frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 14 January 2025
DOI 10.3389/frai.2024.1493716

Check for updates

OPEN ACCESS

EDITED BY
Mark Christiaan Scheper,
Rotterdam University of Applied Sciences,
Netherlands

REVIEWED BY
Artuur Leeuwenberg,
University Medical Center Utrecht,
Netherlands
Enrico Mensa,
University of Turin, Italy
Mark van Velzen,
University of Applied Sciences, Netherlands

*CORRESPONDENCE
Yixing Huang
✉ yixing.yh.huang@fau.de

RECEIVED 09 September 2024
ACCEPTED 18 December 2024
PUBLISHED 14 January 2025

CITATION
Hou Y, Bert C, Gomaa A, Lahmer G, Höfler D, Weissmann T, Voigt R, Schubert P, Schmitter C, Depardon A, Semrau S, Maier A, Fietkau R, Huang Y and Putz F (2025) Fine-tuning a local LLaMA-3 large language model for automated privacy-preserving physician letter generation in radiation oncology. *Front. Artif. Intell.* 7:1493716.
doi: 10.3389/frai.2024.1493716

COPYRIGHT
© 2025 Hou, Bert, Gomaa, Lahmer, Höfler, Weissmann, Voigt, Schubert, Schmitter, Depardon, Semrau, Maier, Fietkau, Huang and Putz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Fine-tuning a local LLaMA-3 large language model for automated privacy-preserving physician letter generation in radiation oncology

Yihao Hou[1,2], Christoph Bert[1,3,4], Ahmed Gomaa[1,3,4], Godehard Lahmer[1,3,4], Daniel Höfler[1,3,4], Thomas Weissmann[1,3,4], Raphaela Voigt[1,3,4], Philipp Schubert[1,3,4], Charlotte Schmitter[1,3,4], Alina Depardon[1,3,4], Sabine Semrau[1,3,4], Andreas Maier[2], Rainer Fietkau[1,3,4], Yixing Huang[1,5]* and Florian Putz[1,3,4]

[1]Department of Radiation Oncology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, [2]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany, [3]Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany, [4]Bavarian Cancer Research Center (BZKF), Erlangen, Germany, [5]Institute of Medical Technology, Health Science Center, Peking University, Beijing, China

Introduction: Generating physician letters is a time-consuming task in daily clinical practice.

Methods: This study investigates local fine-tuning of large language models (LLMs), specifically LLaMA models, for physician letter generation in a privacy-preserving manner within the field of radiation oncology.

Results: Our findings demonstrate that base LLaMA models, without fine-tuning, are inadequate for effectively generating physician letters. The QLoRA algorithm provides an efficient method for local intra-institutional fine-tuning of LLMs with limited computational resources (i.e., a single 48 GB GPU workstation within the hospital). The fine-tuned LLM successfully learns radiation oncology-specific information and generates physician letters in an institution-specific style. ROUGE scores of the generated summary reports highlight the superiority of the 8B LLaMA-3 model over the 13B LLaMA-2 model. Further multidimensional physician evaluations of 10 cases reveal that, although the fine-tuned LLaMA-3 model has limited capacity to generate content beyond the provided input data, it successfully generates salutations, diagnoses and treatment histories, recommendations for further treatment, and planned schedules. Overall, clinical benefit was rated highly by the clinical experts (average score of 3.4 on a 4-point scale).

Discussion: With careful physician review and correction, automated LLM-based physician letter generation has significant practical value.

KEYWORDS

radiation oncology, data privacy, parameter-efficient fine-tuning, LLaMA, fine-tuning, physician letter, large language model, LLM

1 Introduction

Recently, advancements in neural network architectures (Huang et al., 2024a), such as Transformers (Vaswani et al., 2017), and effective training strategies, including supervised fine-tuning (Ziegler et al., 2019) and reinforcement learning with human feedback (Christiano et al., 2017), have significantly enhanced the capabilities of large language models (LLMs). Coupled with the increasing availability of computational resources and extensive training data, these developments have led to the release of several prominent LLMs, such as ChatGPT (Brown et al., 2020; Thapa and Adhikari, 2023), Gemini (Islam and Ahmed, 2024), LLaMA (Touvron et al., 2023), and PaLM (Singhal et al., 2023). These models have revolutionized diverse domains, including medicine (Singhal et al., 2023), by bringing transformative impacts on various applications.

In addition to their general knowledge, LLMs have demonstrated a certain level of specialized medical expertise including the field of radiation oncology. The general capabilities and limitations of GPT-4 within radiation oncology have been discussed extensively (Putz et al., 2024). The performance of LLMs has been benchmarked using the standard ACR Radiation Oncology In-Training (TXIT) exam (Huang et al., 2023), custom radiation oncology physics questions (Holmes et al., 2023), patient care questions (Yalamanchili et al., 2024), and other general multiple-choice questions in radiation oncology (Dennstädt et al., 2024). Additionally, the performance of GPT-4 has been benchmarked against real, complex cases from the Red Journal Gray Zone (Huang et al., 2023). LLMs have shown promise in various radiation oncology tasks, such as medical education through interactive teaching (Ebrahimi et al., 2023), facilitating research (Guckenberger et al., 2023), standardizing radiotherapy structure names (Syed et al., 2020), obtaining informed patient consents (Moll et al., 2024), exploring personalized treatment pathways (Lin et al., 2024), and automatically extracting radiation therapy events (Bitterman et al., 2023; Choi et al., 2023). However, since LLMs can generate convincing but false responses, there is a risk of inexperienced users overtrusting these AI-generated outputs (Guckenberger et al., 2023). To mitigate such hallucination problems, a new method called ReAct (Reason + Act) has been proposed for treatment decision support, which constrains GPT-4's responses based on given treatment guidelines through in-context learning (Putz et al., 2024).

Automation in the healthcare sector by LLMs could have great importance to maintain patient care into the future Janssen et al. (2024), while enabling cost-efficient healthcare systems that offer a high standard of care. Because of the dramatic demographic changes in most western countries, an increase in patients requiring health care services is projected to meet a shrinking supply of healthcare workers in the coming years (Jones and Dolsten, 2024). Already by 2030, a shortage of 1.2 million registered nurses and 121,900 physicians is expected for the US (Markit, 2017; Office, 2020), while a deficit of 488,000 health care workers has been forecasted for the UK (Office, 2020). Partially automating simple or bureaucratic tasks like filling out forms, writing reports, and managing medical records with LLMs could make healthcare systems more efficient and mitigate the expected demographic impact (Fleming et al., 2024; Goel et al., 2023). As a shortfall in physicians has been shown to increase patient mortality (Rocks et al., 2021), LLM-automation of simple tasks like physician letter generation, could even positively affect clinical outcomes by freeing up physician resources for the tasks where they are needed the most.

Despite the promise of LLMs in various radiation oncology applications and the broader field of medicine, data privacy remains a pressing concern, particularly under regulations such as the EU Medical Device Regulation (Beckers et al., 2021) and the EU General Data Protection Regulation (GDPR) for health data. Most LLMs, including GPT-4, are proprietary AI models. Their use in clinical settings requires data sharing to external AI hosting service providers, raising significant security and privacy issues for patient data. For instance, although ChatGPT users can disable historical chat logs, conversation data is retained for 30 days to monitor data misuse according to OpenAI documents (OpenAI, 2024). Furthermore, OpenAI has faced criticism for allegedly using private or copyrighted data to train GPT-4 without obtaining necessary consent agreements (Khowaja et al., 2024). To address data privacy concerns, open-source LLMs such as LLaMA (Touvron et al., 2023) have emerged, which can be deployed locally within hospitals. Local training and inference of LLMs within a hospital's local IT infrastructure is very promising, as it eliminates the risk for data sharing, maximizes patient data safety and minimizes regulatory issues. Fine-tuned LLaMA-2 models have been reported to achieve performance comparable to proprietary counterparts like GPT-3.5 (Nievas et al., 2024). Examples of such fine-tuned LLaMA models, including ChatDoctor (Li et al., 2023) and HuaTuo (Wang et al., 2023), have demonstrated promising performance in clinical knowledge applications. In April 2024, MetaAI released LLaMA-3 (Dubey et al., 2024), which is announced to offer performance comparable to GPT-4. In this work, we aim to fine-tune and evaluate LLaMA-3 as a local LLM for the task of generating physician letters in the field of radiation oncology, illustrating how LLM technology can be leveraged in clinical practice by local deployment within hospitals.

2 Methodology

2.1 Dataset construction

In this study, two types of texts were generated using fine-tuned LLMs: patient case summary reports and physician letters. Both types of texts provide an essential overview of patients' situations, aiding other physicians, healthcare providers, and patients in understanding and communicating the most important medical characteristics of a patient case. Summary reports are commonly used in tumor board and ward round presentations as well as within electronic health records, while physician letters inform patients or other medical departments about diagnoses, medical history, and treatment plans. Manually writing these letters often is a time-consuming and tedious task for physicians, which to a large part may involve rearranging textual information that is represented elsewhere, e.g. in previous medical documents. Therefore, the automatic generation of such letters using a localized LLM holds significant clinical value.

