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

Multimodal Large Language Models for Patient-Specific Radiotherapy Constraint Generation in NSCLC

EMORY WINSHIP CANCER INSTITUTE

Ignacio R. Bartol, PhD [1], Anees H. Dhabaan, PhD[2], Mohammad K. Khan, Ph.D, MD[2*], and Shaheen A. Dewji, PhD[3*],

(1) Nuclear and Radiological Engineering and Medical Physics Programs, Georgia Institute of Technology, Atlanta, GA,

(2) Department of Radiation Oncology and Winship Cancer Institute, Emory University, Atlanta, GA,

Georgia Tech College of Engineering
Nuclear and Radiological Engineering and Medical Physics

Purpose: This work assesses a multimodal Large Language Model (MedGemma) for generating planning objectives in standard fractionation (60Gy/30fx or SBRT indications) lung cancer radiotherapy. Unlike static vendors' templates, the proposed system dynamically provides objectives based on patients Electronic Health Records (HER) (i.e. comorbidities, previous radiation) and applying physical constraints. The model fuses EHR and geometric organ structure distance-based features to inform objective generation.

Methods: We implemented a Retrieval-Augmented Generation (RAG) pipeline. Inputs included: (1) Anonymized EHR text. (2) Precomputed minimum boundary-to-boundary and center-to-center distances for all segmented organs in CT/RTSTRUCT. (3) A vector database of clinical trials (e.g., RTOG 0617). The model generated planning objectives (soft constraints) for four endpoints: Lung V20Gy, Mean Lung Dose, Spinal Cord Max, and Esophagus Mean. Validation used a cohort of 50 patients: 10 real cases with objectives extracted from delivered clinical plans (Ground Truth), and 40 synthetic cases generated by geometric variation of trial protocols. We performed a pooled analysis comparing model outputs to physician ground truth using Spearman's rank correlation and endpoint-specific Mean Absolute Error (MAE). We applied structure-specific Linear-Quadratic (LQ) modeling for EQD2.

Results: In the ground truth comparison (n=40 endpoints), RAG+MedGemma achieved a pooled Spearman correlation of $\rho = 0.98$ ($p < 0.001$). Endpoint-specific MAEs were Lung V20Gy: 1.6 % (pp); Mean Lung: 0.9 Gy; Cord Max: 1.2 Gy; Esophagus Mean: 1.8 Gy. For patients with re-irradiation or comorbidities (40% of cohort), the model recommended objectives tightened by a mean of 22% relative to standard guidelines. The safety audit revealed 0% numeric hallucination (non-existent guidelines), with 98% of objectives rated as logically accurate based on the EQD2 summation.

Conclusion: Although tested on a limited cohort, MedGemma demonstrated high agreement with ground truth on adapting planning objectives based on multimodal data. Integration of explicit spatial attributes and LQ-model logic anchored the model, mitigating hallucinations. Together, these advances yield potential for a risk-adaptive decision-support framework.

CONTACT

Shaheen Azim Dewji
Georgia Institute of Technology
Tel: 404-894-5800
E-mail: shaheen.dewji@gatech.edu

Purpose

To evaluate the **feasibility** of a **physics-aware multimodal RAG** workflow for generating structured, **TPS-export-ready planning** constraints for NSCLC radiotherapy using de-identified EHR, 4DCT-derived anatomy, geometric target–OAR features, retrieved clinical evidence, and deterministic safety guardrails.

Outputs are structured for clinician/physicist review and RTPLAN integration. **The system is not evaluated as autonomous clinical decision-making software.**

Study Design & Cohort

Real retrospective cases reported separately from protocol-derived synthetic stress scenarios.

