Automated Extraction of Unstructured Post-SBRT Toxicity Data from Radiology Reports using Large Language Models

Justin Pijanowski[1], Yakout Mezgueldi[1], Alan Lee[1], Drew Moghanaki[1], Ricky R Savjani[1], James Lamb[1]

UCLA Health®

2026 JULY 19–22 VANCOUVER, BC JOINT AAPM | COMP MEETING

Background

- Biological variability between patients (age, genetics, and co-morbidities) makes radiotherapy treatment response hard to predict
- Population-based dose-volume constraints can overestimate or underestimate toxicity risk for patients, motivating the development of personalized normal tissue complication probability models
- Clinically meaningful outcomes are often within unstructured EHR text
- Large Language Models (LLMs) can transform unstructured clinical text into structured data formats to be used for downstream analysis
- **Our work uses LLMs to extract the following clinically relevant endpoints from free-note text: locoregional progression, distant progression, radiation-induced rib fractures and radiation fibrosis**

Methods

- Lung cancer patients treated with lung SBRT at our institution were retrospectively identified from an internal radiation oncology database
- Patients were included if treated with SBRT for primary lung cancer and excluded if treated for metastases or had a prior non-lung malignancy
- A rulebook was created under the guidance of a thoracic radiation oncologist that defined our clinical endpoints and decision criteria
- Development set (n=60) labeled by two graders using consensus scorings
- Validation set (n=100) labeled by three graders and the report-truth scorings were determined by using majority vote among the three graders

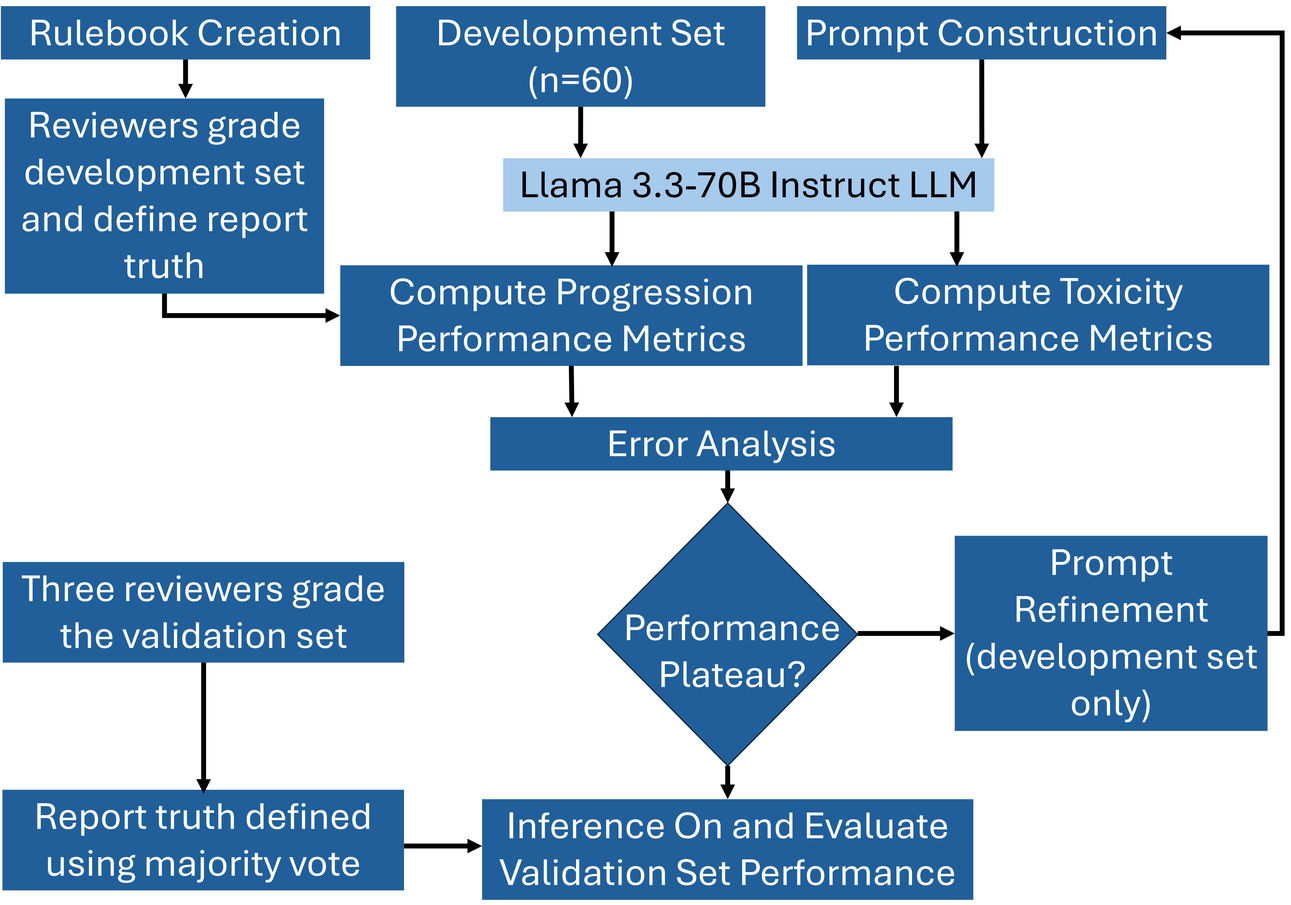

Figure 1: Flowchart detailing the prompt-engineering workflow

- Llama 3.3-70B Instruct LLM was served through Nvidia Nim container
- LLM classifications were compared against human-derived report truth
- The progression endpoints were analyzed using a three-class classification {yes, no, maybe} using a one-vs-rest approach
- Radiation induced toxicities were analyzed as binary-outcomes {yes, no}

