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

Automated Classification of Radiation Oncology Safety Events Using Large Language Models (LLMs)

A novel approach to streamline reporting and enable retrospective analysis

Qiongge Li, PhD[1,2] Jian Liu, PhD[3,4,5], Xing Li, PhD[6], Wei Nie, PhD[1,2], and Jiajin Fan PhD[1]
[1] Department of Radiation Oncology, Inova Schar Cancer Institute, Fairfax, VA
[2] School of Medicine, University of Virginia, Charlottesville, VA
[3] Department of Radiation Medicine, Northwell, New Hyde Park, NY
[4] Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY
[5] Physics and Astronomy, Hofstra University, Hempstead, NY
[6] Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX

ABSTRACT

Purpose: To evaluate the feasibility of using a large language model (LLM) to automatically classify radiation oncology safety events from free-text narratives, and to outline an AI-assisted reporting framework.

Methods: 788 incidents (Jan 2021–Sep 2025) from our institutional reporting system were classified by GPT-5 across seven dimensions via iterative prompt refinement. Output was validated by three independent, blinded expert physicists using Cohen's κ.

Results: Model–reviewer agreement (mean κ = 0.31) was comparable to inter-reviewer agreement (mean κ = 0.42), approaching expert-level performance. Communication failures were the most common failure mode (~40%).

Conclusions: LLM-assisted classification can turn unanalyzed free-text reports into structured, analyzable data with no added reporter burden—a proof-of-concept for AI-enabled incident learning.

CONTACT

Qiongge Li, PhD
Department of Radiation Oncology, Inova Schar Cancer Institute, Fairfax, VA
Email: Qiongge.li@inova.org

This work is under review of Practical Radiation Oncology (PRO) since Feb 2026

INTRODUCTION

Incident reporting underpins a strong safety culture, but its value depends on whether the data can be analyzed.

Our institutional platform, SafetyAlways, prioritizes ease of reporting: staff write a free-text narrative and only minimal structured fields are captured. As a result, 788 incidents accumulated over ~5 years sit essentially unanalyzed.

Structured systems (RO-ILS, SAFRON) capture rich data, but only by asking reporters to classify each event by hand, a disincentive to report in a busy clinic.

Key insight: the information needed for classification already lives in the narrative. What is missing is the extraction. We propose using an LLM to extract structured fields directly from the free text the reporter already writes.

Study goals:

- Demonstrate feasibility of multi-dimensional LLM classification from narrative alone
- Surface hidden patterns: common failure modes and workflow vulnerabilities
- Design an AI-assisted, human-in-the-loop reporting system

METHODS AND MATERIALS

Data: 788 de-identified radiation oncology safety reports (Jan 2021–Sep 2025).

Model: GPT-5 (OpenAI) via API, temperature = 0 for deterministic output. Zero-shot, instruction-based prompting; one stateless call per report.

Seven dimensions: failure mode, severity, treatment type, discoverer role, occurred & discovered workflow stage, and an exclusion flag. Four dimensions permit multiple labels.

Prompt refinement: refined on 50 incidents over 19 reasoning patterns, then frozen. The 50-incident refinement set and 80-incident validation set were strictly separate—reported agreement is held-out.

Validation: three independent, blinded expert medical physicists classified 80 incidents; pairwise Cohen's κ per dimension. After exclusions, 650 records remained for retrospective analysis.

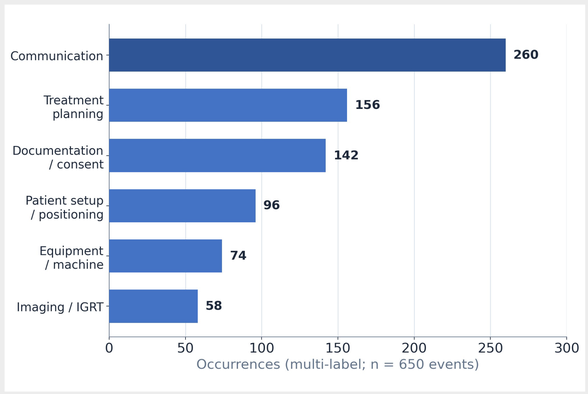

Figure 1. Failure mode distribution (n = 650; multi-label).

RESULTS

Expert-level agreement. Model–reviewer agreement (mean κ = 0.31) was of the same order as agreement among the reviewers themselves (mean κ = 0.42). On average the model's label fell within the experts' group-accepted set in 67.9% of comparisons. (as shown in Table 1)

Per-dimension, not pooled. Agreement ranged from 96.2% (inclusion) and 85.3% (treatment type) down to 39.0% (severity). Severity was the weakest dimension for the model and humans alike (inter-reviewer κ = 0.22).

Multi-label artifact. Exact-set scoring penalizes partial overlaps, so fields that legitimately take multiple labels score lower: occurred workflow (κ≈0.62) > discovered workflow (≈0.40) > discoverer role (≈0.31) > failure mode (≈0.24).

Communication dominates. Communication failures were the most common failure mode (260 occurrences, ~40%), followed by treatment planning (156) and documentation/consent (142). (as shown in Figure 1)

A reporting gap. 54.5% of narratives did not state who detected the event; the model abstains rather than guess—pointing to a concrete reporting-template improvement.

