Zero-Shot Large Language Model Agents for Fully Automated Radiotherapy Treatment Planning

Dongrong Yang[1], **Xin Wu**[1], **Yibo Xie**[1], **Xinyi Li**[2], **Qiuwen Wu**[1], **Jackie Wu**[1], **Yang Sheng**[1*]
[1]Department of Radiation Oncology, Duke University, Durham, NC, USA
[2]Department of Radiation Oncology, University of California Irvine, Irvine, CA, USA
dongrong.yang@duke.edu, yang.sheng@duke.edu

Abstract

Radiation therapy treatment planning is an iterative, expertise-dependent process, and the growing burden of cancer cases has made reliance on manual planning increasingly unsustainable, underscoring the need for automation.In this study, we propose a workflow that leverages an large language model (LLM)-based agent to navigate inverse treatment planning for intensity-modulated radiation therapy (IMRT). The LLM agent was implemented to directly interact with a clinical treatment planning system (TPS) to iteratively extract intermediate plan states and propose new constraint values to guide inverse optimization. The agent's decision-making process is informed by current observations and previous optimization attempts and evaluations, allowing for dynamic strategy refinement. The planning process was performed in a zero-shot inference setting, where the LLM operated without prior exposure to manually generated treatment plans and was utilized without any fine-tuning or task-specific training. The LLM-generated plans were evaluated on twenty head-and-neck cancer cases against clinical manual plans with key dosimetric endpoints' statistics analyzed and reported. LLM-generated plans achieved comparable organ-at-risk (OAR) sparing relative to clinical plans, while demonstrating improved hot spot control ($D_{\max}$: 106.5% vs. 108.8%) and superior conformity (conformity index: 1.18 vs. 1.39 for boost PTV; 1.82 vs. 1.88 for primary PTV). This study demonstrates the feasibility of a zero-shot, LLM-driven workflow for automated IMRT treatment planning in a commercial TPS. The proposed approach provides a generalizable and clinically applicable solution that could reduce planning variability and support broader adoption of AI-based planning strategies.

1 Introduction

Radiotherapy remains an essential component of modern cancer management, with evidence-based models indicating that approximately 50–70% of patients should receive at least one course during their disease trajectory, representing millions of new cases globally each year [1, 2]. Global projections forecast a substantial escalation in demand through 2050, further intensifying pressure on already constrained planning resources [3]. In parallel, workforce analyses highlight that this strain extends directly to treatment planning personnel: a national survey of the UK dosimetrist workforce reported planning as the predominant professional responsibility, coupled with recruitment challenges and training bottlenecks that threaten service capacity [4]. Complementary U.S. data document persistent shortages in the medical physics workforce, a core contributor to treatment planning, with implications for plan turnaround times and quality assurance [5]. Similar trends have been observed across

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

Europe, where the ESTRO-HERO study identified widespread staffing deficits within radiotherapy departments [6]. In the context of IMRT and volumetric modulated arc therapy (VMAT), manual planning remains labor-intensive and prone to substantial inter-planner variability in target coverage and organ-at-risk (OAR) sparing, even within the same institution [7, 8].

Consequently, automation of treatment planning has been a long-standing priority within the radiation oncology community, aimed at mitigating workload pressures and reducing inter-planner variability. Broadly, current approaches can be categorized into four paradigms: knowledge-based planning (KBP), which leverages historical plan libraries to predict achievable dose–volume histograms (DVHs) to guide optimization [9–14]; protocol-based planning, which applies standardized objective sets and optimization priorities to generate consistent plans [15–18]; multi-criteria optimization (MCO), which allows planners to navigate trade-offs between competing objectives [19–22]; and reinforcement learning (RL), in which artificial agents iteratively adjust plan parameters based on dosimetric feedback [23–28]. Each approach offers distinct advantages: KBP can embed institutional expertise, protocol-based planning enforces consistency, MCO provides transparent trade-off control, and RL enables adaptive decision-making. However, all suffer from limitations that restrain broad clinical adoption: KBP often require large, high-quality labeled datasets; protocol-based methods lack flexibility for complex or atypical anatomies; MCO demands substantial planner engagement and expertise; and RL approaches can be computationally intensive and require expert-crafted reward functions.

As a result, despite decades of progress, few auto-planning solutions have achieved universal applicability in routine clinical workflows, underscoring the need for more adaptable and generalizable solutions. Large language models (LLMs) have progressed rapidly in recent years, with successive generations achieving remarkable improvements in reasoning and generalization through increased scale and training data diversity [29, 30]. While early LLMs struggled with tasks outside their training distribution, scaling studies have shown that beyond certain thresholds, these models exhibit abrupt gains in capability, including zero-shot and few-shot learning [31–33]. Such capabilities are particularly relevant to radiation therapy treatment planning, a highly specialized domain with complex decision-making requirements and scarce publicly available training data. By leveraging zero-shot and few-shot capabilities, an LLM provided with explicit clinical objectives and integrated into the treatment planning system (TPS) could adapt its reasoning to diverse clinical scenarios without the need for extensive retraining. This learning paradigm offers a distinct advantage over existing approaches, as it does not rely on large expert-labeled datasets or highly engineered domain-specific systems, thereby facilitating broader generalization and transferability once an operational workflow is established.

In this study, we designed and implemented a feasible workflow for leveraging an LLM agent to generate radiation therapy treatment plans entirely in a zero-shot manner, without reliance on prior plans. To capitalize on the LLM's general reasoning capabilities, the complex treatment planning task was decomposed into domain-agnostic subtasks, guided by clinical objectives and broadly applicable planning principles. The agent autonomously extracted intermediate plan states (e.g., DVHs, objective function losses, and dose–volume objectives), analyzed these states using arithmetic and trend-based reasoning, and iteratively proposed updated optimization objectives. Feasibility was demonstrated in 20 head-and-neck (HN) IMRT cases, with key dosimetric endpoints compared against corresponding clinical plans.