Results

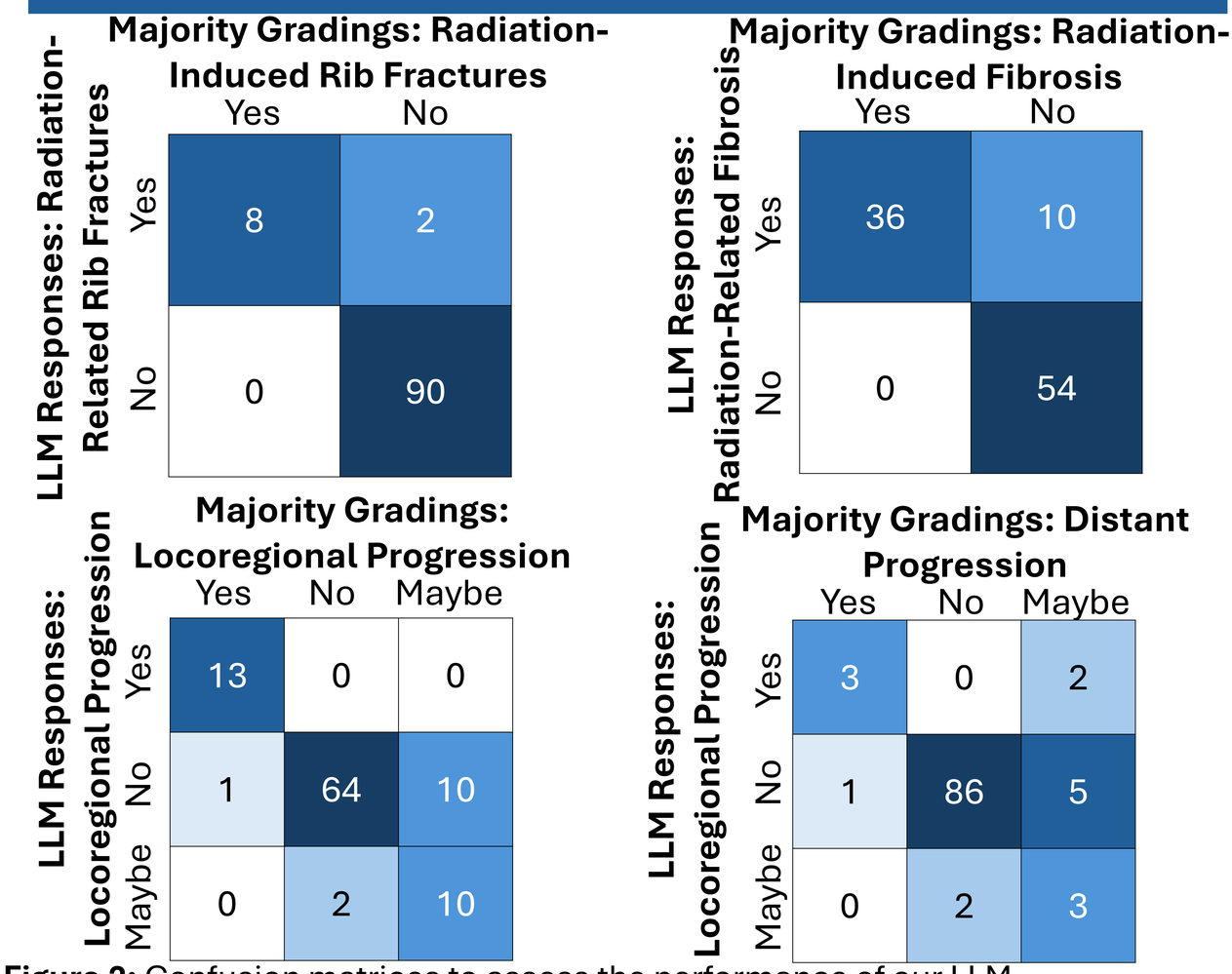

Figure 2: Confusion matrices to assess the performance of our LLM

Category	LLM Extracted Note Excerpt	Majority Grading	LLM Classification
Success: Rib Fracture	"Interval development of focal deformity and linear lucency of the left anterior second rib that may represent pathologic fracture in the context of radiation osteonecrosis"	Yes radiation-induced rib fracture	Yes radiation-induced rib fracture
Failure: Locoregional	"New ill-defined ground glass ...likely of inflammatory etiology and clinical correlation and interval follow-up is recommended to document resolution."	No locoregional	Maybe locoregional

Table 1: Example LLM Classifications

Discussion

- LLM demonstrated the strongest performance for radiation-induced toxicities, where report language was overall more explicit
- Fibrosis classification produced several false positives, since the LLM struggled to differentiate fibrosis and other post-treatment findings
- Progression classification was most reliable when reports clearly described new lesions or explicit concern for progression
  - Performance declined when reports contained ambiguous language
- Emphasizes the need for human review for LLM 'maybe' classifications
- Future work includes using a patient's entire set of notes for progression analysis and developing personalized predictive models

Acknowledgements
This research was fully supported by the Department of Radiation Oncology at the University of California, Los Angeles

jpijanowski@mednet.ucla.edu