10 real de-identified NSCLC cases

40 protocol-derived synthetic stress scenarios

40 real-case endpoint comparisons (10×4)

160 synthetic objective-level safety outputs

- **Reference values:** achieved DVH metrics from delivered clinical plans (not physician-intent ground truth).
- Motion management: **4DCT** · Techniques: **IMRT, SBRT**
- Real cohort includes **4 re-irradiation cases**.
- **Synthetic scenarios were <u>not</u> counted as real patients.**

Endpoints analyzed

*Four displayed endpoints, **analyzed separately each point including synthetic data.***

Endpoint	Metric	Unit	Tol.
Lung V20Gy	V20Gy	pp	2.0
Mean lung dose	Mean dose	Gy	1.5
Cord D0.03cc	D0.03cc (Dmax/D0.1cc tracked)	Gy	2.0
Mean esophageal dose	Mean dose	Gy	2.5

Reference: delivered-plan achieved DVH metric. Lung V20Gy in percentage points (pp); dose endpoints in Gy. Structure conventions locked before submission.

Ablation & Baselines

Performance and safety behavior were compared across institutional template, deterministic guideline lookup, geometry-only, EHR-only, RAG without LLM generation, LLM without deterministic rule engine, and full physics-aware RAG workflow.

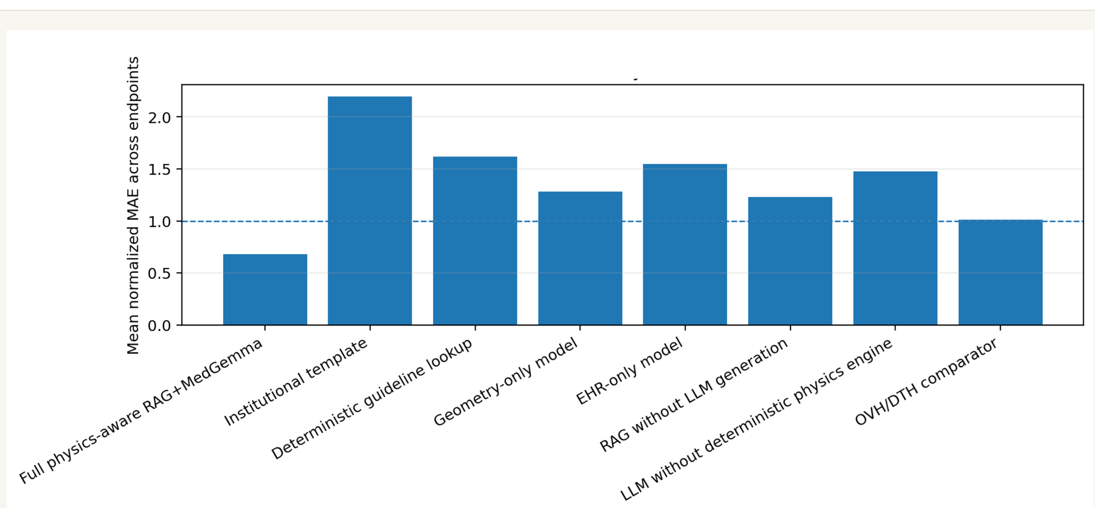

Fig 6. Full workflow (geometry + EHR + RAG + LLM + deterministic checks) vs. reduced variants; deterministic physics/rule validation contributes the safety-critical gains.

Workflow Architecture

4D CT + RTSTRUCT + De-ID + QC → Geometry features + OVH/DTH + distances → EHR + prior RT + structured extraction. → RAG evidence: top-7 best score → Deterministic rules: Units + EQD2 + hard constraints → MedGemma drafting + JSON constraints. → Schema validation - clinician reviews. → TPS RTPLAN export.

Fig 6. Deterministic physics/safety checks are separated from the multimodal LLM.

LLM ROLE
Evidence summarization & structured drafting. Model: MedGemma 3 4B + LoRA, top-7 retrieval

DETERMINISTIC ROLE
Unit validation, EQD2/BED checks, JSON-schema validation, dose constraints check.

