RO-LLaMA: Generalist LLM for Radiation Oncology via Noise Augmentation and Consistency Regularization

Kwanyoung Kim*[1], Yujin Oh*[1], Sangjoon Park*[2,3], Hwa Kyung Byun[4], Jin Sung Kim†[2,5], Yong Bae Kim†[2], Jong Chul Ye†[1],
Kim Jae Chul Graduate School of AI, KAIST[1], Department of Radiation Oncology, Yonsei University College of Medicine[2], Institute for Innovation in Digital Healthcare, Yonsei University[3], Department of Radiation Oncology, Yongin Severance Hospital[4], Oncosoft Inc.[5]
{cubeyoung, yujin.oh, jong.ye}@kaist.ac.kr, {depecher89, khbuyn05, jinsung, ybkim3}@yuhs.ac

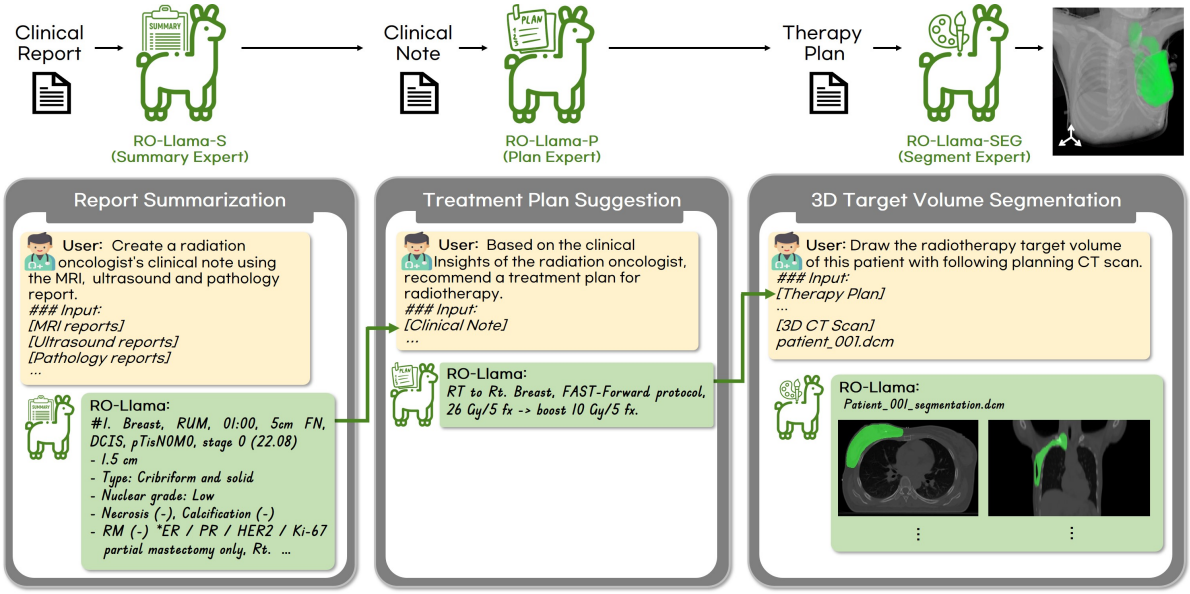


Figure 1. RO-LLaMA as a generalist large language model (LLM) in the field of radiation oncology. The model seamlessly covers various tasks such as clinical report summarization, radiation treatment plan suggestion, and plan-guided target volume segmentation. All the patient information and any potential personally identifiable information are de-identifiied, and the clinical data usage is approved by IRB.

Abstract

Recent advancements in Artificial Intelligence (AI) have profoundly influenced medical fields, by providing tools to reduce clinical workloads. However, most AI models are constrained to execute uni-modal tasks, in stark contrast to the comprehensive approaches utilized by medical professionals. To address this, here we present RO-LLaMA, a versatile generalist large language model (LLM) tailored for the field of radiation oncology. This model seamlessly covers a wide range of the workflow of radiation oncologists, adept at various tasks such as clinical report summarization, radiation treatment plan suggestion, and plan-guided target volume segmentation. In particular, to maximize the end-to-end performance, we further present a novel Consistency Embedding Fine-Tuning (CEFTune) technique, which boosts LLM's robustness to additional errors at the intermediates while preserving the capability of handling clean inputs, and creatively transform this concept into LLM-driven segmentation framework as Consistency Embedding Segmentation (CESEG). Experimental results on multi-centre cohorts demonstrate our RO-LLaMA's promising performance for diverse tasks with generalization capabilities.

*These authors contributed equally to this work
†Corresponding authors

1. Introduction

Recently, the emergence of a new generation of AI models, known as foundation models, marks a significant departure from previous paradigms [34]. These foundation models are now capable of achieving state-of-the-art performance (SOTA) in a wide range of domains, including tasks such as multi-modal reasoning, image-to-text generation, image captioning, and text-guided image segmentation [3, 10, 12, 25, 27, 31].

These characteristics signify a potential paradigm shift in how AI can be integrated into medical practices, which inherently rely on multi-modal information for comprehensive clinical decision-making. Furthermore, this gives an opportunity of overcoming limitations of now over 500 FDA-approved AI models, which are mainly specialized for a specific task with uni-modal information [21]. Specifically, in contrast to these uni-modal AIs, generalist medical AIs can encompass a holistic understanding of clinical workflows, which can receive and reason a variety of medical data, including imaging modalities, electronic health records, laboratory results, genomics, and even clinical reports [34, 39, 41, 46, 50].

Inspired by this paradigm shift in medical AIs, here we present RO-LLaMA, a prototype of a generalist medical AI model, tailored for the clinical workflow in radiation oncology. RO-LLaMA is adept at performing a variety of clinical tasks: (1) It efficiently summarizes extensive patient histories and examination results into concise but informative clinical notes. Additionally, it is capable of (2) proposing appropriate treatment plans from a clinical expert perspective and (3) delineating radiation target volumes on 3-dimensional (3D) computed tomography (CT) scans consistent with the proposed treatment plans. This multifaceted functionality demonstrates a significant alignment with the expertise of clinical professionals.

In the course of the aforementioned sequential tasks, the inevitable accumulation of errors in generations arises. To address this issue, we firstly explore Noisy Embedding Fine-Tuning (NEFTune) [19] which involves injecting uniform noise into embeddings during the training for our targeted tasks. To further enhance the model's applicability, we introduce a novel consistency regularization method, bolstered by our pioneering Consistency Embedding Fine-Tuning (CEFTune) technique, which add regularization loss to enforce the consistency between the prediction given noisy and clean inputs. Expanding beyond text-related tasks, we apply these concepts to 3D segmentation tasks, resulting in Noisy Embedding Segmentation (NESEG) and Consistency Embedding Segmentation (CESEG). These advancements collectively contribute to a marked improvement in the model's generalization capabilities in both internal and external validation. Our contributions can be summarized as:

- We propose a comprehensive framework, denoted as RO-LLaMa, wherein LLM facilitates the entire workflow of radiation oncology.
- We explore noise augmentation and consistency, and propose a novel training approach, such as CEFTune, NESEG, and CESEG.
- Through experiments on both internal and external datasets, we demonstrated that RO-LLaMA outperforms baseline methods.

2. Related Works

Instruction Fine-tuning. Instruction fine-tuning has been emerged as a pivotal technique for augmenting model responsiveness in the field of natural language processing (NLP). The simplicity and effectiveness of this approach have been introduced in numerous works, with a notable scale-up facilitated by advancements of LLMs such like GPT-3 [2], ChatGPT [36], and GPT-4 [37]. Notably, a pioneering contribution, Self-Instruct [47], involves fine-tuning of foundation models using instruction-output pairs generated from InstructGPT [38]. This methodology, along with similar approaches, has led to diverse language model variants, including Alpaca [42], Vicuna [4], Dolly [8], and a series of LLaMA [43, 45], by exhibiting promising performance across a wide range of tasks. In the medical domain, Chat-Doctor [53], Med-Alpaca [14], PMC-LLaMA [49], and Asclepius [24] have been fine-tuned for clinical question-answering (QA) task. Although these models demonstrate robust performance for diverse QA benchmarks, there is a notable gap in their applicability for the practical clinical workflow in a specific domain, such as radiation oncology, which serves as the targeted focus of this work.

Stabilizing LLM Fine-tuning with Noise. To enhance the robustness of language models against noisy input, various strategies have been explored. Approaches such as SMART [20], FreeLB [57], and R3F [1] employ adversarial training methods, introducing small Gaussian perturbations in embedding dimensions to optimize the model's performance against noise. Aparting from the adversarial training, LSNR [17] takes a different approach by directly optimizing the smoothness of each layer of the language model. More recently, NEFTune [19] improves LLM performance during fine-tuning by simply adding random noise into the embedding vectors during training. In the context of this study, we extend NEFTune by incorporating the concept of consistency regularization. This extension aims to further improve the robustness and generalization capabilities of language models when faced with noisy input.

Language-driven Image Segmentation. Recent research in the field of image segmentation has been emerged to incorporate linguistic ability, such like language-driven

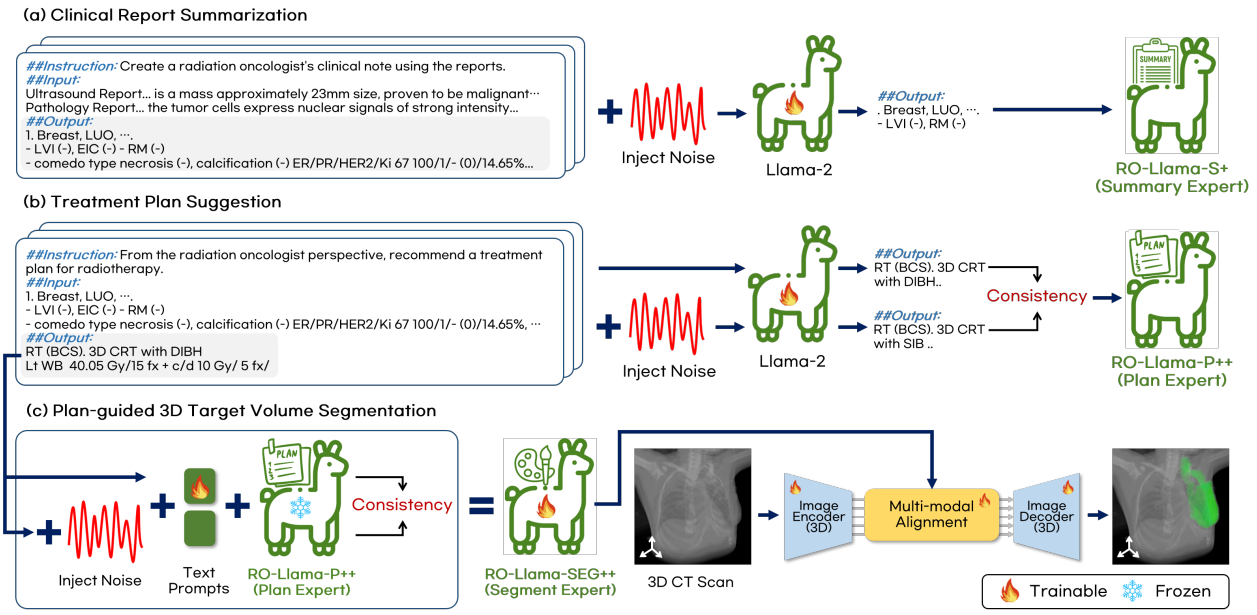


Figure 2. Training schematics of the generalist RO-LLaMA for three different tasks. (a) RO-LLaMA-S+ for clinical note summarization. (b) RO-LLaMA-P++ for radiation treatment plan suggestion. (c) RO-LLaMA-SEG++ for plan-guided target volume segmentation.

semantic segmentation [26], open-vocabulary segmentation [11, 29, 52], referring segmentation [48], and reasoning segmentation [25]. Language-driven segmentation has triggered a paradigm shift in medical domain, where multi-modal knowledge is inevitable. For instance, LViT [28] and ConTEXTualNet [18] introduce text-driven chest X-ray radiography segmentation.

