Uncertainty Estimation in Large Vision Language Models for Automated Radiology Report Generation

Jenny Xu JENNYXU6@STANFORD.EDU
Department of Computer Science, Stanford University, USA

Abstract

The automated generation of free-text radiology reports is crucial for improving diagnosis and treatment in clinical practice. The latest chest X-ray report generation models utilize large vision language model (LVLM) architectures, which demand a higher level of interpretability for clinical deployment. Uncertainty estimation scores can assist clinicians in evaluating the reliability of these model outputs and promoting broader adoption of automated systems. In this paper, we conduct a comprehensive evaluation of the correlation between 16 LLM uncertainty scores and 6 radiology report evaluation metrics across 4 state-of-the-art LVLMs for CXR report generation. Our findings show a strong Pearson correlation, ranging from 0.4 to 0.6 on a scale from -1 to 1, for several models. We provide a detailed analysis of these uncertainty scores and evaluation metrics, offering insights in applying these methods in real clinical settings. This study is the first to evaluate LLM-based uncertainty estimation scores for X-ray report generation LVLM models, establishing a benchmark and laying the groundwork for their adoption in clinical practice.

Keywords: Uncertainty quantification, X-rays report generation, large vision language model (LVLM).

Data and Code Availability We use the MIMIC-CXR-JPG database of chest radiographs (Johnson et al., 2019), which is available on the PhysioNet repository. The code is publicly available: Github repo.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

The automated generation of free-text radiology reports plays a pivotal role in diagnosis and treatment in clinical practice. Automated chest X-ray (CXR) report generation contributes to increased efficiency, enhanced accuracy, and consistency when deployed to support radiologists. Moreover, automated interpretation systems can be integrated into clinical workflows to provide real-time monitoring of patients, particularly in critical care settings. Pioneering radiology report generation methods include CXR-RePaiR (Endo et al., 2021), MedViLL (Moon et al., 2022), and Cls-Gen (Nguyen et al., 2021). With the recent rise of biomedical foundation models, the latest CXR report generation models adopt a fully generative end-to-end LVLM architecture. Notable works include MedVersa (Zhou et al., 2024), LLaVA-Med (Li et al., 2024), and CheXagent (Chen et al., 2024).

The deployment of automated radiology report generation faces several challenges. Radiology reports often vary in format and structure across hospitals, regions, and countries, which complicates the task. Additionally, due to the limited availability of training data, models trained on one dataset may struggle to perform consistently in real-world clinical settings, leading to variability in the quality of generated reports. Incorporating uncertainty quantification scores can help clinicians assess the reliability of model outputs, enhancing trust and accelerating the adoption of these automated technologies.

Previous uncertainty estimation models have been developed for deep learning-based report generation methods (Wang et al., 2024). With the rise of generative biomedical foundation models, it is essential to develop accurate uncertainty quantification methods tailored to LVLMs for automated CXR report generation.

In this paper, we conduct a comprehensive evaluation of the correlation between 16 LLM uncertainty scores and 6 radiology report evaluation metrics across 4 state-of-the-art LVLMs for CXR report generation. The uncertainty scores are divided into single-inference methods and multi-inference methods, with further classification into sample-based and perturbation-based approaches. The 6 evaluation met-

rics, widely used to benchmark radiology report generation models, include domain-specific metrics such as CheXbert, RadCliQ, and RadGraph, as well as general LLM evaluation metrics like BLEU2, BLEU4, and BERTScore (Yu et al., 2023).

Our experiments show that these uncertainty estimation scores exhibit a strong correlation in certain cases, ranging from 0.4 to 0.6 on a scale from -1 to 1.

Our contributions are three-folds:

1. We establish a benchmark to evaluate the uncertainty in LVLMs for CXR report generation. We conduct a comprehensive evaluation of the correlation between 16 LLM uncertainty scores and 6 radiology report evaluation metrics across 4 state-of-the-art LVLMs for CXR report generation, namely MedVersa (Zhou et al., 2024), CheXagent (Chen et al., 2024), XrayGPT (Thawkar et al., 2023), and LLaVA-Med (Li et al., 2024).

2. We observe a strong correlation, ranging from 0.4 to 0.6 on a scale from -1 to 1, in certain cases. Although our correlation results are not perfect (i.e., close to 1 or -1), they offer valuable insights and highlight challenges. LLM-based evaluation metrics contain inherent inaccuracies, and it may be unrealistic to expect a high correlation, such as 0.9, between uncertainty scores and evaluation metrics.

3. We perform ablation studies on using the RadGraph-extracted clinical entities and relations in the computation of single-inference uncertainty scores.

Since our work is the first to evaluate LLM-based uncertainty quantification scores on LVLMs for X-ray report generation, there are no prior benchmarks to compare against. To our knowledge, our work is also the first to establish a benchmark by correlating these uncertainty scores with evaluation metrics. We hope this study establishes a benchmark that motivates the medical AI research community to further explore uncertainty quantification for LVLM models in X-ray report generation and to translate these methods into real clinical practice.

2. Related Works

Uncertainty estimation for LLM and VLM. Uncertainty estimation methods in natural language processing can be broadly classified into calibration confidence-based methods, sampling-based methods, and distribution-based methods (Hu et al., 2023). One widely used technique is conformal prediction, which converts tasks into multi-choice problems and quantifies uncertainty by considering the size of the subset of potential labels (Ye et al., 2024). For instance, (Kostumov et al., 2024) employs conformal prediction to assess over 20 VLMs, focusing on multiple-choice visual question answering (VQA) tasks. In the realm of large language models that generate free-form text, recent approaches to uncertainty estimation directly utilize token-level probability distributions from the outputs (Ahdritz et al., 2024; Huang et al., 2023; Fadeeva et al., 2024). (Kuhn et al., 2023) introduces semantic entropy which incorporates linguistic invariances created by shared meanings.