2 Methods

2.1 Dataset and Planning Environment

To ensure clinical relevance and broad applicability, the proposed workflow was developed and validated on a widely adopted commercial TPS Eclipse™ (version 15.6, Varian Medical Systems, Palo Alto, CA). The LLM-based agent was designed to iteratively adjust treatment planning parameters within the inverse optimization space to meet clinical objectives and enhance plan quality. The agent interfaces directly with the Eclipse TPS via the Eclipse Scripting Application Programming Interface (ESAPI), enabling programmatic access to the treatment planning environment. Through ESAPI, the agent can retrieve intermediate planning states (e.g., DVH metrics, objective function values) and modify inverse planning constraints in a manner similar to that of a human planner. This integration ensures that all interactions occur within the native TPS environment, maintaining consistency with clinical workflows and eliminating discrepancies that may arise from surrogate optimization engines or approximated planning platforms. We validated the feasibility of the proposed method by applying it to IMRT treatment planning for HN cancer, a particularly challenging site due to the close proximity of multiple critical OARs. The substantial anatomical overlap between the planning target volume (PTV) and nearby OARs necessitates complex, non-intuitive trade-off decisions during plan optimization. The agent was tasked with analyzing the planning status and navigating patient-specific trade-offs to optimize plan quality for each case. Twenty HN patients received IMRT treatment in our institution were retrospectively collected with institutional IRB approval. All patients received the same prescription regimen: 70 Gy to the boost planning target volume ($\mathrm{PTV}_{\mathrm{boost}}$) and 44 Gy to the primary PTV ($\mathrm{PTV}_{\mathrm{primary}}$), delivered in 2 Gy per fraction. The clinical plans were manually generated by certified dosimetrist and the dose distribution was reviewed and approved by the attending radiation oncologist prior to treatment delivery. For automated plan generation, the same target and OAR contours, as well as prescription dose levels, were used to ensure consistency and enable direct comparison with the corresponding clinical plans.

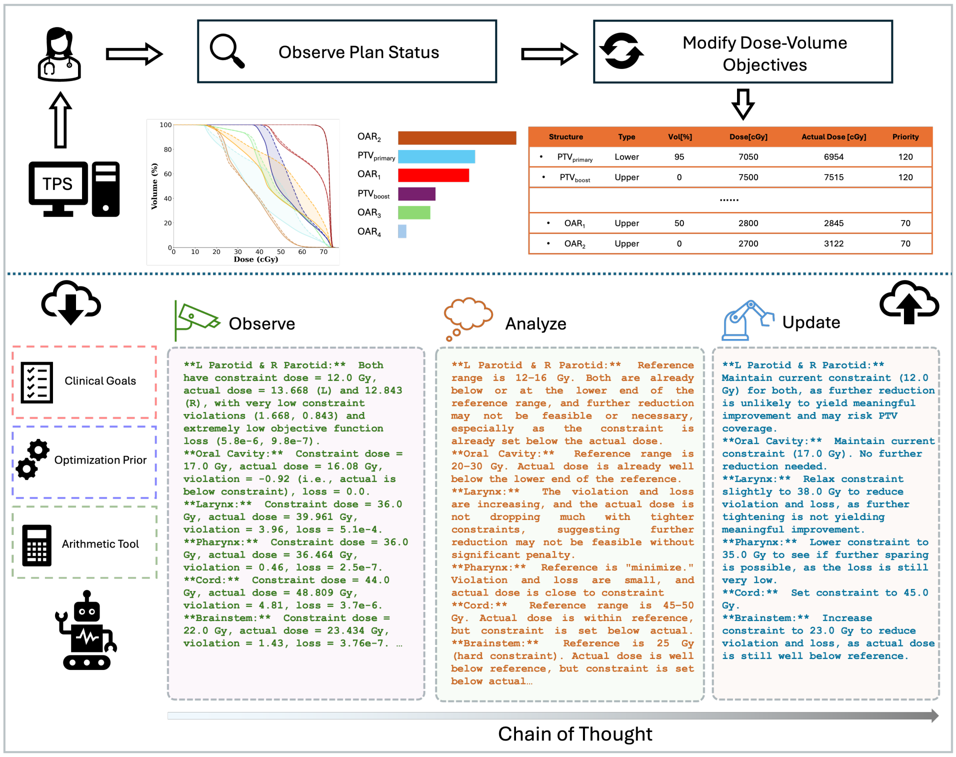
2.2 LLM-based Agentic Workflow for Automatic Inverse Planning

The central concept of the proposed workflow is to leverage an LLM-based agent to iteratively refine optimization objectives during the inverse treatment planning process, with the goal of producing high-quality treatment plans. As shown in Figure 1, the workflow consists of two key components. First, the LLM agent is designed to directly interact with the TPS, allowing it to extract relevant planning information and adjust plan parameters to guide optimization. Second, informed by the current plan status, the agent applies its general reasoning capabilities to propose clinically meaningful modifications that can effectively improve overall plan quality.

To support the LLM's decision-making process, we drew inspiration from standard manual planning workflows. In clinical practice, planners iteratively adjust optimization constraints based on their observations of key dosimetric endpoints extracted from DVH curves, as well as objective function feedback from the optimization engine. The dosimetric endpoints serve as clinically relevant indicators, quantifying how the current dose distribution aligns with prescribed clinical goals. In parallel, the objective function loss is a weight sum of quadratic penalties across all structures and objectives, providing a numerical representation of how much each structure's constraints are being violated, with higher penalties reflecting greater deviation from the specified objectives. These two metrics form the foundation for plan evaluation and constraint adjustment. Based on these metrics, an experienced planner can derive several key insights: (1) the degree of deviation between the current plan and the prescribed clinical goals; (2) whether additional room exists for further plan improvement; and (3) which objectives to adjust and how to do so efficiently to improve plan quality within minimal iterations.

Achieving this level of planning insight requires three core capabilities: (1) arithmetic proficiency to quantify deviations from clinical goals; (2) domain-specific understanding of the optimization system to assess the potential for further improvement; and (3) reasoning ability to interpret trends and current plan status in order to propose targeted constraint adjustments that enhance plan quality effectively and efficiently. Pretrained LLMs possess strong general reasoning capabilities by default but require external support to perform the arithmetic and system-specific evaluation tasks necessary for effective treatment planning [34]. For arithmetic tasks, an arithmetic tool was developed to compute the numerical deviation between current dosimetric endpoint values, clinical goals, and objective constraint settings. Historical data from all prior iterations, including constraints, dosimetric outcomes, and deviations, were compiled and presented to the LLM, enabling trend-based reasoning and informed decision-making.

