Muhammad Usman Manzoor

The Application of Large Language Models for Treatment Recommendations in Oral Cancer

Thesis

School of Technology and Innovations
Master's Thesis - Sustainable and Autonomous Systems

Vaasa 2026

UNIVERSITY OF VAASA
School of Technology and Innovations
Author: Muhammad Usman Manzoor
Title of the Thesis: The Application of Large Language Models for Treatment Recommendations in Oral Cancer
Degree: Master of Science in Computer Science
Programme: Sustainable and Autonomous Systems.
Supervisor: Dr. Mahmoud Elsanhoury (University of Vaasa), Dr. Rasheed Alabi (University of Helsinki)
Year: 2026 **Sivumäärä:** 56

ABSTRACT:

The thesis explores the use of Large Language Models (LLMs) to assist in radiotherapy treatment recommendations for patients with oral cancer, which is a subtype of head and neck cancer (HNC) that involves a complicated process of clinical decision-making. The research thesis examines the manner in which sophisticated LLM-based models could be used to support clinicians through the analysis of structured clinical data and subsequent creation of guideline-congruent treatment recommendations. Four LLMs, namely, BioGPT, Llama, GEMMA, and Meditron were the chosen LLM architectures for radiotherapy treatment recommendations according to the patient characteristics and tumour staging information. In the study, a quantitative experimental design was adopted and anonymised clinical datasets were used. Ordered patient records have been converted to natural language prompts, which can be used to interact with language models. Quantized Low-Rank Adaptation (QLoRA) was used to implement parameter-efficient fine-tuning, which enables the models to acquire task-related knowledge without having to retrain the whole network. To assess the performance of model predictions, typical performance metrics were used, such as accuracy, precision, recall, F1-score and confusion matrix analysis. The findings show that LLM-based architectures are able to process structured clinical information and provide precise treatment advice. It was revealed from the experiments that Llama-3.1 showed the best predicting capacity among all the tested models with an accuracy rate of about 93-94%, whereas BioGPT had good biomedical capabilities and provided predictions with an accuracy rate of about 86-87%. The comparative analysis also revealed that the performance in model architectures differed, thereby demonstrating the significance of domain-specific training and prompt engineering. In general, the results are indicative of the fact that the LLMs can be used as clinical decision-support systems in oncology. Even though such systems cannot substitute professional judgment, they may improve the multidisciplinary decision-making processes, helping clinicians to analyse the patient data and provide evidence-based treatment.

KEYWORDS: Head and Neck Cancer; Oral Cancer; Fine-tuning LLM; Large Language Models, Clinical Decision Support, BioGPT, Llama, Gemma, Meditron, Radiotherapy Prediction, Artificial Intelligence in Oncology.

Contents

Table of Figures

List of Tables

Abbreviations

Abbreviation	Full form
AI	Artificial Intelligence
AJCC	American Joint Committee on Cancer
CSV	Comma-Separated Values
CUDA	Compute Unified Device Architecture
DL	Deep Learning
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HNC	Head and Neck Cancer
HNSCC	Head and Neck Squamous Cell Carcinoma
HPV	Human Papillomavirus
LLM	Large Language Model
MDT	Multidisciplinary Team
ML	Machine Learning
NCI	National Cancer Institute
NLP	Natural Language Processing
OSCC	Oral Squamous Cell Carcinoma
QLoRA	Quantized Low-Rank Adaptation
TNM	Tumour-Node-Metastasis

1 Introduction

1.1 Head and Neck Cancer

Head and Neck Cancer (HNC) is one of the most significant health issues worldwide (Deng et al., 2025). A recent review of precision medicine showed that HNC is the sixth most common cancer in the world, with about 900,000 new cases and 460,000 new deaths each year (Miserocchi et al., 2022). As such, it represents one of the most prevalent types of cancer globally (Miserocchi et al., 2022). According to the National Cancer Institute (NCI), HNC includes the larynx, throat, lips, mouth, and nose, along with salivary glands (NCI, 2025). The site within the HNC includes the oropharynx, hypopharynx, nasopharynx, oral and lip cavities, and larynx, along with salivary glands (Miserocchi et al., 2022). In most cases, HNC arises from the mucosal lining of the upper aerodigestive tract (Dunn et al., 2026; Johnson et al., 2020). Specifically, those that originate from the mucosal epithelium of the oral cavity are referred to as HNSCC (Johnson et al., 2020). HNSCC risk factors encompass exposure to carcinogens, including tobacco and alcohol (Dunn et al., 2026; Johnson et al., 2020). Lifestyle factors and viral risks also influence the etiology of HNSCC. According to Johnson et al., oral cavities and larynx cancers are typically linked to tobacco use, alcohol abuse, or both, whereas pharynx cancers are linked to HPV infection (Johnson et al., 2020). The most common type of HNC is Squamous Cell Carcinoma (Miserocchi et al., 2022). Remarkably, head and squamous cell carcinoma (HNSCC) are aggressive and presented at a late stage (Miserocchi et al., 2022; Jiang et al., 2025). Owing to the presentation of HNSCC at a late stage (Jiang et al., 2025), the treatment should be carefully planned to improve the overall survival of the patients.

