Development and Validation of a GPT-5.2-Based Multi-Agent Radiotherapy Plan Evaluation System for Stereotactic Body Radiotherapy Plans

Caifei Lu[1], Zi Liu[1,2], Xin Yang[1,3] ,Xiaoyan Huang[1,3] , Sijuan Huang[1,3]
[1]State Key Laboratory of Oncology in South China; Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Treatment in Guangdong Province; Guangdong Provincial Clinical Research Center for Malignant Tumors;Sun Yat-sen University Cancer Center， [2]Guangzhou Xinhua University, Guangzhou， [3]Joint Laboratory of Advanced Radiotherapy Technology, Sun Yat-sen University-Guoke Ion (Lanzhou Taiji)

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

ABSTRACT

Purpose: To explore the feasibility of applying large language models (LLMs) for radiotherapy plan (RTP) evaluation to assist radiotherapy plan quality control and clinical decision making.

Methods: This study developed a multi-agent RTP evaluation system (RTPEV) based on GPT-5.2, supporting multimodal inputs including clinical prescriptions, plan PDFs, and plan-related DICOM data. The system incorporates a Retrieval-Augmented Generation (RAG) retriever agent and a dedicated medical knowledge base tailored for RTP, combining clinical guidelines and literature to support the evaluation process. A reviewer agent generates an initial evaluation report, a critic agent performs secondary verification and refinement, and an arbitrator agent outputs the final evaluation reports. Twenty stereotactic body radiation therapy (SBRT) plans from Sun Yat-sen University Cancer Center were retrospectively evaluated. System performance was evaluated using the number of issues detected per case, the agreement of system outputs and physicist review, and quantitative metrics including retrieval evidence relevance, hallucination rate, and review efficiency.

Results: For the 20 plans, the RTPEV system flagged an average of 6.1 issues per case requiring physicist confirmation. After manual review, 91.1% of the AI-flagged uncertain items were considered clinically acceptable. The system identified noncompliant CT slice thickness, mismatched OAR dose constraints, and further stratified identified issues by risk level with corresponding optimization recommendations. Quantitative evaluation showed a hallucination rate of 0.106 and a retrieval evidence relevance score of 0.59, suggesting moderate evidence grounding and stable overall system performance.

Conclusion: The RTPEV multi-agent system demonstrates technical feasibility and clinical auxiliary value in SBRT RTP evaluation, which can stably identify potential problems with a low hallucination rate and support quality control and clinical decision-making for radiotherapy plans.

CONTACT

Sijuan Huang
State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Guangdong Esophageal Cancer Institute, Sun Yat-sen University Cancer Center,Guangzhou, Guangdong 510060, China
huangsijuan@sysucc.org.cn
+86-020287343089

INTRODUCTION

Stereotactic body radiotherapy (SBRT) achieves high local control and low toxicity through large dose-per-fraction, small target volume, and steep dose gradients, but also raises the threshold for plan quality control (QC) to an unprecedented level.

Manual QC of SBRT plans is labor-intensive and error-prone, requiring strict adherence to American Association of Physicists in Medicine (AAPM) TG-101 guidelines and organ-at-risk (OAR) dose limits. Large language models (LLMs) and Retrieval-Augmented Generation (RAG) show promise in clinical decision support but have rarely been applied to radiotherapy plan QC.

We developed the Radiotherapy Plan Evaluation System (RTPEV), a multi-agent system based on GPT-5.2 that can receive clinical prescriptions, plan PDFs, and Digital Imaging and Communications in Medicine (DICOM) data, and output structured, evidence-based evaluation reports to assist physicists in conducting comprehensive and standardized plan reviews.

METHODS AND MATERIALS

This study developed a multi-agent RTPEV based on GPT-5.2, supporting multimodal inputs including clinical prescriptions, plan PDFs, and plan-related DICOM data. The system incorporates a RAG retriever agent and a dedicated medical knowledge base tailored for plan evaluation, combining clinical guidelines and literature to support the evaluation process. A reviewer agent generates an initial evaluation report, a critic agent performs secondary verification and refinement, and an arbitrator agent outputs the final evaluation reports. Twenty SBRT plans from Sun Yat-sen University Cancer Center were retrospectively evaluated. Performance was assessed by the number of issues detected per case, agreement between system outputs and physicist review, and quantitative metrics including retrieval evidence relevance, hallucination rate, and review efficiency.

RESULTS

RTPEV completed the review process for all 20 SBRT plans. The system generated structured reports that summarized plan information, listed potential issues, provided supporting evidence, and gave risk-based suggestions.

On average, RTPEV flagged 6.1 issues per case requiring physicist confirmation across 20 SBRT plans. After manual review, 91.1% of the system-flagged uncertain items were judged clinically acceptable by experienced physicists.

The system identified noncompliant computed tomography (CT) slice thickness (3.0 mm vs. TG-101 preferred ≤ 2 mm), mismatched OAR dose constraints, and low monitor unit (MU)/cGy ratios, among other issues. Each flagged item was classified into a risk category and accompanied by a corresponding optimization recommendation.

The quantitative evaluation showed a hallucination rate of 0.106 and a retrieval evidence relevance score of 0.59, suggesting moderate evidence grounding and stable overall performance. After reviewing the system-generated reports, physicists also identified additional potential plan risks, which supports the complementary value of the system in plan review.