To enable the LLM agent to interpret the optimization environment, relevant domain information about the inverse planning context was encoded into the prompt. This included explanations of key elements such as the meaning and scale of the objective function loss, as well as how constraint deviations relate to potential improvement opportunities. By incorporating this guidance, the agent was better equipped to assess whether a given plan status suggested further room for optimization or, conversely, indicated excessive sparing of certain OARs. Chain-of-thought reasoning [35] was employed to enhance the agent's ability to perform multi-step decision-making, mirroring the logical processes of a human planner. At each iteration, the LLM was prompted to explicitly articulate its reasoning process before proposing new constraint values. This structured reasoning encouraged the



Structure	Type	Vol[%]	Dose[cGy]	Actual Dose [cGy]	Priority
• PTV$_{primary}$	Lower	95	7050	6954	120
• PTV$_{boost}$	Upper	0	7500	7515	120
……					
• OAR$_1$	Upper	50	2800	2845	70
• OAR$_2$	Upper	0	2700	3122	70

L Parotid & R Parotid: Both have constraint dose = 12.0 Gy, actual dose = 13.668 (L) and 12.843 (R), with very low constraint violations (1.668, 0.843) and extremely low objective function loss (5.8e-6, 9.8e-7).
Oral Cavity: Constraint dose = 17.0 Gy, actual dose = 16.08 Gy, violation = -0.92 (i.e., actual is below constraint), loss = 0.0.
Larynx: Constraint dose = 36.0 Gy, actual dose = 39.961 Gy, violation = 3.96, loss = 5.1e-4.
Pharynx: Constraint dose = 36.0 Gy, actual dose = 36.464 Gy, violation = 0.46, loss = 2.5e-7.
Cord: Constraint dose = 44.0 Gy, actual dose = 48.809 Gy, violation = 4.81, loss = 3.7e-6.
Brainstem: Constraint dose = 22.0 Gy, actual dose = 23.434 Gy, violation = 1.43, loss = 3.76e-7. …

L Parotid & R Parotid: Reference range is 12–16 Gy. Both are already below or at the lower end of the reference range, and further reduction may not be feasible or necessary, especially as the constraint is already set below the actual dose.
Oral Cavity: Reference range is 20–30 Gy. Actual dose is already well below the lower end of the reference.
Larynx: The violation and loss are increasing, and the actual dose is not dropping much with tighter constraints, suggesting further reduction may not be feasible without significant penalty.
Pharynx: Reference is "minimize." Violation and loss are small, and actual dose is close to constraint
Cord: Reference range is 45–50 Gy. Actual dose is within reference, but constraint is set below actual.
Brainstem: Reference is 25 Gy (hard constraint). Actual dose is well below reference, but constraint is set below actual…

L Parotid & R Parotid: Maintain current constraint (12.0 Gy) for both, as further reduction is unlikely to yield meaningful improvement and may risk PTV coverage.
Oral Cavity: Maintain current constraint (17.0 Gy). No further reduction needed.
Larynx: Relax constraint slightly to 38.0 Gy to reduce violation and loss, as further tightening is not yielding meaningful improvement.
Pharynx: Lower constraint to 35.0 Gy to see if further sparing is possible, as the loss is still very low.
Cord: Set constraint to 45.0 Gy.
Brainstem: Increase constraint to 23.0 Gy to reduce violation and loss, as actual dose is still well below reference.

Figure 1: LLM-based agentic workflow for automatic inverse planning. The upper panel illustrates the conventional manual planning workflow, where a human planner iteratively reviews intermediate plan states and adjusts dose–volume constraints. The lower panel depicts the proposed LLM-driven agentic workflow, designed to mimic the manual process. Guided by clinical objectives, prior knowledge of optimization systems, and access to computational tools, the LLM leverages its general reasoning capability to analyze plan status and adapt constraints in a human-like manner. At each iteration, structured chain-of-thought reasoning is applied to enhance decision quality and constraint refinement

model to consider clinical trade-offs, assess constraint violations in context, and prioritize adjustments based on trends observed across prior iterations. By decomposing complex planning decisions into interpretable steps, chain-of-thought prompting improved both the transparency and accuracy of the agent's constraint refinement.With these supporting components, the LLM agent was able to perform treatment planning in a zero-shot setting, where no prior treatment plans were used to train or fine-tune the agent.

2.3 Optimization Setup

A total of 10 key structures were included in the optimization process: the primary and boost planning target volumes PTV$_{\text{primary}}$ and PTV$_{\text{boost}}$, along with nine OARs: the left and right parotid glands, oral cavity, larynx, pharynx, spinal cord (including a 5 mm margin), brainstem, and mandible.

For the parotids, oral cavity, larynx, and pharynx, patient-specific median dose objectives were defined by the attending radiation oncologist. These values reflect the physician's clinical guidelines, informed by anatomical considerations and patient-specific factors, and serve as guiding references rather than strict constraints. In some cases, no specific numeric objective was assigned to certain structures; in such instances, the LLM agent was expected to exercise independent qualitative judgment regarding how much sparing could be achieved without compromising target coverage.

In contrast, hard constraints were applied to the spinal cord (with margin), brainstem, and mandible, with explicit maximum dose limits intended to ensure protection of critical structures and minimize the risk of severe toxicity. These constraints were expected to be strictly enforced during optimization.

The clinical objective values for each patient were provided to the LLM agent as guidance. However, the agent was not expected to simply replicate these objectives in the optimization engine. Rather, it was tasked with meeting or improving upon the clinical goals when possible, identifying opportunities for enhanced OAR sparing without compromising target coverage or violating hard constraints.

2.4 Plan Evaluation

Plan quality was quantitatively assessed by comparing the LLM-generated plans with their corresponding manually generated and clinically approved plans. The evaluated endpoints included the plan-wide maximum dose, the median doses to the bilateral parotid glands, oral cavity, larynx, and pharynx, as well as the maximum doses to critical structures such as the spinal cord (including a 5 mm margin), brainstem, and mandible. In addition to these dose-based metrics, conformity and homogeneity of target coverage were also assessed.

The Conformity Index (CI) [36] was computed as the ratio of the volume covered by the prescription dose (V_{Pre}) to the volume of the planning target (V_{PTV}):

$$CI = \frac{V_{\text{Pre}}}{V_{\text{PTV}}}. \tag{1}$$

The Homogeneity Index (HI) [37] was calculated using the formula:

$$HI = \frac{D_2 - D_{98}}{D_{\text{pre}}}, \tag{2}$$

where D_2 and D_{98} represent the doses received by 2% and 98% of the boost PTV, respectively, and D_{pre} is the prescription dose.To evaluate statistical differences between the LLM-generated and clinical plans, a non-parametric Wilcoxon Signed-Rank test was applied to each dosimetric endpoint, using a significance threshold of $p < 0.05$.