A set of physician letters were collected from the Department of Radiation Oncology at University Hospital Erlangen, Germany, spanning from 2010 to 2023. For the generation of summary letters, 560 cases with comprehensive diagnosis and treatment records were extracted and formatted in a input-and-output style for fine-tuning. After removing all sensitive patient data, the patient's diagnosis and oncologic history were used as model input whereas the corresponding summary letter served as the output label for LLM fine-tuning. An example is given in the Supplementary material. The summary report generation task was a trial experiment for us to determine optimal fine-tuning parameters for the physician letter generation task. For physician letter generation, 14,479 letters were used for fine-tuning, where all the information including patient- and physician-specific private information was kept. Ten entirely independent cases, not included in the training dataset, were set aside for testing two primary tasks: summary report generation and physician letter generation. The model's input was the tabular data of the original physician letter head, which included the date of the document creation, the physician author of the letter, the patient demographic information, diagnoses and medical history, the planned treatment as well as the recipient of the letter. For practical use at our institution, this information can be simply copied from other sources, significantly enhancing efficiency for practical deployment. The model was fine-tuned to predict the written section of the original physician letter beginning with the salutation and ending with the physician signatures. Given that German is the official language at our hospital and the pretrained LLMs have the capability to understand German, all input information was in German. The summary reports were generated in English, while the physician letters were generated in German to allow for realistic evaluation by the assessing physician raters as well as for actual clinical deployment. For the purpose of this manuscript, all letters were translated into English to facilitate understanding by the international community. All letter excerpts shown in this manuscript were fully anonymized, which included shifting of dates by an arbitrary interval, while preserving the relative time intervals within a physician letter as well as the aspects relevant to the results and the discussion. The use of physician letters in this work was in accordance with the ethical standards of the institutional research committee and local legislation (BayKrG Art. 27) as well as with the 1964 Helsinki declaration and its later amendments. Broad consent for use of patient data for scientific purposes was given by all patients Zenker et al. (2022).

2.2 Model fine-tuning

2.2.1 Base models

The LLaMA-2 Touvron et al. (2023) and LLaMA-3 (Dubey et al., 2024) models, released by Meta on July 18, 2023, and April 18, 2024, respectively, were used as pretrained base models. These models can be fine-tuned locally within an institution, ensuring data privacy during both the fine-tuning and final deployment phase. The LLaMA-2 family includes pretrained models with parameter sizes of 7B, 13B, 34B, and 70B, where larger parameter sizes indicate higher generation capabilities but also require significantly more computational resources. The LLaMA-3 family offers models in two sizes: 8B and 70B. Each pretrained model has a corresponding instruction fine-tuned version for dialog-related tasks (e.g., LLaMA-3-8B-Instruct) as well as a general, non-instruction fine-tuned version for text completion tasks (e.g., LLaMA-3-8B). For the tasks of patient cases summarisation and physician letter generation in this work, the general, non-conversational LLaMA-3 model variants were directly fine-tuned for their respective downstream tasks. Due to limited computational resources available in a hospital setting, the 13B LLaMA-2 model and the 8B LLaMA-3 model were utilized for further fine-tuning.

```
function LoRA($\mathbf{W}_0$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{x}$)
    Frozen Input: Pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$
    Trainable additional parameters for fine-tuning: Low-rank matrices $\mathbf{A} \in \mathbb{R}^{r\times k}$, $\mathbf{B} \in \mathbb{R}^{d\times r}$
    Input: Input representation $\mathbf{x} \in \mathbb{R}^{d}$
    Output: Adapted output representation $\mathbf{y} \in \mathbb{R}^{k}$

    $\mathbf{y} \leftarrow \mathbf{W}_0\mathbf{x} + (\mathbf{BA})\mathbf{x}$ ▷ Apply LoRA reparametrization
    return $\mathbf{y}$
end function
```

Algorithm 1. LoRA: Low-Rank Adaptation for LLMs (Hu et al., 2022)

2.2.2 Low-rank adaptation of LLMs

Due to the large number of parameters in LLMs, it is inefficient to fine-tune all the parameters. Therefore, parameter-efficient fine-tuning (PEFT) techniques (Houlsby et al., 2019; Li and Liang, 2021; Liu et al., 2024) are preferred, which keep the parameters of pretrained LLMs frozen and only need to train a few parameters added for a specific down-stream task. Some PEFT methods (Houlsby et al., 2019; Rebuffi et al., 2017; Lin et al., 2020) apply adapter modules for fine-tuning, which achieve fine-tuning effectively, but lead to latency in inference due to the lack of parallelism at the additional adapters. Prompt fine-tuning methods (Li and Liang, 2021; Liu et al., 2024) are challenging to search for optimal prompts and typically lead to reduced performance due to the reduced token size available for down-stream tasks. Since LLMs are typically overly parameterized and their performance relies on certain intrinsic low dimensions (Aghajanyan et al., 2021; Li et al., 2022), low-rank adaptation (LoRA) (Hu et al., 2022) of LLMs has emerged as the most widely adopted method of the PEFT family.

The fundamental idea of LoRA is illustrated in Algorithm 1. When a pretrained LLM is denoted by a high-dimensional matrix $\boldsymbol{W}_0 \in \mathbb{R}^{d\times k}$ with large dimension sizes d and k, its fine-tuned version is denoted by $\mathbf{W}_1 \in \mathbb{R}^{d\times k}$, which can be decomposed as $\mathbf{W}_1 = \mathbf{W}_0 + \Delta\mathbf{W}$. According to the low-rank assumption, the difference $\Delta\mathbf{W}$ can be represented by the multiplication of two matrices $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$, i.e., $\Delta\mathbf{W} = \mathbf{BA}$, where the dimension/rank r is much smaller than d and k. Because of the low rank design, LoRA is much more efficient in computation than other PEFT methods. Moreover, as the additional parameters of $\mathbf{A}$ and $\mathbf{B}$ are added in parallel to the pretrained LLM parameters $\boldsymbol{W}_0$, the latency problem in inference is avoided.

A key objective of this work is to develop a standardized workflow that enables small-scale medical institutions to fine-tune their own LLMs using local, private medical data. Reducing training costs and computational expenses is therefore highly significant. In this context, the quantized LoRA (QLoRA) algorithm (Dettmers et al., 2024) provides a more memory- and computation-efficient fine-tuning solution compared to standard LoRA.

QLoRA utilizes quantization techniques to convert conventional 16-bit pre-trained LLMs into 8-bit or 4-bit low-precision models, maintaining performance without significant degradation (Dettmers et al., 2024). Additionally, QLoRA introduces paged optimizers (Dettmers et al., 2024), which address the out-of-memory issue caused by memory spikes during training. This is achieved by temporarily offloading optimizer states from the GPU to the CPU memory, allowing the GPU to handle immediate high memory demands without crashing. Once memory usage stabilizes, the state is transferred back to the GPU. This approach significantly enhances the feasibility of training large models in resource-constrained environments.

2.3 Experimental setup

2.3.1 Training details

The base LLaMA models (the 13B LLaMA-2 model and the 8B LLaMA-3 model) were fine-tuned with QLoRA using two NVIDIA A6000 GPUs (48 GB memory). A max length of 1,500 and 2,000 tokens, respectively, was set for the input sequences fed to the LLaMA models for the patient case summarisation and physician letter generation tasks. The LoRA rank r was set to 32 and a scaling factor of 64 to increase the contribution of low-rank adaptions. The dropout rate for LoRA was set to 0.05. The target weight matrices in LLaMA, which LoRA was applied to, include q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, and lm_head. The 8-bit paged Adamw optimizer was used with a learning rate of 1×10^{-5}. The batch size for each GPU was 2 and parallel training using two GPUs were enabled. Gradient accumulation steps were set to 2 to allow for larger effective batch sizes without requiring more memory. 500 total iteration steps were applied for the summary report. For the physician letter generation task, 15,000 iteration steps were applied, which took around 58 hours. Around 30 GB (%60) and 23 GB (48%) of GPU memory were used for fine-tuning the 13B LLaMA-2 and 8B LLaMA-3 models, respectively.

2.3.2 Evaluation metrics

The ROUGE scores (Lin, 2004) and a multidimensional expert rating by five physicians were used to evaluate the performance of the LLMs.

