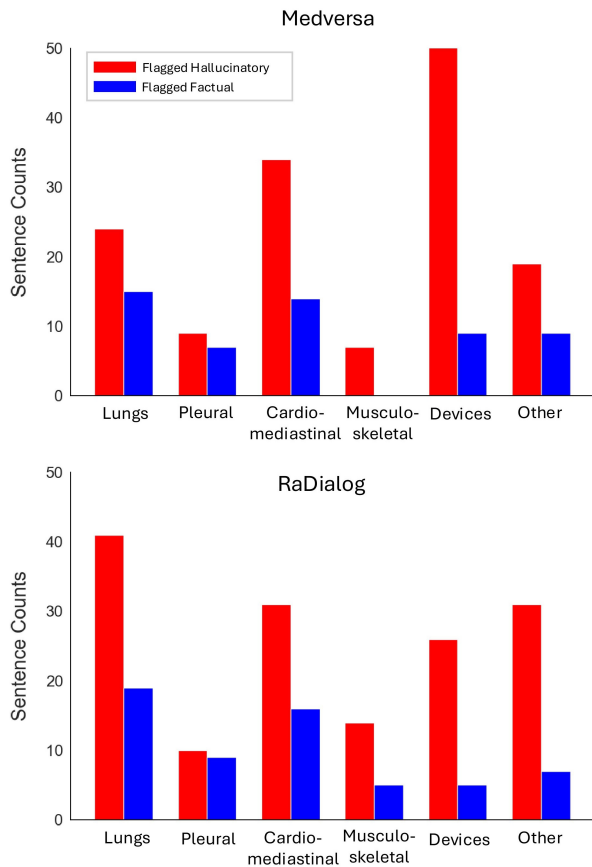


Figure 3: RadFlag consistently flags more hallucinatory sentences than factual sentences, though proportions vary across categories.

The effectiveness of hallucination detection varies across different findings categories and can be seen in Figure 3. Detailed category counts are provided in Appendix E. Both models exhibit strong performance in the “Devices” and “Musculoskeletal” categories, flagging far more hallucinatory than factual statements. However, the sampling method is less effective in the “Pleural,” “Cardiomedastinal,” and “Lungs” categories, where similar numbers of factual

and hallucinated sentences are flagged. These differences across categories may reflect underlying variation in how confident and accurate a model is about different topics. This could be because certain features of chest X-ray images are rare in training data and difficult for VLMs to identify, such as device positions or musculoskeletal conditions.

We evaluate the effect of automatically removing all flagged sentences using RadCliQ-v1 and RadGraph metrics and found the impact on report-level metrics to be relatively minor. This limited effect is likely due to the small number of flagged sentences, as well as the fact that some factual sentences are incorrectly removed in the process. For details on this analysis, please see Appendix F.

5.2. RadFlag Flags Lower-Quality, Hallucinatory Reports

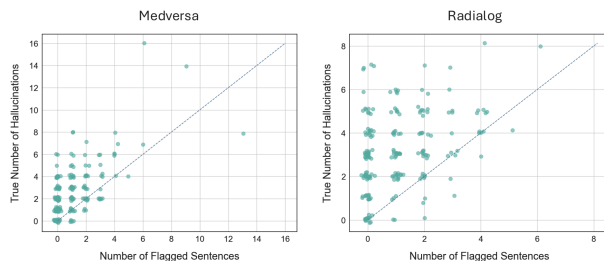


Figure 4: The number of sentences flagged by RadFlag serves as a lower bound for the true number of hallucinations in a report, as the large majority of points fall above the dashed blue line ($y=x$).

Next, we aggregate the flagged sentences in each report and use this as a measure to selectively flag reports that are likely to contain multiple hallucinations. Figure 4 compares the number of flagged sentences and true hallucinations for reports in our test set, and shows that the majority of the reports fall above the ideal $y = x$ line for both models. For example, Medversa-generated reports with 4 flagged sentences contain at least 4 true hallucinations. Therefore, the count of flagged sentences acts as a reliable lower bound for the number of true hallucinations.