2.5 Experimental Design

We utilized two state-of-the-art language models, GPT-4.1 and GPT-4.1-mini, to validate the efficacy of the proposed workflow. GPT-4.1 represents a frontier in OpenAI's large-scale reasoning model development, offering strong performance across complex multi-step reasoning, mathematical analysis, and domain adaptation tasks. GPT-4.1-mini, while a smaller and more computationally efficient variant, retains core reasoning capabilities but with reduced inference cost and latency. The choice of these two models allows us to benchmark performance across different capacity regimes: GPT-4.1 serving as a high-fidelity upper bound for reasoning quality, and GPT-4.1-mini demonstrating the feasibility of deploying the workflow under more practical computational budgets.

In addition, we designed an ablation study to examine the role of optimization priors in enabling successful application of the workflow. Specifically, the LLM-based agent was tested under two configurations: with optimization priors (including the expected numerical ranges of optimization constraints, tunable parameters, and their directional influence on dose distribution) and without such priors. This design allows us to isolate the contribution of priors in structuring the agent's reasoning process and guiding constraint refinement. By comparing the agent's behavior across these conditions, we can assess whether domain-specific grounding is essential for the LLM to successfully conduct automatic inverse planning, as opposed to relying solely on general reasoning ability.

3 Results

3.1 Dosimetric Endpoints' Comparison

The inter-group dosimetric comparison is summarized in Table 1. Both GPT-4.1 and GPT-4.1-mini with access to optimization priors (GPT-4.1-WP and GPT-4.1-mini-WP) generated plans of clinically comparable quality, with minimal variation in target coverage and overall OAR sparing relative to the clinical reference. GPT-4.1-WP achieved the most favorable numerical performance across the majority of evaluated metrics, reflecting its stronger reasoning capacity and efficiency in conducting treatment planning. Importantly, the absence of optimization priors resulted in a marked deterioration in planning performance: both GPT-4.1 and GPT-4.1-mini produced plans with significantly worse OAR sparing in the condition without access to optimization priors (WOP), as indicated by elevated OAR doses, compared with their prior-informed counterparts. Although GPT-4.1-mini-WOP yielded slightly lower plan maximum dose and improved conformity indices, these apparent gains stemmed from inadequate OAR protection and thus represent an unfavorable trade-off from a clinical perspective.

Table 1: Comparison of clinical plans with LLM-generated plans under different configurations. Reported values represent mean (± standard deviation) for key dosimetric metrics across all patients. Results are shown for GPT-4.1 and GPT-4.1-mini, with (WP) and without (WOP) access to optimization priors. $D_{\max}$: the maximum dose within the structure; D_{50}: the median dose within the structure; CI: conformity index; HI: homogeneity index. For each metric, the optimal value is highlighted in bold.

	Clinical	GPT-4.1-WP	GPT-4.1-WOP	GPT-4.1-mini-WP	GPT-4.1-mini-WOP
Plan $D_{\max}$ (Gy)	76.22(±1.44)	74.53(±1.48)	74.17(±1.20)	74.19(±1.07)	**73.87**(±0.93)
Brainstem $D_{\max}$ (Gy)	**22.13**(±6.65)	24.56(±7.21)	27.57(±7.27)	24.21(±6.63)	28.08(±7.26)
Cord + 5mm $D_{\max}$ (Gy)	44.91(±2.82)	**44.46**(±3.47)	48.87(±3.03)	44.58(±3.97)	49.59(±3.06)
Mandible $D_{\max}$ (Gy)	72.06(±6.94)	**70.86**(±6.94)	71.66(±6.69)	71.17(±6.96)	71.62(±6.42)
Left Parotid D_{50} (Gy)	22.66(±11.22)	**19.21**(±3.09)	23.18(±3.97)	21.93(±5.71)	22.99(±3.92)
Right Parotid D_{50} (Gy)	22.52(±10.17)	**20.47**(±3.64)	24.94(±3.75)	20.70(±5.42)	25.42(±5.97)
Oral Cavity D_{50} (Gy)	36.14(±12.44)	34.95(±10.98)	38.48(±9.09)	**33.26**(±11.45)	39.41(±9.88)
Larynx D_{50} (Gy)	33.16(±14.42)	**29.43**(±8.02)	36.24(±9.36)	31.29(±9.96)	37.83(±11.49)
Pharynx D_{50} (Gy)	47.54(±11.50)	**39.85**(±9.62)	49.18(±7.20)	44.37(±9.04)	49.43(±8.34)
$\text{PTV}_{\text{primary}}$ CI	1.88(±0.29)	1.82(±0.17)	1.92(±0.19)	1.83(±0.17)	1.93(±0.17)
$\text{PTV}_{\text{boost}}$ CI	1.39(±0.19)	1.18(±0.10)	1.17(±0.09)	1.17(±0.09)	**1.16**(±0.09)
$\text{PTV}_{\text{boost}}$ HI	0.061(±0.021)	0.062(±0.021)	0.059(±0.020)	0.058(±0.013)	**0.055**(±0.019)

Based on the comparative evaluation, GPT-4.1-WP was selected as the primary model for subsequent analyses, given its consistently superior dosimetric performance and reasoning stability. The distribution of endpoints for clinical plans and GPT-4.1-WP–generated plans are shown in Figure 2. GPT-4.1-WP plans demonstrated lower median values and shorter interquartile ranges for the target conformity index, indicating more consistent target coverage. For parotid sparing, clinical plans showed wider variability, with some cases exhibiting high median doses, whereas GPT-4.1-WP achieved more consistent dose reduction across patients. This discrepancy arises because, in cases with substantial parotid overlap with the target, physicians sometimes omit explicit numerical constraints and instead provide only general guidance to “minimize” dose, leading planners to prioritize the protection of other OARs. In contrast, even without specific parotid constraints, the LLM-agent consistently pursued parotid sparing although moderately and demonstrated reliable improvements across the test cohort. For other OARs, GPT-4.1-WP plans achieved sparing comparable to that of clinical plans, maintaining dose levels within acceptable ranges without compromising target coverage.

3.2 Case Study and Agent Reasoning

The clinical constraints for a representative sample case are summarized in Table 2.

Table 2: Clinical constraints for a representative example case.