Target volume segmentation in the field of radiation oncology is more challenging, due to its intrinsic need for considerations of the clinical aspects beyond the image, such as overall cancer stage, treatment aim, pathologic findings, and so on [5, 6, 15]. Recent research on clinical context-aware breast cancer radiotherapy delineation has demonstrated that the multi-modal AI outperforms the traditional uni-modal AI by a substantial margin, particularly when labeled datasets are scarce [35]. Unlike previous approaches, we incorporate treatment plans generated by LLM directly from clinical reports, which aligns seamlessly with actual clinical workflows in radiaion oncology.

3. Methods

In this section, we provide a detailed methodology for our proposed approach, which is designed for sequential text generation tasks, including summarization and suggestions, as well as text-driven image segmentation, as illustrated in Figure 2. In Sec 3.1, we describe text-related fine-tuning apporaches that yield report summarization and plan suggestion, respectively. In Sec 3.2, we introduce a novel segmentation approach that involves injecting noise and regularizing consistency to the text embedding for improving robustness of plan-guided 3D target volume segmentation.

3.1. LLM Fine-tuning for Clinical Generalist

To realize the generalist LLM with expertise in clinical report summarization and radiation treatment plan suggestion, we conduct instruction fine-tuning for LLaMA2 [45]. Considering the intended objective of each task, we adopt separate strategies to acquire task-specific expertise, namely RO-LLaMA-S (summary expert) and RO-LLaMA-P (plan expert). When the summary expert gets clinical reports as training inputs, the collected reports are raw and unrefined, constituting noisy data for the model. To enhance robustness for the noisy input, we employ Noisy Embedding Fine-Tuning (NEFTune) [19], resulting in RO-LLaMA-S+[†].

In the case of the plan expert, we choose to utilize the train set composed of collected notes instead of generated notes, primarily due to cost considerations. In contrast to training, our model takes the generated notes as input for inference. Consequently, adopting NEFTune is also an effective solution to handle noisy inputs in this task. However, a crucial consideration arises from the nature of the generated notes, which may lie closer to clean inputs (collected notes) or deviate towards noisy inputs. To address this, it is essential to train the model to handle both clean and noisy inputs. To preserve the robustness facilitated by NEFTune while enforcing consistency between the prediction given clean and noisy inputs, we introduce Consis-

[†] '+' , '++' denotes the adoption of NEFTune, and CEFTune, respectively

tency Embedding Fine-Tuning (CEFTune), resulting in RO-LLaMA-P++. More details are as follows.

Revisiting Noise Embedding Fine-Tuning. NEFTune has been recently introduced to enhance the performance of instruction fine-tuning. This approach involves injecting a random noise vector into embeddings during the training process as follows:

$$\mathcal{L}_{\text{NEFTune}}(\theta) = \mathbb{E}_{(\mathbf{x}_{\text{txt}}^{\text{emb}}, \mathbf{y}_{\text{txt}}) \sim D} \mathcal{L}_{\text{ce}}(f_\theta(\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}}), \mathbf{y}_{\text{txt}}),$$
$$\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}} = \mathbf{x}_{\text{txt}}^{\text{emb}} + (\alpha/\sqrt{LC})\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{U}(-1, 1) \tag{1}$$

where $\mathbf{x}_{\text{txt}}^{\text{emb}} \in \mathbb{R}^{B \times L \times C}$ is the embedding of data, B denotes batch size, L is token length, C is embedding dimension, $\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}}$ is perturbed embedding, $f_\theta(\cdot)$ represents the model parameterized by θ, $\mathbf{y}_{\text{txt}}$ is the label of text sample. They demonstrated that the effectiveness of incorporating a random noise vector adjusted based on token length, which yields robust results in most fine-tuning scenarios.

Consistency Embedding Fine-Tuning To bolster the model's capability in handling both noisy and clean inputs, we incorporate a regularization loss to encourage consistency as follows:

$$\begin{aligned} \mathcal{L}_{\text{CEFTune}}(\theta) &= \mathcal{L}_{\text{NEFTune}}(\theta) + \lambda \mathcal{R}(\theta, \theta^-), \\ \text{where } \mathcal{R}(\theta, \theta^-) &= d(\mathcal{T}(f_\theta(\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}})), \mathcal{T}(f_{\theta^-}(\mathbf{x}_{\text{txt}}^{\text{emb}}))) \end{aligned} \tag{2}$$

where θ^- represents the model with stopped gradient, and $d(\cdot, \cdot)$ is used to quantify the discrepancy between $f_\theta(\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}})$ and $f_{\theta^-}(\mathbf{x}_{\text{txt}}^{\text{emb}})$. The function $\mathcal{T}$ serves as either the identity function or performs detokenization to generate sentences. Our objective is to preserve the robustness property of NEFTune while introducing semantic similarity through the integration of consistency between the output of noisy inputs and clean inputs. The metric function d can be any capable of measuring distance in a specific space. However, solely minimizing distance in the same embedding space might weaken the robustness effect and pose limitations on enforcing semantic similarity. In our approach, to capture textual similarity between sentences, we calculate the distance in the feature space of SentenceBERT [40]. *i.e.*, $d(\mathcal{T}(x), \mathcal{T}(y)) = 1 - s(\mathcal{T}(x))^\top s(\mathcal{T}(y))/||s(\mathcal{T}(x))||_1||s(\mathcal{T}(y))||_1$, where $s(\cdot)$ is the projection by using pretrained model such as SentenceBERT [40], $\mathcal{T}$ is set to detokenization operator.

3.2. LLM-driven 3D Target Volume Segmentation

For incorporating the textual information into target volume segmentation framework, we expand the concept of prompt tuning of LLM for context-aware 3D segmentation framework introduced in [35]. However, as the LLM-driven 3D segmentation model gets the generated treatment plan as text conditions which are inevitably noisy from the previous summarization step, we introduce noise augmentation and consistency regularization modules for the target volume segmentation network, namely, Noise Embedding Segmentation (NESEG) and Regularized Consistency Embedding Segmentation (CESEG).

Noise Embedding Segmentation (NESEG). Firstly, by extending aforementioned idea of NEFTune, we inject a random noise vector into input text embeddings to improve the segmentation model robustness for the generated noisy plan as the text condition. The loss function for NESEG can be formulated as:

$$\mathcal{L}_{\text{NESEG}}(\Theta) = \mathbb{E}_{(\mathbf{x}_{\text{img}}, \tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}}, \mathbf{y}_{\text{img}}) \sim D} \mathcal{L}_{\text{ce}}(g_\psi(\mathbf{x}_{\text{img}}; \tilde{\mathbf{y}}_{\text{txt}}^{\text{emb}}), \mathbf{y}_{\text{img}})$$
$$\text{where } \tilde{\mathbf{y}}_{\text{txt}}^{\text{emb}}(\phi) = f_{\theta^*}(\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}}; \mathbf{z}_\phi) \tag{3}$$

where $\Theta = [\psi, \phi]$, $\mathbf{x}_{\text{img}} \in \mathbb{R}^{B \times HWS}$ is the 3D CT scan, B denotes batch size, H, W, and S correspond to height, width, and slice of the CT scan, $\tilde{\mathbf{y}}_{\text{txt}}^{\text{emb}} \in \mathbb{R}^{B \times 1 \times C}$ is the perturbed LLM output embedding, $\mathbf{y}_{\text{img}} \in \mathbb{R}^{B \times HWS}$ is the 3D ground truth segmentation mask, $g_\psi(\cdot)$ represents segmentation expert model, denoted as the RO-LLaMA-SEG+, parameterized by ψ, and $\mathbf{z}_\phi$ is the learnable text prompt.

Consistency Embedding Segmentation (CESEG). We further adapt the consistency regularization module, defined as CESEG, for generalizing our segmentation model to both the generated noisy plan and the clean ground truth plan as for text condition. Apart from the original concept of CEFTune, we modified the consistency regularization module for the multi-modal segmentation task, which combines CEFTune and the text prompt tuning. Once the text prompt-prepended noisy embeddings of treatment plan $\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}}$ are inputted to the frozen LLM, the output embedding is regularized with that of clean embeddings $\mathbf{x}_{\text{txt}}^{\text{emb}}$. The loss function for CESEG is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{CESEG}}(\Theta) &= \mathcal{L}_{\text{NESEG}}(\Theta) + \lambda \mathcal{R}(\phi, \phi^-) \\ \text{where } \mathcal{R}(\phi, \phi^-) &= d(\tilde{\mathbf{y}}_{\text{txt}}^{\text{emb}}(\phi), \mathbf{y}_{\text{txt}}^{\text{emb}}(\phi^-)) \end{aligned} \tag{4}$$

where, ϕ^- represents the learnable text prompts with stopped gradient and $\mathbf{y}_{\text{txt}}^{\text{emb}} = f_{\theta^*}(\mathbf{x}_{\text{txt}}^{\text{emb}}; \mathbf{z}_{\phi^-}) \in \mathbb{R}^{B \times 1 \times C}$ is the LLM embedding given the clean plan as an input.

4. Experiments

4.1. Dataset

To train our model, we collected internal data cohort composed of 5,674 breast cancer patients treated at the Department of Radiation Oncology at Yonsei Cancer Center. For training the plan-guided 3D target volume segmentation network, we utilized multi-modal data from 599 patients with their 3D CT scans and corresponding ground-truth masks. To evaluate the model performance on cross-centre datasets, we further acquired external data cohort

Table 1. The quantitative results for clinical note summarization using different methods. R-1, R-2, R-L denote ROUGE-1, ROUGE-2, and ROUGE-L, respectively.

