Multimodal LLM-Based Quality Assurance for Auto-Contouring: Utility As a Primary Screening Tool

Ryota Tozuka[1], Tomoko Akita[1], Masaki Matsuda[1], Hikaru Tanno[2], Masahide Saito[1], Hikaru Nemoto[1], Noriyuki Kadoya[2], Keiichi Jingu[2] and Hiroshi Onishi[1]

1. Department of Therapeutic Radiology, University of Yamanashi, Chuo, Yamanashi, Japan
2. Department of Radiation Oncology, Tohoku University Graduate School of Medicine, Sendai, Miyagi, Japan

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

ABSTRACT

Purpose: Manual verification of AI-based auto-contouring is labor-intensive and prone to fatigue-related errors. This study evaluated the feasibility of automated visual quality assessment using a multimodal large language model (LLM), Gemini 2.5 Flash, to assess its utility as a clinical primary screening tool to streamline the quality assurance (QA) workflow.

Methods: Twenty male pelvic CT scans from an open dataset were utilized. Two distinct auto-contouring software packages (RatoGuide prototype and clinical syngo.via) were evaluated. auto-contouring results for each slice were exported as PDF images with overlaid contours and input into Gemini 2.5 Flash. The LLM was instructed to rate the contour quality on a 5-point clinical scale (5: Accept; 4: Minor revision; 3: Major revision; 2: Redraw from scratch; 1: Organ not detected). Using evaluations by two board-certified radiation oncologists as ground truth, we calculated Spearman's rank correlation coefficients (ρ) and weighted kappa coefficients (κ). Additionally, scores ≤2 were defined as "Fail" and >2 as "Pass" to calculate sensitivity and specificity for each model to assess screening performance.

Results: Visual assessment by Gemini showed substantial agreement with expert judgment. For RatoGuide and syngo.via, the ρ values were 0.744 and 0.765, respectively, and the κ values were 0.669 and 0.746, respectively, indicating substantial agreement for both software. In terms of screening performance (defining "Fail" as the positive class), RatoGuide and syngo.via achieved sensitivities of 84.2% and 80.0%, and specificities of 93.6% and 95.4%, respectively. Although the LLM showed a tendency to overestimate quality, it exhibited high specificity in identifying clinically acceptable cases.

Conclusion: Gemini 2.5 Flash demonstrated substantial agreement with expert evaluations in image-based auto-contouring quality assessment. While potential overestimation bias (risk of missing "Fail" cases) warrants caution, the observed sensitivity suggests its feasibility as a primary screening QA tool to efficiently filter acceptable contours, thereby reducing the clinical workload.

CONTACT

Ryota Tozuka
Department of Therapeutic Radiology,
University of Yamanashi
1110, Shimokato, Chuo, Yamanashi,
400-0053, Japan
r.tozuka991028@gmail.com

INTRODUCTION

AI-based auto-contouring (AC) improves radiotherapy planning but can generate inaccurate contours, compromising patient safety.
Current quality assurance (QA) relies on time-consuming manual inspection, risking automation bias. Existing automated QA methods using geometric metrics or simple classifiers lack clinical correlation and fail to provide detailed, context-specific explanations for corrections[1].
To overcome these limitations, we developed LAQUA (LLM-based Automated Quality Assurance for Auto-contouring). By leveraging the advanced visual perception and linguistic capabilities of Large Language Models, LAQUA performs fully automated QA, pinpointing specific AC failures and describing the exact nature of required modifications in natural language.

METHODS AND MATERIALS

Twenty male pelvic cases were selected from a publicly available dataset, targeting the bladder, prostate, rectum, and bilateral femoral heads[2]. Auto-contours (AC) were generated using three commercial software platforms under default settings. The AC results were overlaid on CT images and converted into patient-specific PDFs, preserving anatomical relationships and the full field of view. Gemini 2.5 Pro (temperature: 0.1) evaluated these PDFs to output 5-point quality scores and natural language rationales. The prompt used for evaluation is shown below.