CXR report generation and uncertainty estimation. Earlier CXR report generation models frequently utilized pathology classifications within the vision module, as seen in RepsNet (Tanwani et al., 2022) and RaDialog (Pellegrini et al., 2023). Some pioneering models also incorporated image-report retrieval techniques, such as MedViLL (Moon et al., 2022) and CXR-RePaiR (Endo et al., 2021). The latest advancements in medical report generation involve foundation biomedical models that employ fully generative end-to-end LVLM architectures, including MedVersa (Zhou et al., 2024) and CheXagent (Chen et al., 2024). While previous studies on uncertainty-aware report generation for chest X-rays have focused on deep learning-based models (Najdenkoska et al., 2022; Wang et al., 2024), there has been limited research on uncertainty estimation methods for LVLMs in the context of chest X-ray report generation.

3. Methods

3.1. Problem Setup

For each radiographic study, the model inputs are the CXR images associated with the study and a prompt that tells the model to generate a complete radiology report. The LVLM generates the radiology report with a sequence of tokens $Y = [y_1, y_2, ..., y_m]$. For each token, we obtain their probability distribution over the vocabulary. Each of the 16 uncertainty estimation methods aims to calculate a score u regarding the uncertainty of Y. Given the ground truth re-

port G, each of the evaluation metrics calculates a score s to measure the correctness of the generated report. Pearson, Spearman and Kendall's tau correlation coefficients are computed across all 16 uncertainty estimation scores and all 6 evaluation metrics. These correlations vary between -1 and +1 with 0 implying no correlation. A larger absolute value indicates stronger correlation, reflecting the relevancy of the uncertainty estimation score.

Default two-sided test is used in the computation of Pearson, Kendall's Tau and Spearman correlations. Bonferroni correction for multiple comparison tests is used: $0.05 \div (\text{n\_models} \times \text{n\_scores} \times \text{n\_metrics}) = 0.05 \div (4 \times 16 \times 6) = 1.3 \times 10^{-4}$. The significance level is 1.3×10^{-4} after Bonferroni correction.

3.2. Uncertainty Estimation Scores

The 16 uncertainty estimation scores are categorized into three types: single-inference, sample-based, and perturbation-based. An illustration is provided in Appendix A.

Single-inference category The single-inference uncertainty estimation scores are *MaxProb*, *AvgProb*, *MaxEntropy* and *AvgEntropy*. Higher scores indicate higher uncertainty.

Each token generated by the model can be viewed as a classification problem across the entire vocabulary, prompting us to examine the probability distribution of each token over this vocabulary. To calculate uncertainty scores at both the sentence and report levels, we use the four metrics proposed in (Manakul et al., 2023). The LLM generation process employs a greedy strategy with a temperature setting of 0. Let p_{ij} represent the probability of a token at position j in sentence i; the sentence-level uncertainty score is derived by taking either the maximum or average of these probabilities.

$$\text{MaxProb}_i = \max_j(-\log(p_{ij}))$$

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

The entropy H_{ij} for the token at position j in sentence i is calculated based on its distribution over W, where W is a subset of the vocabulary consisting of the top 50 tokens with the highest probabilities.

$$H_{ij} = -\sum_{\tilde{w}\in W} p_{ij}(\tilde{w})\log[p_{ij}(\tilde{w})]$$

$p_{ij}(\tilde{w})$ is the probability of the word $\tilde{w}$ being generated at the j-th position in i-th sentence. The sentence-level entropy based uncertainty estimation scores are calculated as:

$$\text{MaxEntropy}_i = \max_j(H_{ij})$$

$$\text{AvgEntropy}_i = \frac{1}{J}\sum_j(H_{ij})$$

To determine the report-level uncertainty scores, we explore two approaches: calculating the **average** and the **maximum** of all sentence-level uncertainty scores.

Sample-based category Sample-based uncertainty estimation scores utilize the stochastic nature of the generative model, estimating uncertainty based on the divergence among different predictions.

In LLMs, the temperature parameter controls the randomness of the generated output by adjusting the probabilities of various token choices. Lower temperatures make the model more deterministic by favoring high-probability tokens, whereas higher temperatures promote diversity and creativity by allowing less likely tokens to be selected.

As detailed in Appendix A, we perform three stochastic model inferences with a temperature $t > 0$. The Sentence Transformers library (Reimers and Gurevych, 2019) is used to generate vector embeddings for all inference outputs and calculate the cosine similarities for each pair of outputs. Text pairs with higher similarity scores are more semantically related.

We follow (Gal and Ghahramani, 2016), and adapt the variation ratio (VR) and variation ratio for original prediction (VRO) uncertainty estimation metrics to compute *SampleVR* and *SampleVRO*. Let dist() denote the cosine distance between the embeddings of two outputs, p_i be a stochastic inference output, and p_o represent the original report generated without stochasticity.

$$SampleVR = \frac{dist(p_1,p_2) + dist(p_2,p_3) + dist(p_1,p_3)}{3}$$

$$SampleVRO = \frac{dist(p_o,p_1) + dist(p_o,p_2) + dist(p_o,p_3)}{3}$$

Perturbation-based category The stochastic nature of LVLMs can also be activated by altering a

generated token, a process known as test-time augmentation. (Huang et al., 2023) Intuitively, any disturbance in the chain of token generation can influence the following tokens and potentially result in two semantically distinct outputs. We perturb three key points: 1) the token with the highest entropy, 2) the token with the lowest entropy and 3) the token showing the greatest entropy gain from the previous token in the sequence.

As shown in Appendix A, to calculate *MaxVR* and *MaxVRO*, we run three model inferences with a temperature of 0. In the first inference, we replace the token with the highest entropy with the next most probable token among the top 50. In the second inference, we substitute this token with the third highest probability token among the top 50, and in the third inference, we use the fourth highest probability token from the top 50. We then apply the VR and VRO calculations to obtain *MaxVR* and *MaxVRO*.

MinVR, MinVRO, MaxDiff VR, MaxDiff VRO are calculated in a similar manner with different points of perturbation.

Please note that expected calibrated error (ECE) is outside of the scope of this study, since it involves training an auxiliary model to cast the free-form text generation into a binary prediction task.