Inputs: de-identified EHR · 4DCT images · RTSTRUCT contours · prior-RT data · geometric target–OAR features. **Outputs:** RTPLAN patient-specific dose constraints

Structured output

TPS-Ready JSON file with planning objectives with traceable evidence, assumptions, and warnings.


```
{
  "task": "Define planning objectives for 60Gy/30fx NSCLC plan",
  "assumptions": [
    "Patient has Stage IIIA NSCLC",
    "Patient has COPD (GOLD 2)",
     ...
     "Heart is posterior",
    "Spinal cord is posterior",
  ],
  "constraints": [
    {
      "structure": "Lungs (Total)",
      "objective": "V20",
      "limit": {        "value": 30,        "unit": "%"
      },
      "basis": "QUANTEC guidance",
      "rationale": "Risk-informed target for pneumonitis prevention.",
      "citations": [ "DOC002” ]
      ]
    },
    ...
  ],
  "warnings": [
    "Consider cumulative EQD2 for spinal cord if re-irradiation is possible.",
    "Consider patient-specific factors (e.g., ILD) when evaluating constraints."
  ],
  "unknowns": [
    "None"
  ]
}
```


Re-Irradiation Guardrail

If prior RT is available only as prescription dose, the system **did not** treat it as **actual OAR** dose. Prior DICOM dose, contour verification, registration QA, and spatial accumulation were required for cumulative OAR assessment.

Table 1. For re-irradiation cases, remaining EQD was verified using deterministic algorithms and it influenced the reported dose constraints by the LLM adding an additional behavior.

Scenario	Prior EQD2 (Gy_2, $\alpha/\beta=2$)	Behavior
45 Gy/15 fx, **prescription-only**, no cord dose	56.25	**ABSTAIN**
45 Gy/15 fx, **actual** cord D0.03cc	56.25	**ESCALATE**
10 Gy/5 fx, actual cord dose	10.0	**CAP + WARN**
Prior RT documented, **fractions missing**	n/a	**ABSTAIN**

Example LLM: "45 Gy/15 fx ⇒ EQD2 = 56.25 Gy_2 > hypothetical 50 Gy_2 scalar limit. Scalar EQD2 alone is **not sufficient** for clinical re-irradiation; spatial dose accumulation & registration QA required. **7/7 guardrail tests passed.**"

Result: Endpoint Performance

Mean absolute error was evaluated separately for each endpoint using real-patient cases only. Lung V20Gy is reported in percentage points; dose endpoints are reported in Gy. Error bars indicate patient-level clustered bootstrap 95% confidence intervals.

Endpoint	MAE (95% CI)	Bias	In tol.
Lung V20Gy (pp)	1.60 (0.80–2.54)	+1.23	70%
Mean lung dose (Gy)	0.90 (0.70–1.12)	+0.21	90%
Cord D0.03cc (Gy)	1.20 (0.94–1.45)	+0.25	**100%**
Mean esoph. dose (Gy)	1.80 (1.12–2.53)	−0.17	80%

Bland–Altman limits of agreement computed separately per endpoint; units never pooled. Reference = achieved DVH from delivered plans.