Model	Method	Internal Validation (N=60)					External Validation (N=81)				
		R-1 ↑	R-2 ↑	R-L ↑	BERTScore ↑	BARTScore ↑	R-1 ↑	R-2 ↑	R-L ↑	BERTScore ↑	BARTScore ↑
LLaMA2 [45]	-	0.509	0.346	0.508	-0.158	-5.81	0.667	0.467	0.666	-0.128	-5.66
MedAlpaca [14]	-	0.533	0.369	0.532	-0.086	-5.53	0.414	0.328	0.414	-0.05	-5.46
ChatDoctor [53]	-	0.620	0.413	0.619	-0.085	-5.66	0.553	0.428	0.552	0.025	-5.31
Asclepius [24]	-	0.664	0.448	0.664	-0.115	-5.89	0.596	0.416	0.595	-0.118	-5.91
PMC-LLaMA [49]	-	0.400	0.241	0.397	-0.24	-6.07	0.384	0.235	0.378	-0.256	-6.01
ChatGPT	-	0.674	0.468	0.674	-0.021	-5.48	0.656	0.460	0.655	-0.039	-5.59
RO-LLaMA-S	Naive	0.801	0.610	0.800	0.293	-4.03	0.753	0.580	0.752	0.255	-4.22
RO-LLaMA-S+	NEFTune	**0.823**	**0.634**	**0.822**	0.319	-3.94	**0.804**	**0.634**	**0.804**	0.317	-3.92
RO-LLaMA-S++	CEFTune	0.819	0.627	0.817	**0.319**	**-3.92**	0.793	0.631	0.792	**0.334**	**-3.86**

Table 2. Example case for the clinical report summarization task. The red scripts indicate incorrect or redundant information, and inconsistent formatting compared to the label, as evaluated by the board certified clinical expert.

Clinical Report Summarization Example	
User	Using the provided MRI and pathology findings, design a radiation oncologist's clinical overview. # Input: [MRI reports]. . . [Pathology reports] . . .
Ground Truth	#1. Breast, LUM, 11:00, 4cm FN, IDC, pT1cN0M0 - Maximum diameter of invasive tumor: 1.7cm - Maximum diameter of invasive and in situ carcinoma: 1.7cm - H1N2 - LVI (-) - Intraductal carcinoma component - Proportion: 20% - Extensive intraductal component (EIC) : Absent - Type: Cribriform and solid - Nuclear grade: Intermediate - Necrosis: Absent - Location: Central - Calcification: Present (Malignant) - Resection margins - Nipple, upper, outer, and inner margin: Free of carcinoma - ER/PR/Her-2/Ki-67 (90/-/2+, SISH-/11.32)...
Ours	#1. Breast, LUM 11:00, 2cm FN, IDC, pT1N0M0, stage IA - T : 1.7cm (in situ 1.7cm) H1N2 - N : Total (0/5) : SLN (0/4) nonSLN (0/1) - LVI (-) - Intraductal carcinoma component - Proportion: 20% - Extensive intraductal component (EIC) : Absent - Type: Cribriform and solid - Nuclear grade: Intermediate - Necrosis: Absent - Location: Central -Calcification: Present (Malignant) -Resection margins - Nipple, upper, outer, and inner margin: Free of carcinoma. - ER/PR/HER2/Ki67 : 90%/-/2+ SISH-/11.32% Lum B (HER2-) type...
Chat-GPT	#1 Breast, LUM, 11:00, 7cm FN, IDC, icT1N0M0, Stage IA -LUM 07:00 - Invasive ductal carcinoma - ER: 90%, PR: Negative, HER2: Equivocal (2+), Ki-67: 11.32% - Present malignant calcification - Resection margins of nipple, upper, outer, and inner margin: Free of carcinoma - Total LNs (0/5): Left sentinel No1 (0/3), left sentinel No2 (0/1), and left axillary No1 (0/1): Free of carcinoma. . .

Table 3. Details of data for training our generalist RO-LLaMA. "US" denotes Ultrasound. "Path" indicates Pathology.

Task	Input	Output	Internal		External
			Train	Val	Val
Summarization Plan	MRI & US & Path Reports Clinical Note	Clinical Note treatment plan	5,674	60	81
Segmentation	treatment plan & CT Scan	Mask	593	79	81

composed of 81 patients treated at the Department of Radiation Oncology at Yongin Severance Hospital. The detail of dataset is described in Table 3 and Supplementary Material. We included patients who received their initial diagnosis of breast cancer and subsequently underwent radiation therapy following curative surgery, while excluding individuals with recurrent or metastatic breast cancer in both hospitals. This study was approved by the Institutional Review Board (IRB) of each participating hospitals.

4.2. Implementation details

For instruction fine-tuning for RO-LLaMA-S and RO-LLaMA-P, we use the LLaMA-2-7B-Chat [44] model as a baseline. The maximum context length is set at 4096, and the batch size is set to 2. To train RO-LLaMA-SEG, we employ the 3D U-Net [7] using the open-source library MONAI[†] and the LLaMA-2-7B-Chat model initialized with the pre-trained checkpoint of RO-LLaMA-P++. For training RO-LLaMA-SEG, LLM is frozen, while other network parameters including the text prompts are optimized. Further details are deferred to Supplementary Material.

We use the AdamW [33] optimizer for all the tasks, with a learning rate of 5e-5 until reaching 3 epochs for both RO-LLaMA-S and RO-LLaMA-P, and 1e-4 until reaching 100 epochs for RO-LLaMA-SEG, leveraging 4 NVIDIA A6000 GPUs for each task. The hyper-parameter λ for CEFTune is set to 1 in all of tasks, and for consistency regularization we adopt the variants of SBERT [40] which is trained on PubMed[†] dataset, called PubMedBERT.[†]

[†] https://monai.io/
[†] https://www.ncbi.nlm.nih.gov/pubmed/
[†] https://huggingface.co/NeuML/pubmedbert-base-embeddings.

Table 4. The quantitative results for treatment plan suggestion using different methods. R-1, R-2, R-L denote ROUGE-1, ROUGE-2, and ROUGE-L, respectively.

Model	Method	Internal Validation (N=60)						External Validation (N=81)					
		R-1 ↑	R-2 ↑	R-L ↑	BERTScore ↑	BARTScore ↑	MoverScore ↑	R-1 ↑	R-2 ↑	R-L ↑	BERTScore ↑	BARTScore ↑	MoverScore ↑
LLaMA2 [45]	-	0.076	0.037	0.076	-0.349	-7.06	N/A	0.069	0.029	0.068	-0.349	-6.65	N/A
MedAlpaca [14]	-	0.139	0.053	0.138	-0.245	-7.08	-0.089	0.139	0.049	0.138	-0.261	-6.72	-0.084
ChatDoctor [53]	-	0.214	0.082	0.212	-0.164	-7.15	-0.103	0.191	0.065	0.188	-0.167	-6.73	-0.103
Asclepius [24]	-	0.176	0.075	0.175	-0.157	-7.39	-0.166	0.141	0.055	0.141	-0.165	-7.04	-0.171
PMC-LLaMA [49]	-	0.119	0.033	0.118	-0.372	-7.38	-0.084	0.166	0.050	0.164	-0.308	-6.91	-0.128
ChatGPT	-	0.452	0.215	0.446	0.107	-5.76	0.143	0.389	0.193	0.385	0.028	-5.40	0.117
RO-LLaMA-P	Naive	0.633	0.456	0.630	0.361	-4.59	0.405	0.602	0.436	0.599	0.273	-4.01	0.404
RO-LLaMA-P+	NEFTune	0.649	0.479	0.645	0.395	-4.39	0.435	0.599	0.444	0.596	0.284	-3.94	0.410
RO-LLaMA-P++	CEFTune	**0.652**	**0.491**	**0.649**	**0.408**	**-4.34**	**0.446**	**0.606**	**0.450**	**0.604**	**0.285**	**-3.96**	**0.412**

Table 5. Example case for the plan suggestion task compared to baseline methods. The red scripts indicate the incorrect answers, as evaluated by the board certified clinical expert.

Treatment plan Suggestion Example	
User	Suggest a radiotherapy approach using the guidance from the given clinical summary of the radiation oncologist. # Input: [Clinical notes]...
Ground Truth	RT to Lt. whole breast + IMN + SCL + AXL 40.05 Gy / 15 fx → 10 Gy boost (or 48 Gy / 15 fx with IMRT).
Ours	RT to Lt. whole breast, IMN, SCL, AXL 40.05 Gy / 15 fx + 10 Gy boost (IMRT: 48 Gy / 15 fx).
Med-Alpaca	..Breast, LLO 4:00 3cm FN, IDC, icT2N1M0, stage IIB → ypT1cN1M0, stage IIA...
ChatDoctor	. . . The T-stage, N-stage, ER/PR/HER2/Ki67, and Lum A status of the tumor, as well as the patient's chemotherapy and surgical history, can all be used to determine appropriate radiation . . .
Chat-GPT	RT to Left breast, chest wall, L axilla Conventional fractionation 40.05 Gy / 15 fx.

4.3. Evaluation Metric

The evaluation of generated clinical report summaries and treatment plans utilizes common NLP metrics, including ROUGE [30], BERTScore [54], BARTScore [51], and MoverScore [55]. However, recognizing potential limitations in domain specificity, expertise-based metrics, such as GPT-3.5-turbo and a board certified clinical expert evaluation, are incorporated, particularly for the nuanced task of treatment plan suggestion. Adopting concepts from recent evaluation methodologies such as GPT-Score [13] and G-Eval [32], we developed a score rubric curated by clinical experts. The evaluation process includes GPT-3.5-turbo assessing specific examples of outputs based on this rubric. Additionally, clinical expert evaluation involves a comprehensive assessment of all outputs with ground truth label.

To evaluate the 3D target volume segmentation performance using a ground-truth segmentation mask, we measure Dice coefficient (Dice), Intersection over Union (IoU), and 95th percentile of Hausdorff Distance (HD-95) [9] in centimeter (cm) unit.

4.4. Baseline Models

For text-related tasks, to validate the effectiveness of our model, we compare our model with several clinical LLMs serving as baselines, specifically MedAlpaca [14], ChatDoctor [53], Asclepius [24], and PMC-LLAMA [49]. Additionally, we include a comparison with ChatGPT [36] utilizing few-shot in-context learning for both summary and treatment plan suggestion task.

For the segmentation task, we utilize 3D U-Net [7] and ConTEXTualNET [18] as our baseline methods. We further utilize an additional baseline as a variant of RO-LLaMA-SEG, naemly LLaMA-SEG, which is initialized with the checkpoint of vanilla LLaMA-2-7B-Chat model.

4.5. Results

Clinical Report Summarization We report the model performance on the clinical report summarization task in Table 1. Compared to the baselines, our fine-tuned variants of RO-LLaMA-S show significant improvements due to learning specific domain knowledge on all metrics. The qualitative assessment in Table S7 further highlights the effectiveness of our proposed model, providing well-organized content and consistent formatting compared to the ground truth label. While there were minor errors observed in terms of the specific distance measurement from the nipple (red), which appears to stem from the process of integrating ultrasound and MRI images, our model accurately summarizes information in a format consistent with the provided labels. In contrast, Chat-GPT exhibits issues such as omitting essential information and generating hallucinated content (red).

In terms of technical variants, both RO-LLaMA-S+ with

Table 6. The quantitative results for 3D target volume segmentation performance compared to different models.