Table 1 shows that RadFlag successfully flags low-quality reports. For Medversa, there is a considerable disparity in the quality of flagged and accepted reports across all metrics and thresholds. Notably, the 57 reports flagged when $\lambda_2 = 2$ have an average of 4.2 hallucinations per report, while the 151

Model	Quality	Threshold λ_2	n	Average True Hallucinations per Report	RadCliQ-v1	RadGraph Entity Precision	RadGraph Entity Recall	RadGraph Relation Precision	RadGraph Relation Recall
Medversa	Original	–	208	2.5	1.085	0.379	0.275	0.206	0.130
	Accepted	2	151	1.9	1.021	0.410	0.286	0.227	0.137
	Flagged	2	57	4.2	1.290	0.279	0.241	0.137	0.110
	Accepted	3	184	2.1	1.047	0.396	0.279	0.219	0.134
	Flagged	3	24	5.4	1.376	0.249	0.244	0.103	0.101
	Accepted	4	197	2.2	1.050	0.389	0.279	0.216	0.136
	Flagged	4	11	7.8	1.630	0.192	0.195	0.020	0.034
RaDialog	Original	–	208	3.1	1.222	0.308	0.241	0.140	0.103
	Accepted	2	147	2.7	1.188	0.311	0.232	0.142	0.097
	Flagged	2	61	3.6	1.213	0.299	0.262	0.135	0.120
	Accepted	3	183	2.8	1.190	0.312	0.237	0.143	0.101
	Flagged	3	25	4.0	1.233	0.275	0.271	0.119	0.118
	Accepted	4	195	2.9	1.189	0.313	0.239	0.143	0.103
	Flagged	4	13	4.6	1.290	0.269	0.099	0.107	0.336

Table 1: RadFlag splits reports into flagged and accepted sets, where the flagged reports are consistently lower-quality than accepted reports. Higher scores on True Hallucinations and RadCliQ-v1 indicate worse performance, while higher scores on the RadGraph metrics indicate better performance.

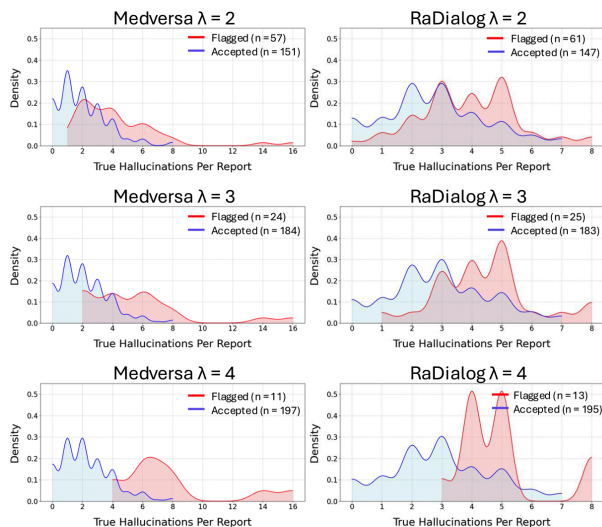


Figure 5: Across models and thresholds, flagged reports tend to have more hallucinations than accepted reports. Nearly all flagged reports contain at least one hallucinatory sentence.

accepted reports have an average of only 1.9 hallucinations per report. We see a similar trend for RaDialog, though the gaps between flagged and accepted reports are less pronounced. Figure 5 shows the distribution of hallucinations in flagged versus accepted reports across both models. According to the true hallucinations metric, RadCliQ-v1, and both RadGraph precision metrics, RaDialog’s accepted reports

are higher-quality than the flagged reports. In some cases, the flagged reports achieve higher RadGraph recall scores, indicating that these reports contain more hallucinations but also offer better coverage of correct content.

Our approach also achieves more reliable differentiation between high and low-quality reports, outperforming baseline entropy methods (see Appendix G). Despite requiring access to per-token probabilities and therefore requiring more technical expertise to implement, we found entropy baselines were not successful in identifying hallucinatory outputs. This is likely because there are many valid ways to express the same radiological idea, thus low probability does not equate to low confidence in the overall claim.

The qualitative examples presented in Figure 6 demonstrate the efficacy of the flagging system in identifying hallucinatory reports from model-generated outputs. This image is included to confirm that the statements are true hallucinations, rather than omissions in the ground truth report.