Over the past few years, treatment interventions have been developed in respect to these patients (Tolley et al., 2023), including targeted therapies and immunotherapies that aim to address the unique challenges posed by HNSCC, such as resistance to conventional treatments and the need for personalised approaches based on Tumour characteristics. Nonetheless, their overall survival has not turned out to be very

satisfactory (Jiang et al., 2025), indicating a need for more effective treatment strategies and interventions to improve patient outcomes. In recent years, machine learning (ML) and deep learning (DL) predictive models have shown promising performance and can enhance the overall management of HNSCC by performing several diagnostic and prognostic tasks in medical oncology, such as improving early detection rates and predicting treatment outcomes in patients with HNSCC (Mäkitie et al., 2023; Alabi et al., 2021a; Alabi et al., 2021c). With the advancements in computing capacity and availability of medical data in several formats, large language models (LLMs) have been trained on large textual data. In recent years, LLMs have shown promising performance in recognising patterns in texts. This ability may be useful in summarising medical texts which may supplement clinical decision-making for personalised medicine aimed at improving the overall management of cancer. Thus, this thesis primarily seeks to investigate the use of LLMs for treatment recommendations in patients with oral cancer.

1.2 Oral cancer

The most common form of oral cancer is the squamous cell carcinoma. Thus, oral squamous cell carcinoma (OSCC), which falls under the broad category of HNSCC, is a cancer that develops in the mucosal lining of the oral cavity and it is by far the most common form of oral cancer, accounting for around 90% of all oral cancers and affecting everyday functions that most people take for granted, including speaking, swallowing, tasting, and appearance (Tan et al., 2023). Lip and mouth cancer ranks as the 16th most prevalent type of cancer worldwide, but the incidence rates have been increasing steadily with each passing year (Kijowska et al., 2024). Annually, there are 389,000 reported new incidences of lip and mouth cancers around the world, and if current trends are allowed to continue, the number will rise by as much as 65% between now and 2050 (Coletta et al., 2024). In the U.S., there are 11.7 cases detected per 100,000 individuals annually while mortality rate is 2.7 per 100,000 (based on cases documented from 2019 to 2023 and mortalities from 2020 to 2024) (NCI, 2025).

The problem is especially a concern given that oral cancer is an easily detectable type of cancer. Unlike tumours buried deep inside the body, abnormalities in the mouth can often be seen or felt during a routine visit to a dentist or doctor (Tan et al., 2023; González-Moles et al., 2022). And yet, most patients walk into their first meeting with an oncologist already carrying advanced disease, not because the cancer was difficult to find, but because nobody looked for it early enough (González-Moles et al., 2022). Once the disease has reached an advanced stage, everything becomes harder. Treatment is more intensive, side effects are more severe, and the chances of surviving the disease drop considerably (González-Moles et al., 2022).

Treatment decisions depend largely on the stage of advancement of the illness. Indeed, the Tumor Nodal Metastasis (TNM) staging, which considers the classification of a tumour according to its size, cancer involvement of the lymph nodes and metastasis elsewhere in the body, forms the foundation of most treatment decision-making frameworks (Suri et al., 2024). Nonetheless, TNM staging was created to offer prognosis for large groups of patients but not considering the uniqueness of each person. For instance, it fails to take into account the general health status, genetic makeup, responsiveness of a person to previous interventions and tolerance to intensive therapies among others (Kang et al., 2025; Alabi et al., 2021c). As a result, there is an increased need for alternative technologies. In recent years, artificial intelligence (AI) subfields such as ML and DL have shown promising results for personalized prognostication (Alabi et al., 2021a-c).

As stated in the previous part, ML and DL have proved themselves to be highly efficient in HNC, and the same applies to the use of these technologies for detecting oral cancer (Alabi et al., 2021a). The early diagnosis of the disease is one of the most crucial aspects of research in the sphere, and it was exactly the limitation that led to greater interest in the application of LLMs in practice (Hao et al., 2025). Thus, this thesis is focused on analysing four different LLMs, namely, BioGPT, Llama, GEMMA, and Meditron, for

generating treatment recommendations, especially, radiotherapy recommendation for patients with oral cancer.

1.3 Objectives

The following are the objectives of the thesis:

1. To explore the use of four main LLMs for treatment recommendations for patients with OSCC. The chosen LLMs are BioGPT, Llama, GEMMA, and Meditron.
2. The highest performing LLM is integrated as a web-based prognostic tool

1.4 Significance and expected contributions

The study is relevant for personalised medicine by leveraging LLMs' ability to analyse complex data and generate human-like text based on the data. These models are intended as ancillary tools for clinical decision-making. Predicting overall survival in patients with cancer is challenging due to different patient-related factors such as age at diagnosis, gender, marital status, pathological grade, tumour characteristics, and available treatment modalities. Traditionally, the American Joint Cancer committee (AJCC) TNM staging system represents an important tool for staging and treatment recommendations. However, the TNM staging has been criticized because it primarily estimates prognosis at the population level. For an individual patient, it is less effective for estimating outcomes due to its inability to consider other tumour- and patient-related risk factors (Alabi et al., 2021c), such as genetic mutations, overall health status, and response to previous treatments. To this end, the LLM considers these factors together to accurately predict patients' outcomes for targeted treatment planning.

2 Literature Review

The current medical treatment of oral cancer is founded on multimodal treatment comprising surgery, radiotherapy and systemic therapy, based on staging, resectability and expected functional results for the individual (Huang & O'Sullivan, 2013). The UK National Multidisciplinary Guidelines on HNC suggest treatment of OSCC is managed by a Multidisciplinary tumor (MDT) board with the purpose of maintaining oncological control and ensuring patient speech, swallowing and eating function are preserved (Homer & Winter, 2024). It is recommended treatment should be selected according to tumour accessibility and preservation of oral function. Oncology control should be maintained by a well-defined and evidence-based multidisciplinary approach (Homer & Winter, 2024; Suri et al., 2024).