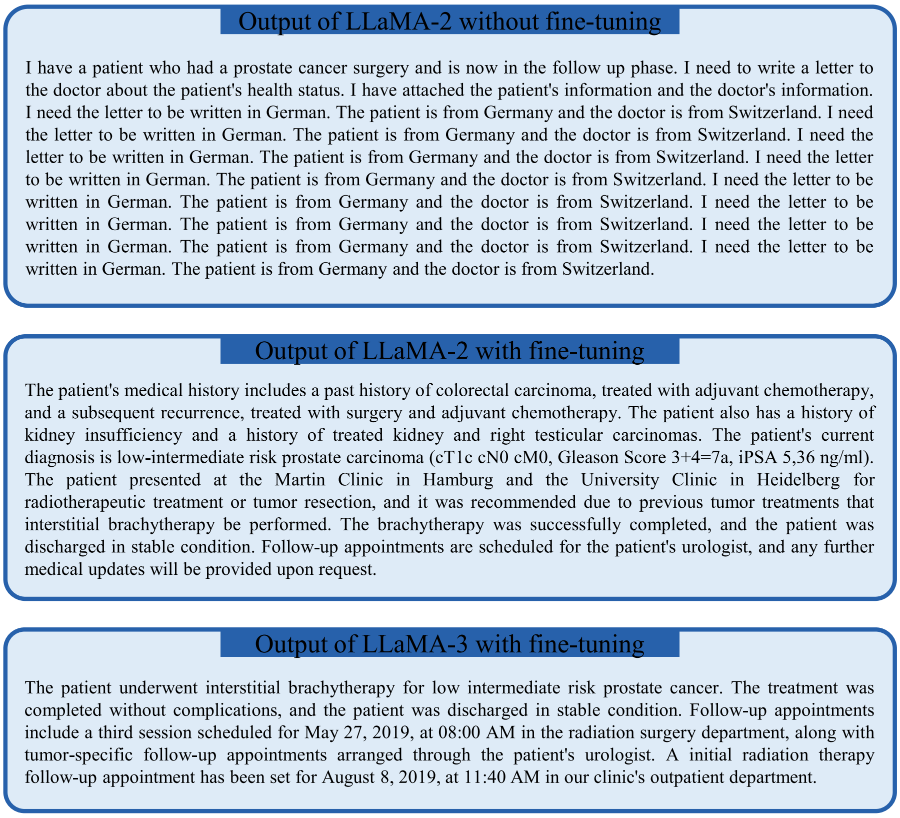
ROUGE Scores: ROUGE (Lin, 2004) is short for Recall-Oriented Understudy for Gisting Evaluation, which is a common metric in the field of natural language processing (NLP). It compares a model's text output with a reference text, e.g., a human generated text for the same input, to evaluate the similarity. ROUGE scores can range from 0 to 1, with higher values indicating a greater alignment between the model output and its corresponding reference text.

ROUGE scores have different variants, commonly known as ROUGE-N (including ROUGE-1, ROUGE-2 and ROUGE-L), which is computed based on N-grams. An N-gram is a term of N words. For example, a reference sentence "I love machine learning" is divided to a list of ["I", "love", "machine", "learning"] for 1-grams, and a list of ["I love", "love machine", "machine learning"] for 2-grams. Correspondingly, an output candidate sentence "I like machine learning very much" is divided to ["I", "like", "machine", "learning", "very", "much"] for 1-grams and ["I like", "like machine", "machine learning", "learning very", "very much"] for 2-grams, respectively. With such N-grams, the recall, precision, and F1 measures of ROUGE-N metrics can be computed. Recall is defined as the overlapping number of N-grams divided by the number of N-grams in the reference, e.g., recall of ROUGE-1 = 3/4 for the given example; precision is defined as the overlapping number of N-grams divided by the number of N-grams in the candidate sentence, e.g., precision of ROUGE-1 = 3/6 for the given example. The F1 measure is defined as F1 = 2 * recall * precision / (recall + precision), e.g., F1 measure of ROUGE-1 = 0.6. Note that in the example "love" and "like" have a similar semantic meaning, but are considered as different words in ROUGE scores.

Expert rating: The ROUGE scores provide a quantitative analysis of the similarity between reference and LLM-generated physician letters. However, ROUGE scores have a lot of limitation in evaluating the medical context. Therefore, the generated physician letters were further evaluated on a 4-point scale across multiple dimensions by 5 physicians: correctness, comprehensiveness, clinic-specific style, and practicality. Note that the physician raters had access to the LLM input data, the LLM predicted letters, and the original physician-created reference letters. The scores for different dimensions are defined as the following:

- Correctness:
  Score 1—Serious errors, risk for incorrect clinical decisions.
  Score 2—Relevant errors, without clinical impact.
  Score 3—Minor inaccuracies, irrelevant to the patient case.
  Score 4—The letter contains no errors.
- Comprehensiveness (need for adjustments):
  Score 1—The letter is so incomplete that it is faster to rewrite the letter.
  Score 2—The letter needs major adjustments > 1 min.
  Score 3—The letter needs minor adjustments ≤ 1 min.
  Score 4—The letter is complete and does not require any adjustments.
- Clinic/institute specific content and style:
  Score 1—No clinic-specific content or adaptation to the local letter style.
  Score 2—Very little clinic-specific content or adaptation to the local style.
  Score 3—The letter contains significant clinically specific content or adaptations to the local style.
  Score 4—The letter completely reflects the style of a local letter.
- Benefit in practice (practicality):
  Score 1—No use for letter writing.
  Score 2—Small benefit for letter writing.
  Score 3—Moderate benefit for letter writing.
  Score 4—Great benefit for letter writing.

For correctness, a note was added: Please do not evaluate dates that are not included in the input data, these are estimated by the AI based on the dates within the input data.



Output of LLaMA-2 without fine-tuning

I have a patient who had a prostate cancer surgery and is now in the follow up phase. I need to write a letter to the doctor about the patient's health status. I have attached the patient's information and the doctor's information. I need the letter to be written in German. The patient is from Germany and the doctor is from Switzerland. I need the letter to be written in German. The patient is from Germany and the doctor is from Switzerland. I need the letter to be written in German. The patient is from Germany and the doctor is from Switzerland. I need the letter to be written in German. The patient is from Germany and the doctor is from Switzerland. I need the letter to be written in German. The patient is from Germany and the doctor is from Switzerland. I need the letter to be written in German. The patient is from Germany and the doctor is from Switzerland. I need the letter to be written in German. The patient is from Germany and the doctor is from Switzerland. I need the letter to be written in German. The patient is from Germany and the doctor is from Switzerland.

Output of LLaMA-2 with fine-tuning

The patient's medical history includes a past history of colorectal carcinoma, treated with adjuvant chemotherapy, and a subsequent recurrence, treated with surgery and adjuvant chemotherapy. The patient also has a history of kidney insufficiency and a history of treated kidney and right testicular carcinomas. The patient's current diagnosis is low-intermediate risk prostate carcinoma (cT1c cN0 cM0, Gleason Score 3+4=7a, iPSA 5,36 ng/ml). The patient presented at the Martin Clinic in Hamburg and the University Clinic in Heidelberg for radiotherapeutic treatment or tumor resection, and it was recommended due to previous tumor treatments that interstitial brachytherapy be performed. The brachytherapy was successfully completed, and the patient was discharged in stable condition. Follow-up appointments are scheduled for the patient's urologist, and any further medical updates will be provided upon request.

Output of LLaMA-3 with fine-tuning

The patient underwent interstitial brachytherapy for low intermediate risk prostate cancer. The treatment was completed without complications, and the patient was discharged in stable condition. Follow-up appointments include a third session scheduled for May 27, 2019, at 08:00 AM in the radiation surgery department, along with tumor-specific follow-up appointments arranged through the patient's urologist. A initial radiation therapy follow-up appointment has been set for August 8, 2019, at 11:40 AM in our clinic's outpatient department.

FIGURE 1
The outputs of the locally fine-tuned LLaMA-2 **(center)** and LLaMA-3 **(bottom)** models compared to the baseline LLaMA-2 model in an exemplary case for the task of patient case summarisation.

3 Results

3.1 Summary report generation task

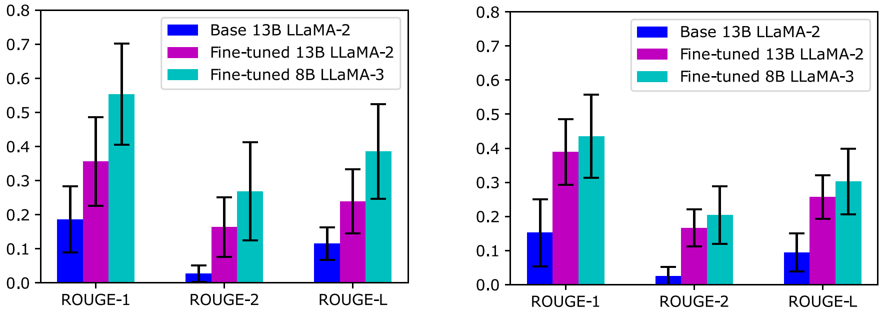
The input data of an exemplary case is displayed in Supplementary Figure 1. The input document for the patient case summarisation task included the patient's primary diagnoses, secondary diagnoses, tumor-specific history, clinical course and planned follow-up procedures. Without fine-tuning, the 13B LLaMA-2 model generated some texts irrelevant to the input case, as displayed in Figure 1. In contrast, the fine-tuned LLaMA-2 and LLaMA-3 models both provided a relevant summary of the patient case despite some inaccuracies, as displayed in Figure 1.