Model	Method	Input Text	Internal Validation (N=79)			External Validation (N=81)		
			Dice ↑	IoU ↑	HD-95 ↓	Dice ↑	IoU ↑	HD-95 ↓
3D U-Net [7]	-	None	$0.802_{\pm 0.121}$	$0.684_{\pm 0.148}$	$8.967_{\pm 8.765}$	$0.689_{\pm 0.170}$	$0.548_{\pm 0.175}$	$25.018_{\pm 9.535}$
ConTEXTualNET† [18]	-	Generated Treatment Plan	$0.810_{\pm 0.117}$	$0.695_{\pm 0.148}$	$7.509_{\pm 7.340}$	$0.696_{\pm 0.150}$	$0.551_{\pm 0.152}$	$21.304_{\pm 5.551}$
LLaMA2-SEG	-		$0.816_{\pm 0.111}$	$0.703_{\pm 0.140}$	$9.420_{\pm 8.597}$	$0.691_{\pm 0.112}$	$0.537_{\pm 0.111}$	$24.033_{\pm 7.105}$
RO-LLaMA-SEG	Naive		$0.825_{\pm 0.101}$	$0.714_{\pm 0.134}$	$5.392_{\pm 7.072}$	$\mathbf{0.775_{\pm 0.149}}$	$0.652_{\pm 0.153}$	$17.623_{\pm 10.141}$
RO-LLaMA-SEG+	NESEG		$0.830_{\pm 0.103}$	$0.720_{\pm 0.127}$	$4.383_{\pm 5.774}$	$0.771_{\pm 0.185}$	$\mathbf{0.653_{\pm 0.177}}$	$28.835_{\pm 8.916}$
RO-LLaMA-SEG++	CESEG		$\mathbf{0.840_{\pm 0.094}}$	$\mathbf{0.733_{\pm 0.120}}$	$\mathbf{4.307_{\pm 5.772}}$	$0.764_{\pm 0.171}$	$0.641_{\pm 0.163}$	$\mathbf{15.422_{\pm 10.725}}$

†modified for 3D target volume segmentation with LLaMA2-7B-Chat as backbone

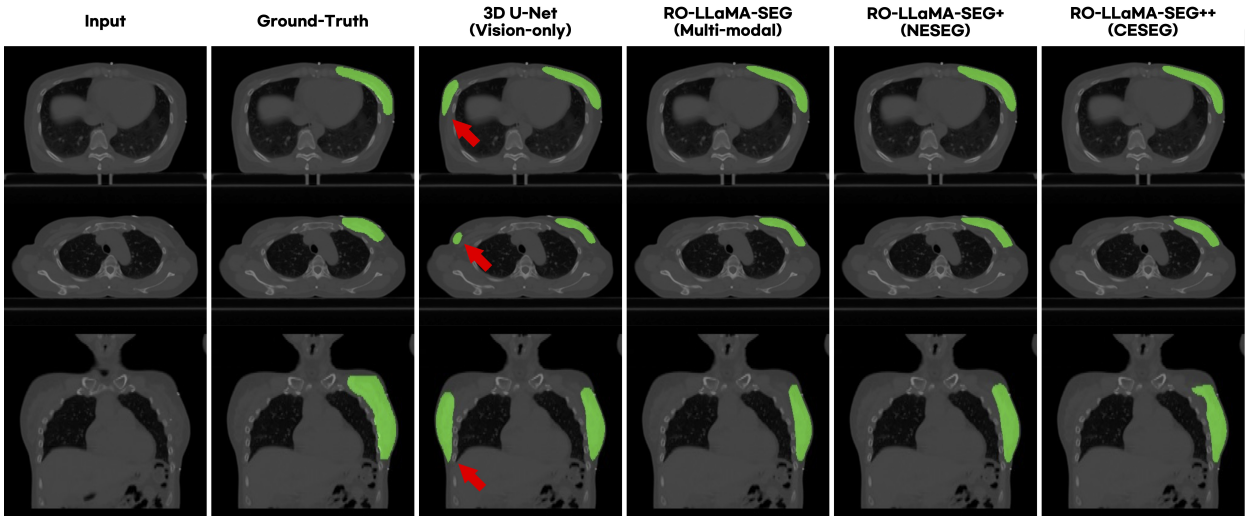


Figure 3. Qualitative comparison of 3D target volume segmentation. Red arrows indicate incorrectly segmented regions.

NEFTune and RO-LLaMA-S++ with CEFTune outperform models trained using the vanilla approach and other clinical expert LLMs. RO-LLaMA+ exhibits superior ROUGE scores for both internal and external datasets, while RO-LLAMA++ outperform in embedding-based metrics. Considering our goal of generating sequentially tailored text data samples for target volume segmentation, we empirically designate the summary expert as RO-LLaMA-S+.

Treatment plan Suggestion As indicated in Table 4, we evaluate the model's performance on treatment plan suggestion task based on generated clinical notes. Similar to the report summarization task, our fine-tuned variants, RO-LLaMA-P, exhibit substantial enhancements in suggestion performance compared to other baselines. Notably, RO-LLaMA-P+ shows inferior performance on the external dataset compared to the vanilla approach. However, our proposed RO-LLaMA-P++ achieves the best performance on both internal and external datasets across all metrics. For a more detailed understanding, examples of the treatment plan suggestion task are provided in Table 5. The proposed RO-LLaMA-P++ (Ours) consistently yields correct answers compared to the labeled ground truth. These results demonstrate the effectiveness of our proposed CEFTune for this specific task. While Chat-GPT has shown success in providing some level of information regarding where to treat, it falls short when it comes to suggesting the specific treatment regions and dose scheme (red).

3D Target Volume Segmentation As reported in Table 6, our proposed plan-guided 3D target segmentation frameworks excel other baseline methods with large gains, specifically remained stable performance across the external validation setting. Notably, compared to the uni-modal 3D U-NET model, our proposed RO-LLaMA-SEG++ shows substantial performance gain of up to 5% and 10% for internal and external validation, respectively. This can be also observed in Figure 3, where the vision-only method fails to segment target volume with incorrect laterality as indicated as red arrows. On the other hand, the proposed RO-LLaMA-SEG++ with CESEG shows the most promising qualitative segmentation performance, aligning consistently with the ground-truth labels as shown in Figure 3. Additional visual results are shown in Supplementary Material.

Table 7. The quantitative results for treatment plan suggestion using various methods, in terms of evaluations by clinical experts and GPT-3.5-turbo.

Model	Treatment Plan Suggestion			
	Internal (N=50)		External (N=50)	
	GPT3.5-turbo	Clinical Expert	GPT3.5-turbo	Clinical Expert
Med-Alpaca [14]	$1.706_{\pm 0.839}$	0	$1.418_{\pm 0.657}$	0
ChatDoctor [53]	$1.739_{\pm 0.787}$	$0.063_{\pm 0.239}$	$1.669_{\pm 0.807}$	$0.220_{\pm 0.418}$
Asclepius [24]	$1.505_{\pm 0.749}$	0	$1.215_{\pm 0.451}$	0
PMC-LLaMA [49]	$1.232_{\pm 0.516}$	0	$1.329_{\pm 0.543}$	0
Chat-GPT	$2.749_{\pm 0.700}$	$2.04_{\pm 0.832}$	$2.720_{\pm 0.723}$	$1.925_{\pm 1.185}$
RO-LLaMA-P++	$\mathbf{3.431}_{\pm \mathbf{0.577}}$	$\mathbf{3.400}_{\pm \mathbf{1.106}}$	$\mathbf{3.103}_{\pm \mathbf{0.526}}$	$\mathbf{3.100}_{\pm \mathbf{1.350}}$

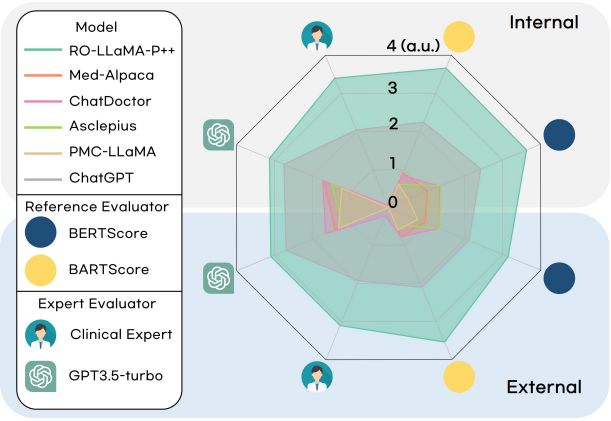


Figure 4. Expertise-based evaluation for treatment plan suggestion with various clinical LLM and ChatGPT.

5. Discussion

5.1. Expert Analysis on Plan Generation

As discussed in Section 4.3, for the treatment plan suggestion task, we comprehensively evaluate the generated plans using expertise-based metrics as indicated in Table 7. In terms of GPT-3.5-tubo evaluation, our RO-LLaMA-P++ outperforms other baseline methods with significant margin, benefiting from the learned knowledge for the field of radiation oncology.

In terms of clinical expert evaluation, most of baseline approaches, except for Chat-GPT, show average score of zero. This indicates that the clinical evaluation is more rigorous than GPT-3.5-turbo's stringency, revealing challenges in generating high-quality treatment plans. Although Chat-GPT achieves meaningful scores, our method notably outperforms all the methods, as shown in Figure 4. These results confirm our model's effectiveness in generating treatment plans aligned with clinical expert assessments, surpassing existing LLMs in the general medical domain.

5.2. Analysis on Consistency Regularization

We further analyze the effect of the proposed consistency module according to the distinct types of input text. As mentioned, both the plan suggestion model and segmentation model utilizes clean ground truth data during training phase, whereas they employ noisy LLM-generated data during inference phase, potentially resulting in a performance gap. As indicated as difference in Table 8, the model's performance without our proposed consistency module significantly degrades for the generated data compared to the ground truth data for both tasks. On the other hand, employing our proposed CEFT and CESEG ensures robust performance for the generated data compared to ground truth data. In particular, for the segmentation task, our model exhibits a performance difference between two distinct input data of less than 1%. The results indicate that our key idea of consistency module effectively alleviates performance degradation when using LLM-generated data as input.

Table 8. Ablation study on consistency regularization with input text variation for treatment plan suggestion and 3D target volume segmentation performance.