2.1 Current treatment approaches for oral cancer

Remarkably, radiotherapy is frequently administered using an intensity-modulated radiotherapy approach and it represents a treatment modality of choice (Huang & O'Sullivan, 2013). According to Huang and O'Sullivan (2013), treatment strategy between surgery and radiotherapy depends upon tumour extent along with expected functional and cosmetic outcome. Huang and O'Sullivan (2013) further indicated that the selection of sole or combined modality is based on various considerations including disease control probability, the anticipated functional and cosmetic outcomes, tumour resectability, patient general condition, and availability of resources and expertise. In more advanced oral cancer, concurrent chemoradiotherapy is now a normal non-surgical organ-sparing approach (Huang & O'Sullivan, 2013). According to Suri et al. (2024), surgery is the primary treatment for early-stage OSCC, while adjuvant radiotherapy or chemoradiotherapy is incorporated for patients at higher risk of recurrence (Suri et al, 2024). In cases where the organ-preserving method is either ineffective or impossible, total surgical resection with reconstruction is still a significant salvage or first-line choice (Huang & O'Sullivan, 2013).

To allow for effective treatment planning and personalised treatment that can facilitate enhanced decision-making, AI-based models have been explored. In a cross-model comparison of Turkish oncology board examinations, Erdat and Kavak (2025) highlights the performance difference in the oncological knowledge. The models Claude 3.5 Sonnet along with ChatGPT 4o demonstrated good performance. On the other hand, Llama-3 and Gemini 1.5 demonstrated poor performance. This means there exists variability in model performance in oncology related tasks. The results generated through AI models in response to user queries suggest that the application of LLMs in education relating to oncology is encouraging, yet they also reveal that the general-purpose models find it hard to reach special reasoning, especially in such a limited field as HNC. The authors found that Llama-3 showed limited accuracy (29.3%) for thoracic tumours and an accuracy of 30.8% for HNC. These findings suggest that there exist limitations in domain specific reasoning within LLM architecture (Lotfian et al., 2025). LLM-based decision support in oncology can be improved by accommodating disease specific parameters.

Some existing literature has compared the LLMs to human experts in oncology decision-making. However limited research highlights specific tumour sites, like OSCC. The first comparison of ChatGPT-4o and Llama3 in otorhinolaryngology, as reported by Buhr et al. (2025), showed good performance. The accuracy score of 84 and 92 per cent was reported. They did not conduct research on real-world diagnostic records. They used simulated cases. Furthermore, LLMs performance in distinguishing between the largely similar oral cavity subtypes was not considered. These models reached a reasonable consensus with human Tumour boards. The researchers highlight the need of supervised clinical integration within LLM recommendations. Thus, leaving open questions regarding whether further fine-tuning can eliminate the difference between human and machine reasoning (Mathew et al., 2024).

Based on this, this thesis evaluates four language models, BioGPT, Llama 3.1 8B, Gemma 3 1B-IT, and Meditron 7B, using structured oral cancer patient records and parameter-efficient fine-tuning with QLoRA. Each model is trained and evaluated using the same

prompt format, standardised treatment labels, and test dataset to compare its ability to recommend radiotherapy or no radiation for oral cancer patients. Models will therefore take the same patient scenarios, based on anonymised oral cancer datasets that contain patient information, basic tumor-related characteristics, and treated parameters for treatment recommendation. The experimental design of the study allows adaptation in real time and reproducibility. With controlled fine-tuning instructions and responses dependent on the model, it is possible to compare the effect of various architectures and mechanisms of fine-tuning on diagnostic accuracy and interpretability. In the study by Erdat and Kavak (2025), which concluded that less specialised training data or smaller models might not be ready yet to be used in the application that demands the use of detailed medical knowledge. This work evaluates influence of structured prompts in establishing reasoning stability across specialised oral cancer scenarios.

2.2 Treatment recommendation challenges

Although several treatment choices exist but there are various constraints that have been found in the treatment of OSCC. For example, functional outcomes, toxicity, and the quality of life may be affected (Ahmadsei et al., 2022). Similarly, Li et al. observe that although there have been improvements in the treatment procedure but still the overall survival rate is low. The treatment is adversely affected as factors like poor targeting and low bioavailability causes low accuracy (Li et al., 2024). Traditionally, the TNM staging forms a cornerstone for treatment recommendations. Despite accuracy of the TNM staging to a reasonable extent still it is not suitable for personalised oncology. Owing to the need to consider other Tumour- and patient-related risk factors, the TNM may not represent an intuitive approach (Alabi et al., 2021c). This indicates the importance of integrating multidimensional clinical variables within decision support framework beyond traditional systems. In this thesis, LLMs are explored as decision-support models that may integrate multiple clinical variables for radiotherapy-related treatment recommendation in a structured experimental setting.

Ah-thiane et al. (2025) compared the performance of LLMs with expert multidisciplinary team choices in oncology (Ah-thiane et al., 2025). They found a high concordance rate. The accuracy of the model was also significant. The results indicate accuracy of 86.6 in tasks treatment recommendations. Guideline-based therapies showed a notably better performance by LLMs, which means that they have a good command of organised clinical reasoning. Thus, there exists a need of controlled implementation of LLM recommendations in complex clinical scenarios. Ahn et al. (2025) show that the application of LLMs in ENT is becoming increasingly popular in clinical decision support, research support, and medical education (Ahn et al., 2025). Large-scale pre-training and instruction tuning can give general-purpose LLMs a benefit over smaller domain-specific models. However, inconsistencies within benchmarking methodologies causes issues within reliable assessment of LLM performance in oral oncology applications. Buhr et al. (2025) made a direct comparison of the recommendations of LLMs with human MDT decisions in HNC oncology and presented concordance rates of 84% with ChatGPT-4o and 92% with Llama 3 (Buhr, 2025). Both models proved useful in identifying the strategies of curative and palliative, and MDT members expressed the likelihood of the benefits of the decision support. The results produced by LLMs closely matched MDT recommendations, however they must be used for assistance rather than replacement of clinicians.