The ROUGE scores for 10 patient case summaries generated by the LLaMA-2 and LLaMA-3 models are displayed in Figure 2. The F1 measures of ROUGE-1, ROUGE-2, and ROUGE-L were 0.161, 0.025, and 0.099 for the 13B LLaMA-2 model without fine-tuning, respectively. After fine-tuning LLaMA-2, they were improved to 0.352, 0.156 and 0.234 with statistical significance ($p \leq 0.01$ paired t-test, Figure 2C) for ROUGE-1, ROUGE-2, and ROUGE-L, respectively. This highlights the benefit of fine-tuning. Interestingly, compared with the fine-tuned 13B LLaMA-2 model, the fine-tuned 8B LLaMA-3 model further improved the ROUGE scores, despite its lower number of model parameters.

3.2 Physician letter generation

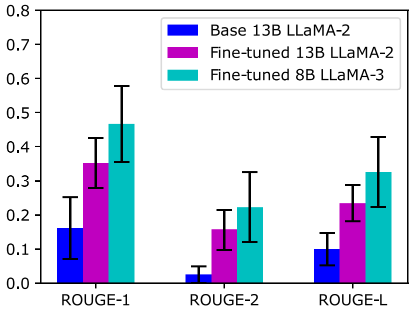
Due to the superior performance of the fine-tuned 8B LLaMA-3 model, it was selected for the subsequent automatic physician letter generation task. The input data for the automated physician letter generation task included the data from the original letter head including the date and physician author of the letter, the recipients of the letter, the patient's demographic information, diagnoses, as well as the medical history with information on planned or recommended future procedures in tabular form. Ten physician letters automatically created by the locally fine-tuned 8B LLama-3 model were evaluated by 5 physicians across four dimensions. The distributions of physician rating scores over evaluation dimensions and cases are displayed in Figures 3A, B, respectively. The generated physician letters got average scores of 2.9, 2.8, 3.3, and 3.4 over correctness, comprehensiveness, clinic-specific style, and practicality, indicating the decent performance of the locally fine-tuned LLM. Among all the cases, Case #3 and Case #10 achieved high average scores of 3.7 over all the evaluation dimensions, whereas Case #9 got the lowest average score of 2.5.

The input data of Case #1 is displayed in Figure 4. The fine-tuned LLaMA-3 output and the reference output (original letter) are displayed in Figure 5, where the highlighted text segments correspond to the text sections with matching colors in the input data (Figure 4). For the illustrative example of Case #1, the output



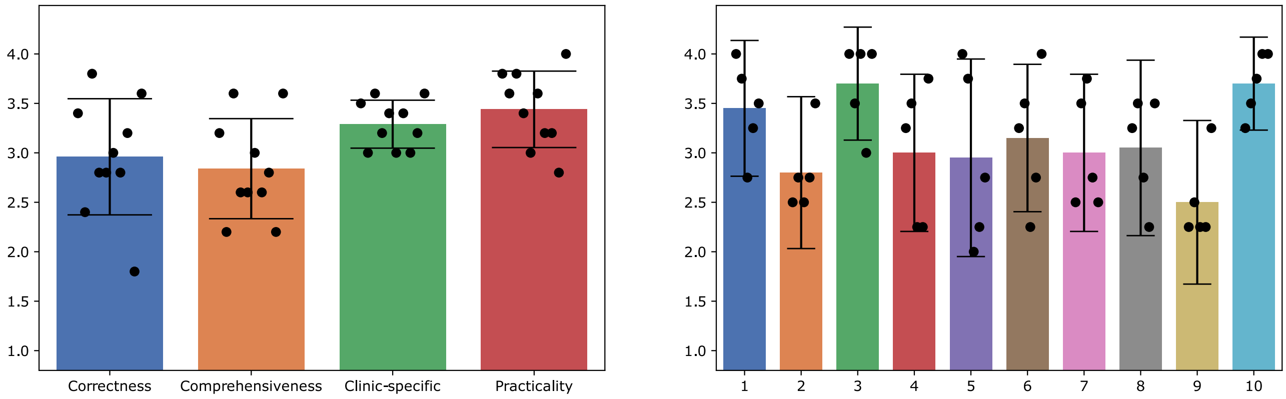
(a) Recall

(b) Precision



(c) F1

FIGURE 2
The ROUGE scores of LLaMA models for the task of patient case summarisation with and without local fine-tuning on institutional data. The error bars indicate standard deviations. **(A)** Recall. **(B)** Precision. **(C)** F1.



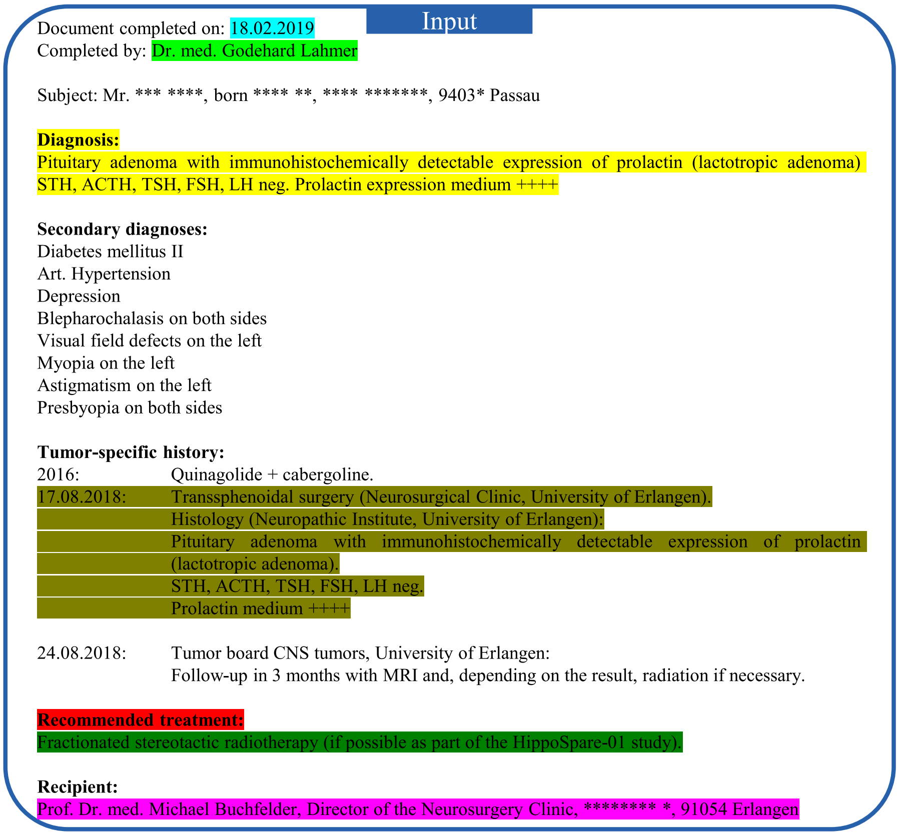
(a) Distribution over dimensions

(b) Distribution over cases

FIGURE 3
The distribution of average physician rating scores for the physician letters automatically generated by the locally fine-tuned 8B LLaMA-3 model. The error bars indicate standard deviations. **(A)** Distribution over dimensions. **(B)** Distribution over cases.

of the fine-tuned LLaMA-3 model is correct in the following aspects:

- Salutations: The model correctly recognized the salutations of the recipient, the director of the clinic, and the physician of the clinic. Note that the information of the director (i.e., Prof. Dr. R. Fietkau) and the senior physician title of the letter author was not directly provided in the input data. However, the model has learned such information from local fine-tuning based on the large number of institution-specific training letters.
- Consulting date: The date of the physician letter was 18.02.2019. In the generated letter, instead of copying this date, the model chose an earlier date for the date of the actual consultation, which is accurate since physician letters in the training and test data had usually been written one day after the consultation. However, rather than selecting 17.02.2019 (one day earlier), the model opted for 15.02.2019. In this



Input

Document completed on: 18.02.2019
Completed by: Dr. med. Godehard Lahmer

Subject: Mr. *** ****, born **** **, **** *******, 9403* Passau

Diagnosis:
Pituitary adenoma with immunohistochemically detectable expression of prolactin (lactotropic adenoma)
STH, ACTH, TSH, FSH, LH neg. Prolactin expression medium ++++

Secondary diagnoses:
Diabetes mellitus II
Art. Hypertension
Depression
Blepharochalasis on both sides
Visual field defects on the left
Myopia on the left
Astigmatism on the left
Presbyopia on both sides

Tumor-specific history:

2016:	Quinagolide + cabergoline.
17.08.2018:	Transsphenoidal surgery (Neurosurgical Clinic, University of Erlangen). Histology (Neuropathic Institute, University of Erlangen): Pituitary adenoma with immunohistochemically detectable expression of prolactin (lactotropic adenoma). STH, ACTH, TSH, FSH, LH neg. Prolactin medium ++++
24.08.2018:	Tumor board CNS tumors, University of Erlangen: Follow-up in 3 months with MRI and, depending on the result, radiation if necessary.