4. Experiments

4.1. Dataset

We use the MIMIC-CXR dataset (Johnson et al., 2019), a large and publicly accessible dataset collected at the Beth Israel Deaconess Medical Center in Boston, MA. It comprises 377,110 chest X-rays corresponding to 227,835 radiology studies. The dataset was fully deidentified, and the protected health information was removed. Since we take the pre-trained weights for the four large vision language models and since they haven't been trained on the official test split, we only utilize the official test split in our experiments. The free-text radiology report preprocessing followed the steps in CXR-RePair (Endo et al., 2021). Further, we remove indication, comparison and any information related to patient history from the ground truth reports. Our final test set includes 1000 studies and corresponding reports. We evaluate the model output at the study level and each study contains one or multiple chest X-rays. These 1000 studies include in total 1592 chest X-ray images.

Table 1: Large vision language models in our study.

Models	Vision Module	Language Module
MedVersa	Swin Transformer	Custom LLM
CheXagent	EVA-CLIP-g	Mistral
XrayGPT	MedClip	Vicuna
LLaVA-Med	LLaVA	LLaVA

4.2. Large Vision Language Models

The four models in our study are listed in Table 1.

1. **MedVersa** (Zhou et al., 2024) is a state-of-the-art GMAI model that supports multimodal outputs, inputs and on-the-fly task specification for 9 different medical image interpretation tasks. On the radiology report generation task, MedVersa is the top on the ReXrank leaderboard (Lab, 2024).

2. **CheXagent** (Chen et al., 2024) is an instruction-tuned Foundation Model capable of analyzing and summarizing CXRs. The model also places among the top models on the ReXrank leaderboard.

3. **LLaVA-Med** (Li et al., 2024) is a vision-language conversational assistant that can answer open-ended research questions of biomedical images. The model is built by adapting LLaVA to the biomedical domain.

4. **XrayGPT** (Thawkar et al., 2023) is a conversational medical vision-language model that can analyze and answer open-ended questions about CXRs.

4.3. Evaluation Metrics

Generated reports are evaluated against the MIMIC-CXR ground-truth reports on the study level using the following evaluation metrics. (Yu et al., 2023)

1. **BLEU2, BLEU4**: Computes n-gram overlap with brevity penalty.

2. **BERTScore**: Uses the contextual embeddings from a BERT model to compute the similarity of two text sequences.

3. **CheXbert**: CheXbert automatic labeler is used to predict the presence or absence of 14

pathologies from machine-generated and human-generated radiology reports. CheXbert vector similarity computes the cosine similarity between these indicator vectors of 14 pathologies.

4. **RadGraph F1**: Computes the overlap in clinical entities and relations that RadGraph extracts from machine- and human-generated reports.

5. **RadCliQ**: This composite metric combines evaluations of BLEU and RadGraph F1.

BLEU and BERTScore are general natural language metrics for measuring the similarity between machine-generated and human-generated texts. CheXbert vector similarity and RadGraph F1 are metrics designed to measure the correctness of clinical information. (Yu et al., 2023) finds that BERTScore and RadGraph F1 are the metrics with the two highest alignments with radiologists.

4.4. Correlation Experiment with Filtering

In this experiment, for each combination of evaluation metric and uncertainty score, the 1000 data points are sorted based on their evaluation metric. The correlations are calculated between each evaluation metric and uncertainty score by progressively removing x% of data points from the median evaluation metric, where x ranges from 0 to 90. The results of this experiment are presented in Section 5.2 and Fig. 1. The results are not presented in Table 2.

Motivation for this experiment: For each pair of evaluation score and uncertainty score, we made a scatter plot and observed that most of the 1000 data samples are concentrated in the middle range in terms of evaluation scores, as shown in Appendix D. They are also concentrated in the middle range in terms of uncertainty scores. Thus, our motivation for the experiments in this section is about calculating the correlations if we only retain samples with more extreme evaluation scores (Fig. 1) or uncertainty scores (Fig. 2).

5. Results and Analysis

5.1. Analysis across Models and Evaluation Metrics

We present the Pearson correlation coefficients between uncertainty scores and the evaluation metrics for the CXR report generation models in Table. 2. For RadCliQ evaluation metric, a lower metric score is better. For all other evaluation metrics, a higher metric score is better. Thus, we present results for -RadCliQ in order to standardize the direction of the correlation coefficients. The Pearson correlation ranges from -1 to +1, with 0 indicating no correlation. Positive correlations show that a higher uncertainty score correlates with better evaluation score, while negative correlations show that a lower uncertainty score correlates with better evaluation score. The absolute value shows the strength of correlation, which helps determine whether the uncertainty estimation can effectively predict the model's performance on a specific evaluation metric. For each evaluation metric and model, we analyze the strength of the correlations across all uncertainty scores.

The strength of correlations. Table 2 shows that most p-values are below 1.3×10^{-4} which is the significance level after Bonferroni correction, demonstrating statistical significance. For MedVersa, the strongest correlation is 0.38 between *SampleVRO* and BertScore. For CheXagent, the highest correlation is 0.49 between *MaxVRO* and -RadCliQ. XrayGPT shows its strongest correlation at 0.20 between BertScore and *SampleVRO*, while LLaVA-Med has its highest correlation of 0.28 between BLEU4 and *MaxEntropy*. Although our correlation results are not perfect (i.e., close to 1 or -1), they offer valuable insights and highlight challenges. Since our work is the first to evaluate LLM-based uncertainty quantification scores on LVLMs for X-ray report generation, there are no prior benchmarks to compare against. LLM-based evaluation metrics contain inherent inaccuracies, and it may be unrealistic to expect a high correlation, such as 0.9, between uncertainty scores and evaluation metrics.

Influence of models. Table 2 reveals that the uncertainty scores showing the strongest correlation with each evaluation metric vary across different models. For instance, with -RadCliQ as the evaluation metric, *SampleVRO* exhibits the highest correlation of 0.34 for MedVersa, while *MaxVRO* shows the highest correlation of -0.49 for CheXagent.