Consistency	Input Text	Treatment Plan Suggestion		Target Segmentation	
		BERTScore ↑	BARTScore ↑	Dice ↑	IoU ↑
✗	Ground Truth	0.602	-3.19	$0.860_{\pm 0.073}$	$0.761_{\pm 0.103}$
	Generated	0.361	-4.60	$0.825_{\pm 0.101}$	$0.714_{\pm 0.134}$
	Difference ↓	0.241	1.41	0.035	0.047
✓	Ground Truth	0.593	-3.29	$0.847_{\pm 0.088}$	$0.743_{\pm 0.111}$
	Generated	0.408	-4.35	$0.840_{\pm 0.094}$	$0.733_{\pm 0.120}$
	Difference ↓	**0.185**	**1.06**	**0.007**	**0.010**

6. Conclusion

In this work, we introduce RO-LLaMA, a versatile and general-purpose foundation model tailored for radiation oncology. Addressing limitations in current medical AI models confined to specific tasks, RO-LLaMA demonstrates proficiency in diverse tasks: clinical report summarization, radiation treatment plan suggestion, and plan-guided 3D target volume segmentation, mirroring real-world clinical workflows. Another key contribution of this work is the introduction of consistency technique into both text and segmentation task. Results from multi-center cohort datasets confirm RO-LLaMA's promising performance and noteworthy generalization capabilities across diverse tasks. These findings mark a significant stride toward developing a versatile AI model, hinting at the potential for a generalist medical AI model in radiation oncology.

Limitation The used dataset is currently confined to patients with their initial diagnoses, necessitating an expansion of the scope to cover diverse patient scenarios.

Ethical Statement All the patient information and any potential personally identifiable information from used datasets are de-identifiied, and clinical data usage is approved by the Institutional Review Board (IRB).

References

[1] Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*, 2020. 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 2

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 2

[5] Min Seo Choi, Byeong Su Choi, Seung Yeun Chung, Nalee Kim, Jaehee Chun, Yong Bae Kim, Jee Suk Chang, and Jin Sung Kim. Clinical evaluation of atlas-and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiotherapy and Oncology*, 153:139–145, 2020. 3

[6] Seung Yeun Chung, Jee Suk Chang, Min Seo Choi, Yongjin Chang, Byong Su Choi, Jaehee Chun, Ki Chang Keum, Jin Sung Kim, and Yong Bae Kim. Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. *Radiation Oncology*, 16(1):1–10, 2021. 3

[7] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 5, 6, 7, 1

[8] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. 2

[9] William R Crum, Oscar Camara, and Derek LG Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, 25(11):1451–1461, 2006. 6

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[11] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip, 2023. 3

[12] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. 2

[13] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023. 6

[14] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023. 2, 5, 6, 8

[15] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018. 3

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2, 4

[17] Hang Hua, Xingjian Li, Dejing Dou, Cheng-Zhong Xu, and Jiebo Luo. Noise stability regularization for improving bert fine-tuning. *arXiv preprint arXiv:2107.04835*, 2021. 2

[18] Zachary Huemann, Junjie Hu, and Tyler Bradshaw. Contextual net: A multimodal vision-language model for segmentation of pneumothorax. *arXiv preprint arXiv:2303.01615*, 2023. 3, 6, 7

[19] Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, et al. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*, 2023. 2, 3

[20] Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*, 2019. 2

[21] Geeta Joshi, Aditi Jain, Sabina Adhikari, Harshit Garg, and Mukund Bhandari. Fda approved artificial intelligence and machine learning (ai/ml)-enabled medical devices: An updated 2022 landscape. *medRxiv*, pages 2022–12, 2022. 2

[22] Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023. 1

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[24] Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. Publicly shareable clinical large language model built on synthetic clinical notes. *arXiv preprint arXiv:2309.00237*, 2023. 2, 5, 6, 8

[25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2, 3

[26] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2

[28] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: Language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2023. 3

[29] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip, 2023. 3

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[32] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, may 2023. *arXiv preprint arXiv:2303.16634*. 6

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[34] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 2

[35] Yujin Oh, Sangjoon Park, Hwa Kyung Byun, Jin Sung Kim, and Jong Chul Ye. Llm-driven multimodal target volume contouring in radiation oncology, 2023. 3, 4, 1

[36] OpenAI. Chatgpt. *OpenAI Blog*, 2021. 2, 6

[37] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2

[38] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155*, 13, 2022. 2

[39] Pranav Rajpurkar and Matthew P Lungren. The current and future state of ai interpretation of medical images. *New England Journal of Medicine*, 388(21):1981–1990, 2023. 2

[40] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 4, 5

[41] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022. 2

[42] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. 2

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 5

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 3, 5, 6

[46] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023. 2

[47] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 2

[48] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 3

[49] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023. 2, 5, 6, 8

[50] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 2

[51] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021. 6

[52] Sukmin Yun, Seong Hyeon Park, Paul Hongsuck Seo, and Jinwoo Shin. Ifseg: Image-free semantic segmentation via vision-language model, 2023. 3

[53] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023. 2, 5, 6, 8

[54] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 6

[55] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*, 2019. 6

[56] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022. 1

[57] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019. 2

RO-LLaMA: Generalist LLM for Radiation Oncology via Noise Augmentation and Consistency Regularization

Supplementary Material

1. Supplementary Section

In this supplemental documents, we present:

- An algorithmic descriptions of our proposed Consistency Embedding Fine Tuning (CEFT) and Consistency Embedding Segmentation (CESEG) in Section 2.
- An explanation of the network architecture for the segmentation model in Section 3.
- Results of ablation studies for both text-related tasks and target volume segmentation tasks in Section 4 and Section 5.
- Detailed information about the dataset, including preprocessing and acquisition procedures, in Section 6.
- Examples of the entire workflow in Section 8.
- Additional comparison results for all tasks in Section 7.
- Detailed descriptions of the evaluation for treatment suggestion, including examples of prompts for GPT-3.5-turbo, in Section 9.

2. Algorithm of CEFTune and CESEG

Algorithm 1 and Algorithm 2 details the training procedure with CEFTune and CESEG, respectively.

Algorithm 1: CEFTune: **C**onsistency **E**mbedding **F**ine **tun**ing

Input : $\{X_i^{\text{txt}}, Y_i^{\text{txt}}\}_{i=1}^{N} \sim D$ tokenized dataset, large language model $f_\theta(\cdot)$, detokenization operator $\mathcal{T}(\cdot)$, the metric which calculate the cosine similiary $d(\cdot,\cdot)$, the hyper-parameter for the noise intensity and loss, α, and λ, repectively. The cross entropy loss $\mathcal{L}_{\text{ce}}$, the embedding layer $\text{emb}(\cdot)$, $\theta^- \leftarrow \text{stopgrad}\ (\theta)$;

repeat
- Sample $(\mathbf{x}_{\text{txt}}, \mathbf{y}_{\text{txt}}) \sim D$;
- $\mathbf{x}_{\text{txt}}^{\text{emb}} \leftarrow \text{emb}(\mathbf{x}_{\text{txt}})$;
- $\hat{\mathbf{y}}_{\text{txt}} \leftarrow f_{\theta^-}(\mathbf{x}_{\text{txt}}^{\text{emb}})$;
- $\epsilon \sim \mathcal{U}(-1, 1)$;
- $\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}} \leftarrow \mathbf{x}_{\text{txt}}^{\text{emb}} + (\alpha/\sqrt{LC})\epsilon$;
- $\tilde{\mathbf{y}}_{\text{txt}} \leftarrow f_\theta(\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}})$;
- $\mathcal{L}(\theta) \leftarrow \mathcal{L}_{\text{ce}}(\hat{\mathbf{y}}_{\text{txt}}, \mathbf{y}_{\text{txt}}) + \lambda d(\mathcal{T}(\tilde{\mathbf{y}}_{\text{txt}}), \mathcal{T}(\hat{\mathbf{y}}_{\text{txt}}))$

until Stopping criteria met/max iterations;

Algorithm 2: CESEG: **C**onsistency **E**mbedding **Seg**mentation

Input : $\{X_i^{\text{img}}, X_i^{\text{txt}}, Y_i^{\text{img}}\}_{i=1}^{M} \sim D$, the entire image segmentation model $g_\Theta(\cdot)$, where $\Theta = [\psi, \phi]$, which consist of segmenation model g_ψ, and learnable text prompt z_ϕ. The frozen large language model $f_{\theta^*}(\cdot)$, output embedding of f_{θ^*} $\mathbf{y}_{\text{txt}}^{\text{emb}}$, the hyper-parameter for the noise intensity and loss α and λ, respectively, the metric which calculate the cosine similiary $d(\cdot,\cdot)$, the cross entropy loss $\mathcal{L}_{\text{ce}}$, the embedding layer $\text{emb}(\cdot)$, $\phi^- \leftarrow \text{stopgrad}\ (\phi)$,

repeat
- Sample $(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}}, \mathbf{y}_{\text{img}}) \sim D$;
- $\mathbf{x}_{\text{txt}}^{\text{emb}} \leftarrow \text{emb}(\mathbf{x}_{\text{txt}})$;
- $\mathbf{y}_{\text{txt}}^{\text{emb}} \leftarrow f_{\theta^*}(\mathbf{x}_{\text{txt}}^{\text{emb}}; \mathbf{z}_\phi)$;
- $\epsilon \sim \mathcal{U}(-1, 1)$;
- $\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}} \leftarrow \mathbf{x}_{\text{txt}}^{\text{emb}} + (\alpha/\sqrt{LC})\epsilon$;
- $\tilde{\mathbf{y}}_{\text{txt}}^{\text{emb}} \leftarrow f_{\theta^*}(\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}}; \mathbf{z}_{\phi^-})$;
- $\mathcal{L}(\Theta) \leftarrow \mathcal{L}_{\text{ce}}(g_\psi(\mathbf{x}_{\text{img}}; \tilde{\mathbf{y}}_{\text{txt}}^{\text{emb}}), \mathbf{y}_{\text{img}}) + \lambda d(\tilde{\mathbf{y}}_{\text{txt}}^{\text{emb}}, \mathbf{y}_{\text{txt}}^{\text{emb}})$

until Stopping criteria met/max iterations;

3. Architecture of RO-LLaMA-SEG++

The schematic of our RO-LLaMA-SEG++ is illustrated in Figure S1. We employed the 3D U-Net [7] for the image module. For the text module, we adapted the idea of text prompt tuning to transfer the linguistic capability of the LLM for realizing multi-modal context-aware segmentation framework, by following the concepts introduced in [22, 35, 56]. We employed pre-trained RO-LLaMA-P++ as for the LLM module within the text module. For the multi-modal alignment module which uses both self-attention and cross-attention mechanisms in an interactive manner (image-to-text and text-to-image), by following the concept of promptable segmentation borrowed from Segment Anything Model (SAM) [23]. During training, we let the entire parameters of LLM frozen, while updating the image module and the text prompts.

