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

Large Language Model-Guided Triage of Incident Learning System Forms In Radiation Oncology

Shayan Siddiqui, Stephen Gardner, Chengyin Li, Brett Miller, Natalie Baughan, Michelle Dickinson, Aharon Feldman, Anthony Doemer, Benjamin Movsas, Kundan Thind

Department of Radiation Oncology, Henry Ford Health – Cancer Detroit, Michigan

ABSTRACT

We evaluated agreement between LLMs and expert reviewers in triaging radiation oncology incident learning system (ILS) forms with regards to three clinically relevant dimensions (workflow process step, severity, and dosimetric impact), with the goal of improving efficiency and effectiveness of routine ILS review/follow-up for larger institutions.

Thirty-nine ILS forms from a single institution (2024-2025) were independently reviewed by three expert observers and three open-source LLMs. Each reviewer classified forms by process step, severity, and dosimetric impact. Inter-rater reliability was assessed using Fleiss' Kappa for nominal data and weighted Fleiss' Kappa with quadratic weights for ordinal data. LLM performance was evaluated via Cohen's Kappa against each expert, comparison of mean LLM-expert agreement to mean expert-expert agreement, and accuracy against majority expert consensus.

LLMs demonstrated moderate agreement with the three expert reviewers across all dimensions, with performance within the range of inter-expert variability. These initial findings suggest that LLM-assisted triage is a viable approach for scaling ILS forms review and providing consistent initial screening in radiation oncology quality improvement programs, pending validation on larger datasets.

CONTACT

Shayan Siddiqui
Henry Ford Health
2800 W Grand Blvd, Detroit, MI 48202
ssiddiq8@hfhs.org
248-825-9222

INTRODUCTION

Henry Ford Health's ILS, launched in 2011, provides a platform for staff to submit process improvement (PI) forms. These submitted PI forms include suggestions, near misses, and errors; all forms are reviewed by local team leads for further evaluation as warranted. In total, over 8,000 forms have been logged. This study compared LLM and expert agreement in triaging radiation oncology ILS forms by workflow step, severity, and dosimetric impact, aiming to improve ILS review efficiency at larger institutions.

METHODS AND MATERIALS

Three experts and three open-source LLMs (GPT-oss120b, Llama3.1:70b, Qwen2.5:32b) independently classified 39 ILS forms (2024–2025) by:

1. Process step (9 categories: Simulation, Contouring, Treatment Planning, Patient Setup, Imaging, Treatment Delivery, Physics Check, Physician Review, Therapy Check)
2. Severity (Low/Medium/High),
3. Dosimetric impact (Low/Medium/High).

Inter-rater reliability used Fleiss' Kappa (weighted for ordinal data), while LLM performance was measured via Cohen's Kappa against each expert, LLM-expert vs. expert-expert agreement, and accuracy against majority consensus.

RESULTS

Expert agreement across the three reviewers (Table 1):

- Process step Fleiss' κ = 0.28 (fair)
- Severity (weighted) κ = 0.34 (fair)
- Dosimetric impact (weighted) κ = 0.51 (moderate).

LLM performance was comparable to inter-expert variability (Table 2):

Process step mean LLM-expert κ = 0.40 vs. mean expert-expert κ = 0.30
Severity mean LLM-expert κ = 0.33 vs. mean expert-expert κ = 0.37
Dosimetric impact mean LLM-expert κ = 0.34 vs. mean expert-expert κ = 0.52.

GPT-oss120b was the best performing model, based on weighted kappa for expert majority vote across process step, severity, dosimetric impact (κ=0.403,0.448,0.435) respectively.

DISCUSSION

This study addresses a critical bottleneck in radiation oncology quality improvement programs: the time-intensive manual review and triage of incident learning system (ILS) forms.

We demonstrate that large language models (LLMs) can achieve moderate agreement (κ = 0.41-0.45) with expert medical physicists across three clinically relevant triage dimensions, with performance matching or exceeding inter-expert variability.

This validation enables automated initial screening of ILS forms, allowing expert reviewers to focus on high-risk cases requiring detailed investigation.

The study also reveals that inter-expert agreement on ILS form categorization is only fair to moderate (κ = 0.28-0.51), highlighting need for standardized evaluation criteria. Our findings provide quantitative evidence supporting the feasibility of LLM-assisted triage in radiation oncology quality assurance.

Table 1. Inter-Rater Agreement Among Three Expert Reviewers

Dimension	Fleiss' Kappa	95% CI*	Agreement
Process Step	0.282	0.17-0.40	Fair
Severity (weighted)	0.344	0.19-0.48	Fair
Dosimetric Impact (weighted)	0.514	0.36-0.64	Moderate

*Bootstrap 95% confidence intervals (1000 iterations). Agreement interpretation: < 0.00 = Poor, 0.00-0.20 = Slight, 0.21-0.40 = Fair, 0.41-0.60 = Moderate, 0.61-0.80 = Substantial, 0.81-1.00 = Almost Perfect (Landis & Koch, 1977).

Table 2. LLM Performance Relative to Inter-Expert Variability

Dimension	Mean Expert-Expert κ	Mean LLM-Expert κ	Result
Process Step	0.30	0.40	LLM > Expert
Severity	0.37	0.33	LLM ≈ Expert
Dosimetric Impact	0.52	0.34	Expert > LLM

Mean Expert-Expert κ represents average pairwise agreement across all three expert pairs. Mean LLM-Expert κ represents average agreement between each LLM and each of the three experts.

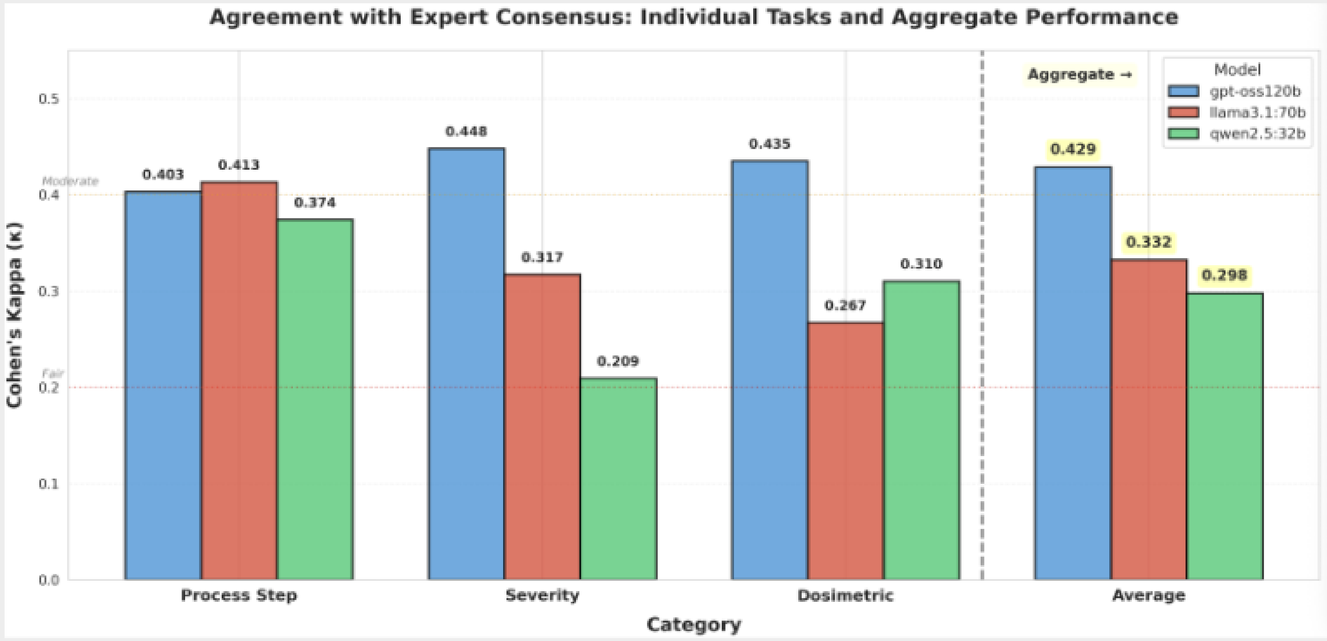

Chart 1. Individual and aggregate performance of LLMs in agreement with expert consensus for ILS form triage.

CONCLUSIONS

LLMs demonstrated moderate agreement with the three expert reviewers across all dimensions, with performance within the range of inter-expert variability.

These initial findings suggest that LLM-assisted triage is a viable approach for scaling ILS forms review and providing consistent initial screening in radiation oncology quality improvement programs, pending validation on larger datasets.

REFERENCES

1. Ford EC, et al. Consensus recommendations for incident learning database structures in radiation oncology. Med Phys. 2012 Dec;39(12):7272-90. doi: 10.1118/1.4764914. PMID: 23231278.
2. Mutic S, et al. Event (error and near-miss) reporting and learning system for process improvement in radiation oncology. Med Phys. 2010 Sep;37(9):5027-36. doi: 10.1118/1.3471377. PMID: 20964222.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com