OAR	Right Parotid	Left Parotid	Oral Cavity	Larynx	Pharynx	Cord+5 mm	Brainstem	Mandible
Dose (Gy)	30–35	16	35	25–30	Minimize	45	25	70
Volume(absolute or %)	Median	Median	Median	Median	–	Max	Max	Max

The constraints are heterogeneous in nature: some are expressed as absolute limits, others as ranges, while certain structures have no explicit numerical constraints. For these latter cases, the directive “minimize” indicates that the planner should achieve the lowest dose reasonably achievable. By providing these clinical constraints to the agent, we aim for it to interpret and utilize them in a manner similar to experienced human planners, treating them as reference guidelines and a starting point rather than the ending point, while ultimately striving to achieve optimal plan quality.

The progression of optimization objective adjustments is illustrated in Figure 3. At Step 0, the LLM initialized the optimization by selecting constraint values close to the clinical goals to accelerate

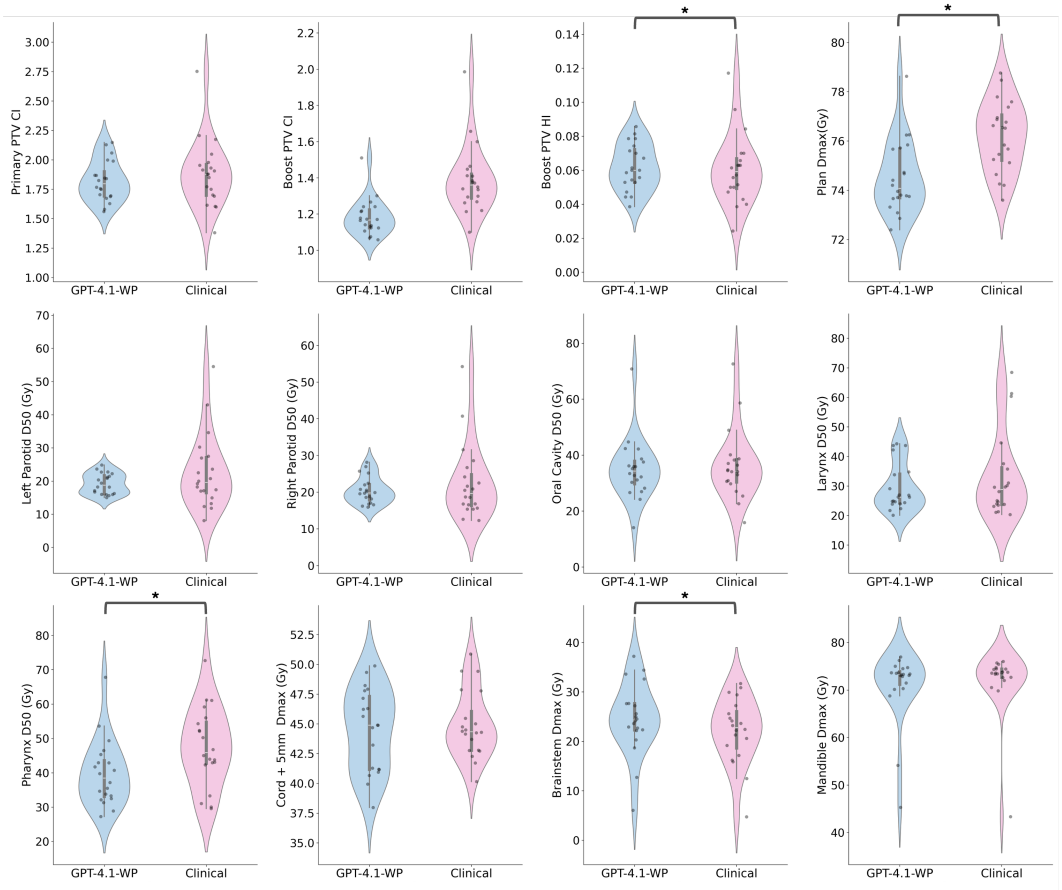


Figure 2: Distribution of dosimetric endpoints for GPT-4.1-WP–generated plans compared with clinical plans. CI: conformity index; HI: homogeneity index; D_{50}: median dose. Asterisks ($*$) indicate statistically significant differences between groups ($p < 0.05$).

convergence. For structures with range-based constraints, the agent chose values at the boundary. For the pharynx, which lacked an explicit numerical constraint, the LLM selected a starting point of 45 Gy median dose, reasoning that this value was "well below the current dose, but not so low as to risk infeasibility." After applying the first set of constraints, the pharynx showed a favorable dose response, prompting the agent to further tighten sparing. In general, the LLM adopted a strategy of using larger step sizes in the early stages to probe the sparing potential of each structure, followed by smaller step sizes in later stages for fine-tuning and to avoid oversparing. For the mandible, the clinical goal was $D_{\max} < 70$ Gy. The agent attempted to lower the optimization constraint to 58 Gy in pursuit of the clinical objective; however, the attained dose plateaued around 70 Gy and was accompanied by a large increase in the objective function loss. From this history, the LLM reasoned that "previous steps show that lowering the constraint further increases violation and loss, and actual dose is not decreasing much. This suggests that further sparing is difficult and may compromise PTV coverage." Consequently, the agent relaxed the mandible constraint to preserve target coverage. A similar trend was observed for the brainstem and cord+5 mm.

Both the attained dose trajectories and the DVH variations confirm that the LLM agent effectively and efficiently improved plan quality within only a few optimization steps, guided by strong and interpretable reasoning.Planning was performed on a workstation with an Intel Xeon CPU and 32 GB RAM, completing in under 5 minutes, significantly faster than manual planning.

The isodose distribution of the LLM-generated plan and the clinical plan is shown in Figure 4. Both plans achieved adequate coverage of the boost PTV and primary PTV, as evidenced by the 70.0 Gy and 44.0 Gy isodose lines in the axial and sagittal views. In the axial slices of the first column,