In the specific context of oral cancer, Karuppan Perumal et al. (2025) highlight that AI-driven clinical decision support systems hold genuine promise for improving the quality and consistency of early diagnosis and treatment planning. Their review notes that improvements in patient survival remain fundamentally tied to how quickly and accurately the disease is detected and treated, an area where LLM-based tools are increasingly being explored (Hao et al., 2025). Similarly, Alabi et al. (2021b) demonstrated that deep ML applied to OSCC pathological images achieved diagnostic accuracies of between 77.89% and 97.51%, establishing a strong evidence base upon which LLM-based approaches can now build.

2.3 Comparing LLMs for treatment recommendations

This thesis investigates whether fine-tuned LLMs can support radiotherapy-related treatment recommendation for oral cancer patients. Having a prior model for treatment recommendation can facilitate positive outcomes. This is quite important considering the complex nature of the management of oral cancer. Oncology treatment planning tasks with LLMs have shown growth as shown by recent studies. The article by Sarker (2024) highlights the ability of LLMs in assisting within complex clinical environment. Despite the promising potentials of LLMs, there are concerns, such as hallucinations and ethical concerns, that should be carefully considered and addressed prior to implementation (Chen et al., 2025a). The study by Benary et al. (2023) investigated the role of LLMs in enhanced clinical decision-making by comparing treatment recommendations by LLMs with those of a human expert molecular Tumour Board (Benary et al., 2023). The authors engaged ten fictional and clinically realistic cancer cases and evaluated four LLMs based on the number of treatments as well as their relevancy that they generated.

The experiment suggested that a single LLM showed low overlapping with expert advice, as the F1-scores of 0.04, -0.19, and that the combination of the products of several models showed insignificant enhancement in performance (F1 = 0.29). Despite the situation where, in some cases, LLMs suggested some novel or potentially useful treatment options, their recommendations were not often supported with enough clinical reasoning and could be readily recognised as the AI form of recommendation. This study highlighted the limitations of LLM-generated treatment recommendations in oncology. Their results emphasize that recommendations of the treatment that are generated by LLM are frequently overly reliant on prompt formulation and do not necessarily consider clinical individual limitations and this may lead to uncritical implementation in high-stakes clinical care settings. The finding of this research indicates that there is a need of supervision over LLM recommendations before use in clinical applications.

Although Chen et al. (2025b) acknowledge the practical constraint associated with clinical deployment. Even though the authors state that LLM can be used to complement decision support systems by extracting and synthesising information about electronic health records and biomedical literature, they point out that, nevertheless, such applications are not well-validated. The key challenge exists in terms of evaluation framework specific to disease specific oncology related applications (Dusin et al., 2023). In the context of oral cancer specifically, Hao et al. (2025) conducted a systematic review. LLM integrations in cancer decision-making reported an average overall accuracy of 76.2% across cancer types. While this evidence base is encouraging, the authors caution that performance varied considerably across different cancer types and clinical tasks, reinforcing the need for disease-specific evaluation of LLMs in oral oncology settings.

2.4 Research gap and contribution summary

The literature survey highlights the performance gap that exists corresponding to general purpose LLMs. This exists especially in case of specialized oncological tasks. This was reflected in detection of HNC using Llama-3 that yield accuracy of 30.8% only (Lotfian et al., 2025). This indicates that general purpose LLMs generally have limited domain specific precision. Furthermore, it also indicates that they have limited application in clinical decision making. Although previous studies have examined LLMs in oncology education, tumour board simulation, and general treatment recommendation tasks, limited research has focused specifically on oral cancer treatment recommendation using disease-specific structured patient records. Existing studies often evaluate general-purpose models using simulated cases, examination-style questions, or broad oncology tasks, while fewer studies compare biomedical and general-purpose LLMs under the same dataset, prompt structure, fine-tuning method, and evaluation metrics (Benary et al., 2023; Buhr et al., 2025; Hao et al., 2025).

This thesis addresses this gap by showing that fine tuning LLMs on structured oral cancer datasets can significantly improve performance of prediction. The models are evaluated using standardised patient records, binary treatment labels, and parameter-efficient

fine-tuning with QLoRA. The contribution of this work is not to replace clinical decision-making, but to examine whether fine-tuned general-purpose or biomedical LLMs can provide consistent decision-support outputs in a controlled oral cancer prediction task. This supports future research on safe, interpretable, and clinically supervised AI-assisted treatment planning.

3 Methodology

The study design is aimed at a comparative evaluation of general-purpose and medical-adapted LLMs in terms of radiotherapy treatment recommendations. Rather than using just a single LLM, this research aims to demonstrate comparison between multiple LLMs than just a single-model assessment approach. The chapter describes the research design and rationale, research philosophy, flow of proposed work, data collection, label standardisation, prompt engineering, model selection, parameter-efficient fine-tuning using QLoRA, model training procedure, inference on the test dataset, performance evaluation, comparative model analysis, experimental environment, reproducibility and implementation structure, reliability and validity, ethical and data protection considerations, and integration into a web-based prognostic tool. Such an approach ensures a holistic way of assessing the quality of LLMs. The focus on the rationale of model choice, the fine-tuning discipline, prompt engineering rigor, and the clinically based assessment make the outcomes of the current study at once scientific and practical.