Additionally, the direction of the correlation—whether positive or negative—depends on the model. In the single-inference categories, for MedVersa, all uncertainty scores demonstrate a negative correlation with evaluation metrics, indicating that

Table 2: Pearson correlation coefficients between uncertainty scores and evaluation metrics for X-ray report generation performance across 4 models. The Pearson correlation varies beteween -1 and +1 with 0 implying no correlation. For each pair of evaluation metric and report generation model, the strongest correlations are ranked and highlighted as *top-1*, *top-2* and *top-3*. To improve the readability of the table, only the p-values of the top highlighted correlations are shown in parentheses. For all evaluation metrics, a higher metric score is better. Note that we compare with the -RadCliQ because higher -RadCliQ is better. (Gray shading is only for readability).

Model	Uncertainty Category	Uncertainty Scores	Evaluation Metrics					
			BERTScore	CheXbert	-RadCliQ	RadGraph	BLEU2	BLEU4
MedVersa	Single-inference (average)	MaxProb	-0.20	-0.14	-0.22	-0.23 (1e-14)	-0.15	-0.22
		AvgProb	-0.22	-0.12	-0.22	-0.22	-0.16	-0.26 (6e-18)
		MaxEntropy	-0.20	-0.15 (5e-5)	-0.22	-0.25 (7e-16)	-0.17	-0.24
		AvgEntropy	-0.21	-0.12	-0.21	-0.21	-0.17	-0.27 (5e-18)
	Single-inference (max)	MaxProb	-0.34 (2e-28)	-0.13	-0.27 (3e-18)	-0.18	-0.15	-0.20
		AvgProb	-0.34	-0.11	-0.25	-0.16	-0.15	-0.19
		MaxEntropy	-0.36 (2e-31)	-0.15 (2e-6)	-0.30 (5e-22)	-0.23	-0.19 (2e-6)	-0.24
		AvgEntropy	-0.34 (3e-28)	-0.12	-0.26	-0.17	-0.16	-0.21
	Sample-based	SampleVR	0.23	0.17 (2e-6)	0.22	0.19	0.21 (5e-11)	0.23
		SampleVRO	0.38 (5e-35)	0.20 (4e-9)	0.34 (2e-27)	0.28 (1e-20)	0.29 (1e-20)	0.30 (1e-21)
	Perturbation-based	MaxVR	0.09	0.05	0.08	0.07	0.11	0.08
		MaxVRO	0.14	0.06	0.11	0.07	0.12	0.10
		MinVR	-0.10	0.01	-0.036	-0.01	-0.1	-0.12
		MinVRO	-0.07	0.002	-0.03	0.004	-0.08	-0.11
		MaxDiffVR	0.10	0.09	0.11	0.09	0.10	0.09
		MaxDiffVRO	0.11	0.06	0.09	0.07	0.10	0.07
CheXagent	Single-inference (average)	MaxProb	0.09	-0.03	0.05	0.08	-0.02	-0.08
		AvgProb	0.14	0.06	0.145	0.22	0.02	-0.04
		MaxEntropy	0.08	-0.07	0.01	0.003	-0.05	-0.11
		AvgEntropy	0.18	0.09	0.18	0.25	0.05	-0.01
	Single-inference (max)	MaxProb	0.10	0.12	0.15	0.28	0.26	0.12
		AvgProb	0.24	0.22	0.29	0.40	0.32	0.19
		MaxEntropy	0.07	0.08	0.11	0.23	0.21	0.08
		AvgEntropy	0.27	0.24	0.32	0.43 (2e-45)	0.34	0.21
	Sample-based	SampleVR	0.11	0.13	0.13	0.08	0.07	0.08
		SampleVRO	0.12	0.12	0.13	0.11	0.11	0.10
	Perturbation-based	MaxVR	0.28	0.27	0.33	0.37	0.29	0.19
		MaxVRO	0.45 (4e-51)	0.37 (4e-34)	0.49 (5e-60)	0.48 (2e-59)	0.41 (1e-41)	0.29 (3e-20)
		MinVR	0.26	0.27	0.30	0.26	0.31	0.24
		MinVRO	0.47 (6e-43)	0.36 (2e-32)	0.45 (8e-51)	0.44 (2e-48)	0.46 (1e-52)	0.34 (2e-28)
		MaxDiffVR	0.31 (4e-24)	0.32 (1e-24)	0.36 (1e-32)	0.37	0.41 (1e-40)	0.31 (8e-23)
		MaxDiffVRO	0.29	0.24	0.30	0.30	0.36	0.28
XrayGPT	Single-inference (average)	MaxProb	-0.06	-0.15(3e-6)	-0.14	-0.12 (2e-4)	-0.02	-0.07
		AvgProb	-0.03	-0.14	-0.11	-0.08	-0.02	-0.08
		MaxEntropy	0.01	-0.13	-0.09	-0.1	-0.01	-0.05
		AvgEntropy	-0.03	-0.13 (2e-5)	-0.11	-0.09	-0.02	-0.08
	Single-inference (max)	MaxProb	-0.12 (1e-4)	-0.13	-0.15(3e-6)	0.09	-0.04	-0.09 (4e-3)
		AvgProb	-0.02	-0.08	-0.06	-0.03	-0.03	-0.04
		MaxEntropy	-0.12(1e-4)	-0.16 (2e-7)	-0.17(1e-7)	-0.11(7e-4)	-0.07 (3e-2)	-0.09 (4e-3)
		AvgEntropy	-0.03	-0.09	-0.07	-0.03	-0.03	-0.04
	Sample-based	SampleVR	0.09	0.1	0.12	0.08	0.04	0.05
		SampleVRO	0.20 (2e-10)	0.12	0.18 (1e-8)	0.13 (5e-5)	0.10 (2e-3)	0.10 (9e-4)
	Perturbation-based	MaxVR	0.02	0.02	0.03	0.02	-0.03	-0.02
		MaxVRO	0.04	0.02	0.04	0.03	-0.01	-0.01
		MinVR	-0.05	-0.05	-0.07	-0.1	-0.04	-0.05
		MinVRO	-0.01	-0.02	-0.03	-0.06	-0.04	-0.03
		MaxDiffVR	-0.02	-0.05	-0.04	-0.02	-0.03	-0.03
		MaxDiffVRO	-0.03	-0.02	-0.02	-0.01	-0.05 (1e-1)	-0.06
LLaVA-Med	Single-inference (average)	MaxProb	0.16	0.07 (2e-2)	0.13 (5e-5)	0.11 (3e-4)	0.20 (1e-10)	0.22 (4e-12)
		AvgProb	-0.16	0.04	-0.05	-0.02	0.11	0.11
		MaxEntropy	0.14	0.10 (2e-3)	0.13 (3e-5)	0.08	0.27 (6e-18)	0.28 (8e-19)
		AvgEntropy	-0.09	0.06 (5e-2)	-0.01	-0.01	0.20 (6e-10)	0.19 (7e-10)
	Single-inference (max)	MaxProb	0.19 (3e-9)	-0.02	-0.09	0.14 (6e-6)	0.02	0.01
		AvgProb	-0.19 (1e-9)	0.03	-0.08	-0.05	0.12	0.12
		MaxEntropy	0.27 (1e-17)	0.03	0.16 (1e-6)	0.09 (5e-3)	-0.02	-0.04
		AvgEntropy	-0.14	0.04	-0.05	-0.06	0.17	0.18
	Sample-based	Sample-VR	0.03	0.03	0.05	0.07	-0.02	-0.001
		Sample-VRO	0.02	0.05	0.06	0.05	-0.06	-0.03
	Perturbation-based	MaxVR	0.14	0.04	0.10	0.07	-0.08	-0.07
		MaxVRO	0.05	0.02	0.06	0.07	-0.10	-0.13
		MinVR	-0.05	-0.03	-0.04	-0.02	-0.05	-0.10
		MinVRO	-0.04	0.002	-0.02	-0.01	-0.04	-0.07
		MaxDiffVR	0.14	0.04	0.10	0.07	-0.08	-0.07
		MaxDiffVRO	0.05	0.02	0.06	0.07	-0.10	-0.13