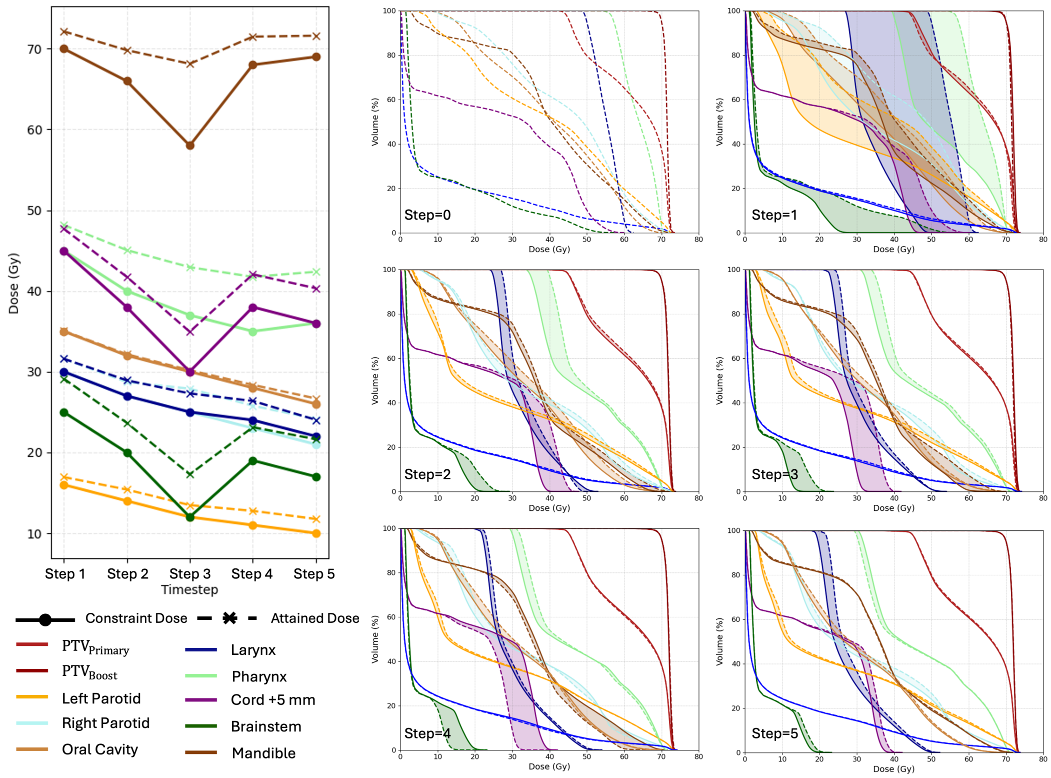


Figure 3: Planning log for the example case. Left panel: trajectories of dose constraints (solid lines) and attained dosimetric endpoints (dashed lines with markers) across optimization steps. Right panels: evolution of DVHs. Step 0 shows the initial plan optimized with PTV constraints only. In subsequent steps, dashed lines indicate the DVHs from the previous step, solid lines represent the updated DVHs, and the shaded regions highlight inter-step DVH changes.

effective spinal cord sparing is evident in the GPT-4.1-WP plan: the 33 Gy isodose line encircles the cord, highlighting a sharp dose fall-off around this critical structure. A similar pattern of dose fall-off and sparing is observed for the larynx in the axial views of the second column.

In the coronal views (third column), both the LLM-generated and clinical plans demonstrate good conformity of high-dose regions to the target volumes, with the 70.0 Gy and 73.5 Gy isodose lines closely wrapping around the boost PTV. The dose fall-off toward adjacent OARs is well preserved in both plans. In the GPT-4.1-WP plan, the parotid glands exhibit a steeper gradient of intermediate isodose lines, suggesting potential for improved sparing without compromising PTV coverage. Overall, the alignment of isodose lines between the two plans indicates that the LLM agent was capable of reproducing clinically acceptable dosimetric trade-offs. While target coverage was comparable between the clinical and GPT-4.1-WP plans, the LLM-generated plan demonstrated sharper dose gradients around the spinal cord and larynx.

4 Discussion

In this study, the experiments were conducted within a widely adopted commercial treatment planning system (Eclipse™, Varian Medical Systems). In contrast to prior work [38] that relied on in-house research platforms, embedding the agent directly into a clinical-grade system enhances both generalizability and translational potential. Furthermore, by constraining the agent to the same information available to human planners and restricting its actions to the parameter adjustment space routinely used in practice, we maximized clinical applicability and interpretability.

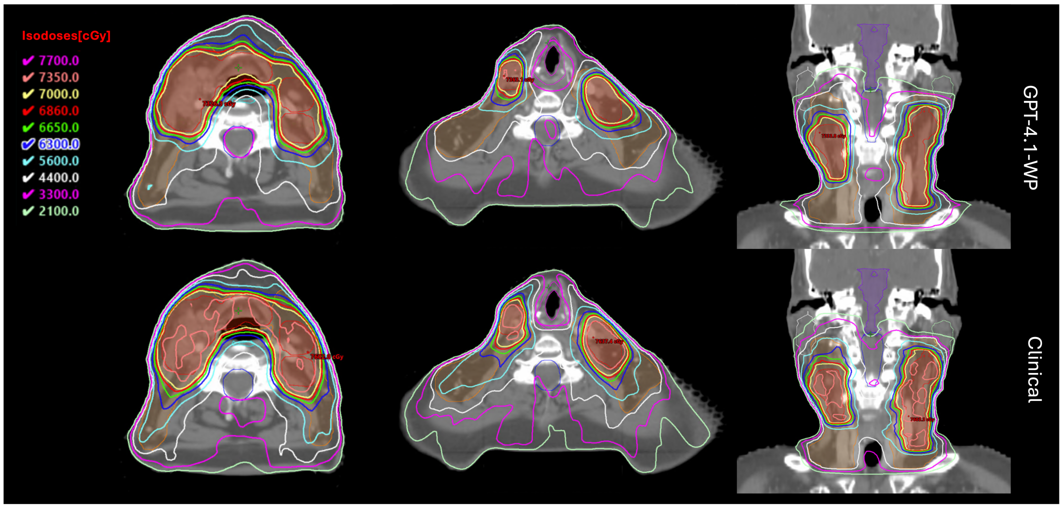


Figure 4: Comparison of isodose distributions between the GPT-4.1-WP–generated plan (left) and the clinical reference plan (right). Red segment: boost PTV; orange segment: primary PTV.

A notable strength of this work is that the experiments were performed entirely in a zero-shot manner. This is an important step toward generalizable AI-driven planning, particularly for centers where access to large, high-quality training datasets may be limited. Wang et al. [39] have previously demonstrated a few-shot LLM-based planning approach in lung and cervical cancer, showing feasibility when prior plans are provided. Our work advances this paradigm by eliminating the need for historical plans altogether, thereby reducing bias from training data size and quality.