Table 1. Per-dimension validation agreement (Model vs. consensus; Cohen's κ).

Dimension	Within set	Model κ	Inter κ
Treatment type	85.3%	0.46	0.38
Occurred workflow	76.6%	0.62	0.63
Failure mode	79.5%	0.23	0.51
Discoverer role	50.0%	0.31	0.49
Discovered workflow	48.7%	0.40	0.41
Severity	**39.0%**	**0.14**	**0.22**

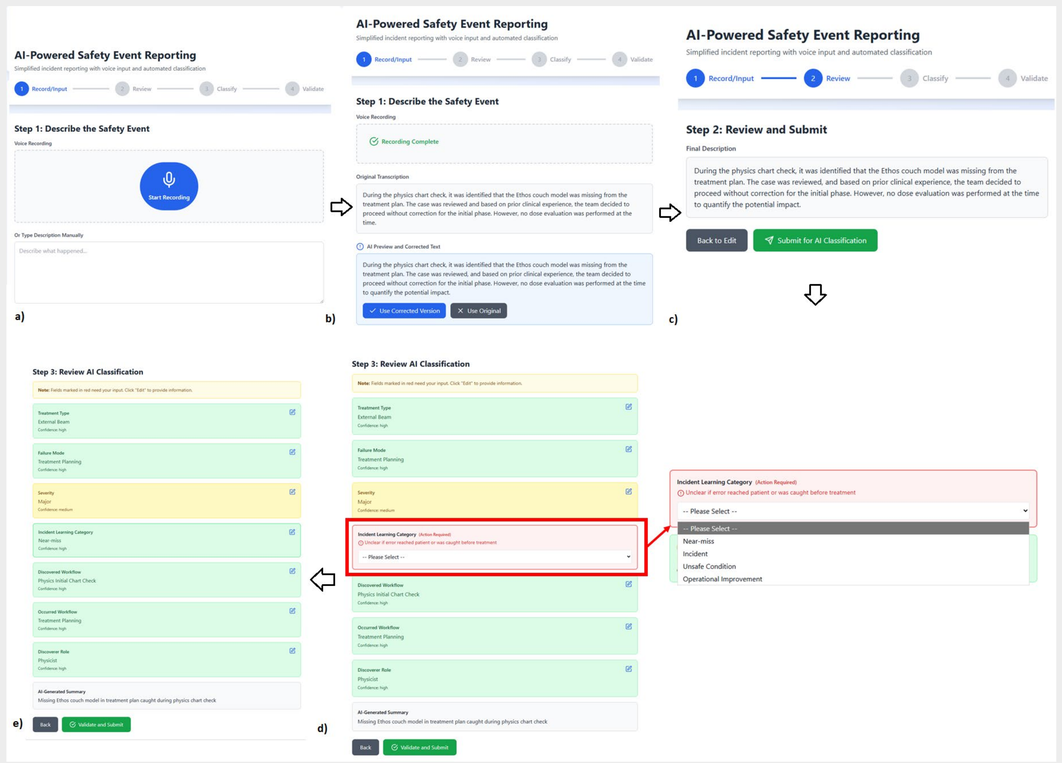

Figure 2. Future AI assisted safety reporting system

DISCUSSION

Bounded by narrative quality. Where cues are clear, model and experts converge; where the narrative underdetermines the answer, both are limited equally. Reliability is bounded chiefly by report quality—not model capability.

Severity is service-quality by design. Our scale weights service-delivery failures (delays, rescheduling) as Major even without physical harm, explaining the high Major rate (43.8%). We propose splitting severity into separate clinical-risk (AAPM TG-100–aligned) and service-quality axes.

Efficient and accessible. All 788 reports were classified in ~7 hours of unattended computation at ~$20—versus an estimated ~40 hours of expert time for a single rater.

Proposed system. As in Figure 2, a human-in-the-loop workflow: the reporter narrates (typed or voice), the LLM classifies, prompts the reporter only where uncertain, and the reporter validates before submission—requiring HIPAA-compliant hosting and automated PHI de-identification.

CONCLUSIONS

We demonstrate a LLM (GPT 5.0) classifies radiation oncology safety events automatically across seven dimensions from narrative alone, reaching agreement comparable to expert inter-reviewer agreement.

REFERENCES

1. Hoopes DJ, et al. RO-ILS: Radiation Oncology Incident Learning System. Pract Radiat Oncol. 2015;5(5):312–318.
2. Eric C Ford, et al. Consensus recommendations for incident learning database structures in radiation oncology. Med Phys., 39(12):7272–7290, 2012. doi: 10.1118/1.4764914.
3. Mullen T, et al. Interrater reliability of a near-miss risk index for incident learning systems. Pract Radiat Oncol. 2016;6(6):429–435.
4. Huq MS, et al. The report of Task Group 100 of the AAPM. Med Phys. 2016;43(7):4209–4262.
5. Frewen H, et al. Failure mode and effects analysis in radiation oncology. J Med Imaging Radiat Oncol. 2018.