3.1 Research design and rationale

The research design implemented in the study is a quantitative and comparative experimental design focused on evaluating LLMs within a controlled clinical reasoning task. The proposed project aims to determine whether LLMs can support treatment recommendations for oral cancer by analysing heterogeneous clinical information. Unlike conventional ML approaches that process numerical feature, this study conceptualises LLMs as text-based reasoning systems that can be used to support clinicians by organising information, promoting factors of interest, and proposing treatment options. A comparative design was selected in order to make sure that the disparity in model performances could be linked to architectural design and training data, and not the experimental bias. The same clinical cases, prompt structures and evaluation criteria were used to evaluate all the models, to facilitate fair and reproducible comparison. The design is a direct response to limitations found in the existing literature,

with most studies testing a single model in isolation or testing generic NLP benchmarks that are not specifically for medical applications.

3.2 Research philosophy

The study has a positivist philosophy as a cornerstone because it is assumed that clinically significant patterns can be objectively learnt based on data. Philosophical viewpoint provided above is suitable for AI application in the field of oncology studies where the quality of model functioning is evaluated based on objective criteria, such as accuracy and predictability. However, the pragmatism concepts should also be considered in terms of safety and feasibility of clinical application since the implementation of the model into the real life is more about confidence and reliability than merely technical performance.

3.3 Flow of proposed work

This section (Figure 1) explains methodological approach used in proposing work. The crucial issues that are discussed in this section involve designing, implementing, tuning, and assessing transformer-based LLM system. Proposed approach is used to predict radiation therapy requirements for patients suffering from oral cancer and combines structured clinical data with instruction fine-tuning strategies. In addition, it involves parameter-efficient tuning approaches to create an effective and computationally efficient prediction system. Overall methodological frameworks given within following Figure 1. The methodology follows a systematic workflow with different phases. These include dataset preparation, label standardisation, prompt engineering, model selection, QLoRA-based fine-tuning, supervised training, inference generation, and comparative evaluation. Four transformer architectures including BioGPT, Llama-3.1-8B, Gemma-3-1B, and Meditron-7B were fine-tuned to test performance of each on oral cancer dataset. The benchmarking using identical preprocessing pipelines was applied to ensure experimental consistency. These models were selected for evaluation due to their advance architecture and complimentary capabilities in biomedical language

understanding and clinical reasoning performance. The implementation of each model is done in Python and mentioned in Appendix A.

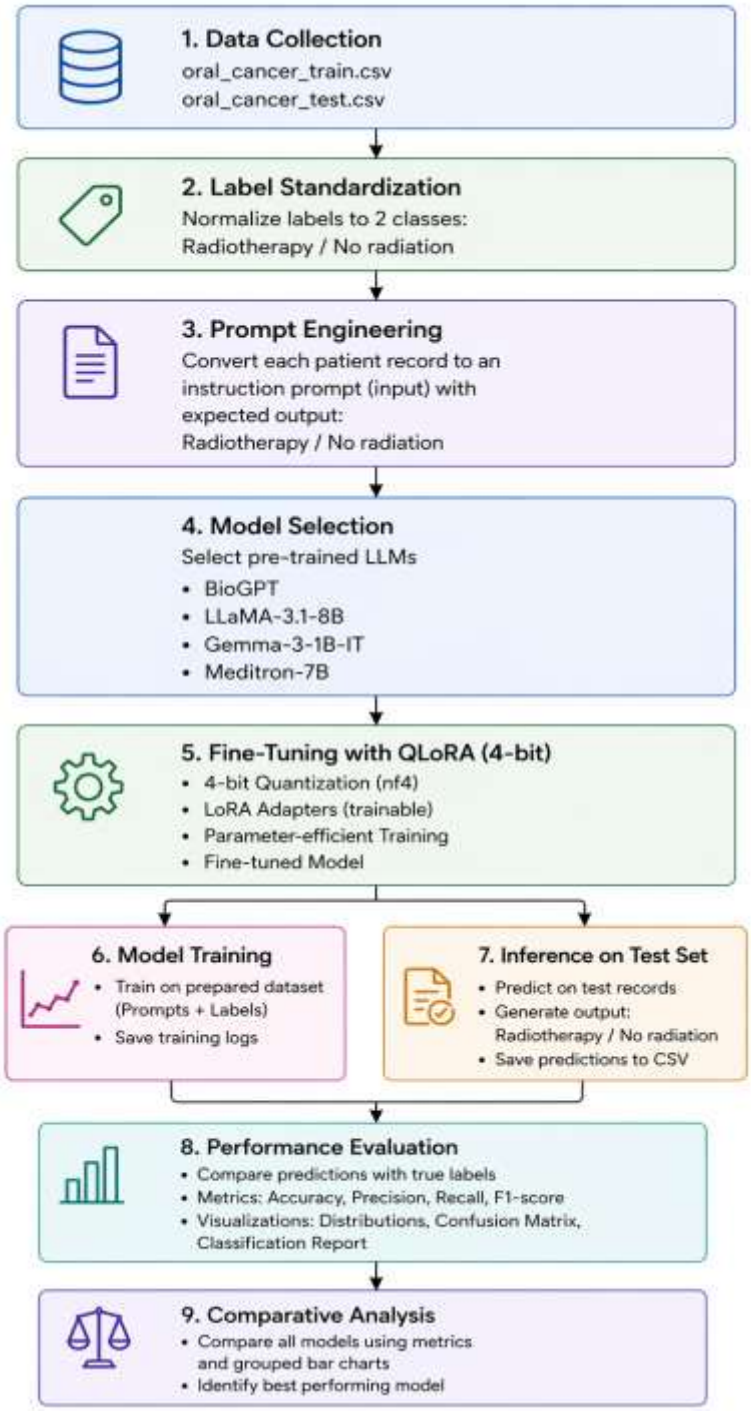

Figure 1: Flow of proposed work.