The results also underscore a critical insight: while LLMs exhibit strong general reasoning capabilities, their clinical utility depends heavily on the information provided. One key component is the interpretation of clinical constraints. In our institution, clinical constraints are determined by attending physicians based on patient assessment and prior experience. These constraints often represent reference values rather than strictly achievable endpoints. For an LLM agent to make clinically relevant decisions, it must be guided to interpret these objectives as flexible reference points rather than absolute targets, mirroring the approach of experienced human planners. Importantly, practices may vary across institutions depending on local clinical guidelines and physician preferences, and thus institution-specific instructions must be provided to ensure that LLM-driven decisions remain clinically relevant.

A clear understanding of the optimization engine is also essential for successful treatment planning. For example, since the Eclipse engine is driven by a quadratic loss function, effective optimization typically requires setting objectives lower than the desired dose to create a driving force for sparing. The extent of this offset depends on available dosimetric trade-offs and is not intuitive without prior planning experience. Such "hidden rules" are not encoded in the LLM a priori, yet are critical for clinically meaningful outcomes. Clear and structured provision of this knowledge is therefore essential to enable effective autonomous planning, as demonstrated by our results.

By validating this zero-shot workflow in a commercial TPS, we show that LLMs can be deployed across diverse institutions without large training datasets or custom expert systems. This could help alleviate the burden on high-volume centers and provide advanced planning support to smaller centers with limited resources. Nonetheless, this work remains a pilot study. Its efficacy in other disease sites, planning modalities, and clinical environments requires further investigation. Future work will extend this framework to additional tumor sites to validate its robustness and to establish the generalizability of LLM-guided planning beyond head-and-neck cancer.

5 Conclusion

In this study, we proposed a clinical ready workflow that leverages an LLM-based agent to perform inverse treatment planning for IMRT within a commercial TPS in a zero-shot setting. Our results demonstrate that the LLM agent can efficiently generate treatment plans with consistent quality, underscoring its potential as a clinically applicable tool for autonomous radiotherapy planning.

References

[1] M. B. Barton, S. Jacob, J. Shafiq, K. Wong, S. R. Thompson, T. P. Hanna, and G. P. Delaney, "Estimating the demand for radiotherapy from the evidence: a review of changes from 2003 to 2012," *Radiother Oncol*, vol. 112, no. 1, pp. 140–4, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24833561

[2] G. Delaney, S. Jacob, C. Featherstone, and M. Barton, "The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines," *Cancer*, vol. 104, no. 6, pp. 1129–37, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16080176

[3] H. Zhu, M. L. K. Chua, I. Chitapanarux, O. Kaidar-Person, C. Mwaba, M. Alghamdi, A. Rodriguez Mignola, N. Amrogowicz, G. Yazici, Z. Bourhaleb, H. Mahmood, G. M. Faruque, M. Thiagarajan, A. Acharki, M. Ma, M. Harutyunyan, H. Sriplung, Y. Chen, R. Camacho, Z. Zhang, and M. Abdel-Wahab, "Global radiotherapy demands and corresponding radiotherapy-professional workforce requirements in 2022 and predicted to 2050: a population-based study," *Lancet Glob Health*, vol. 12, no. 12, pp. e1945–e1953, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/39401508

[4] N. Blackler, K. E. Bradley, C. Kelly, S. Murphy, C. Cross, and M. Kirby, "A national survey of the radiotherapy dosimetrist workforce in the uk," *Br J Radiol*, vol. 95, no. 1139, p. 20220459, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/36063424

[5] W. D. Newhauser, D. A. Gress, M. D. Mills, D. W. Jordan, S. G. Sutlief, M. C. Martin, and E. Jackson, "Medical physics workforce in the united states," *J Appl Clin Med Phys*, vol. 23 Suppl 1, no. Suppl 1, p. e13762, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/36705248

[6] Y. Lievens, N. Defourny, M. Coffey, J. M. Borras, P. Dunscombe, B. Slotman, J. Malicki, M. Bogusz, C. Gasparotto, C. Grau, H. Consortium, A. Kokobobo, F. Sedlmayer, E. Slobina, P. Coucke, R. Gabrovski, M. Vosmik, J. G. Eriksen, J. Jaal, C. Dejean, C. Polgar, J. Johannsson, M. Cunningham, V. Atkocius, C. Back, M. Pirotta, V. Karadjinovic, S. Levernes, B. Maciejewski, M. L. Trigo, B. Segedin, A. Palacios, B. Pastoors, C. Beardmore, S. Erridge, G. Smyth, and R. Cleries Soler, "Radiotherapy staffing in the european countries: final results from the estro-hero survey," *Radiother Oncol*, vol. 112, no. 2, pp. 178–86, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25300718

[7] S. L. Berry, A. Boczkowski, R. Ma, J. Mechalakos, and M. Hunt, "Interobserver variability in radiation therapy plan output: Results of a single-institution study," *Pract Radiat Oncol*, vol. 6, no. 6, pp. 442–449, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27374191

[8] K. Kubo, H. Monzen, K. Ishii, M. Tamura, Y. Nakasaka, M. Kusawake, S. Kishimoto, R. Nakahara, S. Matsuda, and T. Nakajima, "Inter-planner variation in treatment-plan quality of plans created with a knowledge-based treatment planning system," *Physica Medica*, vol. 67, pp. 132–140, 2019.

[9] L. Yuan, Y. Ge, W. R. Lee, F. F. Yin, J. P. Kirkpatrick, and Q. J. Wu, "Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in imrt plans," *Med Phys*, vol. 39, no. 11, pp. 6868–78, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23127079

[10] D. Nguyen, X. Jia, D. Sher, M.-H. Lin, Z. Iqbal, H. Liu, and S. Jiang, "3d radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture," *Physics in medicine & Biology*, vol. 64, no. 6, p. 065020, 2019.

[11] D. Nguyen, T. Long, X. Jia, W. Lu, X. Gu, Z. Iqbal, and S. Jiang, "A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning," *Sci Rep*, vol. 9, no. 1, p. 1076, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30705354

[12] X. Li, J. Zhang, Y. Sheng, Y. Chang, F. F. Yin, Y. Ge, Q. J. Wu, and C. Wang, "Automatic imrt planning via static field fluence prediction (aip-sffp): a deep learning algorithm for real-time prostate treatment planning," *Phys Med Biol*, vol. 65, no. 17, p. 175014, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32663813

[13] X. Li, Y. Sheng, Q. J. Wu, Y. Ge, D. M. Brizel, Y. M. Mowery, D. Yang, F. Yin, and Q. Wu, "Clinical commissioning and introduction of an in-house artificial intelligence (ai) platform for automated head and neck intensity modulated radiation therapy (imrt) treatment planning," *Journal of Applied Clinical Medical Physics*, p. e14558, 2024.