higher uncertainty scores correspond to lower model performance. However, for CheXagent, we observe a majority of positive correlation.

Overall, we find that correlations are stronger for MedVersa and CheXagent compared to LLaVA-Med and XrayGPT. MedVersa and CheXagent are notably high-ranked on the RexRank leaderboard.

Comparison across uncertainty score categories.

Sample-based and perturbation-based uncertainty scores outperform single-inference scores in three out of four LVLMs. For MedVersa and XrayGPT, *SampleVRO* achieves the highest correlations across nearly all six evaluation metrics, while for CheXagent, *MaxVRO* and *MinVRO* lead in top correlations. For LLaVA-Med, single-inference uncertainty scores lead in top correlations.

Within the multi-inference categories, including sample-based and perturbation-based methods, VRO uncertainty scores consistently outperform their VR counterparts. This is evident from the performance difference between *SampleVR* and *SampleVRO* for MedVersa and XrayGPT. While VR (variation ratio) scores utilize stochastic inference outputs, VRO (variation ratio original) scores also incorporate the original deterministic inference output. Our findings suggest that combining stochastic inference outputs with the original deterministic output provides a more accurate prediction of uncertainty.

In the single-inference categories, *MaxProb* and *MaxEntropy* generally perform better than *AvgProb* and *AvgEntropy*, respectively, across all six evaluation metrics. Specifically, for BERTScore, taking the maximum of sentence-level uncertainty scores yields better results than averaging these scores.

The sign of the correlation varies across categories. For single-inference categories, MedVersa and XrayGPT show a negative correlation with evaluation metrics, indicating that a lower uncertainty score correlates with better model performance. For the sample-based category, all four models show a majority of positive correlations. For the perturbation-based category, CheXagent shows a positive correlation with evaluation metrics.

5.2. Achieving a strong correlation of 0.6

This experiment is described in Section 4.4 and the results are presented in Fig. 1 for MedVersa. We aim to convey that if the report has a high or low evaluation score, the uncertainty score will be more informative for the users.

We observe a notable increase in correlation strength across Pearson, Spearman, and Kendall's Tau correlations, with MedVersa reaching a high correlation of 0.6 in some cases. All Pearson, Spearman and Kendall's Tau coefficients increase in strength as we compute with a more balanced dataset. Spearman and Pearson correlations even reach 0.6 or -0.6 for BERTScore, RadCliQ and BLEU2, while having a p-value below 0.01, showing statistical significance. Appendix C shows the same experiment on CheXagent, where Spearman correlations show greater strength and achieves a correlation above 0.6 in some cases.

The data consistently indicates that Pearson correlation is stronger than Kendall's Tau correlation across all evaluation scores. This suggests that for MedVersa, a linear relationship model is more appropriate for assessing the performance of uncertainty scores than a ranked data model.

We also assess whether a similar trend emerges when removing data points based on their uncertainty scores from the middle range. In Fig. 2, we present results for MedVersa using BertScore as the evaluation metric, with the X-axis representing the percentage of data points removed. The correlation coefficients consistently strengthen with this approach.

5.3. Ablation Studies

For single-inference scores, we conduct ablation studies on RadGraph-based uncertainty quantification scores. These scores are calculated using only tokens related to clinical entities and relations extracted from RadGraph.