3.4 Data collection

The collective structure of oral cancer datasets is stored in CSV format, and it begins with the experimental workflow. Two datasets were used in this study. These include oral_cancer_train.csv and oral_cancer_test.csv. Training dataset was used to fine-tune the transformer models. The testing dataset on the other hand was reserved exclusively for evaluating prediction performance. These datasets contain clinically approved characteristics describing tumour characteristics, treatment indicators, and diagnostic variables associated with oral cancer management. These characteristics altogether support binary classification of treatment recommendations into radiotherapy and none-radiotherapy categories. Before training, preprocessing routines verify dataset integrity, which is implemented in Appendix A, ensuring that all models received identical input structures.

3.5 Model selection

The proposed work compares four LLMs that were chosen to reflect various degrees of generality and medical specialisation. These models include BioGPT, Llama 3.1, GEMMA 3, and Meditron-7b. These models have been divided to cover two general purpose models and two medical-specific models. The reason for the selection of these LLMs is that they either represent a general-purpose model or have been trained specifically on biomedical or clinical data. Such a mix makes it possible to comparatively assess the effects of various training goals on clinical knowledge and the quality of the recommendations of treatment for patients with oral cancer. The comparison of these models with each other enables comparison between general-purpose and medical-specific models in their predictive abilities of the radiotherapy treatment recommendations by using patient clinical features (Figure 2). This comparison highlights the impact of different model architectures on predictive performance

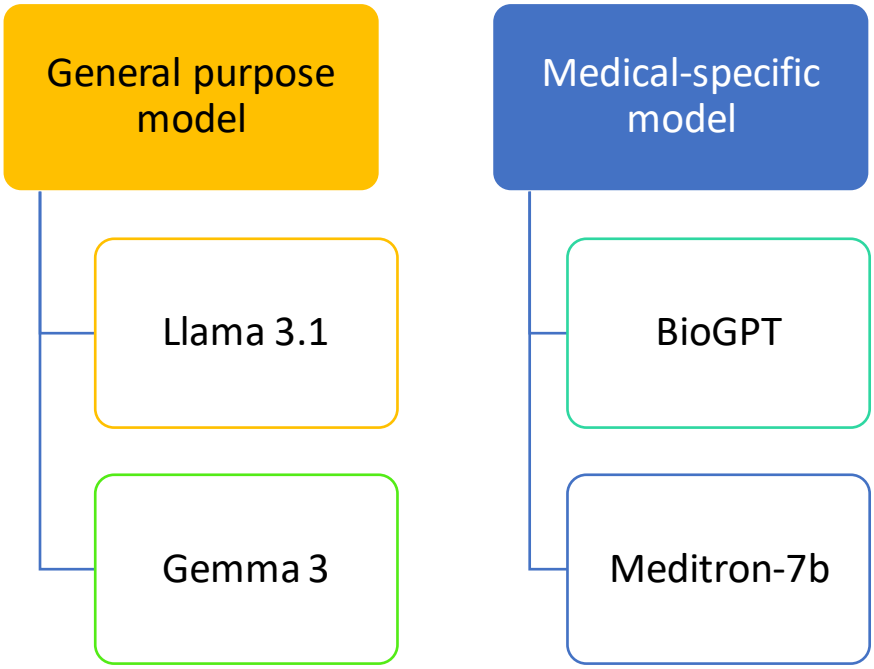

Figure 2: Types of General Purpose and Medical Specific Models

3.5.1 Model Parameters

The model parameters include max length sequence, batch size, and dropout rate. Max sequence length used within this model is 256 and batch size of 4-16. The dropout rate of 0.05. Overall, this configuration is used for avoiding overfitting and controlling context windows for clinical predictions.

3.5.2 General-purpose LLMs

This work compares a collection of general-purpose LLMs, such as Llama 3.1 and GEMMA 3. They are general-purpose LLMs that are sophisticated for reasoning tasks. Both general-purpose and biomedical NLP models are driven by the complementary properties of the two models. General-purpose LLMs are good at high-level clinical reasoning, reading between the lines and understanding guidelines, which are key to generating coherent and clinically meaningful treatment advice. Nevertheless, their accuracy might be reduced when dealing with specialised biomedical terms. Biomedical NLP models overcome this limitation by offering a correct and consistent processing of the domain-specific clinical language.

3.5.3 Llama

This is a general-purpose transformer-based model which is not specifically trained over biomedical literature and was created by Meta. However, it has strong adaptability and can be trained over biomedical data set to make accurate decisions regarding health conditions (Phadtare et al., 2026). This model is selected for comparative analysis and evaluates the validation and accuracy of other models in clinical decision support tasks. Anisuzzaman et al. in 2025 suggested that LLMs can be fine-tuned over medical datasets to improve early detection along with risk assessment. Furthermore, its domain specific adaption makes LLMs more accurate in clinical decision making (Anisuzzaman et al., 2025).