As illustrated in Figure S1(c) for the text module, we introduced a total of N-text prompts $\{v^n|_{n=1}^N\}$, where each $v^n \in \mathbb{R}^{K\times D}$ is composed of K vectors with the dimension

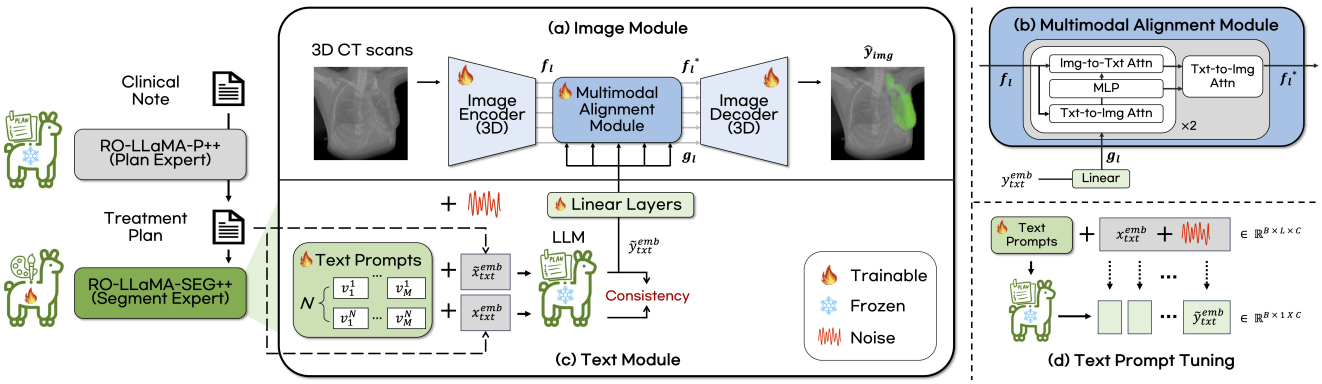


Figure S1. Detailed schematics of RO-LLaMA-SEG++ for plan-guided 3D target volume segmentation task.

D. These learnable vectors were randomly initialized, and then consistently prepended to each of perturbed embedded treatment plan $\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}} \in \mathbb{R}^{B\times L\times C}$. Here, the final prompted text input t can be formulated as follows:

$$t = \{v_1^n, v_2^n, ..., v_K^n, \tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}}\}. \tag{5}$$

Then, using the prompted text input t, the frozen LLM resulted the plan embeddings $\mathbf{y}_{\text{txt}}^{\text{emb}} \in \mathbb{R}^{B\times 1\times C}$ as for the last token output, as illustrated as in Figure S1(d). To align the text module output with the image module features, $\mathbf{y}_{\text{txt}}^{\text{emb}}$ was projected to become $g_l \in \mathbb{R}^{b\times 1\times Ch_l}$ which have the identical dimension with that of each f_l through layer-wise linear layer, where $f_l \in \mathbb{R}^{H_l W_l S_l \times Ch_l}$, where f_l is the l-th layer output of the image module, H_l, W_l, and S_l correspond to height, width, and slice of the image embeddings, and Ch_l is the intermediate channel dimension of each l-th layer output. We adapted CESEG to regularize consistency between the noisy input and the clean input by following Equation (4).

In Figure S1(a) and (b) for the image module, the projected plan embeddings g_l were self-attended and crossly-attended with the image features f_l for each layer through the multimodal alignment module, and the multi-modal embeddings $f_l{}^*$ were inputted to the decoder part of the image module to yield final prediction $\hat{y}_{\text{img}} \in \mathbb{R}^{B\times HWS}$.

4. Ablation Study on Therapy Plan Suggestion

To investigate the effectiveness of our proposed methods, we conduct the ablation studies for treatment plan suggestion task, as shown in Table S1.

Consistency Regularization Loss We manipulate the objective function in Equation 3 to analyze the effect of consistency regularization. By default, we use $f_{\theta^-}(\mathbf{x}_{\text{txt}}^{\text{emb}})$ as the target (referred to as "clean teacher"). For comparison, we set $f_{\theta^-}(\tilde{\mathbf{x}}_{\text{txt}}^{\text{emb}})$ as the teacher and $f_\theta(\mathbf{x}_{\text{txt}}^{\text{emb}})$ as the student (referred to as "noisy teacher"). We observe that the clean teacher significantly outperforms the noisy teacher. By restricting the divergence of the distribution of students, the clean teacher yields enhanced performance compared to the noisy teacher.

Effect of SBERT Module We study the effect of the SBERT module in enforcing consistency between clean and noisy input. Our model, equipped with the medical domain-specific SBERT variant, PubMedBERT, outperforms models trained on general language. This demonstrates the effectiveness of incorporating domain-specific knowledge.

Analysis of Noise Intensity (α) We analyze the effect of noise intensity by varying α from 5 to 15. Smaller noise intensities result in better performance on the internal dataset, while larger noise intensities improve performance on the external dataset. However, excessively large noise intensity leads to a significant drop in performance on the external dataset. Based on empirical results, we adopt α set to 5 as the baseline.

Analysis of Model Backbone We conduct experiment using the larger backbone 13B LLaMA-2 version and observed that there are no significant differences compared to the 7B model in terms of performance. Considering cost-effectiveness and performance, we select the 7B model as the default configuration.

5. Ablation Study on Target Segmentation

To optimize treatment target segmentation performance, we conduct the ablation studies, as shown in Table S2.

Analysis of LLM Tuning Methods We compare the different LLM tuning methods to find optimal performance, where NESEG nor CESEG module is not activated. We analyze various methods including no tuning, low-rank adaptation (LoRA) fine-tuning [16] and prompt tuning methods

with a single or multiple text prompts. From the experimental results, the text prompt tuning shows the most reliable performance across the external validation setting. Additionally, the use of single text prompt shows a remarkable improvement in performance compared to utilizing multiple text prompts.

Analysis of Noise Intensity (α) We analyze the effect of noise intensity by varying α from 5 to 20, where CESEG module is not activated. We empirically find that α as 10 yields the best performance. We further train the hyperparameter α as a learnable parameter, however, yielding inferior performance compared to employing the fixed noise.

6. Details of Training Dataset

For data preparation, we collected internal training & validation data from patients treated at Yonsei Cancer Center between January 2009 and December 2022. For external validation, we further collected data from patients from Yongin Severance Hospital between January 2018 and December 2022. We confirmed that the external test set was not overlapped with the training dataset.

For pre-processing image data for training RO-LLaMA-SEG++, all the 3D chest CT scans and target volumes were re-sampled with identical voxel spacing of $1.0 \times 1.0 \times 3.0$ mm^3. The intensity values of the CT scans were truncated between -1,000 and 1,000 of Hounsfield unit (HU), and linearly normalized between 0 and 1.0. When training, a 3D patch with a spatial dimension of $384 \times 384 \times 128$ pixels was randomly cropped. When inference, the entire 3D CT scans was tested using sliding windows with the identical spatial dimension of the 3D patch to that used for training.

7. Additional Qualitative Results

In this section, we provide the additional qualitative results on all tasks as shown in Table S3, Table S4 and Figure S2.

In the context of report summarization tasks involving cases with multiple (two or more) tumor masses, as shown in in Table S3, our model adeptly summarized the existence of two tumors in the first line of the summary, despite minor inaccuracies in pinpointing the exact locations of the tumors. In contrast, ChatGPT incorrectly summarized these cases as having only one tumor mass in the first line. Additionally, ChatGPT exhibited instances of hallucination, such as inappropriately mentioning tumor sizes in the first line, erroneously including post-therapy (yp) stages for patients who did not receive chemotherapy, and excessively condensing key information such as the presence of the lymph node metastasis, making it challenging to discern. Overall, unlike our model which aligns closely with the ground truth in its summaries, ChatGPT demonstrated significant issues, including the generation of erroneous hallucination and the failure to comprehensively encapsulate necessary details.

In the evaluation of treatment plan suggestion capabilities in Table S4, a notable observation was that only our model and ChatGPT were able to offer somewhat meaningful suggestions related to radiation therapy, with other models falling short. ChatGPT, drawing upon its built-in knowledge base, seemed to propose radiation therapy plans that were partially relevant. However, it exhibited a fundamental misunderstanding of the radiation dosing strategies, frequently misinterpreting hypofractionation as conventional fractionation. A significant limitation in ChatGPT's approach was its inability to incorporate all regional lymph node areas into the treatment's target volume, often disproportionately focusing on specific areas, such as the axillary lymph nodes. Additionally, ChatGPT struggled to distinguish between the treatment of the breast and the chest wall, leading to flawed treatment plan suggestions in detail. This limitation likely arises from ChatGPT's lack of a holistic understanding of radiation oncology, unlike our domain-specific model.

Figure S2 provides qualitative comparisons of the target volume segmentation performance for breast cancer patients. In Figure S2(a) and (b), despite the ground-truth label posing target volume on the one side of the breast, the vision-only 3D U-Net incorrectly segment contours not only the target volume but also the outside of the target breast. Moreover, RO-LLaMA-SEG+ with NESEG module yield noisy segmentation output as indicated as red arrows. RO-LLaMA-SEG without any proposed module shows reasonable perfomormance, yet sometimes undersegments the treatment target volume. In contrast, the our proposed RO-LLaMA-SEG++ with CESEG module accurately contours the breast and regional lymph nodes that need to be treated.

8. Examples of Entire Workflow

We present examples of the entire workflow as indicated in Table S5 and Table S6. RO-LLaMA can take input as a pair of MRI report and Pathology report or another pair of Ultrasound report and Pathology report. Our model seamlessly serves as an assistant for radiation oncologists by conducting report summarization, treatment plan suggestion, and 3D Target Volume Segmentation.

9. Prompt for Evaluation of Plan Suggestion

In this paper, we design prompts for evaluating generated treatment plans, including a reference-guided score rubric curated by clinical expert, as shown in Table S7. Similar to previous GPT-based evaluation methods, we tailor the prompts combining the specific knowledge of clinical experts. By using this prompt, we evaluate the results of all the clinical LLMs using GPT-3.5-turbo.

Table S1. Component analysis of RO-LLaMA-P framework on treatmet plan suggestion.

Component	Condition	Ours	Internal Validation (N=60)			External Validation (N=81)		
			BERTScore ↑	BARTScore ↑	MoverScore ↑	BERTScore ↑	BARTScore ↑	MoverScore ↑
Teacher Module	Noisy teacher		0.369	-4.52	0.405	0.264	-4.03	0.394
	Clean teacher	✓	0.408	-4.34	0.446	0.285	-3.93	0.412
SBERT Moudle	SBERT		0.402	-4.40	0.425	0.267	-4.05	0.392
	PubMedBERT	✓	0.408	-4.34	0.446	0.285	-3.93	0.412
Noise Intensity (α)	5	✓	0.408	-4.34	0.446	0.285	-3.93	0.412
	10		0.403	-4.34	0.435	0.291	-3.86	0.419
	15		0.400	-4.35	0.445	0.270	-3.97	0.401
Backbone	13B		0.392	-4.42	0.427	0.262	-4.03	0.389
	7B	✓	0.408	-4.34	0.446	0.285	-3.93	0.412

Table S2. Component analysis of RO-LLaMA-SEG framework on target volume segmentation performance.