RadGraph (Jain et al., 2021) provides a dataset of clinical entity and relation annotations for radiology reports, defining a novel schema for extracting clinically relevant information. We apply single-inference uncertainty estimation methods to the clinical entities and relations extracted from the RadGraph Benchmark (Jain et al., 2021), a Deep Learning model designed for this purpose. We investigate whether the token probabilities associated with clinical entities and relations exhibit a stronger correlation with uncertainty scores. As shown in Appendix A, we compute uncertainty scores — such as *MaxProb, AvgProb, MaxEntropy*, and *AvgEntropy* — using only tokens like "low," "lung," and "volumes," which are extracted clinical relations and entities. However, as detailed in Appendix B, RadGraph-based uncertainty scores

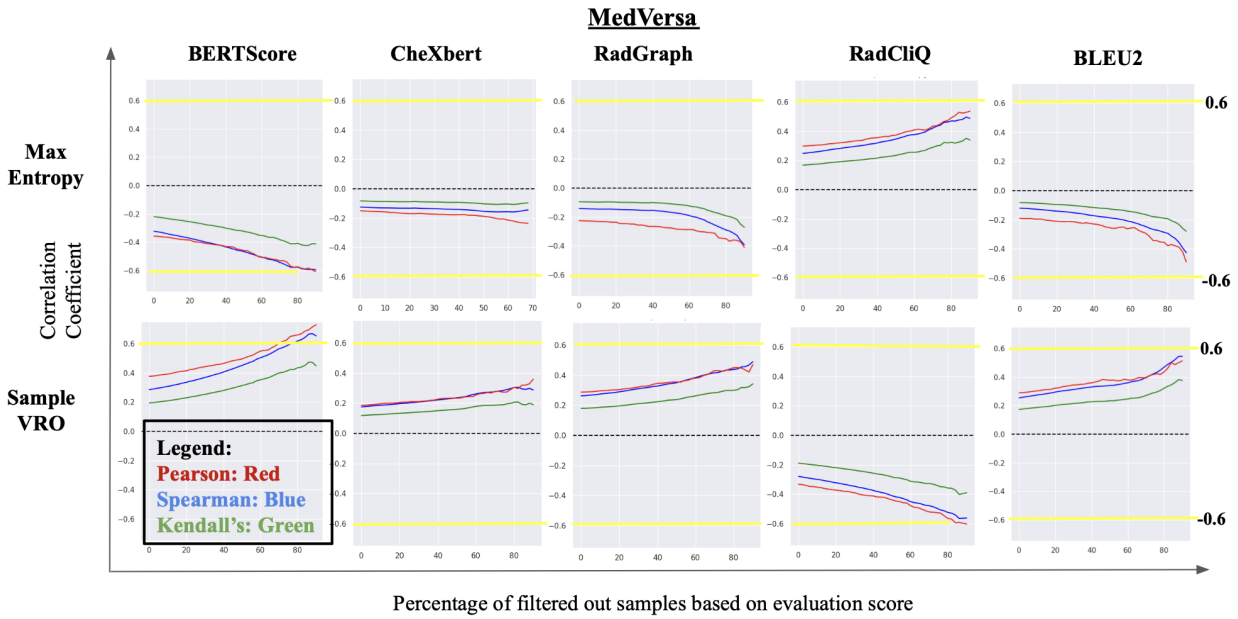

Figure 1: Experiments run on Medversa as described in Section 4.4 and 5.2.

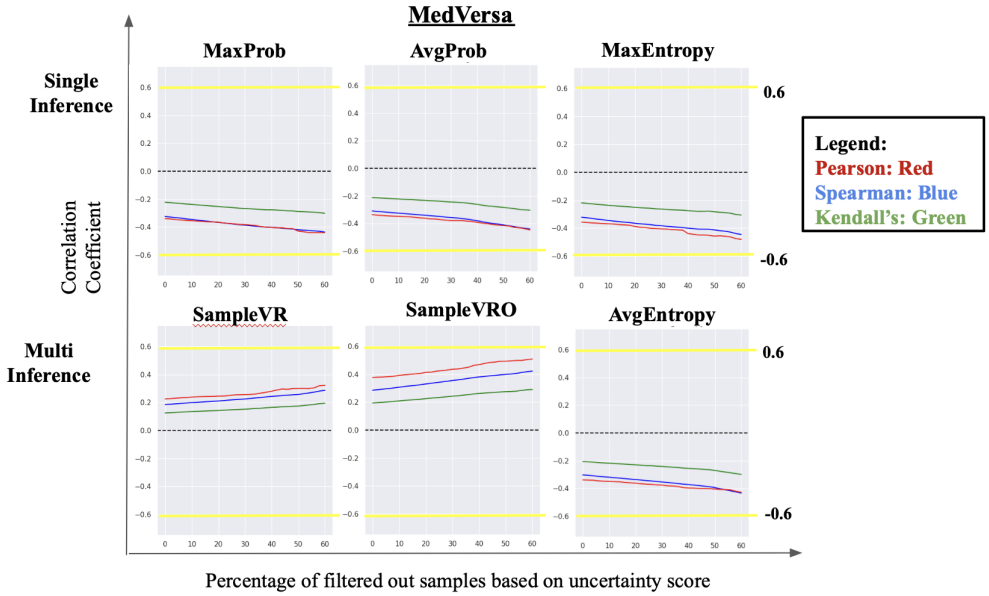

Figure 2: Experiments run on MedVersa for BertScore as described in Section 5.2. The correlations are calculated by progressively removing x% of data points from the median uncertainty scores.

do not demonstrate improved correlation compared to the original uncertainty scores.

6. Discussion: Clinical Impact

Since our work is the first to evaluate LLM-based uncertainty quantification scores on LVLMs for X-ray report generation, there are no prior benchmarks to compare against. Our work is also the first to establish a benchmark by correlating these uncertainty scores with evaluation metrics. Although our correlation results are not perfect (i.e., close to 1 or -1), they offer valuable insights and highlight challenges. LLM-based evaluation metrics contain inherent inaccuracies, and it may be unrealistic to expect a high correlation, such as 0.9, between uncertainty scores and evaluation metrics.

Section 5.1 contains a comparison across uncertainty score categories. Sample-based and perturbation-based uncertainty scores outperform single-inference scores in three out of four LVLMs. For MedVersa and XrayGPT, *SampleVRO* achieves the highest correlations across nearly all six evaluation metrics, while for CheXagent, *MaxVRO* and *MinVRO* lead in top correlations. For LLaVA-Med, single-inference uncertainty scores lead in top correlations.