3.5.4 Gemma

Gemma-3 is lightweight LLM created by Google deep mind. It was designed especially for providing efficient performance that reduces computational complexity as compared to other models. This model, however, is not exclusively trained over biomedical datasets. However, it can be fine-tuned corresponding to healthcare data sets and patient level summarisation can be derived. This model is useful in evaluating the tradeoff between the model size and computational efficiency. This model provides knowledge extraction along with NLP deployment support within medical field (Kasa et al., 2026). The study conducted by Anil et al. in 2023 showed that overfitting and underfitting in LLMs like Gemma can be handled by optimised parameter tuning (Anil & Singh, 2023). This is applied in fine tuning of Gemma.

3.5.5 Medical-specific LLMs

There exit medical-specific LLMs as well including BioGPT and Meditron. These models are designed to process biomedical and clinical text more accurately. General-purpose LLMs may not be that accurate as compared to medical specific LLMs. These models are pretrained on domain-specific datasets. These include PubMed articles, clinical notes,

treatment guidelines, and electronic health records (Saha et al., 2026). The training on these datasets enables LLMs to better understand medical terminology and context.

3.5.6 BioGPT

BioGPT is transformer based LLM which was created by Microsoft research. This model was created exclusively in biomedical literature which is sourced from PubMed abstracts. This model can understand specialised medical terminologies. It can be able to assess biomedical relationships as well as clinical reasoning patterns (Luo et al., 2022). This model was selected as it has the advantage after reading medically coherent responses that are aligned with the scientific technology.

3.5.7 Meditron

Meditron is an open-source medical LLM. This model is designed for clinical reasoning along with healthcare interpretation. This model is derived from Llama-2 and trained using curated clinical datasets. The main objective associated with this model is to improve safety and reliability decision making within medical environment. It is pre-trained and useful for guideline aligned responses for medical field (Chen et al., 2024).

3.6 Label standardisation

To convert heterogeneous treatment annotations into a consistent binary classification format label standardisation is performed. Set feature variables were normalised to two categories: “Radiotherapy” and “No Radiotherapy”. It guarantees compatibility with the instruction-tuned transformers and minimizes potential confusion in the supervised learning process. The categorical inconsistencies and missing data points were sorted out using structured filtering techniques in the data processing module as specified in Appendix A. Normalisation of the predicted labels contributed to consistency in achieving convergence through different LLM architectures, ensuring that the classification was consistent. Normalisation was useful in connecting the structured data points in the datasets with instruction-led responses.

3.7 Prompt engineering

Prompt engineering formed a central component of the methodology. It is necessary as LLM-based instruction tuning requires textual task representation rather than conventional numerical feature matrices. Each patient record was converted into a structured instruction-response format. It explicitly described the prediction objective. For each instance of dataset, clinical attributes were embedded within a standardised instruction template. They requested the model to determine whether radiotherapy was required. The expected output was restricted to a binary textual response. These correspond to the standardised labels defined earlier. The prompt construction logic implemented for each transformer model is defined within Appendix A. This transformation allowed structured tabular medical data to be interpreted by LLM architecture as contextual reasoning tasks. However, it does not represent conventional classification inputs. The models learned semantic associations between clinical indicators as a result. The treatment recommendations more effective using prompt engineering as compared to feature-vector-only pipelines.

3.8 Parameter-efficient fine-tuning using QLoRA

Fine-tuning large transformer architectures usually require significant computational resources. To address this issue, QLoRA was applied to enable memory-efficient parameter updates while preserving predictive capability. QLoRA has quantising pretrained model weights into 4-bit representations which reduces training complexity, while introducing trainable QLoRA adapter matrices within attention layers (Dettmers et al., 2023). Instead of updating all model parameters, only lightweight adapter components are optimised during training. This strategy significantly reduces GPU memory requirements and training time without compromising classification performance. The QLoRA configuration used in this study including quantisation setup, adapter injection, and optimiser initialisation is implemented in Appendix A. By applying identical adapter configurations across all the transformer architectures guarantee fairness in comparative evaluation.

3.9 Model training procedure

Following dataset preparation and adapter configuration, supervised fine-tuning was performed using instruction-formatted clinical data. During training, the transformer models learned relationships through gradient-based optimisation. The learned relationship was between patient attributes and treatment recommendations. Each training pipeline consists of loading pretrained tokenizer weights, applying 4-bit quantisation, inserting QLoRA adapters, and executing supervised fine-tuning using labelled prompts. this process saved Model checkpoints generated during this process for later inference and evaluation. Overall, this training mechanism helps in reducing the training loss.

3.10 Inference on the test dataset

After completing fine-tuning, each LLM evaluated using the independent testing dataset to measure generalisation capability. During inference, the trained adapters were reloaded. It will be applied to unseen patient records formatted using the same prompt structure adopted during training. Each model generated predicted treatment recommendations corresponding to radiotherapy or no radiotherapy classes. These predictions were exported into structured CSV files for further evaluation. The inference pipelines used to generate these outputs are included in Appendix A. Using a separate test dataset ensured that reported results reflect real predictive performance rather than memorisation of training samples.

3.11 Performance evaluation

Several classification metrics were employed to calculate the model performance in order to assess its predictive quality comprehensively. For assessing the accuracy of prediction correctness, accuracy was utilised, while for assessing the reliability of prediction, precision was utilised. Recall assessed the ability of each model to correctly identify patients requiring treatment, and the F1-score provided a balanced measure combining precision and recall. The confusion matrix and classification report were

created to depict the distribution of predictions and misclassification patterns. The process was done for each LLM in Appendix A, where predictions are combined from all transformer models and performance measures are calculated comparatively. The use of multiple metrics guaranteed reliable assessment of the predictive accuracy of models without using one particular metric.

