Radiation Oncology

RESEARCH Open Access

Feasibility study of automatic radiotherapy treatment planning for cervical cancer using a large language model

Shuoyang Wei[1], Ankang Hu[2,3], Yongguang Liang[1], Jingru Yang[1], Lang Yu[1], Wenbo Li[1], Bo Yang[1*] and Jie Qiu[1]

Abstract

Background Radiotherapy treatment planning traditionally involves complex and time-consuming processes, often relying on trial-and-error methods. The emergence of artificial intelligence, particularly Large Language Models (LLMs), surpassing human capabilities and existing algorithms in various domains, presents an opportunity to automate and enhance this optimization process.

Purpose This study seeks to evaluate the capacity of LLMs to generate radiotherapy treatment plans comparable to those crafted by human medical physicists, focusing on target volume conformity and organs-at-risk (OARs) dose sparing. The goal is to automate the optimization process of radiotherapy treatment plans through the utilization of LLMs.

Methods Multiple LLMs were employed to adjust optimization parameters for radiotherapy treatment plans, using a dataset comprising 35 cervical cancer patients treated with volumetric modulated arc therapy (VMAT). Customized prompts were applied to 5 patients to tailor the LLMs, which were subsequently tested on 30 patients. Evaluation metrics included target volume conformity, dose homogeneity, monitor units (MU) value, and OARs dose sparing, comparing plans generated by various LLMs to manual plans.

Results With the exception of Gemini-1.5-flash, which faced challenges due to hallucinations, Qwen-2.5-max and Llama-3.2 produced acceptable VMAT plans in 16.3 ± 5.0 and 9.8 ± 2.1 min, respectively, outperforming an experienced human physicist's time cost of about 20 min. The average conformity index (CI) for Qwen-2.5-max plans, Llama-3.2 plans, and manual plans on the test set were 0.929 ± 0.007, 0.928 ± 0.007, and 0.926 ± 0.007, respectively. The average homogeneity index (HI) was 0.058 ± 0.006, 0.059 ± 0.005, and 0.065 ± 0.006, respectively. While there was a significant difference in target volume conformity between LLM plans and manual plans, OARs dose sparing showed no significant variations. In lateral comparisons among different LLMs, no statistically significant differences were observed in the PTV dose, OARs dose sparing, and target volume conformity between Qwen-2.5-max and Llama-3.2 plans.

Conclusions Through an assessment of LLM-generated plans and clinical plans in terms of target volume conformity and OARs dose sparing, this study provides preliminary evidence supporting the viability of LLMs for optimizing

*Correspondence:
Bo Yang
yb1632@163.com

Full list of author information is available at the end of the article

radiotherapy treatment plans. The implementation of LLMs demonstrates the potential for enhancing clinical workflows and reducing the workload associated with treatment planning.

Keywords Radiotherapy, Treatment planning, Large language model, Automation

Introduction

Radiotherapy treatment planning stands as a pivotal element within the realm of radiotherapy procedures. A well-crafted radiotherapy treatment plan is paramount in ensuring that the planning target volume (PTV) receives an optimal radiation dose while simultaneously minimizing exposure to surrounding organs at risk (OARs). Crafting such a plan poses an intricate optimization conundrum with multiple conflicting objectives [1–6]. Traditionally, medical physicists rely on a labor-intensive trial-and-error methodology to iteratively fine-tune optimization parameters, such as the dose goals for different OARs, to achieve an acceptable treatment plan. This conventional process is not only time-consuming but also computationally intensive, leading to inefficiencies in the treatment planning workflow [7–10].

To enhance the efficiency of treatment planning, researchers have proposed diverse methodologies, including automated rule implementation and reasoning (ARIR), knowledge-based planning (KBP), and multicriteria optimization (MCO). ARIR employs pre-set rules within the treatment planning system (TPS) to automate beam setup and optimize treatment plans based on dose-volume histograms (DVH) [11, 12]. KBP involves creating a repository of plans serving as templates for various cancer types, allowing for the identification of the most analogous plan for a new case, thereby reducing the requisite number of iterations [13–15]. In the MCO approach, physicists simultaneously generate multiple anchor plans, each optimizing a single DVH criterion of an OAR to achieve optimal sparing without compromising PTV dosimetric criteria, forming a Pareto surface with a spectrum of optimal plans across different dosimetric criteria. This approach allows physicists to navigate through various Pareto-optimal plans to select the most suitable one based on their preferences [16–18]. MCO has been successfully implemented in TPS such as RayStation (RaySearch, Stockholm, Sweden) and Eclipse (Varian, Palo Alto, USA).

With the advent of artificial intelligence (AI), novel avenues for efficient treatment planning have emerged. One notable approach involves leveraging AI to predict fluence maps, convertible into multi-leaf collimator (MLC) sequences for intensity modulated radiotherapy (IMRT) plans using tools like the Eclipse scripting API (ESAPI) [19, 20]. For VMAT plans, researchers predict dose distributions from CT images and input them into the TPS for further optimization [21, 22]. McIntosh et al. proposed a voxel-based dose mimicking algorithm to convert the predicted dose distribution to a complete treatment plan, achieving a fully automated treatment planning [23–26].

Seeking further automation, researchers have explored reinforcement learning techniques [27, 28]. Some studies have utilized neural networks to generate and adjust optimization parameters iteratively, enhancing plan optimization efficiency [29]. Reinforcement learning networks have been integrated into TPS, allowing for the generation of machine-executable plans directly [30, 31]. However, training neural networks for reinforcement learning demands substantial data and computational resources, additionally, a separate network must be trained for each specific cancer type, presenting challenges for clinical application [32].

Large language models (LLMs) have found broad applications across diverse domains, including radiotherapy. While LLMs have facilitated knowledge dissemination to patients and aided in medical record documentation, their potential in radiotherapy physics remains underexplored [33–39]. Some researchers have studied the role of LLM in dose prediction. Dong et al. proposed DoseGNN for dose prediction, in which LLM was used to process prescription information to enhance network performance [40]. During radiotherapy treatment planning, physicists evaluate dosimetric parameters provided by the TPS to adjust optimization parameters. Current LLMs possess analysis and reasoning capabilities and are trained by a large number of standard corpora, allowing them to be customized to master some skills with minimal data. Therefore, LLMs can be attempted to perform iterative optimization of radiotherapy plans to improve the automation of the optimization process with a small amount of training data. Liu et al. attempted to use LLM to optimize treatment plans for prostate cancer, showing the potential of LLM [32].

In this study, we investigated the feasibility of leveraging LLMs for automating radiotherapy treatment planning, focusing on cervical cancer plans. By utilizing multi-modality LLMs to evaluate and adjust plans iteratively, we aim to assess the efficacy of LLMs in optimizing treatment plans compared to manual optimization by experienced physicists.

Materials and methods

Research framework

The research framework for this study is illustrated in Fig. 1. To optimize treatment planning using LLM, we developed a virtual plan designer utilizing multimodal LLM. This system is capable of evaluating plans