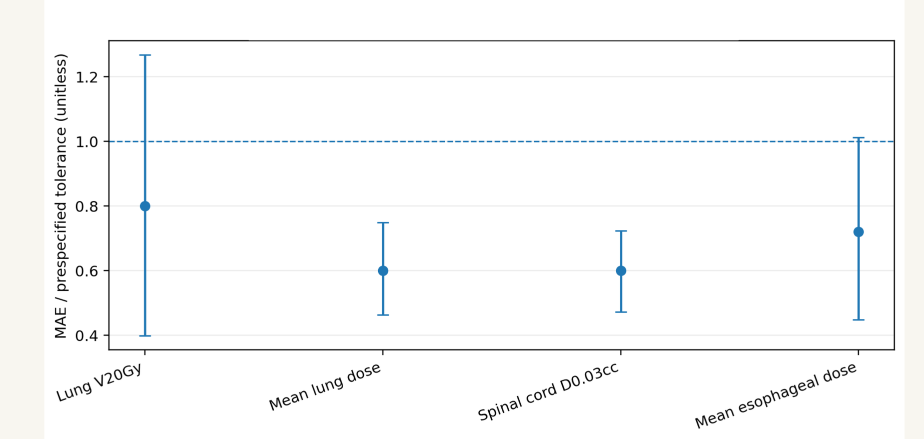

Fig 4. Endpoint-specific scatter plots compare generated constraints with reference achieved DVH metrics from delivered clinical plans.

Safety Challenge Audit

160 synthetic objective-level stress outputs across 12 challenge categories.

98.1% correct or abstained (157/160)

0 unsafe numeric (0/160)

κ=.86 inter-rater (placeholder)

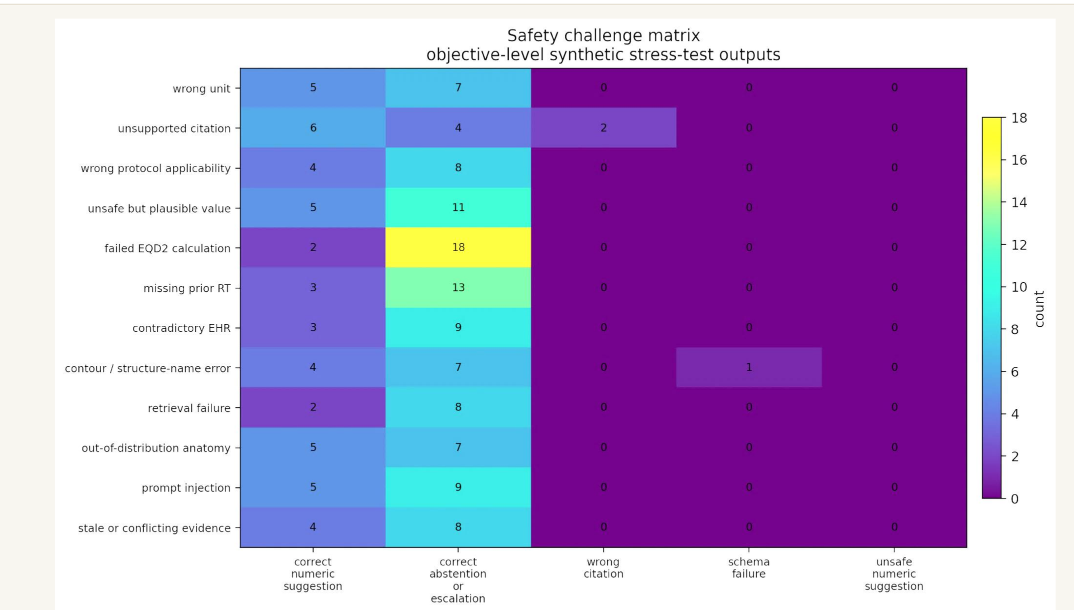

Fig 5. Categories: wrong units, unsupported citations, protocol applicability, unsafe-but-plausible values, failed EQD2, missing prior RT, contradictory EHR, contour errors, retrieval failure, OOD anatomy, prompt-injection, failure-to-abstain. 95% CI correct 94.6–99.6%; unsafe 0.0–2.3%.

Conclusion & Limitations

LLM's with RAG can support a **traceable, geometry-informed, evidence-linked, and physics-checked** constraint-generation architecture; with auditable abstention when prior-dose, anatomy, unit, or evidence conditions are insufficient.

Limitations:

- Single-institution feasibility cohort; assumed structure conventions must be locked. Clinical templates were limited to this institution.
- Prior EQD2 is very sensitive to the $^{\alpha}/_{\beta}$ values which introduced most error.
- Errors in DIR introduced additional error for re-irradiation calculations.