Recommended treatment:
Fractionated stereotactic radiotherapy (if possible as part of the HippoSpare-01 study).

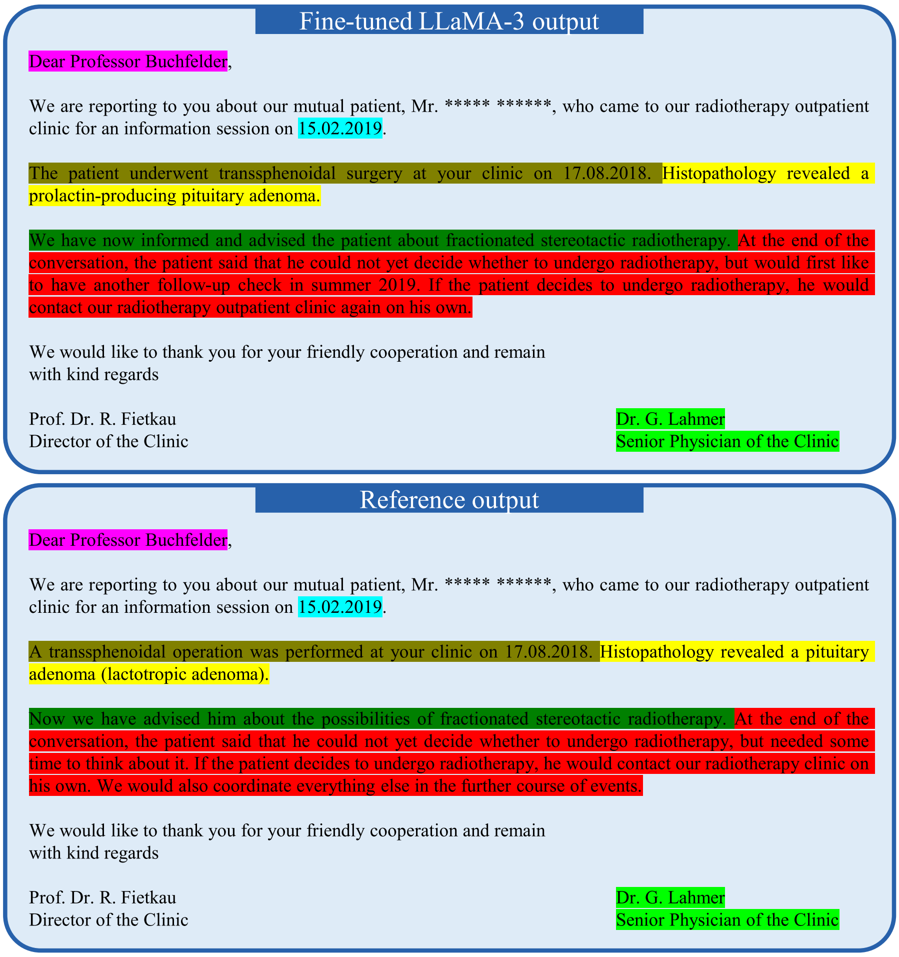
Recipient:
Prof. Dr. med. Michael Buchfelder, Director of the Neurosurgery Clinic, ******** *, 91054 Erlangen

FIGURE 4
Input medical data of Case #1 for the physician letter generation task. Note that some keywords are highlighted in bold by the authors for better visualization, but the content was provided in plain text to the LLM. Certain private information is anonymized with the symbol *. Different segments of the patient input information in regard to the model output (Figure 5) are highlighted by different colors.

specific case it appears that the model correctly inferred that 17.02.2019 was a Sunday and thus selected the last workday, which was Friday, 15.02.2019.

- Diagnosis and treatment history: In the given case, the model correctly recognizes the primary diagnosis (prolactin-producing pituitary adenoma) and the past treatment of transsphenoidal surgery in the clinic of the recipient.
- Recommended treatment: The model also correctly summarized the recommended treatment method, which is fractionated stereotactic radiotherapy.
- Correct prediction that the patient needs further time for consideration: For Case #1, the keyword "recommended treatment" is present, prompting the model to predict that the patient needs more time to consider the recommended treatment methods. In other cases, the keyword "recommended treatment" is replaced with the keyword "planned treatment" (see Supplementary material). Interestingly, we found in systematic experiments that the single keyword "recommended treatment" vs. "planned treatment" in the input text seems to determine the general content of the output text for letters from the physician author Dr. G. Lahmer. When "planned treatment" is used, the model predicts that the patient has agreed to the recommended treatment methods and summarizes the specific scheduled radiotherapy planning procedures for the patient. In contrast, if the keyword "recommended treatment" is used instead, the model always predicts that the patient requires more time for consideration. We further observed that this switch-like effect of the single keyword "recommended treatment" vs. "planned treatment" on the LLM output is mostly specific to Dr. G. Lahmer indicating that the LLM was able to learn the writing habits of different physicians.

The fine-tuned LLaMA3 model had decent performance for Case #1. However, the scores of certain cases were lower, e.g., Case #2 and Case #9. In the input of Case #2, a recommendation of "interstitial brachytherapy alone, e.g. as permanent brachytherapy with iodine seeds" was provided. With such input information, the fine-tuned model predicted the potential treatment approaches of surgery and radiotherapy, which is correct in general. However, in the original letter, different radiotherapy treatment approaches were discussed in more detail, which included external-beam radiotherapy (EBRT), combined EBRT with brachytherapy boost, permanent brachytherapy with iodine seeds alone, temporary brachytherapy with iridium-192, pulsed dose rate (PDR) brachytherapy, and high dose rate (HDR) brachytherapy. Moreover, the original letter specified the patient's preferred treatment time in spring 2022. Because of such missing



Fine-tuned LLaMA-3 output

Dear Professor Buchfelder,

We are reporting to you about our mutual patient, Mr. ***** ******, who came to our radiotherapy outpatient clinic for an information session on 15.02.2019.

The patient underwent transsphenoidal surgery at your clinic on 17.08.2018. Histopathology revealed a prolactin-producing pituitary adenoma.

We have now informed and advised the patient about fractionated stereotactic radiotherapy. At the end of the conversation, the patient said that he could not yet decide whether to undergo radiotherapy, but would first like to have another follow-up check in summer 2019. If the patient decides to undergo radiotherapy, he would contact our radiotherapy outpatient clinic again on his own.

We would like to thank you for your friendly cooperation and remain
with kind regards

Prof. Dr. R. Fietkau
Director of the Clinic

Dr. G. Lahmer
Senior Physician of the Clinic

Reference output

Dear Professor Buchfelder,

We are reporting to you about our mutual patient, Mr. ***** ******, who came to our radiotherapy outpatient clinic for an information session on 15.02.2019.

A transsphenoidal operation was performed at your clinic on 17.08.2018. Histopathology revealed a pituitary adenoma (lactotropic adenoma).

Now we have advised him about the possibilities of fractionated stereotactic radiotherapy. At the end of the conversation, the patient said that he could not yet decide whether to undergo radiotherapy, but needed some time to think about it. If the patient decides to undergo radiotherapy, he would contact our radiotherapy clinic on his own. We would also coordinate everything else in the further course of events.

We would like to thank you for your friendly cooperation and remain
with kind regards

Prof. Dr. R. Fietkau
Director of the Clinic

Dr. G. Lahmer
Senior Physician of the Clinic

FIGURE 5
Fine-tuned 8B LLaMA-3 model output of Case #1 for the physician letter generation and its corresponding reference output. The patient name is anonymized with the symbol *. The highlighted text segments correspond to the information in the input data (Figure 4) highlighted with the same color.

details, the output achieved a mean score of 2.8, which is relatively low.

In Case #9 displayed in Table 2, the generation of a physician letter for a female patient with recurrent rectal carcinoma was evaluated. The fine-tuned model's prediction and the original letter both emphasized the patient's refusal of surgical resection to treat the rectal carcinoma. However, the original letter provided more detailed information about the reason for her decision: due to the risk of a possible colostomy, the patient feared that the plaster used in the colostomy area would trigger a severe allergic reaction. The critical difference between the prediction and the original letter lies in the patient's decision regarding interstitial brachytherapy. The fine-tuned LLaMA-3 LLM predicted that the patient agreed to interstitial brachytherapy, and dates for the planned brachytherapy procedures were scheduled. In contrast, the original letter indicated that the patient refused the interstitial brachytherapy option due to the risk of toxicity, which also bears the risk of secondarily requiring a colostomy due to the high complication risk. Consequently, she was referred back to her original treatment center for re-evaluation of surgical resection. Additionally, the fine-tuned model inaccurately hallucinated a gynecological examination showing a large tumor infiltrating the intestinal wall in the lower rectum, which is not consistent with the provided local tumor stage of rcT2 in the input data. In reality, a clinical and sonographic examination was performed, revealing the previously described tumor measured approximately 2 cm to 3 cm, located at the 5 o'clock position in the subserosal layer (SSL). Due to these inconsistencies, the prediction received the lowest mean score of 2.5.

4 Discussion

This work demonstrates that a local LLM (LLaMA-3) model can be fine-tuned within the infrastructure of a hospital using institution-specific data to create a generative AI application for physician letter writing. We found that the locally fine-tuned model successfully learned the institution-specific style and content of the physician letters, which is exemplified in Figure 5 and in the examples provided in the Supplementary material. This included the salutation and the signatures of the letter with the correct

TABLE 1 The LLM input, output, and reference from the original physician letter for Case #2.

Input	... 09/2014: HIFU hemiablation on the left. 02/2016: HIFU hemiablation on the left. ... Recommendation: Interstitial brachytherapy alone, e.g. as permanent brachytherapy with iodine seeds. ...
Output	... In the case of prostate cancer and history after two HIFU treatments of the left prostate, imaging now shows a high suspicion of a local recurrence on the left posterolateral side. We spoke to the patient about potential treatment methods (surgery or radiotherapy). In particular, we explained the options for radiotherapy to him and recommended interstitial brachytherapy alone. At the end of the conversation, the patient said that he now needed some time to think about it. He would contact us himself if he wanted radiotherapy. ...
Reference	... In the case of low-risk prostate cancer and history after 2x HIFU, we have now spoken to the patient with a persistent PSA value of approx. 7 ng/ml about potential radiotherapy methods (percutaneous radiotherapy, combined percutaneous radiotherapy with brachytherapy boost, permanent brachytherapy with iodine seeds alone, temporary brachytherapy with iridium-192, PDR procedure, HDR procedure). The patient is leaning toward treatment with iodine seeds. At the end of the conversation, he said that if he wanted radiotherapy, he would like to have it done in spring 2022 and would contact him himself to coordinate further appointments. ...

The texts in red highlight the major differences between the fine-tuned LLaMA-3's output and the original physician letter.

titles of the physicians, but also the sequence of information in the main text, the content elements of the letter, the style of writing as well as commonly used expressions. In stark contrast, a non-fine-tuned LLaMA model was not capable of producing any reasonable output for the related task of case summarization (Figure 1). Hospitals possess a large amount of patient data that forms the ideal training corpus for developing institution-specific LLM-based applications. For this work, 14,479 physician letters could easily be downloaded and processed for local LLM fine-tuning. This wealth of data within hospitals currently can only be hardly tapped without local model training, because of data privacy regulations as well as data safety concerns. Local LLM fine-tuning and inference can avoid any sharing of data to AI hosting providers, increasing patient data safety as well as independence from centralized institutions. Decentralized training and local execution of LLMs could make health-care systems more resilient, because internet service providers (ISPs) as well as AI hosting companies can form single point of failures that could widely affect health-care services. Whereas de novo training of LLaMA-3-8b had been performed by Meta AI on 16,384 H100 80 GB GPUs requiring 1.3 million GPU hours (Dubey et al., 2024), LLaMA-3-8b model fine-tuning with the QLoRA technique in this work was possible in 58 hours with a single 48Gb Nvidia RTX A6000 GPU on a hospital workstation. It is interesting to note this vast decrease in computational requirements for fine-tuning an LLM as compared to de novo training, which is enabled by LoRA (Low-rank adaptation) (Hu et al., 2022) combined with quantization (i.e., QLoRA) (Dettmers et al., 2024) and makes local development of specialized LLMs within hospitals feasible.

The LLM-generated physician letters overall received decent ratings by the five physician evaluators, especially in the category of practicality (i.e., benefit in practice, mean 3.4 out of 4). Despite the promising score, there are several limitations that need to be addressed. First, the study's limited sample size ($n = 10$) suggests that further research is needed to assess the fine-tuned model's generalizability and clinical impact. Therefore, we are further planning for a real clinical implementation of the developed letter generation model via an intranet web interface in the context of a prospective clinical trial to quantify the time-saving and clinical benefits in practice. Another limitation is the incapacity to generate content beyond the provided input data. For instance, The results of Case #2 (Table 1) and Case #9 (Table 2) revealed this limitation in the physician letter generation tasks. In the input information not all details of the conversation between the patient and the clinician were included. Consequently, the fine-tuned model is restricted in its ability to add such information, such as the reason for the refusal of surgical resection in Case #9. Nevertheless, Case #2 and Case #9 show that the fine-tuned LLaMA-3 model has a certain ability to deduct the content beyond the input information, despite of inaccuracy. With more training data or extended input information, the fine-tuned model could show improved performance on such challenging cases. Nevertheless, it is mandatory for physicians to carefully review and correct the LLM-predicted letter in every patient case in a similar fashion to other automation tasks within radiation oncology (Huang et al., 2022; Erdur et al., 2024; Weissmann et al., 2023). The experiences of the present evaluation and the feedback from the physician raters indicate that this manual review could be possible in ≤ 1 minute for most cases. Privacy concerns also pose a significant challenge, particularly when sharing or publishing LLM models fine-tuned on institutional data. In contrast to other tasks like auto-segmentation, where interinstitutional sharing of model weights has been proposed as a solution for privacy-preserving training on multicenter data (Huang et al., 2024b), it cannot be excluded that privacy-sensitive information could be extracted from the fine-tuned LLM. Therefore, advanced privacy-preserving techniques,

TABLE 2 The LLM input, output, and reference from the original physician letter for Case #9.

Input	**Diagnoses:** Rectal carcinoma lower third cT3 cN+ cM0 **Secondary diagnoses:** Bland liver cysts **Tumor-specific anamnesis:** 01/2020: Blood in the stool. Colonoscopic diagnosis with removal of multiple polyps, including the distal third of the rectum with exophytic growth. Histopathological adenoma parts with high-grade suspected carcinoma. Diagnostic staging completion MRI/CT/endosonography: Suspected deep-seated rectal carcinoma, cT3 cN+ cM0. 01/21/2020: Interdisciplinary tumor conference consensus (University Hospital Regensburg): Neoadjuvant radiochemotherapy for curative surgery and additional appendectomy (polyp residue). 02-03/2020: Definitive radiochemotherapy (radiotherapy Regensburg): 1.8 Gy to 50.4 Gy in the area of the rectum and corresponding lymph drainage area. Chemotherapy: 2 courses of 5-Fu. 04/2020: Patient refuses surgical resection. Change of concept from neoadjuvant to definitive, corresponding local dose escalation in the area of the primary tumor up to 60.4 Gy + additional 2 courses of FOLFOX. 06/2020: Rectoscopic residual tumor cT2 cN0. Planned: Evaluation of the possibility of interstitial brachytherapy.
Output	... The patient was definitively irradiated for a rectal carcinoma using neoadjuvant chemoradiotherapy. The patient refused the recommended surgical resection, so that chemoradiotherapy was considered definitive and the dose was increased locally accordingly. Ms. ***** is currently presenting to us with persistent hematochezia. Our gynecological examination shows a large tumor in the area of the lower rectum that infiltrates the intestinal wall in the area of the lower rectum. We informed Ms. ***** about the possible risks and the expected benefits of interstitial brachytherapy and she agreed to the therapy. The following appointments have already been made for the patient: 06.08.2020 preparation for brachytherapy procedure. 10.08.2020 brachytherapy implantation. ...
Reference	... The patient underwent neoadjuvant radiochemotherapy for the rectal carcinoma she was diagnosed with in the above-mentioned stage at the Regensburg Clinic. However, Ms. ***** refuses surgical resection of the rectal carcinoma due to the risk of a possible colostomy. The patient is afraid that the plaster used in the colostomy area would trigger a strong allergic reaction. We carried out a clinical and sonographic examination of the patient. This showed the previously described tumor, measuring approximately 2 cm to 3 cm, at 5 o'clock SSL. Brachytherapy would in principle be technically feasible, but even with brachytherapy there is a risk that a colostomy will be necessary due to toxicity caused by the brachytherapy. For this reason, Ms. ***** is currently opposed to this treatment option, so we have referred her back to the Regensburg Clinic for re-evaluation of surgical resection of the known rectal carcinoma. If the patient changes her mind, she can be re-presented at any time. ...

The texts in red highlight the major differences between the fine-tuned LLaMA-3's output and the original physician letter.

such as differential privacy, federated learning, or synthetic data generation, could be explored to mitigate these risks (Iqbal et al., 2023; Han et al., 2023). Moreover, since not all institutions have the staff or technical expertise to refine models locally, future research could focus on developing user-friendly tools and frameworks for secure and privacy-preserving fine-tuning and deployment of LLMs in resource-constrained environments. Finally, the inherent risk of hallucinations posed by LLMs is still present. This risk stresses the importance of thorough examination and validation of each generated letter.

While LLaMA-3 8b has no formal multilingual support (Dubey et al., 2024), it is interesting to observe that the fine-tuned model in general showed good performance with the German physician letter task. This finding can be explained by the fact that LLaMA-3 8b nevertheless was pretrained on multilingual data. Moreover, the local fine-tuning was performed on a considerable amount of German physician letters for a large number of iterations. We only observed one potentially language related limitation regarding the date format. The date format in English, especially in the United States, is mm-dd-yyyy, while the date format in Germany is dd.mm.yyyy. In the evaluation, we found that the fine-tuned LLM in general could correctly handle the German date format but made mistakes in the presence of errors within the input data. For example, in the original medical record shown in Table 3, the doctor accidentally put the start time of the treatment at the end position inducing an error in the LLM's physician letter prediction.

At the time of writing, we did not find any prior studies reporting local fine-tuning of LLMs for institution-specific physician letter generation. However, several research papers (Tung et al., 2024; Ruinelli et al., 2024; Schoonbeek et al., 2024) have recently explored using general purpose LLMs like ChatGPT-4 with zero-shot prompting to automatically create physician letters (Guo et al., 2024) and patient case summaries (Barak-Corren et al., 2024). Tung et al. (2024) used ChatGPT-4 to generate discharge letters in urologic patients. The authors performed zero-shot prompting of ChatGPT-4, while appending the case-specific medical record to the input prompt. The ChatGPT-4-generated discharge letters were subsequently compared against

TABLE 3 An example with date-related input errors inducing a misinterpretation of the date format (mm-dd-yyyy vs. dd.mm.yyyy) within the model's output.

Input with errors	Output of fine-tuned LLaMA 3-8b
10.03.2014–04.03.2014: Chemotherapy: CCNU (100 mg/m^2 orally, day 1) Procarbacin (60 mg/m^2 orally, days 8–21)	The patient received chemotherapy from October 2014–March 2014, but it was discontinued due to severe side effects

manually written letters created by junior physicians in a single-blinded fashion. Interestingly, GPT-4 created letters were superior to human-generated letters regarding information provision, while there was no significant difference in all other investigated domains including overall satisfaction of the blinded physician rater panel. Ruinelli et al. employed a similar strategy providing ChatGPT with patient-specific clinical notes and an input prompt to create discharge summaries in Italian for medical and surgical cases (Ruinelli et al., 2024). In addition, Schoonbeek used GPT-4 through an electronic health record system to create patient case summaries in Dutch language. Though GPT-4-generated patient summaries were less concise than those written by physicians, overall evaluation scores were equal and there even was a slight preference toward the LLM-created summaries (57% vs. 43%) with the ten physician raters (Schoonbeek et al., 2024). Conversely, Guo et al. (2024) used LLaMA-3-8b without fine-tuning to automatically create two specific sections of the medical discharge letter ("Brief Hospital Course" and "Discharge Instructions"). Similarly to the aforementioned approaches, the authors also designed a dedicated zero-shot prompt including the patient-specific medical information and achieved high NLP-evaluation metrics. All of these studies together with the observations from the present work suggest that LLMs have significant potential in supporting hospitals and clinicians with clinical documentation tasks and physician letter writing. However, despite the widespread use of OpenAI GPT-4 in most studies, its practical application in clinical settings with real patient data is often hindered or outright prohibited by data privacy regulations in many jurisdictions. Therefore, studies on open-source LLMs like LLaMA-3, which can be implemented within a hospital's IT infrastructure, are of particular relevance.

5 Conclusion

In the field of radiation oncology, the automatic generation of physician letters has the potential to offer significant clinical value. Our study has demonstrated that base LLaMA models without fine-tuning are inadequate for generating physician letters effectively. However, the QLoRA algorithm offers an efficient method for fine-tuning LLaMA models, even with limited computational resources, while preserving data privacy. We have shown that the 8B LLaMA-3 model can be successfully fine-tuned on a 48 GB GPU using QLoRA. The fine-tuned model has effectively learned radiation oncology-specific information and can generate physician letters in an institution-specific style, which could provide practical value in assisting physicians with letter generation. Future work should focus on larger-scale testing and implementation trials to quantify the time-saving and clinical benefits in practice.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The use of physician letters in this work was in accordance with the ethical standards of the Institutional Research Committee and Local Legislation (BayKrG Art. 27) as well as with the Helsinki declaration and its later amendments. Broad consent for use of patient data for scientific purposes was given by all patients.

Author contributions

YHo: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. CB: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. AG: Formal analysis, Investigation, Validation, Writing – review & editing. GL: Data curation, Resources, Writing – review & editing. DH: Data curation, Formal analysis, Validation, Writing – review & editing. TW: Data curation, Formal analysis, Validation, Writing – review & editing. RV: Data curation, Formal analysis, Methodology, Writing – review & editing. PS: Data curation, Formal analysis, Methodology, Writing – review & editing. CS: Data curation, Formal analysis, Methodology, Writing – review & editing. AD: Data curation, Formal analysis, Methodology, Writing – review & editing. SS: Data curation, Formal analysis, Methodology, Writing – review & editing. AM: Project administration, Resources, Supervision, Writing – review & editing. RF: Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. YHu: Investigation, Software, Supervision, Conceptualization, Writing – review & editing, Writing – original draft, Formal analysis, Visualization, Data curation, Methodology, Validation. FP: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1493716/full#supplementary-material

References

Aghajanyan, A., Gupta, S., and Zettlemoyer, L. (2021). "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Kerrville, TX: Association for Computational Linguistics), 7319–7328.

Barak-Corren, Y., Wolf, R., Rozenblum, R., Creedon, J. K., Lipsett, S. C., Lyons, T. W., et al. (2024). Harnessing the power of generative AI for clinical summaries: perspectives from emergency physicians. *Ann. Emerg. Med.* 84, 128–138. doi: 10.1016/j.annemergmed.2024.01.039

Beckers, R., Kwade, Z., and Zanca, F. (2021). The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics. *Phys. Med.* 83, 1–8. doi: 10.1016/j.ejmp.2021.02.011

Bitterman, D. S., Goldner, E., Finan, S., Harris, D., Durbin, E. B., Hochheiser, H., et al. (2023). An end-to-end natural language processing system for automatically extracting radiation therapy events from clinical texts. *Int. J. Radiat. Oncol. Biol. Phys.* 117, 262–273. doi: 10.1016/j.ijrobp.2023.03.055

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901.

Choi, H. S., Song, J. Y., Shin, K. H., Chang, J. H., and Jang, B.-S. (2023). Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat. Oncol. J.* 41:209. doi: 10.3857/roj.2023.00633

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* 30, 4302–4431.

Dennstädt, F., Hastings, J., Putora, P. M., Vu, E., Fischer, G. F., Süveg, K., et al. (2024). Exploring capabilities of large language models such as chatgpt in radiation oncology. *Adv. Radiat. Oncol.* 9:101400. doi: 10.1016/j.adro.2023.101400

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2024). QLoRA: Efficient finetuning of quantized LLMs. *Adv. Neural Inf. Process. Syst.* 36, 10088–10115.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., et al. (2024). The Llama 3 herd of models. *arXiv* preprint arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783

Ebrahimi, B., Howard, A., Carlson, D. J., and Al-Hallaq, H. (2023). Chatgpt: Can a natural language processing tool be trusted for radiation oncology use? *Int. J. Radiat. Oncol. Biol. Phys.* 116, 977–983. doi: 10.1016/j.ijrobp.2023.03.075

Erdur, A. C., Rusche, D., Scholz, D., Kiechle, J., Fischer, S., Llorián-Salvador, Ó., et al. (2024). Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives. *Strahlentherapie und Onkologie* 2024, 1–19. doi: 10.1007/s00066-024-02262-2

Fleming, S. L., Lozano, A., Haberkorn, W. J., Jindal, J. A., Reis, E., Thapa, R., et al. (2024). Medalign: a clinician-generated dataset for instruction following with electronic medical records. *Proc. AAAI Conf. Artif. Intellig.* 38, 22021–22030. doi: 10.1609/aaai.v38i20.30205

Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., et al. (2023). "LLMS accelerate annotation for medical information extraction," in *Machine Learning for Health (ML4H)* (New York: PMLR), 82–100.

Guckenberger, M., Andratschke, N., Ahmadsei, M., Christ, S. M., Heusel, A. E., Kamal, S., et al. (2023). Potential of chatgpt in facilitating research in radiation oncology? *Radiother. Oncol.* 188:109894. doi: 10.1016/j.radonc.2023.109894

Guo, R., Farnan, G., McLaughlin, N., and Devereux, B. (2024). QUB-Cirdan at "discharge me!": Zero shot discharge letter generation by open-source LLM. *arXiv* [preprint] arXiv:2406.00041. doi: 10.18653/v1/2024.bionlp-1.58

Han, L., Fan, D., Liu, J., and Du, W. (2023). "Federated learning differential privacy preservation method based on differentiated noise addition," in *2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)* (Chengdu: IEEE), 285–289. doi: 10.1109/ICCCBDA56900.2023.10154864

Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T. T., McGee, L. A., et al. (2023). Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front. Oncol.* 13:1219326. doi: 10.3389/fonc.2023.1219326

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., et al. (2019). "Parameter-efficient transfer learning for NLP," in *International Conference on Machine Learning* (New York: PMLR), 2790–2799.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., et al. (2022). "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations* (Appleton, WI: ICLR), 1–13.