[Prompt]
You are an expert in radiation therapy. I will provide PDFs of the contouring for a prostate cancer patient.
Evaluate the clinical usability of each Structure image (bladder, prostate, rectum, and bilateral femoral heads) based on the following 5-point scale.

[Scale]
5: Optimal; clinically usable without modification.
4: Acceptable; clinically usable despite minor deviations.
3: Suboptimal; requires revisions for clinical use.
2: Unacceptable; requires complete redrawing.
1: Unacceptable; organ not recognized or completely wrong.

[Output Format]
Respond ONLY in JSON list format.
[
{"roi_name": "Structure Name", "clinical_score": 1-5, "reason": "Specific reason"}
]

The evaluation comprised two stages. First, the concordance between LLM-derived scores and expert consensus for screening performance was assessed using Spearman's correlation coefficient (ρ) and weighted kappa (κ). Second, radiation oncologists visually evaluated the LLM-generated rationales based on four metrics: Accuracy of Error Detection, Absence of Hallucination, Clinical Relevance, and Anatomical Understanding.

Figure 1. Workflow of the proposed system

Score	Criteria
5	**Optimal;** clinically usable without modification
4	**Acceptable;** clinically usable despite minor deviations
3	**Suboptimal;** requires revisions for clinical use
2	**Unacceptable;** requires complete redrawing.
1	**Unacceptable;** organ not recognized or completely wrong

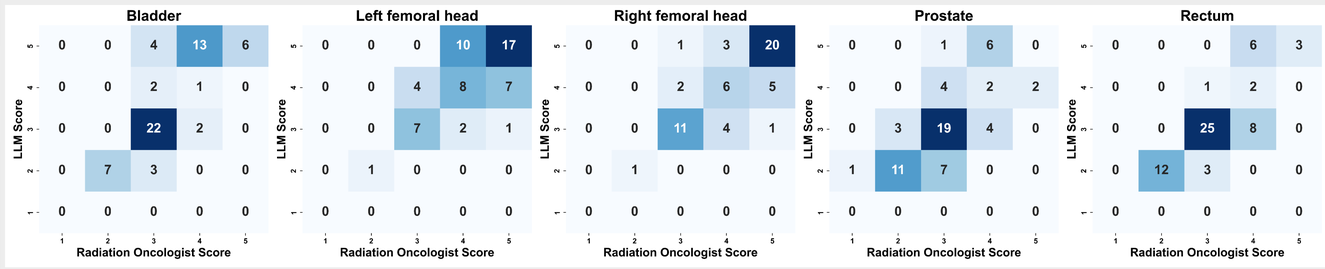
Table 1. Qualitative evaluation criteria for auto-contours

RESULTS

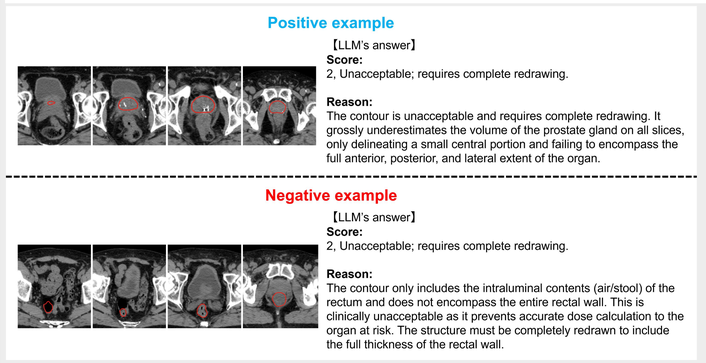
The LLM evaluations demonstrated strong agreement with expert judgments, as organ-specific analyses revealed moderate or higher correlations across all evaluated structures. The highest agreement was observed in the rectum (Spearman's rank correlation = 0.835, quadratic weighted kappa = 0.804), while the left femoral head showed the lowest correlation (Spearman's = 0.567, kappa = 0.639).

Regarding clinical screening performance, the LLM effectively detected inadequate auto-contours. When defining an "adequate" contour as a score of ≥3 (meaning complete redrawing is unnecessary), the bilateral femoral heads achieved a perfect sensitivity and specificity of 1.000. Using a stricter "adequate" threshold of ≥4 (usable but requiring modification), the rectum demonstrated the highest sensitivity (0.976), and the left femoral head exhibited the highest specificity (0.933).

In the qualitative evaluation of the LLM's natural language rationales, the overall mean score was 1.70 ± 0.48. The LLM accurately pinpointed error locations and reasons, with 155 of 291 evaluated outputs achieving the maximum score of 2 across all predefined criteria.



↑Figure 2. Confusion matrices comparing the LLM scores (y-axis) and the radiation oncologist scores (x-axis) for the auto-contouring results



↑Figure 3. Specific examples of organ evaluation outputs generated by the LLM.

	Spearman's ρ	Quadratic κ
Bladder	0.814	0.711
Left femoral head	0.567	0.639
Right femoral head	0.730	0.744
Prostate	0.712	0.667
Rectum	0.835	0.804

↑Table 2. Correlation between large language model evaluations and radiation oncologist judgments by organ

	Accuracy of Error Detection	Absence of Hallucination	Clinical Relevance	Anatomical Understanding
Bladder	1.52 ± 0.54	1.70 ± 0.50	1.73 ± 0.48	1.77 ± 0.43
Left femoral head	1.65 ± 0.48	1.67 ± 0.48	1.72 ± 0.45	1.86 ± 0.35
Right femoral head	1.63 ± 0.49	1.65 ± 0.52	1.74 ± 0.44	1.87 ± 0.34
Prostate	1.50 ± 0.50	1.78 ± 0.42	1.80 ± 0.40	1.83 ± 0.38
Rectum	1.42 ± 0.62	1.70 ± 0.50	1.60 ± 0.59	1.82 ± 0.43

↑Table 3. Mean visual evaluation scores of large language model performance by organ (mean ± SD)

DISCUSSION

This study demonstrates the novel application of "LLM-as-a-Judge" in medical imaging through the LAQUA system, automating radiotherapy auto-contouring (AC) quality assurance. Unlike conventional methods relying on geometric metrics, LAQUA provides detailed, natural language feedback specifying exact contour error locations. Our findings revealed strong correlations between LLM evaluations and expert judgments, alongside robust clinical screening performance for detecting inadequate contours. Qualitatively, the LLM achieved a high mean score for accurately describing necessary modifications, effectively mitigating automation bias within a human-in-the-loop framework[3]. However, LAQUA occasionally exhibited hallucinations due to a lack of specialized clinical knowledge. Integrating Retrieval-Augmented Generation (RAG) with contouring guidelines could resolve this. Limitations include reliance on a small dataset and 2D PDF conversions. Despite these challenges, LAQUA shows immense potential as a primary screening tool.

CONCLUSIONS

In conclusion, this study developed and evaluated the LAQUA system, a fully automated QA system for AC powered by an LLM. Overall, the system's evaluations demonstrated a strong correlation with expert assessments, suggesting its potential utility as an effective primary screening tool.

REFERENCES

1. Sheng J, et al. Beyond Pixel Agreement: Large Language Models as Clinical Guardrails for Reliable Medical Image Segmentation. arXiv preprint 2025.
2. Kanwar A, et al. Stress-testing pelvic autosegmentation algorithms using anatomical edge cases. Phys Imaging Radiat Oncol. 2023.
3. Gu J, et al. A survey on LLM-as-a-Judge. The Innovation. 2026.

CONFLICT OF INTEREST

Noriyuki Kadoya and Ryota Tozuka owned stock in AiRato Inc., while Masahide Saito, Hiroshi Onishi, and Keiichi Jingu received grants from AiRato Inc.