6. Discussion

This study is the first to propose a black-box method for mitigating hallucinations in radiology reports. We introduce RadFlag and evaluate its efficacy when identifying hallucinations produced by two high-performing models. Our approach is capable of flagging individual hallucinated sentences and problematic reports with high precision.

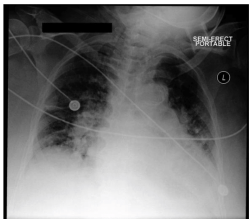
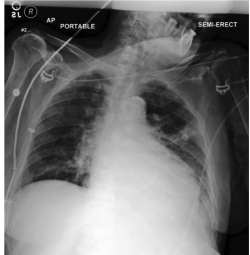
	<p>Predicted Report: The patient has been extubated. The NG tube is been removed. The right IJ central line is present, tip overlying the SVC/RA junction. There is cardiomegaly, with prominence of the hila, and upper zone redistribution, consistent with CHF. There is bibasilar atelectasis. There is blunting of the left costophrenic angle, consistent with a small left effusion. There is a small amount of subcutaneous emphysema in the neck. No pneumothorax.</p>	<p>Ground Truth: Previous moderately severe pulmonary edema has improved. Severe cardiomegaly and pulmonary artery dilatation are chronic. Poor aeration at the base of the left lung is persistent, either atelectasis or pneumonia. Pleural effusion is presumed, but not large.</p>
	<p>Predicted Report: The heart is mildly enlarged. There is pulmonary vascular redistribution. There is volume loss. There is infiltrate in both lower lungs. There is a small left effusion. There is a nondisplaced left lateral sixth rib fracture.</p>	<p>Ground Truth: Portable AP semi-upright view of the chest was provided. Patient is quite rotated to the left which limits the evaluation. The heart is enlarged which could in part reflect leftward rotation. There is a left pleural effusion which is small in size. There is a small area of consolidation in the left upper lobe which could represent a small focus of pneumonia. Mild pulmonary edema is present. Atherosclerotic calcifications at the aortic knob noted. Bony structures are intact.</p>

Figure 6: Qualitative examples from Medversa (top row) and RaDialog (bottom row) reports with 4+ hallucinations. Colored text shows radiologist-verified hallucinations: red for those flagged by RadFlag, blue for those missed.

While our method shows promise in detecting hallucinations, there is room for further improvement. We acknowledge that our reliance on GPT-4 for both entailment scoring and hallucination detection may introduce correlated errors that could inflate our accuracy metrics. While we validated GPT-4’s entailment scores with clinicians, future work would benefit from more extensive clinical validation of the hallucination detection process itself. Our findings suggest that category-specific thresholds (λ_1) for flagging errors could improve performance, given observed variations in our method’s performance across categories. Furthermore, complementary methods could be developed to identify other errors such as omissions, providing a more comprehensive assessment of reports. Finally, an implicit assumption in our current method is that AI models exhibit higher confidence when correct and lower confidence when incorrect. Future methods could test how well this assumption holds for specific models, making it possible to identify models that would most benefit from sampling-based methods.

Acknowledgments

We would like to thank Dr. Sara Gomez Villegas and Dr. Jung Oh Lee for their clinical expertise and help in evaluating GPT-4’s performance through entailment labeling.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2023. URL <https://arxiv.org/abs/2208.02814>.
- Anonymous. Hallucination mitigating for medical report generation. In *Submitted to ACL Rolling Review - June 2024*, 2024. URL <https://openreview.net/forum?id=X9h5qaKqQT>. under review.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2024. URL <https://arxiv.org/abs/2404.18930>.
- Oishi Banerjee, Hong-Yu Zhou, Subathra Adithan, Stephen Kwak, Kay Wu, and Pranav Rajpurkar. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation, 2024. URL <https://arxiv.org/abs/2406.06496>.

- Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. URL <https://arxiv.org/abs/2406.04449>.
- Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation, 2024. URL <https://arxiv.org/abs/2401.12208>.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.
- Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.
- Florin C Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Eli Gibson, RS Vishwanath, Abishek Balachandran, James M Balter, Yue Cao, Ramandeep Singh, et al. Quantifying and leveraging predictive uncertainty for medical image assessment. *Medical Image Analysis*, 68:101855, 2021.
- Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models, 2023. URL <https://arxiv.org/abs/2307.10236>.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- Rajpurkar Lab. Rexrank leaderboard. <https://rajpurkarlab.github.io/rexrank/>, July 2024.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models, 2024a. URL <https://arxiv.org/abs/2402.00253>.
- Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. Uncertainty estimation and quantification for llms: A simple supervised approach, 2024b. URL <https://arxiv.org/abs/2404.15993>.
- Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL <https://aclanthology.org/2023.emnlp-main.557>.
- Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation, 2024. URL <https://arxiv.org/abs/2305.15852>.
- Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance, 2023. URL <https://arxiv.org/abs/2311.18681>.
- Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR, 2022.
- Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.00718. URL <http://dx.doi.org/10.1109/CVPR52729.2023.00718>.

- Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.652. URL <https://aclanthology.org/2023.acl-long.652>.
- Yixin Wang, Zihao Lin, Zhe Xu, Haoyu Dong, Jie Luo, Jiang Tian, Zhongchao Shi, Lifu Huang, Yang Zhang, Jianping Fan, et al. Trust it or not: Confidence-guided automatic radiology report generation. *Neurocomputing*, 578:127374, 2024.
- Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.84. URL <https://aclanthology.org/2021.acl-long.84>.
- Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention, 2024. URL <https://arxiv.org/abs/2405.01563>.
- Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.
- Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. Luq: Long-text uncertainty quantification for llms, 2024. URL <https://arxiv.org/abs/2403.20279>.
- Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J. Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation, 2024. URL <https://arxiv.org/abs/2405.07988>.

Appendix A. Entailment Prompts

The prompts used for our method can be found in Figures A1 and A2. All prompting was done using GPT-4.

Prompt:

You are an AI radiology assistant helping process reports from chest X-rays. You are given a finding (F) from a chest X-ray and a list of reports (R1-R10). Your job is to determine if the ground truth report entails the finding, partially entails the finding, or do not entail the finding. Here are the definitions for entailment, partially entail, and do not entail:

Completely Entailed: A positive finding is entailed if it is clearly mentioned in the report. It is also entailed if the larger finding (atelectasis in both lungs) in the report includes the finding (atelectasis in the left lung). A negative finding is entailed if it is not mentioned in the report because we can assume that negative findings are not always explicitly stated. Note that negative findings are "normal findings" that include lungs are clear, heart is normal size, contours are normal, heart is not enlarged; even if they're not stated explicitly if it can be inferred that they're normal, then it is entailed.

Not Entailed: A positive finding is not entailed when the report doesn't mention the finding at all, or explicitly mentions the complete opposite finding. A negative finding is not entailed when the report mentions the complete opposite finding.

Partially Entailed: A positive finding is somewhat entailed if it is close to correct, but not quite—for example, there is a difference in the severity or location of the finding. A device is partially entailed if there are similar (but not the same) kinds of devices in the same position in the reports. Lastly, if there are any priors present in the finding that are absent from the report, it is also somewhat entailed.

Here are some examples:

#1 **Finding:** "There is no pleural effusion." **Report:** "There is a pneumothorax. There is no focal consolidation." **Score:** Completely entailed, because it is a negative finding not mentioned in the report.

#2 **Finding:** "The lungs are clear." **Report:** "No focal consolidation, pleural effusion, or evidence of pneumothorax is seen. There is no overt pulmonary edema." **Score:** Completely entailed, because "lungs are clear" is a negative finding and the report doesn't mention anywhere that the lungs ARE NOT CLEAR (which only happens if atelectasis, consolidations, pneumonia, or opacities etc. are present).

The end of every report will have no additional positive findings. Other negative findings are ALWAYS entailed.

Return a json of the format: "E": [], "P": [], "N": [], where the list contains the report number that falls in each category. For example, if a finding F is completely entailed by reports 1, 2, and 3, and not entailed by reports 4, 5, and 6, and partially entailed with respect to reports 7, 8, 9, and 10, then the dictionary returned is "E": [1, 2, 3], "P": [7, 8, 9, 10], "N": [4, 5, 6]. Note that the length of all E, N, and P MUST ADD UP to 10.

Figure A1: Prompt used for entailment score function $E(s, \{r_1, \dots, r_n\})$ to compute entailment score.