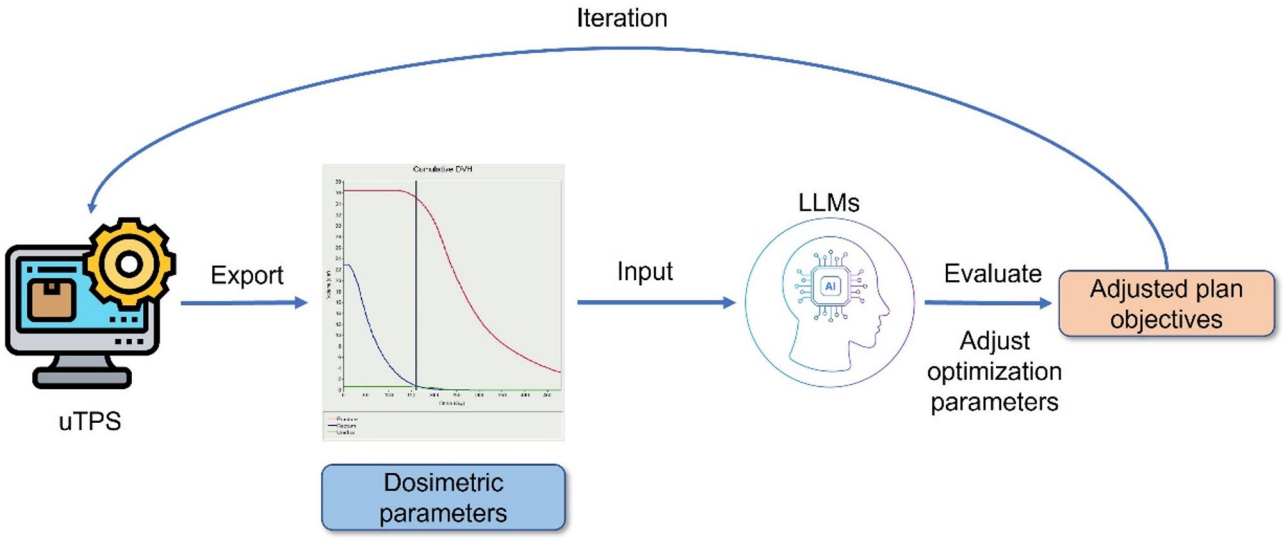

Fig. 1 The workflow of using LLM for radiotherapy treatment planning

and adjusting optimization parameters. Following each optimization iteration, dosimetric parameters for the PTV and OARs are input into the LLM for assessment. Adjusted optimization parameters are then generated based on this evaluation and fed back into the TPS for subsequent iterations until the PTV and OARs dose criteria are met.

Dataset

The dataset comprises 35 cervical cancer patients who underwent VMAT at our hospital between September 2023 and August 2024. Patients with cervical cancer were included, the upper boundary of the PTV was below the kidney, and those with lymph node metastasis were excluded. OARs considered in the study encompassed the bladder, small intestine, rectum, bone marrow, spinal cord, as well as the left and right femoral heads. Each patient's data included CT images delineating the PTV and OARs, along with clinically approved manual plans.

According to our preliminary research, it takes less than 5 patient cases to customize an LLM. In this study, 5 of the 35 patients were allocated to the training set for customizing LLMs in radiotherapy treatment planning. The remaining 30 patients constituted the test set, where LLM autonomously conducted treatment planning. To assess the generalization ability of LLMs in generating radiotherapy plans with varying target locations, two patients with PTVs extending to the inguinal region were included in the test set.

The prescribed dose for the PTV in cervical cancer was 36.0 Gy (1.8 Gy per fraction × 20 fractions), covering 95% of the PTV. A coplanar two-arc VMAT plan with gantry angles ranging from 0° to 360° was employed for the cervical cancer plans.

TPS and plan optimization

In this work, the treatment plan optimization and dose calculation were performed using uTPS (United Imaging Healthcare Co., Ltd., Shanghai, China). Upon OARs delineation, auxiliary structures including R2 (expand PTV by 0.5 cm and create a ring of 2.5 cm), R3 (expand PTV by 1.0 cm and create a ring of 2.0 cm), and R4 (expand PTV by 1.5 cm and create a ring of 1.5 cm) were created to aid in plan generation. Plan templates were customized via the clinical protocol manager in uTPS, with initial plan objectives empirically set based on historical plan metrics and clinical expertise.

Various optimization objectives such as maximum dose, minimum dose, DVH (Dose-Volume Histogram) parameters, and mean dose were selected in uTPS. During optimization, dose goals for OARs were adjusted while maintaining consistent weights. Plan objectives for target volumes, OARs, and auxiliary structures along with their initial dose goals are detailed in Table 1.

After each iteration, uTPS could provide a "constraint target value" based on the gap between the dose goal and the current dose for each optimization objective. The constraint target value is expressed in scientific notation and ranges from [0.00E + 0, +∞]. The smaller the value, the closer the current dose is to the dose goal, indicating that this optimization objective is not strictly limited.

In the clinical planning process, on the premise of not affecting the target volume conformity, it is generally believed that when the constraint target value is between 1.00E-2 and 5.00E-2, the optimization objectives of the small intestine, rectum, and bladder will be strictly limited. For other OARs, to avoid overly strict restrictions, resulting in poor target volume conformity or high dose of important OARs like small intestine, it is generally

Table 1 Plan objectives and initial dose goals for PTV, oars, and auxiliary structures in cervical cancer treatment plan optimization

Structure	Plan objective	Weight	Initial dose goal
PTV	Minimum dose	100	3760 cGy
	DVH parameters	100	3770 cGy for 99% volume
	Maximum dose	100	3780 cGy
Bladder	Mean dose	5	3000 cGy
Rectum	Mean dose	5	3300 cGy
Small Intestine	Mean dose	5	2050 cGy
Spinal cord	Maximum dose	3	2400 cGy
Bone marrow	Mean dose	3	2400 cGy
Femoral head left	Mean dose	1	2000 cGy
Femoral head right	Mean dose	1	2000 cGy
R2	Maximum dose	5	2880 cGy
R3	Maximum dose	5	2520 cGy
R4	Maximum dose	5	2160 cGy

considered that when the constraint target value is between 1.00E-4 and 1.00E-2, the degree of restriction is appropriate. This criterion provides a reference for planning optimization using LLMs.

In this study, the photon energy was set at 6MV. The collimator angles of the two arcs were both set at 0°. The Fluence Map Optimization (FMO) algorithm was used for VMAT treatment plan optimization [41]. For dose calculation, the collapsed cone algorithm was used with a dose calculation grid of 2 mm [42]. In the optimization process, two terminating criteria were introduced for the optimization iteration: [1] A maximum iteration number of 20 was reached; [2] Once the PTV dose is greater than the prescription dose and the constraint target values of OARs are in their ideal ranges.

LLMs and prompt engineering

Several LLMs were utilized as virtual physicists for plan optimization, including Qwen-2.5-Max (Alibaba cloud, Hangzhou, China), Gemini-1.5-Flash (Google, Mountain View, CA), and Llama-3.2 (Meta AI, Menlo Park, CA). Parameters fed into the LLMs comprised current doses, dose goals, constraint target values, ideal ranges of constraint target values for OAR optimization objectives, and PTV D95 values.

In our clinical practice, the prescription dose is specified using the PTV D95 metric. For the cases in this study, the prescribed dose for the PTV D95 is 3600 cGy. According to clinical practice, to optimize the dose distribution while avoiding excessive hot spots or cold spots, we initially set a PTV D95 target between 1.00 and 1.02 times the prescribed dose. This approach allows for greater flexibility in achieving optimal dose conformity and coverage while preventing substantial deviations from the prescribed dose. Following the initial optimization, we perform dose normalization to adjust the PTV D95 to exactly 3600 cGy.

During customization, we first introduced the main goal of this task to LLM, which was to increase the PTV dose greater than the prescription while keeping the constraint target value of OARs within the ideal interval. Subsequently, the adjustment strategy was introduced, that is, when the constraint target value is greater than/less than its ideal range, the dose goal should be increased/decreased, while if the constraint target value is 0, the dose goal should be adjusted to less than the current dose. In supplementary material, we describe in detail the specific methods of LLM customization through prompt engineering, as well as the inputs and outputs of LLM in the planning optimization of a sample case. For 5 patients in the training set, after each iteration, we input the dosimetric parameters of PTV and OARs to LLM, allowing it to conduct evaluation and adjust plan objectives. We then provided the LLM with the evaluation opinions of human physicists and the adjustment scheme, enabling the LLMs to gradually master the decision-making process of human physicists. For the test set patients, adjusted objectives from LLMs were directly input into uTPS for optimization without manual intervention, until the terminating criteria were met.

To enhance LLMs' adaptive capabilities, prior adjustment strategies and dosimetric parameter changes were considered before each optimization objective assessment.

Planning evaluation

Clinical plans were developed and approved for all patients by experienced medical physicists using uTPS through inverse optimization. We collected these clinical plans as the benchmark to verify the feasibility of using LLM to optimize radiotherapy plans. All plans were normalized to ensure that 95% of PTV coverage met the prescription dose. Target volume conformity, OARs dose sparing, and Monitor Unit (MU) values were calculated and compared between clinical and LLM plans.