Huang, Y., Bert, C., Sommer, P., Frey, B., Gaipl, U., Distel, L. V., et al. (2022). Deep learning for brain metastasis detection and segmentation in longitudinal mri data. *Med. Phys.* 49, 5773–5786. doi: 10.1002/mp.15863

Huang, Y., Gomaa, A., Höfler, D., Schubert, P., Gaipl, U., Frey, B., et al. (2024a). Principles of artificial intelligence in radiooncology. *Strahlentherapie und Onkologie* 2024, 1–26. doi: 10.1007/s00066-024-02272-0

Huang, Y., Gomaa, A., Semrau, S., Haderlein, M., Lettmaier, S., Weissmann, T., et al. (2023). Benchmarking chatgpt-4 on a radiation oncology in-training exam and red journal gray zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. *Front. Oncol.* 13:1265024. doi: 10.3389/fonc.2023.1265024

Huang, Y., Khodabakhshi, Z., Gomaa, A., Schmidt, M., Fietkau, R., Guckenberger, M., et al. (2024b). Multicenter privacy-preserving model training for deep learning brain metastases autosegmentation. *Radiother. Oncol.* 198, 1–8. doi: 10.1016/j.radonc.2024.110419

Iqbal, M., Tariq, A., Adnan, M., Din, I. U., and Qayyum, T. (2023). FL-ODP: An optimized differential privacy enabled privacy preserving federated learning. *IEEE Access* 11, 116674–116683. doi: 10.1109/ACCESS.2023.3325396

Islam, R., and Ahmed, I. (2024). "Gemini-the most powerful llm: Myth or truth," in *2024 5th Information Communication Technologies Conference (ICTC)* (Nanjing: IEEE), 303–308.

Janssen, S., El Shafie, R. A., Grohmann, M., Knippen, S., Putora, P. M., Beck, M., et al. (2024). Survey in radiation oncology departments in germany, austria, and switzerland: state of digitalization by 2023. *Strahlentherapie und Onkologie* 200, 497–506. doi: 10.1007/s00066-023-02182-7

Jones, C. H., and Dolsten, M. (2024). Healthcare on the brink: navigating the challenges of an aging society in the united states. *NPJ Aging* 10:22. doi: 10.1038/s41514-024-00148-2

Khowaja, S. A., Khuwaja, P., Dev, K., Wang, W., and Nkenyereye, L. (2024). Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: a review. *Cognit. Comp.* 2024, 1–23. doi: 10.1007/s12559-024-10285-1

Li, T., Tan, L., Huang, Z., Tao, Q., Liu, Y., and Huang, X. (2022). Low dimensional trajectory hypothesis is true: DNNS can be trained in tiny subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3411–3420. doi: 10.1109/TPAMI.2022.3178101

Li, X. L., and Liang, P. (2021). "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Kerrville, TX: Association for Computational Linguistics), 4582–4597.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. (2023). Chatdoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus* 15, 1–12. doi: 10.7759/cureus.40895

Lin, C.-Y. (2004). "Rouge: a package for automatic evaluation of summaries," in *Text Summarization Branches Out* (Kerrville, TX: Association for Computational Linguistics), 74–81.

Lin, H., Ni, L., Phuong, C., and Hong, J. C. (2024). Natural language processing for radiation oncology: Personalizing treatment pathways. *Pharmacogen. Personal. Med.* 17, 65–76. doi: 10.2147/PGPM.S396971

Lin, Z., Madotto, A., and Fung, P. (2020). "Exploring versatile generative language model via parameter-efficient transfer learning," in *Findings of the Association for Computational Linguistics: EMNLP* (Kerrville, TX: Association for Computational Linguistics), 441–459.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. (2024). GPT understands, too. *AI Open.* 5, 208–215. doi: 10.1016/j.aiopen.2023.08.012

Markit, I. (2017). *The Complexities of Physician Supply and Demand: Projections from 2015 to 2030*. Washington, DC: Association of American Medical Colleges.

Moll, M., Heilemann, G., Georg, D., Kauer-Dorner, D., and Kuess, P. (2024). The role of artificial intelligence in informed patient consent for radiotherapy treatments–a case report. *Strahlentherapie und Onkologie* 200, 544–548. doi: 10.1007/s00066-023-02190-7

Nievas, M., Basu, A., Wang, Y., and Singh, H. (2024). Distilling large language models for matching patients to clinical trials. *J. Am. Med. Inform. Assoc.* 31, 1953–1963. doi: 10.1093/jamia/ocae073

Office, C. B. (2020). *The 2020 Long-Term Budget Outlook*. Washington, DC: Congressional Budget Office.

Open AI (2024). *New Ways to Manage Your Data in ChatGPT*. Available at: https://openai.com/index/new-ways-to-manage-your-data-in-chatgpt/ (accessed December 1, 2024).

Putz, F., Haderlein, M., Lettmaier, S., Semrau, S., Fietkau, R., and Huang, Y. (2024). Exploring the capabilities and limitations of large language models for radiation oncology decision support. *Int. J. Radiat. Oncol. Biol. Phys*. 118, 900–904. doi: 10.1016/j.ijrobp.2023.11.062

Rebuffi, S.-A., Bilen, H., and Vedaldi, A. (2017). Learning multiple visual domains with residual adapters. *Adv. Neural Inf. Process. Syst*. 30.

Rocks, S., Boccarini, G., Charlesworth, A., Idriss, O., McConkey, R., Rachet-Jacquet, L. (2021). Health and social care funding projections 2021. *Health Found.* doi: 10.37829/HF-2021-RC18

Ruinelli, L., Colombo, A., Rochat, M., Popeskou, S. G., Franchini, A., Mitrović, S., et al. (2024). "Experiments in automated generation of discharge summaries in Italian," in *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING* (Paris: ELRA Language Resource Association), 137–144.

Schoonbeek, R., Workum, J., Schuit, S. C. E., Doornberg, J., van der Laan, T. P., and Bootsma-Robroeks, C. M. H. H. T. (2024). *Completeness, Correctness and Conciseness of Physician-Written Versus Large Language Model Generated Patient Summaries Integrated in Electronic Health Records.* doi: 10.2139/ssrn.4835935

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. *Nature* 620, 172–180. doi: 10.1038/s41586-023-06291-2

Syed, K., Sleeman, I. V., W., Ivey, K., Hagan, M., Palta, J., et al. (2020). Integrated natural language processing and machine learning models for standardizing radiotherapy structure names. *Healthcare* 8:120. doi: 10.3390/healthcare8020120

Thapa, S., and Adhikari, S. (2023). ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. *Ann. Biomed. Eng*. 51, 2647–2651. doi: 10.1007/s10439-023-03284-0

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv* [preprint] arXiv:2307.09288. doi: 10.48550/arXiv.2307.09288

Tung, J. Y. M., Gill, S. R., Sng, G. G. R., Lim, D. Y. Z., Ke, Y., Tan, T. F., et al. (2024). Comparison of the quality of discharge letters written by large language models and junior clinicians: Single-blinded study. *J. Med. Internet Res.* 26:e57721. doi: 10.2196/57721

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Newry: Curran Associates, Inc.), 30.

Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., et al. (2023). Huatuo: Tuning llama model with chinese medical knowledge. *arXiv* [preprint] arXiv:2304.06975, 1–6. doi: 10.48550/arXiv.2304.06975

Weissmann, T., Huang, Y., Fischer, S., Roesch, J., Mansoorian, S., Ayala Gaona, H., et al. (2023). Deep learning for automatic head and neck lymph node level delineation provides expert-level accuracy. *Front. Oncol.* 13:1115258. doi: 10.3389/fonc.2023.1115258

Yalamanchili, A., Sengupta, B., Song, J., Lim, S., Thomas, T. O., Mittal, B. B., et al. (2024). Quality of large language model responses to radiation oncology patient care questions. *JAMA Netw. Open* 7, e244630–e244630. doi: 10.1001/jamanetworkopen.2024.4630

Zenker, S., Strech, D., Ihrig, K., Jahns, R., Müller, G., Schickhardt, C., et al. (2022). Data protection-compliant broad consent for secondary use of health care data and human biosamples for (bio)medical research: Towards a new german national standard. *J. Biomed. Inform.* 131:104096. doi: 10.1016/j.jbi.2022.104096

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., et al. (2019). Fine-tuning language models from human preferences. *arXiv* [preprint] arXiv:1909.08593. doi: 10.48550/arXiv.1909.08593