[14] D. Yang, C. Murr, X. Li, S. Yoo, R. Blitzblau, S. Mcduff, S. Stephens, Q. J. Wu, Q. Wu, and Y. Sheng, "Understanding and modeling human-ai interaction of artificial intelligence tool in radiation oncology clinic using deep neural network: a feasibility study using three year prospective data," *Physics in Medicine and Biology*, 2024.

[15] C. Boylan and C. Rowbottom, "A bias-free, automated planning tool for technique comparison in radiotherapy-application to nasopharyngeal carcinoma treatments," *Journal of Applied Clinical Medical Physics*, vol. 15, no. 1, pp. 213–225, 2014.

[16] I. Xhaferllari, E. Wong, K. Bzdusek, M. Lock, and J. Z. Chen, "Automated imrt planning with regional optimization using planning scripts," *Journal of applied clinical medical physics*, vol. 14, no. 1, pp. 176–191, 2013.

[17] H. Yan, F. Yin, H. Guan, and J. H. Kim, "Fuzzy logic guided inverse treatment planning," *Medical physics*, vol. 30, no. 10, pp. 2675–2685, 2003.

[18] J. Krayenbuehl, M. Di Martino, M. Guckenberger, and N. Andratschke, "Improved plan quality with automated radiotherapy planning for whole brain with hippocampus sparing: a comparison to the rtog 0933 trial," *Radiation Oncology*, vol. 12, pp. 1–7, 2017.

[19] S. Breedveld, D. Craft, R. van Haveren, and B. Heijmen, "Multi-criteria optimization and decision-making in radiotherapy," *European Journal of Operational Research*, vol. 277, no. 1, pp. 1–19, 2019.

[20] B. S. Müller, H. A. Shih, J. A. Efstathiou, T. Bortfeld, and D. Craft, "Multicriteria plan optimization in the hands of physicians: a pilot study in prostate cancer and brain tumors," *Radiation Oncology*, vol. 12, pp. 1–11, 2017.

[21] J. Wala, D. Craft, J. Paly, A. Zietman, and J. Efstathiou, "Maximizing dosimetric benefits of imrt in the treatment of localized prostate cancer through multicriteria optimization planning," *Medical Dosimetry*, vol. 38, no. 3, pp. 298–303, 2013.

[22] S. Ghandour, O. Matzinger, and M. Pachoud, "Volumetric-modulated arc therapy planning using multicriteria optimization for localized prostate cancer," *Journal of applied clinical medical physics*, vol. 16, no. 3, pp. 258–269, 2015.

[23] C. Shen, L. Chen, Y. Gonzalez, and X. Jia, "Improving efficiency of training a virtual treatment planner network via knowledge-guided deep reinforcement learning for intelligent automatic treatment planning of radiotherapy," *Med Phys*, vol. 48, no. 4, pp. 1909–1920, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33432646

[24] Y. Gao, C. Shen, X. Jia, and Y. Kyun Park, "Implementation and evaluation of an intelligent automatic treatment planning robot for prostate cancer stereotactic body radiation therapy," *Radiother Oncol*, vol. 184, p. 109685, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/37120103

[25] J. Zhang, C. Wang, Y. Sheng, M. Palta, B. Czito, C. Willett, J. Zhang, P. J. Jensen, F. F. Yin, Q. Wu, Y. Ge, and Q. J. Wu, "An interpretable planning bot for pancreas stereotactic body radiation therapy," *Int J Radiat Oncol Biol Phys*, vol. 109, no. 4, pp. 1076–1085, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33115686

[26] D. Yang, X. Wu, X. Li, R. Mansfield, Y. Xie, Q. Wu, Q. J. Wu, and Y. Sheng, "Automated treatment planning with deep reinforcement learning for head-and-neck (hn) cancer intensity modulated radiation therapy (imrt)," *Physics in Medicine and Biology*, 2024.

[27] D. Yang, X. Li, S. Yoo, R. Blitzblau, S. McDuff, S. Stephens, P. Segars, Q. Wu, F. Yin, and Y. Sheng, "A reinforcement learning approach to automate breast radiation therapy treatment planning using electronic compensation (ecomp)," *International Journal of Radiation Oncology, Biology, Physics*, vol. 120, no. 2, p. S63, 2024.

[28] W. T. Hrinivich and J. Lee, "Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning," *Med Phys*, vol. 47, no. 12, pp. 6140–6150, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33070336

[29] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nat Med*, vol. 29, no. 8, pp. 1930–1940, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/37460753

[30] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, and Z. Dong, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, 2023.

[31] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, and D. Metzler, “Emergent abilities of large language models,” *arXiv preprint arXiv:2206.07682*, 2022.

[32] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, vol. 1, no. 3, p. 3, 2020.

[33] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” *arXiv preprint arXiv:2109.01652*, 2021.

[34] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, and S. Anadkat, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.

[35] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[36] L. Feuvret, G. Noël, J.-J. Mazeron, and P. Bey, “Conformity index: a review,” *International Journal of Radiation Oncology* Biology* Physics*, vol. 64, no. 2, pp. 333–342, 2006.

[37] T. Kataria, K. Sharma, V. Subramani, K. Karrthick, and S. S. Bisht, “Homogeneity index: An objective tool for assessment of conformal radiation treatments,” *Journal of medical physics*, vol. 37, no. 4, pp. 207–213, 2012.

[38] S. Liu, O. Pastor-Serrano, Y. Chen, M. Gopaulchan, W. Liang, M. Buyyounouski, E. Pollom, Q. T. Le, M. Gensheimer, P. Dong, Y. Yang, J. Zou, and L. Xing, “Automated radiotherapy treatment planning guided by gpt-4vision,” *Phys Med Biol*, vol. 70, no. 15, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/40664228

[39] Q. Wang, Z. Wang, M. Li, X. Ni, R. Tan, W. Zhang, M. Wubulaishan, W. Wang, Z. Yuan, Z. Zhang, and C. Liu, “A feasibility study of automating radiotherapy planning with large language model agents,” *Phys Med Biol*, vol. 70, no. 7, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/40073507

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data consist of clinical radiation therapy treatment plans and associated patient records, which cannot be shared publicly due to patient privacy regulations (e.g., HIPAA/GDPR) and institutional review board (IRB) restrictions. Data access is therefore limited to protect patient confidentiality.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.