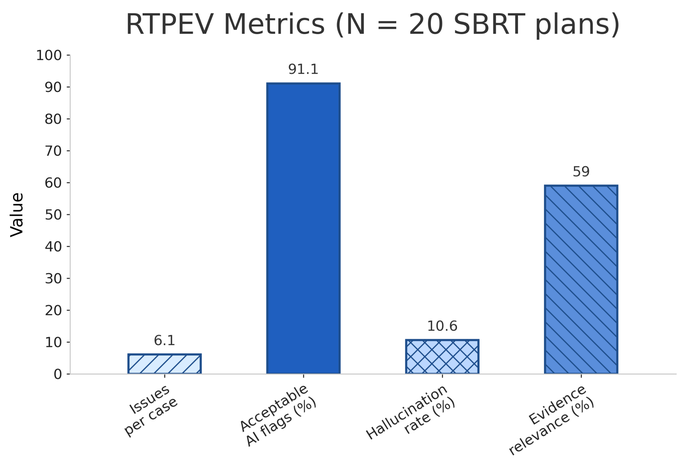

Figure 2. Key quantitative performance metrics across 20 SBRT plans.

DISCUSSION

RTPEV flagged an average of 6.1 items per case. After manual review, 91.1% of flagged uncertain items were clinically acceptable, suggesting a cautious screening strategy. Such sensitivity may aid early QC, but specificity still needs improvement.

The system identified plan-related concerns, including noncompliant CT slice thickness and mismatched OAR dose constraints. It grouped findings by risk level and provided suggestions for review or optimization. These features may help standardize plan evaluation. Physicists found additional risks after reading system reports, suggesting RTPEV may serve as a useful second reader. It does not replace clinical judgment but may help reviewers recheck key details and improve completeness.

However, RTPEV is not ready for independent clinical use. The hallucination rate was 0.106 and the evidence relevance score was 0.59, indicating moderate evidence grounding. Future work should improve the knowledge base, retrieval strategy, and DICOM parsing.

This retrospective study included only 20 SBRT plans from a single institution. Results show feasibility but not broad clinical performance. Future studies should include larger datasets, more disease sites, and prospective workflow testing.

CONCLUSIONS

The RTPEV system demonstrates technical feasibility and clinical auxiliary value for SBRT plan evaluation. It stably identifies potential issues at a low hallucination rate and supports both QC and clinical decision-making for radiotherapy plans. By integrating RAG with structured multi-agent review, RTPEV offers a reproducible framework for automated plan evaluation. Future work will focus on prospective multi-center validation and expansion to additional treatment modalities.

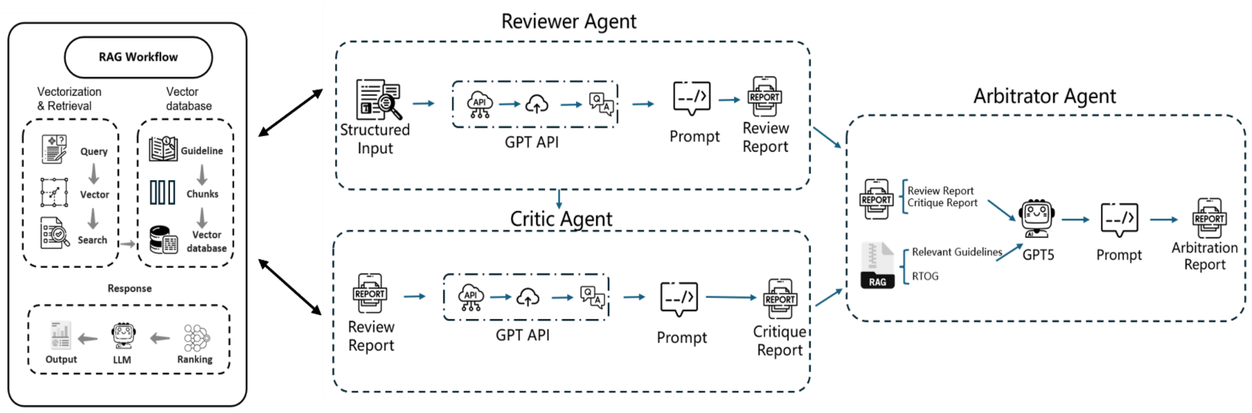

Figure 1. system flowchart.

Table 1. Checklist excerpt from a representative SBRT case evaluation.

Module	Check Item	Reference Standard	Conclusion
Imaging	CT Slice Thickness	TG-101 ≤ 2 mm pref.	TBD
Imaging	Scan Range	≥ 50 mm beyond PTV	Pass
Structure	Rx Match	PTV30Gy/5F+ PTV25Gy/5F	Pass
Dose	Rx Dose	3000/2500 cGy, 5F	Pass
Dose	MU/cGy	SBRT typically > 2	TBD
Plan	Collimator Angle	Avoid sensitive align.	TBD

REFERENCES

[1] Benedict SH, et al. Stereotactic body radiation therapy: the report of AAPM Task Group 101. Med Phys. 2010;37(8):4078-4101.

[2] Wang D, Hu Z, Li Y. MARTP: a multi-agent simulation framework for automated radiation therapy planning based on LLMs. Phys Med Biol. 2026. doi:10.1088/1361-6560/ae6af6

[3] Fu J, Cheng Y, Li Z. Intelligent Support for Radiotherapy: A Review of Clinical Applications for Large Language Models. J Clin Med. 2026;15(7):2531. doi:10.3390/jcm15072531

[4] Öğdü ÇU, Arslanoglu K, Karakose M. An Adaptive Multi-Agent LLM-Based CDSS Integrating Biomedical RAG and Web Intelligence. IEEE Access. 2025. doi:10.1109/ACCESS.2025.3613340