The adoption of the uncertainty quantification methods outlined in this paper should follow a rigorous calibration process for different models, such as evaluating the correlations between the uncertainty scores and evaluation metrics for each specific model on a wide range of datasets. In the Results and Analysis section, we discussed that the strength of correlation varies across models. MedVersa and CheXagent are ranked higher on the RexRank leaderboard (Lab, 2024). The correlations are overall stronger for these models than LLaVA-Med and XrayGPT.

Our results show several insights related to clinical deployment. Through Fig. 1, we aim to convey that if the report has a high or low evaluation score, the uncertainty score will be more informative for the users. Upon calibration of the X-ray report generation model, the results in Fig. 2 reveal that an AI system could be set up so that it automatically generates a draft report only if the uncertainty score is within a threshold, thus avoiding disruption to clinicians' workflows when the probability of the model failing is higher, but helping when the model is correct.

Future steps: We think that evaluating more models, using additional X-ray report generation datasets and computing uncertainty scores at clinical "statement level" are promising future directions.

7. Conclusion

Recent LVLM chest X-rays report generation models achieved unprecedented accuracy (Zhou et al., 2024; Chen et al., 2024). This work is the first to evaluate LLM-based uncertainty quantification scores on LVLMs for X-ray report generation, by correlating these uncertainty scores with evaluation metrics. We hope this study establishes a benchmark that motivates the medical AI research community to further explore uncertainty quantification for LVLM models in X-ray report generation and to translate these methods into real clinical practice.

References

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L Edelman. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*, 2024.

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, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*, 2023.

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. *arXiv preprint arXiv:2307.10236*, 2023.

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, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Vasily Kostumov, Bulat Nutfullin, Oleg Pilipenko, and Eugene Ilyushin. Uncertainty-aware evaluation for vision-language models. *arXiv preprint arXiv:2402.14418*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Rajpurkar Lab. Rexrank leaderboard. https://rajpurkarlab.github.io/rexrank/, July 2024.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26 (12):6070–6080, 2022.

Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. Uncertainty-aware report generation for chest x-rays by variational topic inference. *Medical Image Analysis*, 82:102603, 2022.

Hoang TN Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated generation of accurate\& fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*, 2021.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

Ajay K Tanwani, Joelle Barral, and Daniel Freedman. Repsnet: Combining vision with language for automated medical reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 714–724. Springer, 2022.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

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.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*, 2024.

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.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.

Appendix A. Uncertainty Estimation Methods

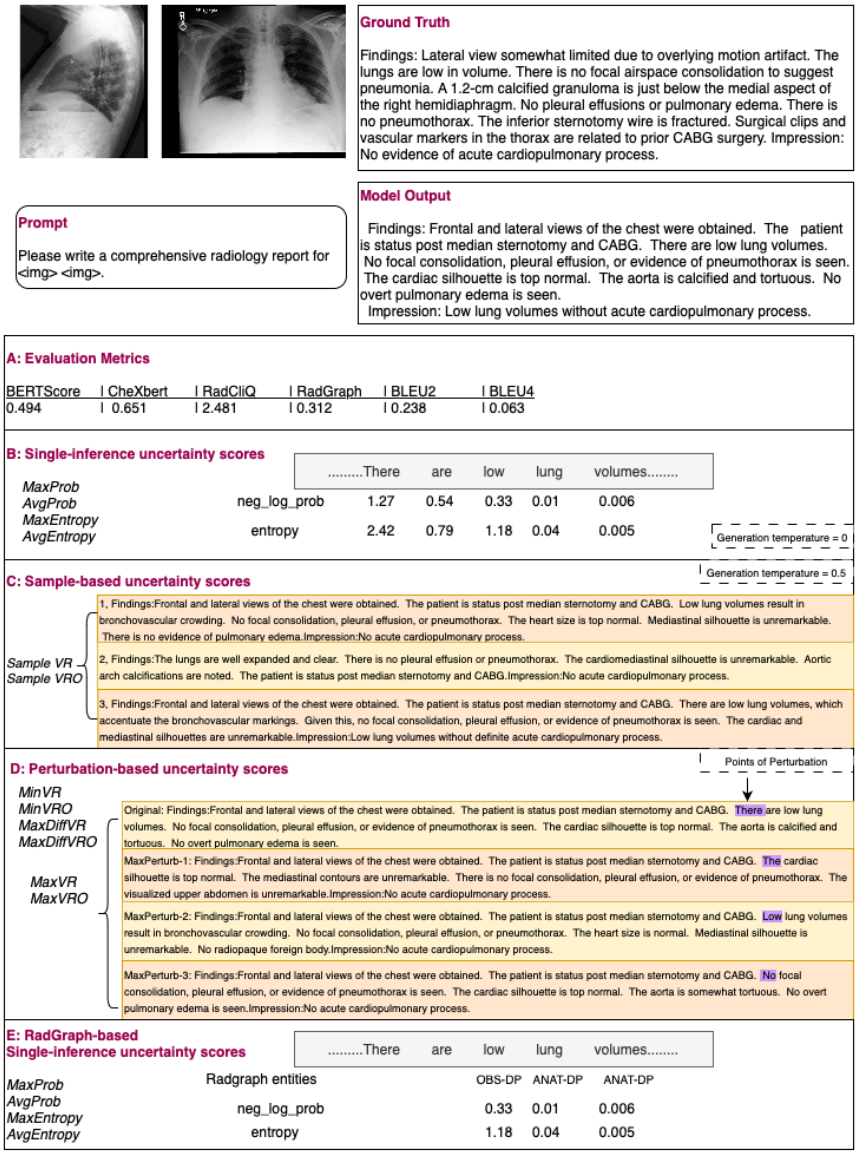

Figure 3: An illustration of uncertainty estimation methods using the MedVersa model. Panel E illustrates RadGraph-based uncertainty quantification scores using single-inference uncertainty scores, which is explained in Section 5.3.

Appendix B. Results for Ablation Studies

Table 3: A comparison of Pearson correlations between the original uncertainty quantification scores and RadGraph-based uncertainty scores. Higher correlations are highlighted in red for each uncertainty score.

Model	Uncertainty Score	Radgraph	Evaluation Metrics					
			BERTscore	CheXbert	-RadCliQ	RadGraph	BlEU2	BLEU4
MedVersa	MaxProb	Baseline	-0.339	-0.126	-0.27	-0.181	-0.149	-0.196
		With Radgraph	-0.317	-0.099	-0.244	-0.172	-0.129	-0.171
	AvgProb	Baseline	-0.335	-0.106	-0.253	-0.161	-0.148	-0.194
		With Radgraph	-0.267	-0.106	-0.206	-0.11	-0.126	-0.166
	MaxEntropy	Baseline	-0.357	-0.149	-0.298	-0.226	-0.191	-0.236
		With Radgraph	-0.323	-0.125	-0.265	-0.195	-0.141	-0.192
	AvgEntropy	Baseline	-0.339	-0.118	-0.262	-0.174	-0.162	-0.211
		With Radgraph	-0.261	-0.113	-0.206	-0.111	-0.134	-0.174
CheXagent	MaxProb	Baseline	0.104	0.119	0.152	0.276	0.26	0.117
		With Radgraph	0.081	0.1	0.125	0.243	0.257	0.12
	AvgProb	Baseline	0.237	0.215	0.289	0.399	0.315	0.185
		With Radgraph	0.196	0.176	0.242	0.352	0.283	0.171
	MaxEntropy	Baseline	0.073	0.083	0.113	0.228	0.207	0.081
		With Radgraph	0.076	0.076	0.108	0.216	0.22	0.097
	AvgEntropy	Baseline	0.268	0.241	0.322	0.426	0.339	0.21
		With Radgraph	0.253	0.226	0.303	0.401	0.315	0.211

Appendix C. Additional Results

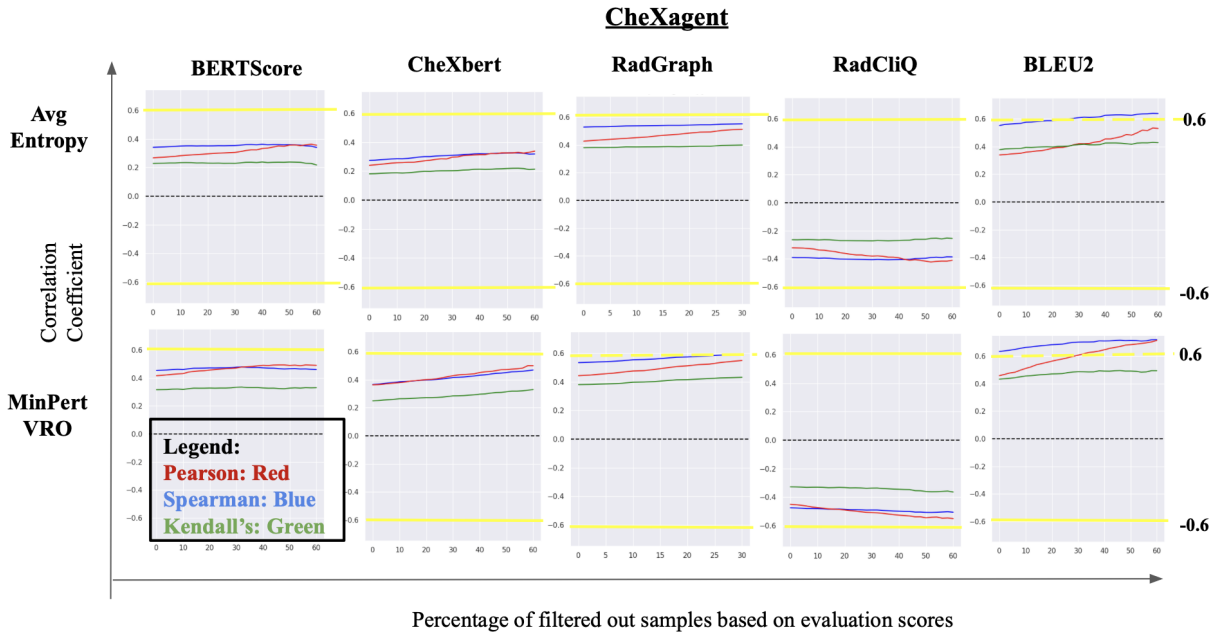

Figure 4: Experiments run on CheXagent as described in Section 4.4. and Section 5.2.

Appendix D. Example Scatter Plots for Correlations

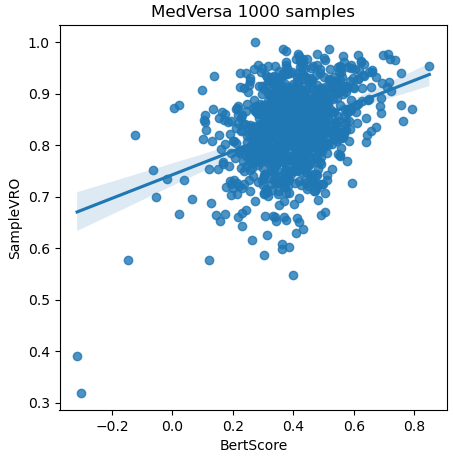

Figure 5: Scatter plot of the SampleVRO uncertainty scores against BertScore evaluation metric for MedVersa.

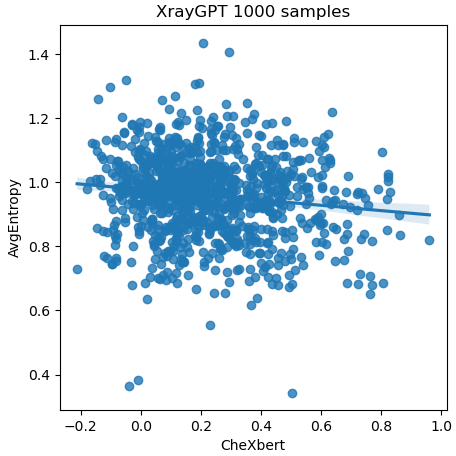

Figure 6: Scatter plot of the AvgEntropy uncertainty scores against CheXbert evaluation metric for XrayGPT.