Component	Condition	Ours	Internal Validation (N=79)			External Validation (N=81)		
			Dice ↑	IoU ↑	HD-95 ↓	Dice ↑	IoU ↑	HD-95 ↓
LLM Tuning Method	None		$0.822_{\pm 0.112}$	$0.712_{\pm 0.137}$	$12.368_{\pm 12.115}$	$0.752_{\pm 0.155}$	$0.622_{\pm 0.161}$	$24.849_{\pm 12.687}$
	LoRA [16]		$0.810_{\pm 0.136}$	$0.699_{\pm 0.157}$	$\mathbf{4.989_{\pm 6.087}}$	$0.759_{\pm 0.162}$	$0.637_{\pm 0.165}$	$\mathbf{16.692_{\pm 9.572}}$
	Text Prompts (p=1)	✓	$\mathbf{0.825_{\pm 0.101}}$	$\mathbf{0.714_{\pm 0.134}}$	$5.392_{\pm 7.072}$	$\mathbf{0.775_{\pm 0.149}}$	$\mathbf{0.652_{\pm 0.153}}$	$17.623_{\pm 10.141}$
	Text Prompts (p=2)		$0.822_{\pm 0.105}$	$0.711_{\pm 0.134}$	$7.030_{\pm 9.640}$	$0.757_{\pm 0.139}$	$0.627_{\pm 0.151}$	$24.640_{\pm 11.620}$
Noise Intensity (α)	5		$0.821_{\pm 0.096}$	$0.707_{\pm 0.121}$	$5.917_{\pm 6.125}$	$0.738_{\pm 0.144}$	$0.601_{\pm 0.145}$	$\mathbf{22.320_{\pm 11.894}}$
	10	✓	$\mathbf{0.830_{\pm 0.103}}$	$\mathbf{0.720_{\pm 0.127}}$	$\mathbf{4.383_{\pm 5.774}}$	$\mathbf{0.771_{\pm 0.185}}$	$\mathbf{0.653_{\pm 0.177}}$	$28.835_{\pm 8.916}$
	20		$0.826_{\pm 0.105}$	$0.716_{\pm 0.129}$	$5.575_{\pm 6.144}$	$0.732_{\pm 0.141}$	$0.593_{\pm 0.143}$	$26.090_{\pm 10.644}$
	Learnable		$0.828_{\pm 0.090}$	$0.715_{\pm 0.113}$	$5.024_{\pm 6.071}$	$0.757_{\pm 0.172}$	$0.633_{\pm 0.169}$	$23.694_{\pm 12.972}$

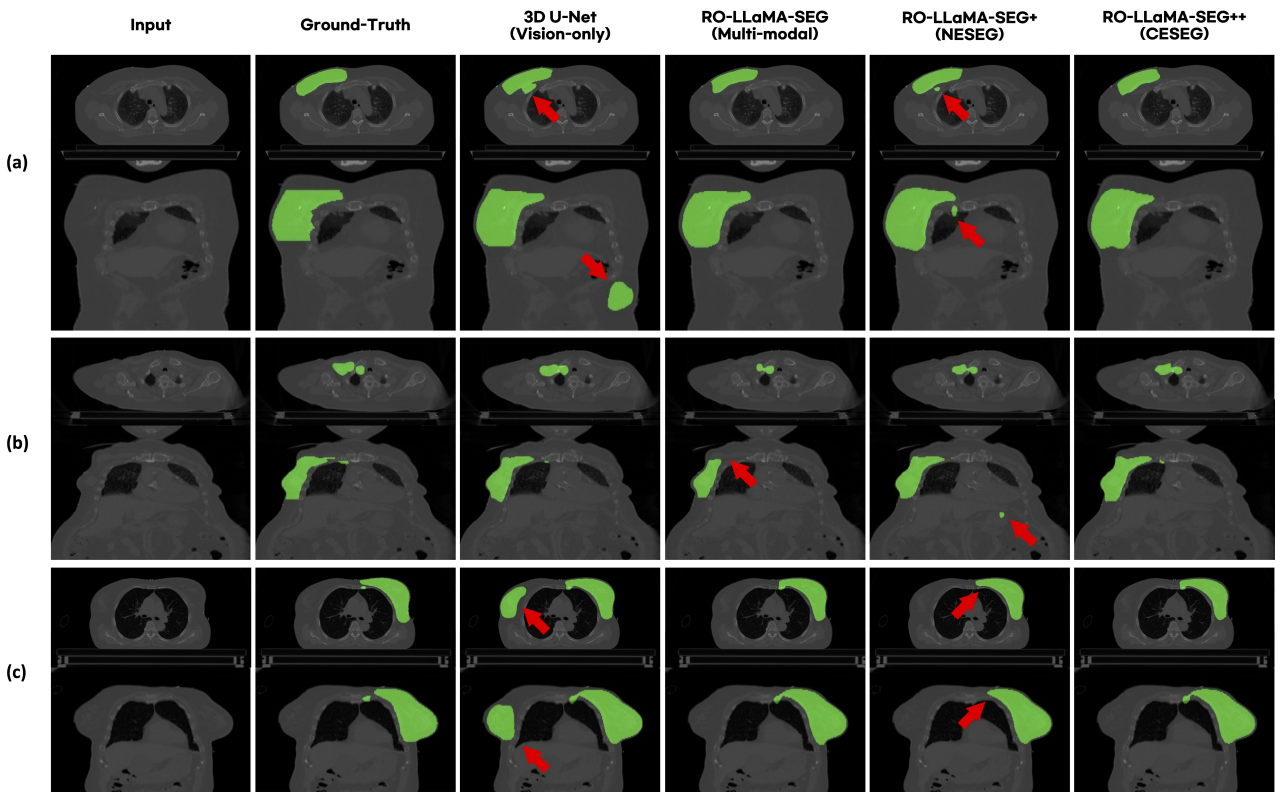


Figure S2. 3D target volume segmentation cases compared to baseline methods. Red arrows indicate incorrectly segmented regions.

Table S3. Example case for the clinical report summarization task. The red scripts indicate the incorrect or redundant information, and inconsistent format compared to label.

Clinical Report Summarization Example	
User	Using the provided MRI and pathology findings, design a radiation oncologist's clinical overview. ## `Input: [MRI report] ... [Pathology report] ...`
Ground truth	# 1. Breast, RLO 08:00 4cm FN, RLM 04:00 2cm FN, IDC, two, pT1cN1aM0, stage IIA - Histologic type: Invasive ductal carcinoma, NOS - Tumor margin: Infiltrative - Maximum diameter of invasive tumor: 1.5cm and 0.4cm - Maximum diameter of invasive and in situ carcinoma: 3.0cm - Nuclear grade: 2 - Histologic grade: II (3-2-1) - LVI (-) PNI (-) - Intraductal carcinoma component - Proportion: 10% - Extensive intraductal component (EIC): Absent - Type: Cribriform - Nuclear grade: Low - Necrosis: Absent - Location: Periphery - Calcification: Present (Malignant abd benign) - Associated benign lesion: Intraductal papilloma, sclerosing adenosis and columnar cell lesion - RM (-) - Overlying skin (superficial margin): Free of ca (safety margin: 1.4cm) - Underlying fascia (deep margin): Free of ca (safety margin: 2.3cm) Lymph nodes, total (2/13): PNE (-, Maximal diameter: 3mm) - Axillary (0/10): Free of ca - Sentinel lymph node1 (1/1), axillary lymph node1 (1/1), axillary lymph node2 (0/1) and axillary lymph node3 (0/0); total (2/3): PNE (-, Maximal diameter: 3mm) - ER, PR, HER-2, Ki-67 (90/90/0/6.5%) Lum A Oncotype Dx: RS 10 CTx omission
Ours	# 1. Breast, RLO 09:00 4cm FN RMC 03:00 3cm FN, IDC, Multifocal, pT1cN1aM0, stage IIA - Histologic type: Invasive ductal carcinoma, NOS - Tumor margin: Infiltrative - Maximum diameter of invasive tumor: 1.5cm and 0.4cm - Maximum diameter of invasive and in situ carcinoma: 3.0cm - Nuclear grade: 2 - Histologic grade: II (3-2-1) - Lymphovascular invasion: Absent - Perineural invasion: Absent - Intraductal carcinoma component - Proportion: 10% - Extensive intraductal component (EIC): Absent - Type: Cribriform - Nuclear grade: Low - Necrosis: Absent - Location: Periphery - Calcification: Present (Malignant abd benign) - Associated benign lesion: Intraductal papilloma, sclerosing adenosis and columnar cell lesion - Resection margins - Overlying skin (superficial margin): Free of carcinoma (safety margin: 1.4cm) - Underlying fascia (deep margin): Free of carcinoma (safety margin: 2.3cm) Lymph nodes, total (2/13): PNE (-), max 3mm - Axillary (0/10): Free of carcinoma - Sentinel lymph node1 (1/1), axillary lymph node1 (1/1), axillary lymph node2 (0/1) and axillary lymph node3 (0/0); total (2/3): Metastatic carcinoma in 2 out of 3 lymph nodes without perinodal soft tissue extension (Maximal diameter: 3mm) - ER/PR/HER2/Ki-67 (90/90/0/6.5%) Lum A Oncotype DX RS: 20 Adj. CTx omission
Chat-GPT	#1. Breast, RLO 08:00, 15mm, IDC, pT1cN1(f)M0, Stage IIA ypT2 0cN0M0 - two (1.5cm and 0.4cm), H2N2 - two of 13 + - LVI (-), DLI (-), HD (-) - Extensive IC: (-) - Calcification (+) - RM (-) - Overlying skin: Free of ca (safety margin: 1.4cm) - Underlying fascia: Free of ca (safety margin: 2.3cm) - ER/PR/HER2/Ki-67 90/90/0/6.5% Lum A

Table S4. Example case for the plan suggestion task compared to baseline methods. The red scripts indicate the incorrect answers.