Target volume conformity was assessed using the Paddick conformity index (CI) and homogeneity index (HI). CI value and HI value were defined as Eqs. (1) and (2), respectively.

$$CI = TV_{RI} \times \frac{TV_{RI}}{V_{RI} \times TV} \tag{1}$$

$$HI = \frac{D_{2\%} - D_{98\%}}{D_{50\%}} \tag{2}$$

In Eq. (1), V_{RI} and TV represent the volume of the reference isodose cloud and the volume of the target, respectively. TV_{RI} represents the intersection of V_{RI} and TV. In

Table 2 Dosimetric parameters selected for cervical cancer plan evaluation

Structure	Dosimetric parameters
PTV	Dmax, D98
Bladder	D50
Rectum	D50
Small intestine	D50
Spinal Cord	D0.1 cc
Bone marrow	D90
Femoral head left	D5
Femoral head right	D5

D5: dose covering 5% of the volume

D50: dose covering 50% of the volume

D90: dose covering 90% of the volume

D98: dose covering 98% of the volume

D0.1 cc: dose covering 0.1 cc of the volume

Eq. (2), D2%, D98%, and D50% represent the dose covering 2%, 98%, and 50% of the volume, respectively. A CI value of approximately 1 indicates better conformity. The closer the HI value is to 0, the better the uniformity of the target volume. OARs dose sparing was evaluated using various dosimetric parameters detailed in Table 2.

To verify the clinical deliverability of the LLM plans, we performed patient-specific quality assurance on all LLM plans and manual plans using Third QA software (United Imaging Healthcare Co., Ltd., Shanghai, China), which has been clinically validated.

Statistical analyses were conducted using the Wilcoxon signed-rank test with a significance set at $P<0.05$ using SPSS version 29.0 (IBM Corp., Armonk, NY, USA) [43].

Results

Plan generation results and iteration time

All generated LLM plans underwent evaluation by three senior medical physicists. All plans generated by Qwen-2.5-max and Llama-3.2 were deemed clinically acceptable for all 30 patients in the test set. However, Gemini-1.5-flash exhibited issues, with only the first two plans meeting the acceptability criteria. In the remaining 28 patients, Gemini-1.5-flash displayed hallucinations, confusing OARs dose sparing across iterations and fabricating data for analysis. Additionally, it occasionally deviated from evaluation criteria, leading to counterproductive optimization adjustments or no adjustments, impeding its ability to generate clinically acceptable plans for some patients.

Regarding iteration time, uTPS required approximately 56 s per iteration, while Qwen-2.5-max, Gemini-1.5-flash, and Llama-3.2 took around 30, 12, and 6 s, respectively, for a single processing and output adjustment cycle. For the test set, automatic planning took on average 16.3 ± 5.0 min and 9.8 ± 2.1 min for Qwen-2.5-max and Llama-3.2, respectively, compared to approximately 20 min for manual planning in our hospital. The LLM-based method demonstrated faster plan generation speed than manual methods, with Qwen-2.5-max averaging 10.6 ± 3.5 iterations and Llama-3.2 averaging 8.5 ± 2.0 iterations for plan optimization.

Treatment planning performance between LLM plans and manual plans

All manual plans and LLM plans by Qwen-2.5-max and Llama-3.2 met clinical prescription requirements for target coverage. Table 3; Fig. 2 provide a quantitative comparison between manual plans and LLM plans in terms of PTV dose, OARs dose sparing, target volume conformity, dose homogeneity, and MU values. Comparative analyses between manual plans and LLM plans revealed that Gemini-1.5-flash plans exhibited suboptimal performance due to hallucinations leading to misguided optimization directions. That is, when the restriction on one OAR in the current plan is strict enough, Gemini-1.5-flash would further limit its dose, and as the limitations of the current plan for an OAR tended to become nonexistent, Gemini-1.5-flash would further relax the limitations until they became nonexistent. Therefore,

Table 3 Dosimetric parameters of manual plans and plans by different LLMs in the test set (Gy)

Structure	Metrics	Manual plans	Qwen-2.5-Max		Llama-3.2		Gemini-1.5-Flash	
		Mean ± std	Mean ± std	*p*-value	Mean ± std	*p*-value	Mean ± std	*p*-value
PTV	Dmax (Gy)	38.61 ± 0.30	38.33 ± 0.42	0.055	38.37 ± 0.28	**0.041**	39.08 ± 0.82	0.277
	D98 (Gy)	35.49 ± 0.07	35.51 ± 0.05	0.25	35.50 ± 0.05	0.45	35.24 ± 0.25	**0.002**
Bladder	D50 (Gy)	30.31 ± 3.36	30.87 ± 2.15	0.804	30.91 ± 2.47	0.804	29.07 ± 2.41	**0.041**
Small Intestine	D50 (Gy)	14.57 ± 3.30	14.25 ± 3.34	0.083	14.26 ± 3.55	0.095	14.78 ± 2.40	0.421
Rectum	D50 (Gy)	34.51 ± 2.65	34.12 ± 2.59	0.135	34.22 ± 2.45	0.135	33.47 ± 2.83	0.095
Bone marrow	D90 (Gy)	12.09 ± 1.72	12.03 ± 1.93	0.762	12.23 ± 1.71	0.489	12.02 ± 2.11	0.847
Spinal cord	D0.1 cc (Gy)	19.87 ± 3.21	20.38 ± 2.44	0.169	20.72 ± 1.73	0.095	20.16 ± 2.92	0.389
Femoral head left	D5 (Gy)	27.13 ± 2.43	27.20 ± 2.37	0.762	27.22 ± 2.03	1.000	27.89 ± 1.57	0.151
Femoral head right	D5 (Gy)	26.86 ± 2.24	26.66 ± 2.05	0.389	26.89 ± 2.13	0.720	27.34 ± 1.39	0.107
	CI	0.926 ± 0.007	0.929 ± 0.007	**0.005**	0.928 ± 0.007	**0.035**	0.915 ± 0.014	**0.007**
	HI	0.065 ± 0.006	0.058 ± 0.006	**0.005**	0.059 ± 0.005	**0.015**	0.077 ± 0.021	0.107
	MU	852.78 ± 76.92	812.65 ± 66.27	0.064	803.95 ± 63.21	**0.004**	815.46 ± 88.18	0.121

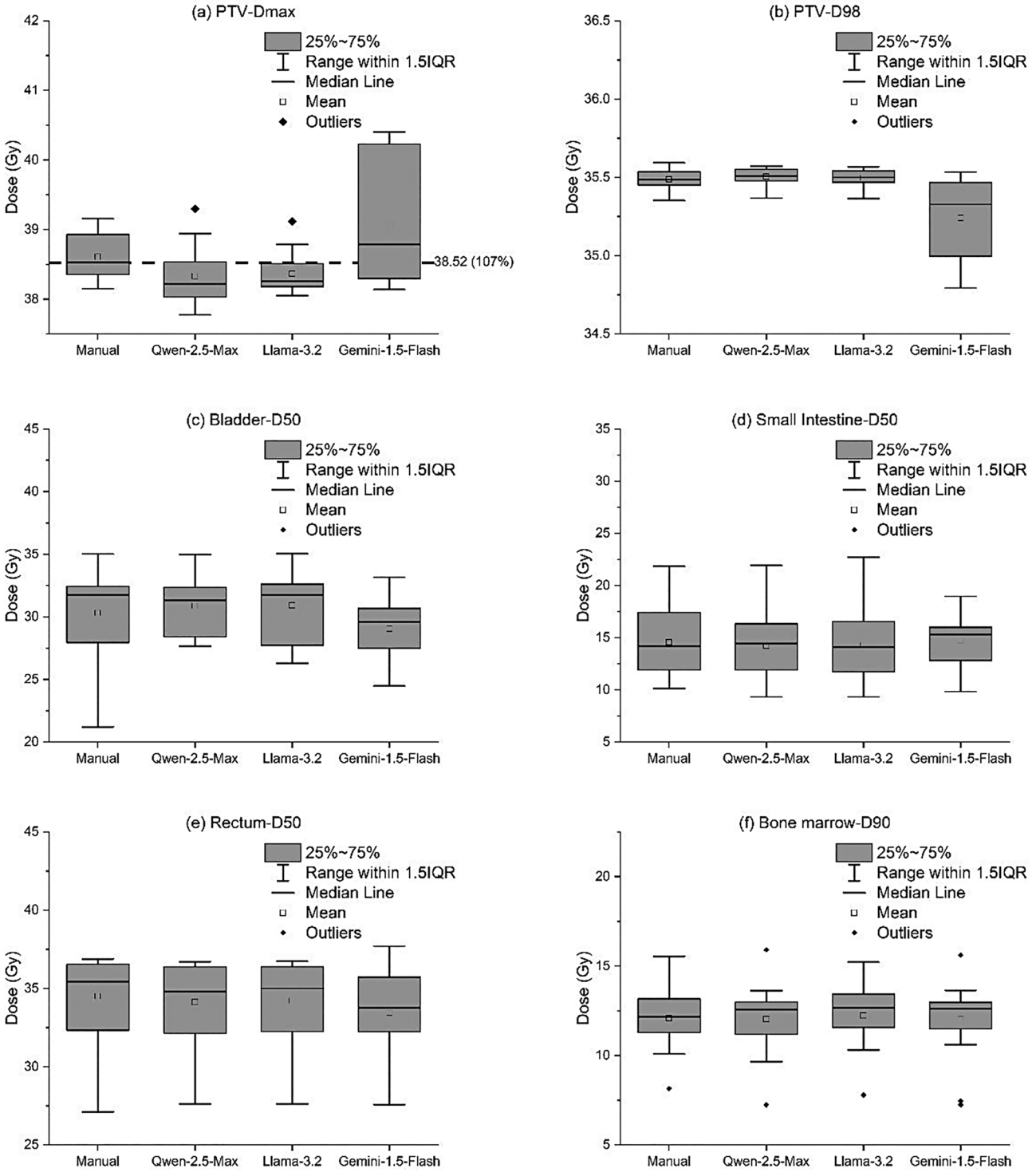

Fig. 2 Boxplots of dosimetric parameters and other plan parameters in the test set

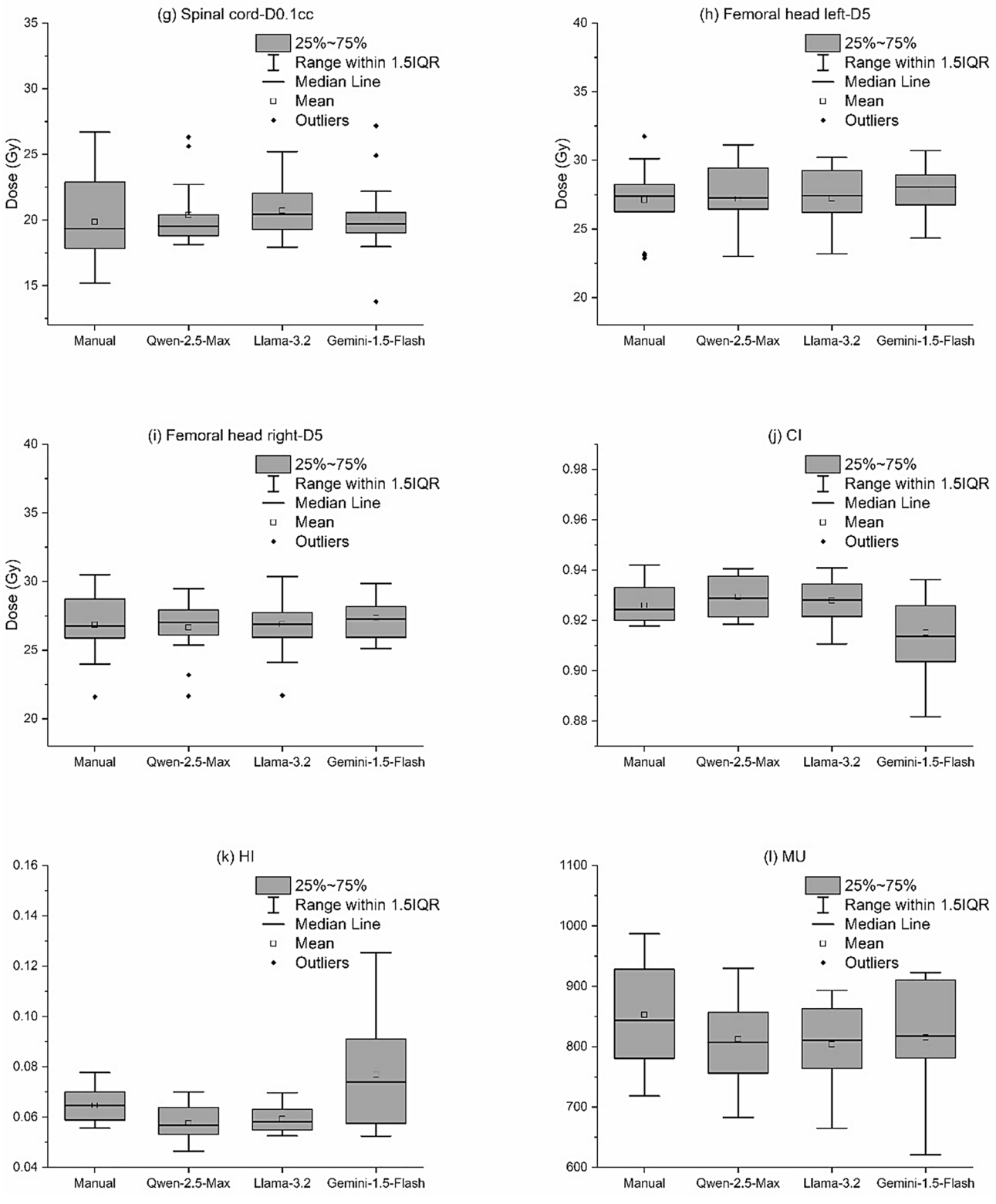

Fig. 2 (continued)

in the Gemini-1.5-flash plan, the dose sparing of some OARs is too limited and the others are ignored. As a result, Gemini-1.5-flash plans showed compromised target conformity, dose homogeneity, and maximum dose control compared to other methods. The CI value was significantly lower than that of manual plans, and the maximum dose of the target volume was higher than that of plans made by other methods, as shown in Fig. 2.

In terms of OARs dose sparing, Qwen-2.5-max and Llama-3.2 plans showed no significant differences from manual plans in most dosimetric parameters. In most cases, the bladder D50 values for LLM and manual plans were comparable. However, one outlier case exhibited a significant difference, with the bladder D50 in the LLM plan being approximately 6 Gy higher than in the manual plan. This outlier contributed to the higher average bladder D50 value observed for LLM plans in Table 3; Fig. 2. In terms of the maximum dose in the target volume, the average maximum dose of Qwen-2.5-max and Llama-3.2 were lower than those of manual plans. Notably, Llama-3.2 demonstrated a statistically significant advantage in maximum dose control within the target volume. Furthermore, Qwen-2.5-max and Llama-3.2 plans exhibited significantly improved HI values and lower MU values compared to manual plans, indicating enhanced target volume conformity, dose homogeneity, and reduced complexity.

Figure 3 shows the isodose distribution of the manual plan, Qwen-2.5-max plan, and Llama-3.2 plan in three different slices of one sample patient. It can be found that in this case, there is no obvious difference in the target coverage across these three plans, with faster dose fall-off in the bladder and rectum areas in LLM plans. Figure 4 shows the comparison of dose-volume histograms between different plans of this patient, which are (a) manual plan and Qwen-2.5-max plan, and (b) manual plan and Llama-3.2 plan. It can be seen that doses to the rectum and bladder in these two LLM plans are lower, and the target volume conformity of LLM plans is better.

To verify the feasibility of using LLM to optimize radiotherapy treatment plans for different patients, we compared OARs dose sparing between manual plans and LLM plans generated by Qwen-2.5-max and Llama-3.2 for 30 test patients. Figure 5 shows the distribution of the

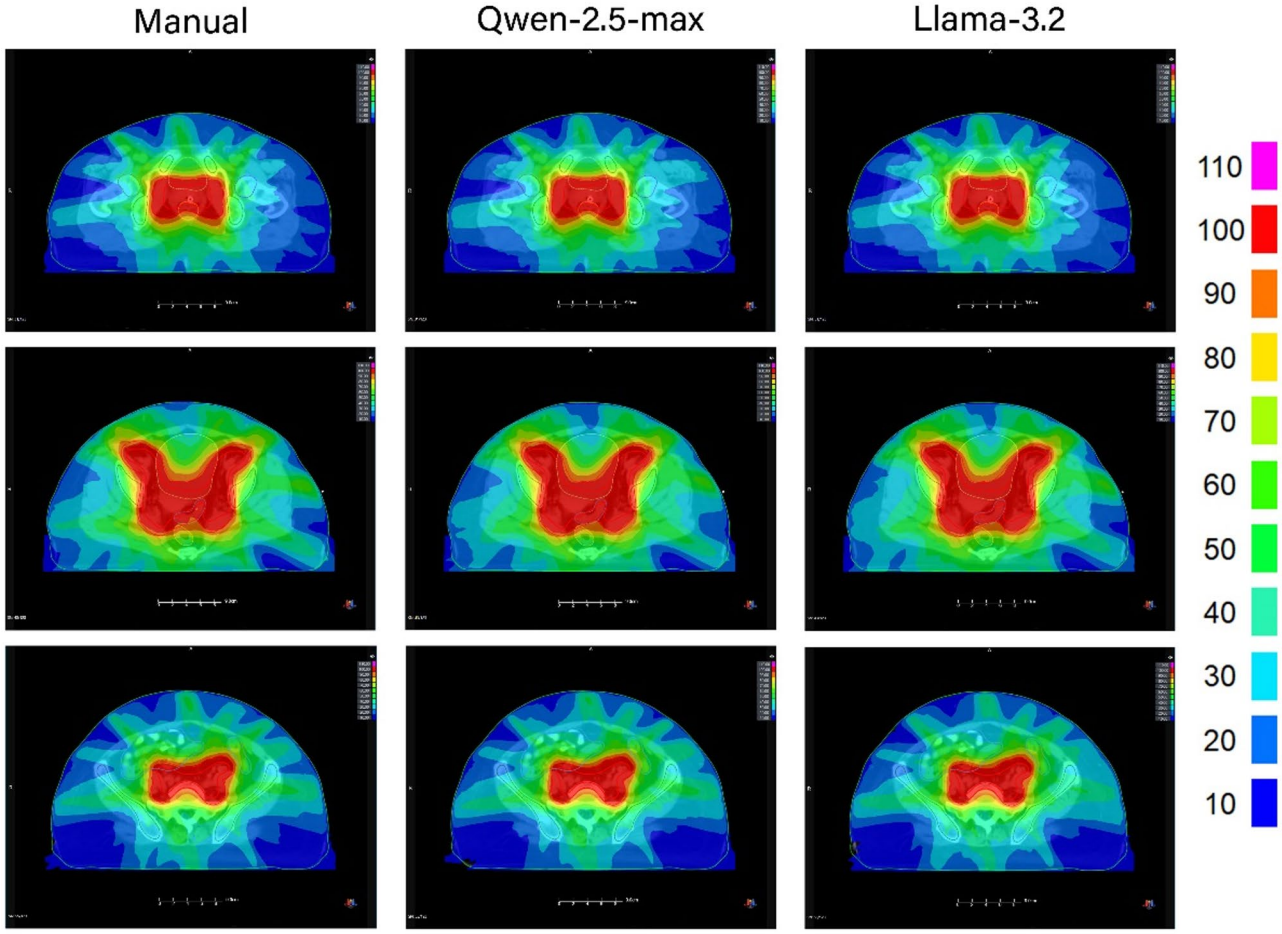

Fig. 3 Isodose distribution of three slices from a sample patient's plans

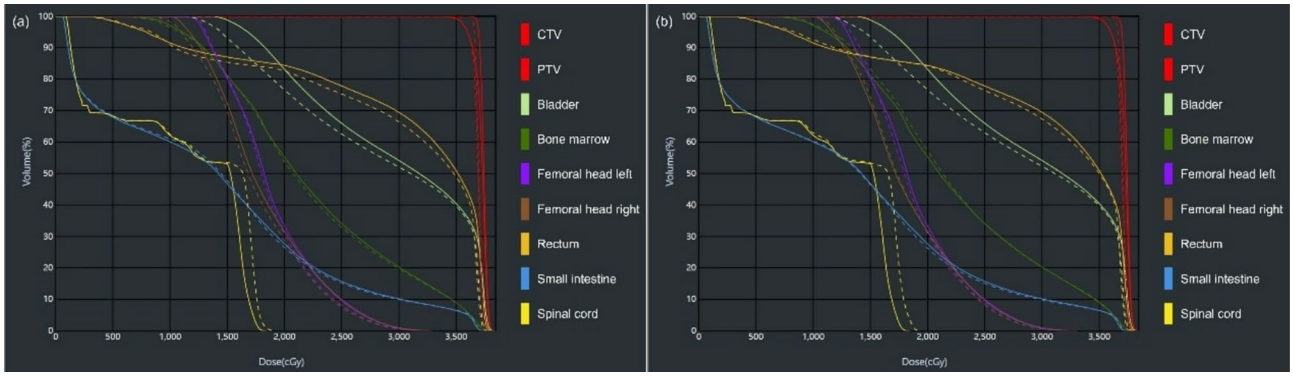

Fig. 4 Comparison of dose-volume histograms between manual plan and different LLM plans (dashed line)

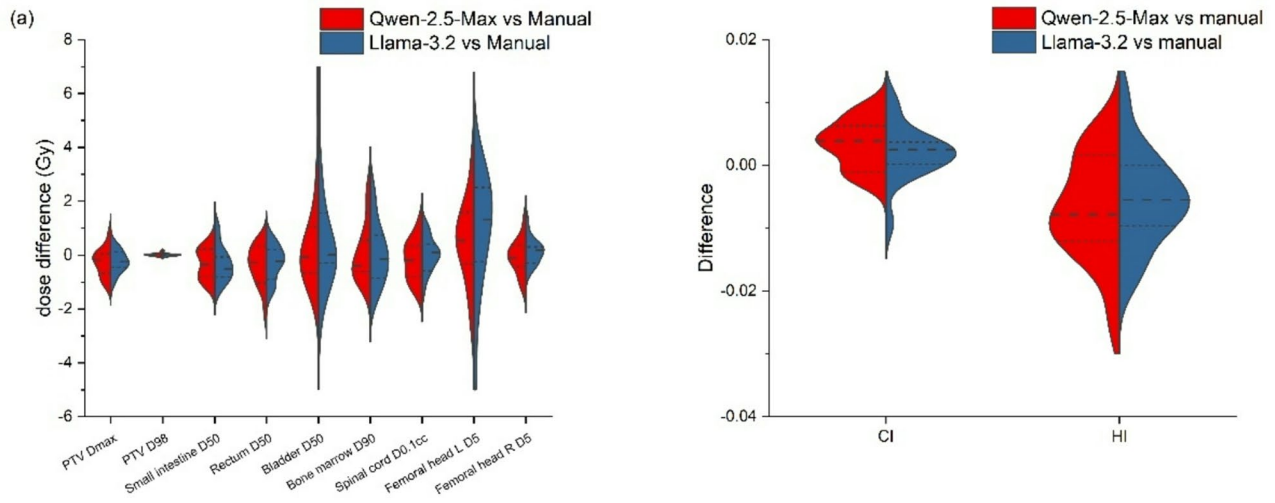

Fig. 5 Difference distributions between manual plan and different LLM plans on test set, (**a**) dosimetric parameters, (**b**) target volume conformity and dose homogeneity

Table 4 Gamma passing rates of manual plans and LLM plans under different criteria

Criteria	Manual plans	Qwen-2.5-Max	Llama-3.2	Gemini-1.5-Flash
3 mm/3%	0.999 ± 0.001	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
2 mm/2%	0.995 ± 0.007	0.999 ± 0.001	0.999 ± 0.001	0.999 ± 0.001

difference on the test set. The results revealed notable advantages in PTV Dmax, Bladder D50, Small Intestine D50, and Bone Marrow D90 with LLMs outperforming manual plans.

Overall, cervical cancer radiotherapy plans generated by Qwen-2.5-max and Llama-3.2 closely matched manual plans in OARs dose sparing, with superior target volume conformity and dose homogeneity. This preliminarily validates the feasibility of utilizing LLM for cervical cancer treatment planning.

Regarding the results of patient-specific quality assurance, when the evaluation criteria are set at 3 mm/3%, the gamma passing rate for all plans exceeds 0.999. When the criteria are tightened to 2 mm/2%, the gamma passing rate for manually optimized plans is greater than 0.99, whereas the gamma passing rate for LLM plans is greater than 0.995. This indicates that the LLM plans exhibit excellent deliverability. Table 4 presents the mean gamma passing rates of plans obtained by different methods under various criteria.

Horizontal comparison across different LLMs

To further compare the differences in the plans generated by different LLMs, we calculated the relative differences in dosimetric parameters between plans generated by Qwen-2.5-max and Llama-3.2, as shown in Table 5. It can be found that the relative differences between plans generated by these two LLMs were minimal, with average relative differences of less than 3% for PTV dose and OARs dose sparing, showing no statistical significance. Despite significantly lower MU values in Llama-3.2 plans, the mean difference between the two LLMs was only 8.7 MU, indicating effective treatment planning and acceptable plan generation capability across different LLMs (Table 5).

Table 5 Comparison of dosimetric parameters between plans by Qwen-2.5-max and Llama-3.2

Structures	Metrics	Qwen-2.5-max	Llama-3.2	Mean relative difference	*p*-value
PTV	Dmax	38.33 ± 0.42	38.37 ± 0.28	0.10%	0.489
	D98	35.51 ± 0.05	35.50 ± 0.05	-0.02%	0.303
Bladder	D50	30.87 ± 2.15	30.91 ± 2.47	-0.10%	0.277
Small Intestine	D50	14.25 ± 3.34	14.26 ± 3.55	0.32%	0.762
Rectum	D50	34.12 ± 2.59	34.22 ± 2.45	0.06%	0.720
Bone marrow	D90	12.03 ± 1.93	12.23 ± 1.71	2.16%	0.095
Spinal cord	D0.1 cc	20.38 ± 2.44	20.72 ± 1.73	2.09%	0.208
Femoral head left	D5	27.20 ± 2.37	27.22 ± 2.03	0.25%	0.847
Femoral head right	D5	26.66 ± 2.05	26.89 ± 2.13	0.87%	0.167
	CI	0.929 ± 0.007	0.928 ± 0.007	-0.15%	0.083
	HI	0.058 ± 0.006	0.059 ± 0.005	3.79%	0.136
	MU	812.65 ± 66.27	803.95 ± 63.21	-1.04%	**0.048**

Discussion

This study delved into the feasibility of leveraging large language models for optimizing radiotherapy treatment plans. The findings revealed that Gemini-1.5-flash struggled to produce clinically acceptable plans due to hallucination effects, while both Qwen-2.5-max and Llama-3.2 successfully generated such plans. Concerning OARs dose sparing, neither the Qwen-2.5-max nor Llama-3.2 plans exhibited statistically significant discrepancies from manual plans. Across the test set, the average D50 values for the small intestine and rectum in the Qwen-2.5-max and Llama-3.2 plans were marginally lower than those in manual plans. However, for the bladder's D50 and spinal cord's D0.1 cc, manual plans slightly outperformed LLM plans, with the bladder D50 averaging 0.5–0.6 Gy lower and the spinal cord D0.1 cc 0.5–0.8 Gy lower. Studies by Okonogi [44] and Jadon [45] suggest that a 0.5 Gy dose variation may not lead to distinct clinical outcomes for the small intestine, rectum, and bladder. Moreover, a spinal cord D0.1 cc of 20 Gy typically does not induce significant spinal cord injury [46]. Thus, the differences in OARs dose sparing between LLM plans and manual plans are not substantial.

In this study, we compared the dose differences observed in our research with findings from previous studies. Kang et al. reported no significant differences between KBP and manual plans in terms of PTV D2%, D98%, and conformity index, though a significant reduction in the homogeneity index was observed in KBP plans. Regarding OAR dose sparing, KBP reduced doses for most OARs, except for the small intestine and the left femoral head [47]. Similarly, Swamidas et al. found that KBP achieved comparable target coverage to manual plans while significantly reducing doses to the spinal cord and femoral heads [48]. Compared to these KBP-based methods, the LLM-based method in our study demonstrated advantages in CI and HI. Regarding OAR dose sparing, the LLM plans resulted in lower doses to the small intestine and rectum than the manual plans. These findings highlight the potential of LLMs in automatic treatment planning. In future research, we plan to explore a variety of other treatment planning methods, including knowledge-based planning (KBP), and compare the dose differences among plans generated by different approaches. This will provide a more comprehensive evaluation of the performance of the LLM-based method and its potential applications in clinical practice.

The initial plan objectives were established by senior medical physicists based on their clinical expertise in this study. The application of these uniform initial objectives enhanced plan optimization efficiency, yielding acceptable results in under 20 iterations across the test set, even in cases with inguinal targets, in which the initial dose goal might be too strict for femoral heads. Specifically, the initial dose objective for the mean dose to the femoral heads was set at 2000 cGy. For typical cervical cancer plans, this dose objective is relatively close to the dose ultimately achieved. During the optimization process, only a few adjustments were typically required to bring the constraint target value into the ideal range. However, for patients with inguinal targets, the mean dose to the femoral heads may reach 2300 to 2400 cGy. Consequently, in the initial iterations, the constraint target value was relatively high, often exceeding 1.00E-1. While such a large deviation from the ideal range is uncommon in other typical plans, the LLM in this study effectively overcame this challenge by making substantial initial adjustments, followed by gradual reductions in adjustment amplitude. This approach allowed the femoral head constraint target value to reach the ideal range after only a few iterations.

For all patients in the test set, Llama-3.2 required fewer than 11 iterations, and Qwen-2.5-max fewer than 18 iterations, demonstrating the method's adaptability for various PTV and OARs locations. Unlike existing AI methods employing dose prediction and reinforcement learning, this approach utilizes fixed initial objectives

without the need for extensive training data, thereby reducing training time and data requirements.

During treatment planning using LLMs, we observed the models' capacity to learn from previous iterations and adjust optimization objectives progressively. The models autonomously adapt the adjustment amplitude of dose goals based on feedback, ensuring convergence towards the ideal constraint target value. In this study, since the initial objectives for most OARs were not strictly limited, after the first few iterations, the constraint target value of some OARs was usually too small or even zero. In this scenario, LLM adjusted the dose goal by a large amplitude (for example, 100 cGy). When the constraint target value was gradually approaching its ideal range, LLM detected this trend without human intervention and actively reduced the adjustment amplitude of the dose objective (such as 50 cGy and 20 cGy) until the constraint target value reached the ideal range. In the following research, we aim to explore the potential for further refinement by allowing the LLM to independently assess the potential for optimization based on dose distribution and other relevant information, rather than relying solely on empirical rules. This approach will enable the LLM to make autonomous decisions regarding dose objective adjustments, potentially leading to improved solutions.

Notably, Qwen-2.5-max and Llama-3.2 outperformed Gemini-1.5-flash in this study, as they effectively interpreted input data and generated reasonable outputs. Conversely, Gemini-1.5-flash exhibited hallucinations, ignored provided iteration results, and occasionally misjudged adjustment directions, leading to subpar performance. The occurrence of hallucinations in LLMs, as noted by Lee et al., may result from excessive repetitive information biasing the model's knowledge memory [49]. In our study, standardized input specifications and prompts might contribute to hallucination issues. Beyond that, the hallucination of LLM may also be caused by factors such as model structure, model size, and decoding algorithm [50, 51].

The relatively large model sizes of Qwen-2.5-max and Llama-3.2 pose challenges for integration and deployment within clinical TPS. Future research should explore employing smaller LLM models for iterative radiotherapy treatment plan optimization. Additionally, investigating more effective prompts, incorporating Retrieval Augmented Generation (RAG) technology [52–54], or experimenting with Controlled Text Generation (CTG) methods [55, 56] could also be explored to enhance the reliability of utilizing smaller LLMs to adjust plan optimization parameters.

In subsequent studies, we will try to leverage LLMs to optimize radiotherapy plans for diverse tumor sites and integrate LLMs with TPS to expand their applicability in radiotherapy plan optimization, ultimately enhancing the automation of radiotherapy treatment planning.

Conclusion

This study delved into the feasibility of employing large language models for automated radiotherapy treatment planning, utilizing various LLMs. Comparative analysis between LLM-generated plans and manual plans revealed that Qwen-2.5-max and Llama-3.2 effectively crafted clinically acceptable cervical cancer plans based on provided prompts. Notably, the LLM-based method exhibited a lower time cost for plan generation compared to the manual method. Regarding dosimetric parameters, no significant disparities were observed in PTV dose and OARs dose sparing between LLM plans and manual plans. LLM plans showcased superior target volume conformity relative to manual plans. In essence, this study offers initial validation for leveraging LLMs in automating treatment planning for cervical cancer. By supplanting manual planning, the LLM-based approach has the potential to alleviate physicists' workload and streamline their processes, thereby augmenting overall clinical efficiency.

Abbreviations

LLM	Large Language Models
OAR	Organ-at-risk
VMAT	Volumetric modulated arc therapy
MU	Monitor unit
CI	Conformity index
HI	Homogeneity index
PTV	Planning target volume
ARIR	Automated rule implementation and reasoning
KBP	Knowledge-based planning
MCO	Multi-criteria optimization
TPS	Treatment planning system
DVH	Dose-volume histogram
AI	Artificial intelligence
MLC	Multi-leaf collimator
IMRT	Intensity modulated radiotherapy
ESAPI	Eclipse scripting API
FMO	Fluence Map Optimization
D5	Dose covering 5% of the volume
D50	Dose covering 50% of the volume
D90	Dose covering 90% of the volume
D98	Dose covering 98% of the volume
D0.1cc	Dose covering 0.1 cc of the volume
RAG	Retrieval Augmented Generation
CTG	Controlled Text Generation

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13014-025-02660-5.

Supplementary Material 1

Author contributions

Paper idea: Shuoyang Wei, Ankang Hu, Bo Yang. The source of datasets: Shuoyang Wei, Yongguang Liang, Jingru Yang, Lang Yu, Wenbo Li. Assessment of radiotherapy treatment plan: Shuoyang Wei, Yongguang Liang, Jingru Yang.

Writing of the paper: Shuoayang Wei, Ankang Hu, Bo Yang, Jie Qiu. All authors read and approved the final manuscript.

Funding
National key research and development program of China, China (No. 2022YFC2404606). National High Level Hospital Clinical Research Funding (No. 2022-PUMCH-B-116).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethical approval
This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013) and was approved by the institutional ethics board of the Peking Union Medical College Hospital (No. I-24PJ0295). Informed consent was obtained from all the patients.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Radiotherapy, Peking Union Medical College Hospital, Beijing 100730, China
[2]Department of Engineering Physics, Tsinghua University, Beijing 100084, China
[3]Key Laboratory of Particle & Radiation Imaging (Tsinghua University), Ministry of Education, Beijing 100084, China

Received: 4 December 2024 / Accepted: 6 May 2025
Published online: 15 May 2025

References

1. Romeijn HE, Ahuja RK, Dempsey JF, Kumar A. A new linear programming approach to radiation therapy treatment planning problems. Oper Res. 2006;54(2):201–16.
2. Breedveld S, Craft D, Van Haveren R, Heijmen B. Multi-criteria optimization and decision-making in radiotherapy. Eur J Oper Res. 2019;277(1):1–19.
3. Yu Y, Zhang J, Cheng G, Schell M, Okunieff P. Multi-objective optimization in radiotherapy: applications to stereotactic radiosurgery and prostate brachytherapy. Artif Intell Med. 2000;19(1):39–51.
4. Nelms BE, Robinson G, Markham J, Velasco K, Boyd S, Narayan S, et al. Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems. Practical Radiation Oncol. 2012;2(4):296–305.
5. Craft DL, Halabi TF, Shih HA, Bortfeld TR. Approximating convex Pareto surfaces in multiobjective radiotherapy planning. Med Phys. 2006;33(9):3399–407.
6. Kyroudi A, Petersson K, Ozsahin E, Bourhis J, Bochud F, Moeckli R. Exploration of clinical preferences in treatment planning of radiotherapy for prostate cancer using Pareto fronts and clinical grading analysis. Phys Imaging Radiation Oncol. 2020;14:82–6.
7. Slotman BJ, Vos PH. Planning of radiotherapy capacity and productivity. Radiother Oncol. 2013;106(2):266–70.
8. Das IJ, Moskvin V, Johnstone PA. Analysis of treatment planning time among systems and planners for intensity-modulated radiation therapy. J Am Coll Radiol. 2009;6(7):514–7.
9. Eschwège MT, Francois. Conformal radiotherapy and intensity-modulated radiotherapy: clinical data. Acta Oncol. 2000;39(5):555–67.
10. Hong TS, Craft DL, Carlsson F, Bortfeld TR. Multicriteria optimization in intensity-modulated radiation therapy treatment planning for locally advanced cancer of the pancreatic head. Int J Radiation Oncology* Biology* Phys. 2008;72(4):1208–14.
11. Wang C, Zhu X, Hong JC, Zheng D. Artificial intelligence in radiotherapy treatment planning: present and future. Technol Cancer Res Treat. 2019;18:1533033819873922.
12. Meyer P, Biston MC, Khamphan C, Marghani T, Mazurier J, Bodez V, et al. Automation in radiotherapy treatment planning: examples of use in clinical practice and future trends for a complete automated workflow. Cancer/Radiothérapie. 2021;25(6):617–22.
13. Portik D, Clementel E, Krayenbühl J, Bakx N, Andratschke N, Hurkmans C. Knowledge-based versus deep learning based treatment planning for breast radiotherapy. Phys Imaging Radiation Oncol. 2024;29:100539.
14. Cao W, Gronberg M, Olanrewaju A, Whitaker T, Hoffman K, Cardenas C, et al. Knowledge-based planning for the radiation therapy treatment plan quality assurance for patients with head and neck cancer. J Appl Clin Med Phys. 2022;23(6):e13614.
15. Momin S, Fu Y, Lei Y, Roper J, Bradley JD, Curran WJ, et al. Knowledge-based radiation treatment planning: A data-driven method survey. J Appl Clin Med Phys. 2021;22(8):16–44.
16. Craft DL, Hong TS, Shih HA, Bortfeld TR. Improved planning time and plan quality through multicriteria optimization for Intensity-Modulated radiotherapy. Int J Radiation Oncology*Biology*Physics. 2012;82(1):e83–90.
17. Miguel-Chumacero E, Currie G, Johnston A, Currie S. Effectiveness of Multi-Criteria Optimization-based Trade-Off exploration in combination with rapid-plan for head & neck radiotherapy planning. Radiat Oncol. 2018;13(1):229.
18. Müller BS, Shih HA, Efstathiou JA, Bortfeld T, Craft D. Multicriteria plan optimization in the hands of physicians: a pilot study in prostate cancer and brain tumors. Radiat Oncol. 2017;12(1):168.
19. Li X, Zhang J, Sheng Y, Chang Y, Yin F-F, Ge Y, et al. Automatic IMRT planning via static field fluence prediction (AIP-SFFP): a deep learning algorithm for real-time prostate treatment planning. Phys Med Biol. 2020;65(17):175014.
20. Vandewinckele L, Willems S, Lambrecht M, Berkovic P, Maes F, Crijns W. Treatment plan prediction for lung IMRT using deep learning based fluence map generation. Physica Med. 2022;99:44–54.
21. Tsang DS, Tsui G, McIntosh C, Purdie T, Bauman G, Dama H, et al. A pilot study of machine-learning based automated planning for primary brain tumours. Radiat Oncol. 2022;17(1):3.
22. Wang N, Fan J, Xu Y, Yan L, Chen D, Wang W, et al. Clinical implementation and evaluation of deep learning-assisted automatic radiotherapy treatment planning for lung cancer. Physica Med. 2024;124:104492.
23. McIntosh C, Welch M, McNiven A, Jaffray DA, Purdie TG. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. Phys Med Biol. 2017;62(15):5926–44.
24. Nguyen D, Jia X, Sher D, Lin M-H, Iqbal Z, Liu H, et al. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. Phys Med Biol. 2019;64(6):065020.
25. Borderias-Villarroel E, Huet Dastarac M, Barragán-Montero AM, Helander R, Holmstrom M, Geets X, et al. Machine learning-based automatic proton therapy planning: impact of post-processing and dose-mimicking in plan robustness. Med Phys. 2023;50(7):4480–90.
26. Zeverino M, Fabiano S, Jeanneret-Sozzi W, Bourhis J, Bochud F, Moeckli R. Enhancing automated right-sided early-stage breast cancer treatments via deep learning model adaptation without additional training. Med Phys. 2025;52(5):3280–97.
27. Li C, Guo Y, Lin X, Feng X, Xu D, Yang R. Deep reinforcement learning in radiation therapy planning optimization: A comprehensive review. Physica Med. 2024;125:104498.
28. Pu G, Jiang S, Yang Z, Hu Y, Liu Z. Deep reinforcement learning for treatment planning in high-dose-rate cervical brachytherapy. Physica Med. 2022;94:1–7.
29. Shen C, Chen L, Jia X. A hierarchical deep reinforcement learning framework for intelligent automatic treatment planning of prostate cancer intensity modulated radiation therapy. Phys Med Biol. 2021;66(13):134002.
30. Hrinivich WT, Bhattacharya M, Mekki L, McNutt T, Jia X, Li H, et al. Clinical VMAT machine parameter optimization for localized prostate cancer using deep reinforcement learning. Med Phys. 2024;51(6):3972–84.
31. Gao Y, Shen C, Jia X, Park YK. Implementation and evaluation of an intelligent automatic treatment planning robot for prostate cancer stereotactic body radiation therapy. Radiother Oncol. 2023;184:109685.
32. Liu S, Pastor-Serrano O, Chen Y, Gopaulchan M, Liang W, Buyyounouski M, et al. Automated radiotherapy treatment planning guided by GPT-4Vision. arXiv preprint. 2024. https://doi.org/10.48550/arXiv.2406.15609
33. Floyd W, Kleber T, Carpenter DJ, Pasli M, Qazi J, Huang C, et al. Current strengths and weaknesses of ChatGPT as a resource for radiation oncology patients and providers. International Journal of Radiation Oncology* Biology* Physics; 2023.
34. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483–e.

35. Liao W, Liu Z, Dai H, Xu S, Wu Z, Zhang Y, et al. Differentiating ChatGPT-generated and human-written medical texts: quantitative Study. JMIR Med Educ. 2023;9(1):e48904. https://doi.org/10.2196/48904
36. Liu Z, Wang P, Li Y, Holmes J, Shu P, Zhang L, et al. Radonc-gpt: A large Language model for radiation oncology. arXiv preprint 2023. https://doi.org/10.48550/arXiv.2309.10160
37. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. 2023;9(10):1437–40.
38. Putz F, Haderlein M, Lettmaier S, Semrau S, Fietkau R, Huang Y. Exploring the capabilities and limitations of large Language models for radiation oncology decision support. Int J Radiat Oncol Biol Phys. 2024;118(4):900–4.
39. Wu DJ, Bibault J-E. Pilot applications of GPT-4 in radiation oncology: summarizing patient symptom intake and targeted chatbot applications. Radiother Oncol. 2024;190:109978.
40. Dong Z, Chen Y, Gay H, Hao Y, Hugo GD, Samson P, et al. Large-language-model empowered 3D dose prediction for intensity-modulated radiotherapy. Med Phys. 2025;52(1):619–32.
41. Romeijn HE, Ahuja RK, Dempsey JF, Kumar A, Li JG. A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. Phys Med Biol. 2003;48(21):3521.
42. Hasenbalg F, Neuenschwander H, Mini R, Born EJ. Collapsed cone Convolution and analytical anisotropic algorithm dose calculations compared to VMC + + Monte Carlo simulations in clinical cases. Phys Med Biol. 2007;52(13):3679.
43. Argyrous G. Statistics for research: with a guide to SPSS. Stat Res. 2011:1–608.
44. Okonogi N, Fukahori M, Wakatsuki M, Ohkubo Y, Kato S, Miyasaka Y, et al. Dose constraints in the rectum and bladder following carbon-ion radiotherapy for uterus carcinoma: a retrospective pooled analysis. Radiat Oncol. 2018;13:1–9.
45. Jadon R, Higgins E, Hanna L, Evans M, Coles B, Staffurth J. A systematic review of dose-volume predictors and constraints for late bowel toxicity following pelvic radiotherapy. Radiat Oncol. 2019;14:1–14.
46. Ryu S, Jin JY, Jin R, Rock J, Ajlouni M, Movsas B, et al. Partial volume tolerance of the spinal cord and complications of single-dose radiosurgery. Cancer: Interdisciplinary Int J Am Cancer Soc. 2007;109(3):628–36.
47. Kang Z, Fu L, Liu J, Shi L, Li Y. A practical method to improve the performance of knowledge-based VMAT planning for endometrial and cervical cancer. Acta Oncol. 2022;61(8):1012–8.
48. Swamidas J, Pradhan S, Chopra S, Panda S, Gupta Y, Sood S, et al. Development and clinical validation of knowledge-based planning for volumetric modulated Arc therapy of cervical cancer including pelvic and Para aortic fields. Phys Imaging Radiation Oncol. 2021;18:61–7.
49. Lee K, Ippolito D, Nystrom A, Zhang C, Eck D, Callison-Burch C, et al. Deduplicating training data makes Language models better. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 8424–8445, Dublin, Ireland. Association for Computational Linguistics.
50. Lee N, Ping W, Xu P, Patwary M, Fung PN, Shoeybi M, et al. Factuality enhanced Language models for open-ended text generation. Adv Neural Inf Process Syst. 2022;35:34586–99.
51. Rawte V, Chakraborty S, Pathak A, Sarkar A, Tonmoy S, Chadha A, et al. The troubling emergence of hallucination in large Language models–an extensive definition, quantification, and prescriptive remediations. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2541–2573, Singapore. Association for Computational Linguistics.
52. Martino A, Iannelli M, Truong C, eds. Knowledge Injection to Counter Large Language Model (LLM) Hallucination. European Semantic Web Conference; 2023: Springer.
53. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive Nlp tasks. Adv Neural Inf Process Syst. 2020;33:9459–74.
54. Ram O, Levine Y, Dalmedigos I, Muhlgay D, Shashua A, Leyton-Brown K, et al. In-context retrieval-augmented Language models. Trans Association Comput Linguistics. 2023;11:1316–31.
55. Kumar S, Malmi E, Severyn A, Tsvetkov Y. Controlled text generation as continuous optimization with multiple constraints. Adv Neural Inf Process Syst. 2021;34:14542–54.
56. Dathathri S, Madotto A, Lan J, Hung J, Frank E, Molino P, et al. Plug and play Language models: A simple approach to controlled text generation. arXiv preprint. 2019. https://doi.org/10.48550/arXiv.1912.02164

Publisher's note