Prompt: Return a json of the format: {"status": "entailed"}, where the status contains completely entailed, partially entailed, or not entailed. For example, if a finding F is not entailed by the report, return {"status": "not entailed"}. If there is partial entailment, return {"status": "partially entailed"}.

Figure A2: Prompt used for ground truth scoring function $g(s, r)$ to compute True Hallucinations. The entailment scheme used in this prompt mirrors the one in Figure A1

Appendix B. Clinician Entailment Labels

Table A1 verifies that GPT-4 entailment scores align with clinician labels.

	Clinician Entailed	Clinician Not Entailed
GPT-4 Entailed	279	47
GPT-4 Not Entailed	42	179

Table A1: Clinician entailment labels align well with our GPT-4 Scoring Function.

Appendix C. Category Definitions

We categorize each sentence into one of six medical findings to assess how well RadFlag flags hallucinations in each category. These categories were developed in consultation with experienced radiologists to ensure comprehensive coverage of common chest X-ray observations. Below is a list of keywords that comprise each category. If sentences contain keywords from multiple categories, we assign them to the first matching category based on the following priority order: Devices, Cardiomedastinal, Lungs, Musculoskeletal, Pleura, and Other. We use regular expressions to search for these keywords when going through each sentence:

- **Devices:** PICC, Endotracheal, Nasogastric, Tube, Catheter, Pacemaker, Stent, NG, ET, ETT, Clip, Staple, Coil, ICD, LVAD, RVAD, Tracheostomy, Valve, Plate, Device, Drain, Gastric, Electrode, Lead, Port-a-cath, Cath, Clamp, Defibrillator, Internal Jugular Line, Subclavian Line, Hickman Line, Right Atrial Line, Broviac Line, Pacer, Tip, IJ, NJ, Wires, SVC, Dobbhoff, Intubated, Pump, Port, Extubate, Nasojejunal, Enteric, Impella, Mitraclip, Sternotomy, Suture, Subclavian, IABP, Balloon Pump, Filter, IVC Filter, TAVR, Stent Graft, Pigtail, Large Bore, Chest Tube, Epidural, Stimulator, Fixation, ORIF
- **Cardiovascular:** Hila, Hilar, Hilum, Lymph, Lymphadenopathy, Hypertension, Artery, Arteries, Mediastinal, Aorta, Hernia, Aortic, Mediastinum, Hiatal Hernia, Vascular Plethora, Pneumomediastinum, Cardiomegaly, Pneumopericardium, Pericardial Effusion, Pericardial, Cardio, Cardiac, Heart, Cardium, CHF, Redistribution, Congest, Engorge, Fluid Overload, Vascular, Fluid, Pneumomediastinum, Tension, Air, Shift, Pericardial Effusion
- **Lung:** Inflate, Volume, Bronchovascular, Diaphragm, Emphysema, Hyperinflated, COPD, Low, Hypoinflated, Well Expanded, Overinflated, Flattening, Aerated, Blebs, Cyst, Mass, Nodule, Lesion, Pneumonia, ARDS, Hazy, Haze, Atelectasis, Opacity, Opacit, Consolidation, Edema, Interstitial, Hemidiaphragm, Vessel, Infiltrate, Clear, Fibrosis, Scarring, Scar

- **Musculoskeletal:** Bones, Bony, Demineralized, Compression, Degenerative, Osseous, Fracture, Fractures, Rib, Spine, Scoliosis, Kyphosis, Osteopenia, Osteoporosis, Deformity, Deformities, Listhesis, Shoulder, Dislocation, Clavicle, Scapula, Humeral, Height, Loss, Stable, Unstable, Anterolisthesis, Retrolisthesis, Burst, Extension, Displaced, Distracted, Angulated, Angulation
- **Pleural:** Pneumothorax, Pneumoth, Effusion, Pleural, Thickening, Scarring, Pleural Mass, Costophrenic, Plaque, Blunting, Blunted, Silhouetting

Appendix D. Alpha Comparison for Sentence-Level Flagging

Using α values of .02 and .05, we arrive at λ_1 values of 4 and 6 for MedVersa and 2 and 4 for RaDialog. Below is a comparison of performance at these two values of α .

Table A2: Comparison of Model Performance

Model	α	Flag Halluc.	Flag Factual	Accept Factual	Accept Halluc.
Medversa	0.02	84	19	887	438
	0.05	147	54	852	375
RaDialog	0.02	84	24	687	560
	0.05	153	61	650	491

Appendix E. Category-Level Breakdowns

The following charts show the breakdown of flagged hallucinatory, flagged factual, accepted factual, and accepted hallucinatory statements across both models at $\alpha = 0.05$, or $\lambda_1 = 6$ for Medversa and $\lambda_1 = 4$ for RaDialog.

Table A3: Summary of Medversa Findings at $\lambda = 6$ ($\alpha = 0.05$)

Finding	Flag Halluc.	Flag Factual	Accept Factual	Accept Halluc.
Lungs	24	15	99	187
Pleura	9	7	63	327
Cardiomediastinal	34	14	100	216
Musculoskeletal	7	0	4	50
Devices	54	9	87	33
Other	19	9	22	39

Table A4: Summary of RaDialog Findings at $\lambda = 4$ ($\alpha = 0.05$)

Finding	Flag Halluc.	Flag Factual	Accept Factual	Accept Halluc.
Lungs	41	19	160	158
Pleura	10	9	60	251
Cardiomediastinal	31	16	103	134
Musculoskeletal	14	5	16	32
Devices	26	5	79	39
Other	31	7	73	36

Appendix F. Effect of Removing Flagged Sentences on Report-level Metrics

Table A5 shows that removing flagged sentences from reports does not seem to change overall report-level metric performance. The * indicates adjusted report-level metrics from removing empty reports (reports with all sentences flagged) that inflate RadCliQ-v1 scores.

Table A5: Performance Metrics (Findings Level) for Medversa and RaDialog

Model	Method	Avg. RadCliQ-v1	Avg. RadGraph Entity Precision	Avg. RadGraph Entity Recall	Avg. RadGraph Relation Precision	Avg. RadGraph Relation Recall
Medversa	Original	1.085	0.379	0.275	0.206	0.130
	$\alpha = 0.02$	1.084	0.390	0.268	0.212	0.127
	$\alpha = 0.05$	1.156	0.395	0.256	0.215	0.120
	$\alpha = 0.05^*$	1.106	0.396	0.257	0.216	0.119
RaDialog	Original	1.22	0.308	0.241	0.140	0.103
	$\alpha = 0.02$	1.20	0.328	0.232	0.152	0.100
	$\alpha = 0.05$	1.21	0.344	0.221	0.163	0.094

Appendix G. Entropy Baselines

To provide a comparison point for our black-box sampling method, we tested two baseline methods used by Manakul et al. (2023) that require access to output probabilities: average negative log probability and average entropy. In an LLM’s response, p_{ij} represents the probability of generating the j -th token in the i -th sentence, where J indicates the total token count per sentence. Average negative log probability is calculated as:

$$\text{Avg}(-\log p) = -\frac{1}{J} \sum_j \log p_{ij}$$

Entropy is calculated according to the equation $\mathcal{H}_{ij} = -\sum_{\tilde{w} \in \mathcal{W}} p_{ij}(\tilde{w}) \log p_{ij}(\tilde{w})$. Average entropy is thus calculated as:

$$\text{Avg}(\mathcal{H}) = \frac{1}{J} \sum_j \mathcal{H}_{ij}$$

We ran experiments using a new set of MedVersa generations (temp = 0.1, excluding 11 datapoints where generation failed) on 197 MIMIC-CXR test datapoints, flagging the reports that scored highest on these metrics. Despite requiring access to output probabilities, these baseline methods performed poorly, so hallucinations appear at similar rates in "accepted" and "flagged" sentences.

Table A6: Comparison of different ranking methods and their hallucination detection performance.

Ranking Method	Label	n	Mean True Hallucinations	Median True Hallucinations
Average Entropy	Accepted	187	2.63	2.0
	Flagged	10	3.50	2.5
	Accepted	174	2.59	2.0
	Flagged	23	3.30	3.0
	Accepted	143	2.46	2.0
	Flagged	54	3.20	3.0
Average Negative Log Probability	Accepted	187	2.70	2.0
	Flagged	10	2.10	1.50
	Accepted	174	2.59	2.0
	Flagged	23	3.30	3.0
	Accepted	143	2.46	2.0
	Flagged	54	3.22	3.0