Example 1	
User	Suggest a radiotherapy approach using the guidance from the given clinical summary of the radiation oncologist. # Input: [Clinical notes]...
Ground truth	RT to WB and tumor cavity (SIB) 40.05 Gy / 48 Gy in 15 fx.
Ours	RT to whole breast + tumor cavity (SIB) 40.05 Gy / 48 Gy in 15 fx.
Med-Alpaca	...Breast, RUC, 12:00, 5cm FN, IDC, pT1cN0M0, stage IA...
ChatDoctor	. . . Depending on the patient's individual circumstances, the radiation oncologist may recommend a course of external beam radiation therapy, brachytherapy, or a combination of the two . . .
Chat-GPT	RT to right breast Conventional fractionation Total dose: 50 Gy in 25 fractions.
Example 2	
User	Plan a radiotherapy treatment based on the insights from the radiation oncologist's clinical note. #Input: [Clinical notes]...
Ground truth	RT to Rt. chest wall + AXL Lv 3, IMN, SCL 40.05 Gy / 15 fx.
Ours	RT to Rt. chest wall, AXL Lv 3, IMN, SCL 40.05 Gy / 15 fx.
Med-Alpaca	.. Status post core biopsy of the right breast, performed ...
ChatDoctor	. . . Deliver a dose of 40 Gy in 10 fractions over 4 weeks to the primary tumor, with a margin of 1.1 cm for the overlying skin and 1.2 cm for the underlying fascia. . . .
Chat-GPT	RT to chest wall, IMN, and AXL Conventional fractionation 40.05 Gy / 15 fx.
Example 3	
User	Considering the clinical insights from the radiation oncologist's note, set a treatment plan in terms of radiotherapy. # Input: [Clinical notes]...
Ground truth	RT to Lt. WB and tumor cavity (SIB) 40.05 Gy / 48 Gy in 15 fx (2.67 Gy / 3.2 Gy in daily fx).
Ours	RT to Lt. WB and tumor cavity (SIB) 40.05 Gy / 48 Gy in 15 fx (2.67 Gy / 3.2 Gy daily fx)
Med-Alpaca	.. Lymph nodes, total (0/4): Free of carcinoma - Sentinel lymph node1 (0/1), sentinel lymph node2 (0/1), sentinel lymph node3 (0/1), and axillary lymph node1 (0/1); total (0/4): Free of carcinoma ...
ChatDoctor	. . . Radiotherapy will be administered to the breast, left, 2 o'clock, 4 cm FN IDC, pT1bN0M0, H1N2, PNI (-), LVI (-), EIC (-), Calcification: . . .
Chat-GPT	RT to Left breast Conventional fractionation 40.05 Gy / 15 fx.
Example 4	
User	Considering the clinical insights from the radiation oncologist's note, set a treatment plan in terms of radiotherapy. # Input: [Clinical notes]...
Ground truth	RT to Lt. WB/Cavity, AXL 2, 3, Lv 4 (ESTRO), IMN 40.05 Gy / 48 Gy in 15 fx (2.67 Gy / 3.2 Gy in daily fx).
Ours	RT to Lt. WB/Cavity, AXL 2, 3, Lv 4 (ESTRO), IMN 40.05 Gy / 48 Gy in 15 fx (2.67 Gy / 3.2 Gy daily fx).
Med-Alpaca	.. ER: 20% of the tumor cells express nuclear signals of moderate intensity PR: Negative ...
ChatDoctor	. . . These findings suggest that the patient has an aggressive form of breast cancer and may benefit from additional chemotherapy and targeted therapy. . . .
Chat-GPT	RT to left chest wall and axilla Conventional fractionation 40.05 Gy / 15 fx for chest wall 40.05 Gy / 15 fx for axilla.

Table S5. Example of Entire Workflow in the case of taking MRI report and Pathology report as input.

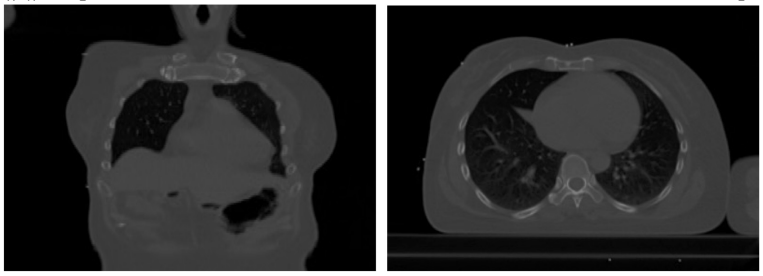
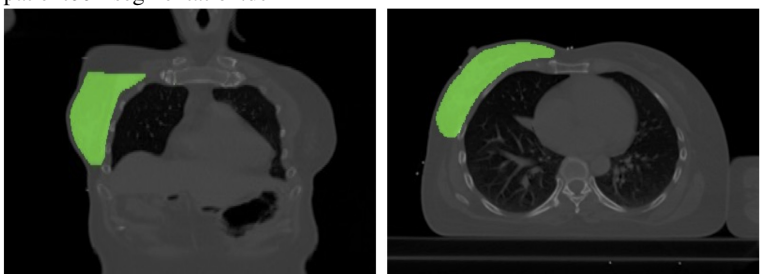
Case 1. MRI report and Pathology report	
User	Using the provided MRI and pathology findings, design a radiation oncologist's clinical overview. ## `Input: [MRI report]... [Pathology report] ...`
RO-LLaMA-S+	#1. Breast, ROC, 09:00, 3cm FN, IDC, icT2N1(f)M0 stage IIB (anatomic), stage IIIB (prognostic) ypT0N0M0 [ic] T: - 39mm size mass, H3N3 - LVI (-) N: - Multiple suspicious lymph nodes at Rt. axilla level 1, 2, 3 (FNA+) *ER/PR/HER2/Ki-67 (-/-/3+/31.48%) HER2 positive (non Lum) [yp] - No residual ca - Fibrosis with hemosiderin laden macrophages and foam cell collection, favoring regressed tumor bed - Total LNs (0/25) : Axillary (0/11), Level II (0/6) : SLN (0/4), axillary (non SLN) (0/4) - Calcification (+, Benign) - RM (-): Nipple, upper, and lower margin: Free of carcinoma (See SF21-04688) * Lymphedema prediction - 2yr risk: 17.70% - 3yr risk: 20.28% - 5yr risk: 22.97%
User	Suggest a radiotherapy approach using the guidance from the given clinical summary of the radiation oncologist. ## `Input:[Generated Clinical Note] ...`
RO-LLaMA-P++	RT to Rt. whole breast + AXL Lv 3 + SCL 48 (SIB) - 40.05 Gy / 15 fx.
User	Draw the radiotherapy target volume of this patient with following planning CT scan. ## `Input:[Generated Treatment Plan] ... [3D CT Scan] patient001.dcm` 
RO-LLaMA-SEG++	patient001_segmentation.dcm 

Table S6. Example of Entire Workflow in the case of taking Ultrasound report and Pathology report as input.

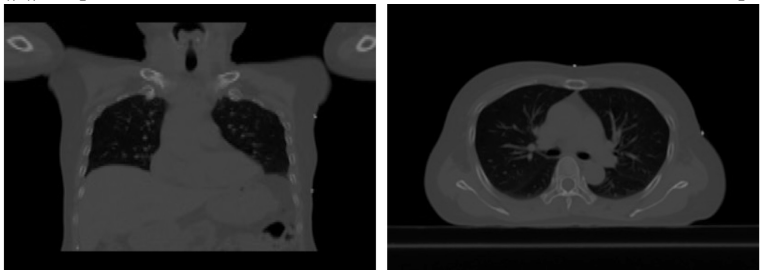
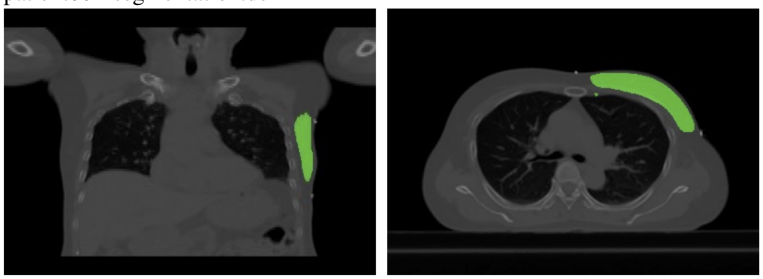
Case 2. UltraSound report and Pathology report	
User	Summarize a clinical summary note as a radiation oncologist might, using the provided Ultrasound and pathology details. ## `Input: [Ultrasound reports ]... [Pathology reports] ...`
RO-LLaMA-S+	#1. Breast, LUO 02:00 2cm FN, IDC, pT1cN0M0, stage IA - Histologic type: Invasive ductal carcinoma - Tumor margin: Infiltrative - Nuclear grade: 2 - Histologic grade: II (3-2-1) - Maximum diameter of invasive tumor: 1.4cm - Maximum diameter of invasive and in situ carcinoma: 2cm - Lymphovascular invasion: Not identified - Perineural invasion: Not identified - Intraductal carcinoma component: Present - Proportion: 35% (Extensive) - Type: Solid / Cribriform - Nuclear grade: Intermediate - Necrosis: Present (Comedo) - Location: Central and periphery - Calcification: Present (Malignant and benign) - Tumor infiltrating lymphocytes, stromal: $< 10\%$ - Resection margin: - Upper, lower, inner and outer margin: Free of ca - Total LNs (0/11): SLN (0/6), non SLN (0/5) ER: Positive (AS 8 = PS 5 + IS 3) PR: Positive (AS 5 = PS 3 + IS 2) HER2: Negative (1+) AR: Positive (AS 7 = PS 5 + IS 2) Ki67 L.I.: 20.8%in tumor cells - ER/PR/HER2/Ki-67 100/100/1+/20.8% Lum B (HER2 negative)
User	Suggest a radiotherapy approach using the guidance from the given clinical summary of the radiation oncologist. ## `Input:[Generated Clinical Note] ...`
RO-LLaMA-P++	RT to Lt. whole breast 26 Gy / 5 fx + 10 Gy / 5 fx boost.
User	Draw the radiotherapy target volume of this patient with following planning CT scan. ## `Input:[Generated Treatment Plan] ... [3D CT Scan] patient002.dcm` 
RO-LLaMA-SEG++	patient002_segmentation.dcm 

Table S7. Example of prompts for evaluation treatment plan suggestion .

We would like to request your feedback on the performance of the response of the assistant displayed below. In the feedback, I want you to evaluate suitability of the proposed radiotherapy plan by comparing it to the ground truth label according to the following score rubric:

Rubric:
1. Radiation therapy dose (expressed in 'Gy') and fraction (expressed as 'fx') are correct.
2. Radiation therapy target volume (expressed in 'RTOG' or 'ESTRO') is correct.
3. Radiation therapy technique (expressed in 'SIB' or 'PMRT' or 'IMRT' or 'BCS') is correct.
4. Radiation therapy orientation (expressed in 'Rt.' or 'Lt.') is correct.
5. Radiation therapy margin (expressed in 'SCL' or 'IMN' or 'AXL') is correct.
6. The format of the Model Response compared to the format of Ground Truth Answer is similar.

Here are abbreviations.
BCS $\rightarrow$ breast conserving surgery
PMRT $\rightarrow$ post mastectomy radiation therapy
IMRT$\rightarrow$ intensity modulated radiation therapy
PBI $\rightarrow$ partial breast irradiation
SIB $\rightarrow$ with simultaneous integrated boost technique
SCL $\rightarrow$ supraclavicular lymph node
IMN $\rightarrow$ internal mammary lymph node
AXL $\rightarrow$ axilla lymph node

Score Rule:
Score 1: The response totally fails to accomplish the requirements of the rubric.
Score 2: The response partially satisfies the requirements of the rubric, but needs major challenges and improvements to satisfy the requirements.
Score 3: The response mainly satisfies the requirements of the rubric, but it lacks some parts compared to the ground truth answer.
Score 4: The response satisfies the requirements of the rubric competitive to the ground truth answer.
Score 5: The response fully satisfies the requirements of the rubric better than the ground truth answer.

Evaluation Steps:
1. Read the Ground Truth Answer and the Model Response carefully.
2. Assign a score from 1 to 5 based on Score Rule.
3. If the Model Response format is too different from the Ground Truth Answer, Assign a score 0.

[Ground Truth Answer Begin]
{ Answer }
[Ground Truth Answer End]

[Model Response Begin]
{ Output }
[Model Response End]

Evaluation Form (scores ONLY, without explanation and should be limited to only one decimal place):

- Average Score for all Rubric: