RadALIGN: Mitigating Hallucinations and Improving Reliability in AI-Generated Radiology Reports

Krish Malik, Researcher, Junior, Parkland High School
Vedant Malik

Introduction

AI is transforming the field of radiology by enhancing multi-page report summarization complimenting radiologists. However, significant challenges with AI generated reports remain like speculative language, inaccuracies, deviations from factual content and lack of deep domain knowledge. LLMs can generate speculative conclusions, under adversarial conditions leading to flawed clinical decisions. RadALIGN framework aims to address these challenges by aligning LLMs, ensuring nonspeculative language, mitigating unwarranted assumptions to produce accurate, reliable and trustworthy radiology reports, thereby reducing hallucinations.

Hypothesis

If model alignment is designed to address issues like speculative language, factual inaccuracies, and compliance gaps with radiology-specific standards, then it will significantly enhance reliability, accuracy, and trustworthiness of AI-generated radiology reports.

Methods

A three-step framework using Indiana University Chest X-ray Radiology reports as input data with Llama3.2 3B Large Language Model (LLM) was implemented:

1. Baseline Generation: Establish baseline AI generated summary under standard conditions without model alignment: Reports generated using pre-trained LLM served as reference to identify factual gaps, clarity, speculative language.
2. Red Teaming: Identify vulnerabilities in AI generated report summaries.
3. Model Alignment: Align the AI based LLM model to generate safer, compliant, and factually accurate summaries.

Results

Model Alignment reduced Perplexity (prediction uncertainty) across all summarized reports, improving fluency and consistency. Perplexity dropped from range of 19.2655–71.0075 in the Baseline Model (unaligned) to 5.1829–6.2896 range with Model Alignment.
Token Diversity (vocab variability) decreased from range of 0.7778–1.0 to 0.5595–0.7099 range, restricting imagination, reducing hallucinations, and ensuring factual outputs. Red Teaming increased both Token Diversity (0.26 to 0.36-0.41) and

Perplexity (1.1 to 2.1-2.6) exposing vulnerabilities in Baseline model generated radiology reports, which were mitigated through Aligning the Model.

Conclusion

Model alignment significantly improved output robustness and factual consistency by reducing perplexity, leading to more confident and factual radiology report summary generation. By addressing vulnerabilities exposed through Red Teaming, Model Alignment reduced variability and hallucinations, ensuring outputs aligned with radiology standards for reliability and safety.

Figure(s)

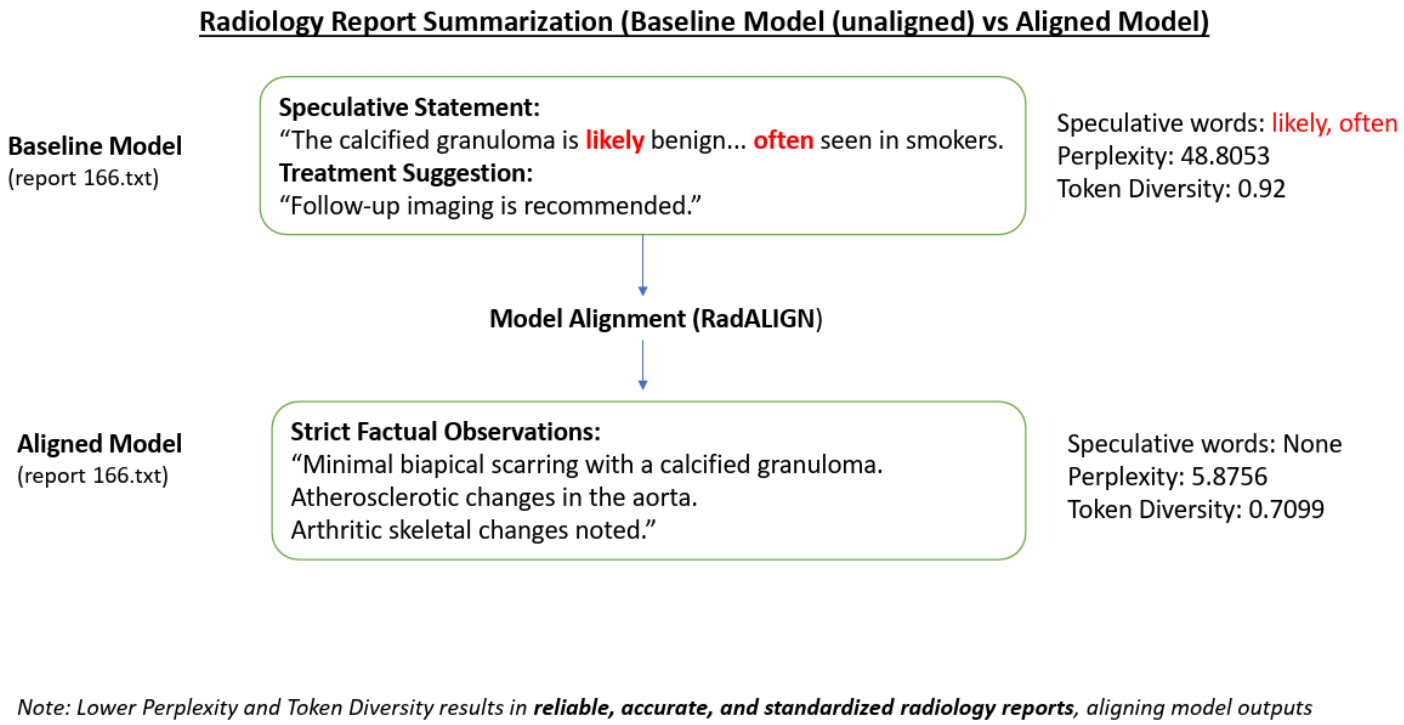
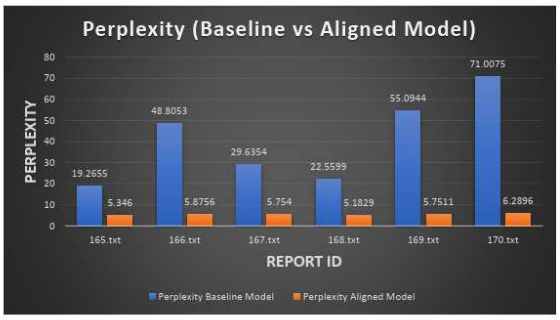
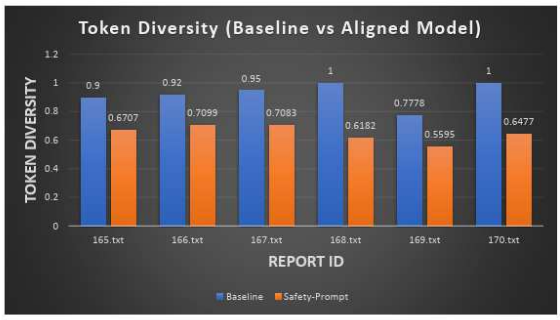


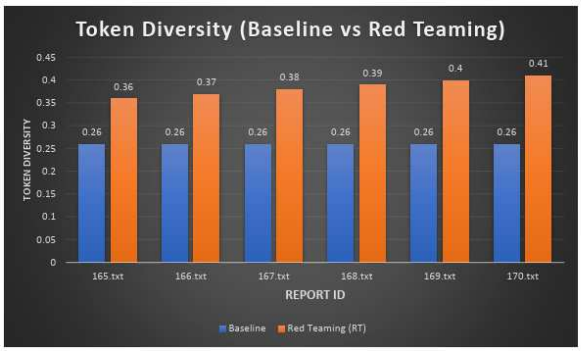
Figure 1. Perplexity and Token Diversity: Baseline vs. Aligned Model and Red Teaming

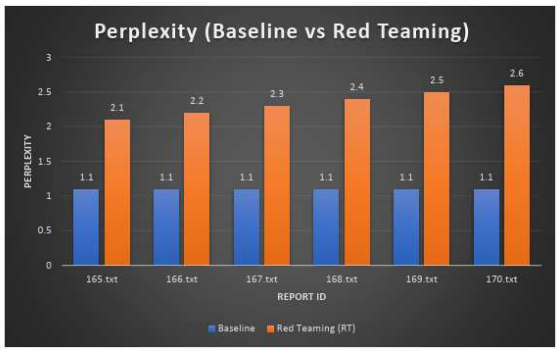




Note:
Perplexity is a measure used to evaluate Large Language Models (LLMs), representing how well a model predicts a sample. Lower perplexity indicates that the model is more confident and accurate in its predictions.

Note:
Token Diversity: 0 (no diversity) and 1 (high diversity





- **Red Teaming** increases Token Diversity and Perplexity, exposing vulnerabilities in radiology report summaries generated by an Large Language Model (LLM)
- **Red Teaming Baseline model** generates varied and uncertain outputs, including speculative content, **exposing vulnerabilities in an LLM and its ability to generate accurate Radiology Reports.**
- **Model alignment resolves vulnerabilities exposed by Red Teaming** by reducing Token Diversity and Perplexity ensuring controlled and factual outputs aligned with radiology standards, thereby reducing hallucinations and speculative language.

Figure 2. Baseline Model vs RadAlign based Model Alignment Output

Keywords

Applications; Artificial Intelligence/Machine Learning; Emerging Technologies