3.12 Comparative model analysis

Comparison studies were conducted to find the best-suited transformer architecture for clinical decision-support predictions. Performance metrics generated during evaluation were aggregated and analysed to identify differences in classification behaviour across models. Ranking of LLM's architectures in terms of their performance was possible using the comparative methodology provided in Appendix A. In addition to the numerical assessment of the models by comparison, it also assisted in the qualitative interpretation of the weaknesses and strengths of the models. Such a systematic benchmarking process ensured that empirical evidence was used to make final model selection and not by architectural assumptions.

3.13 Experimental environment

Oracl_cancer_train and oral_cancer_test datasets were used in the testing of different models. The clinical features present within dataset and used in implementation are given in table 1.

Table 1: Clinical features used in the work

Feature	**Description**
Age	Patient age at diagnosis
Sex	Patient gender
Marital status	Marital status at diagnosis
Primary site	Anatomical location of tumour
Grade	Histological tumour grade

T stage	Tumour size and invasion level
N stage	Lymph node involvement
M stage	Distant metastasis

The train dataset contains 13 attributes or features and there are 7551 records within the dataset. The characteristic of dataset is given in table 2.

Table 2: Train Dataset Characteristics

Parameter	Value
Train Shape	(7551, 13)
Number of Rows	7551
Number of Columns	13
Features	prompt, target
Dataset Type	HuggingFace Dataset

Table 2 indicates first stage of the model training process. The output confirms that the training dataset 7,551 patient records was effectively loaded into the Hugging Face dataset framework, with two main features the prompt, which is the structured clinical input and the target, which is the label of the treatment. The implementation of the suggested methodology was performed within a Jupyter Notebook with Python as the programming language for training the transformers. The stack included PyTorch to load model and conduct the extensive DL training procedures, Hugging Face Transformers for importing the pretrained models, BitsAndBytes for supporting 4-bit quantisation, and PEFT for implementing the QLoRA adapters. Data preprocessing and evaluation were done using Pandas, NumPy, and Scikit-learn. After the fine-tuning stage, the selected models were evaluated on the test data set, which was composed of new patients. All source codes related to the data set preparation, prompt design, QLoRA implementation, training procedure, inference, and evaluation can be found in Appendix A.

3.14 Reproducibility and implementation structure

Reproducibility was ensured through keeping constant preprocessing pipelines, standard prompt templates, adapter setup and using different test sets in each of the models examined. Each step in the process from dataset generation to benchmarking was implemented as a scriptable notebook to facilitate experimentation. The details of the implementation pipeline is provided in Appendix A. The dataset was proprietary to a research institute in Finland.

3.15 Reliability and validity

Reliability in this study was achieved through a consistent experimental setting applied across all models. Training and testing operations involved the use of the same dataset partition, the same prompts format for all models, and the use of identical evaluation metrics to compare models' performance objectively. Moreover, consistency was guaranteed by processing data with similar procedures, including missing value handling and standardising of input variables. The same training and evaluation operations provided the basis of minimising variability to enable the direct comparison of performances of the models. Regarding the validity, clinically relevant input variables, such as demographic and Tumour-related ones, were selected based on their relevance to oncology decision-making practices. Moreover, in order to ensure the relevance of the prediction problem in the clinical field, the predicted values were restricted only to Radiotherapy and No Radiotherapy categories.

As the input features involve TNM staging, Tumour characteristics, among others, the construct validity was guaranteed. Realistic variables and treatments categories also provide an indication that external validity has been achieved in this study, since prediction is close to clinical practices involving risk factors identification. Despite having imbalanced data, no specific class balancing operation was performed during model training. This ensured that the models were trained on the original, unmodified data distribution, but it also introduced a potential source of bias that could affect predictions.

Therefore, model performance was evaluated using precision, recall, F1-score, and confusion matrix analysis in addition to accuracy.

3.16 Ethical and data protection considerations

Ethical adherence was also one of the fundamental elements during the research. No personally identifiable patient information was employed because all data were anonymised. The approach complies with the regulations of data protection, including the GDPR, and adheres to responsible AI principles. The proposed system is clearly presented as a decision-support tool that is meant to support, and not to substitute, professional judgment.

3.17 Integration into a web-based prognostic tool

The best performing model is integrated as a web-based prognostic tool. This is to facilitate testing by other multicentre institutions. Additionally, this is intended to Essentially, the web-based tool will allow clinicians to input either real or systematised patient data for treatment recommendations and evaluate if it reflects the known recommended guidelines. Remarkably, the web-based tool is intended as a clinical decision support system. Therefore, the treatment recommendations should be considered with caution. The final decision regarding treatment recommendation is based on the clinicians' judgement and multidisciplinary tumour board discussions.

4 Results

4.1 Dataset description

The dataset used in this experiment consisted of structured oral cancer patient records stored in CSV format. The data included two main files:

- oral_cancer_train.csv
- oral_cancer_test.csv

One of which is a training file and the other is a test file. The training set contained 7,551 cases (patient records) and the testing set contained 944 cases. In terms of the target outcome (radiotherapy treatment recommendation), the distribution is presented in Figure 3 (5860 no radiation cases, 1691 Radiotherapy cases). However, 15 records were used as sample for testing the models. All treatment labels were categorised into two main categories, namely radiotherapy and non-radiotherapy

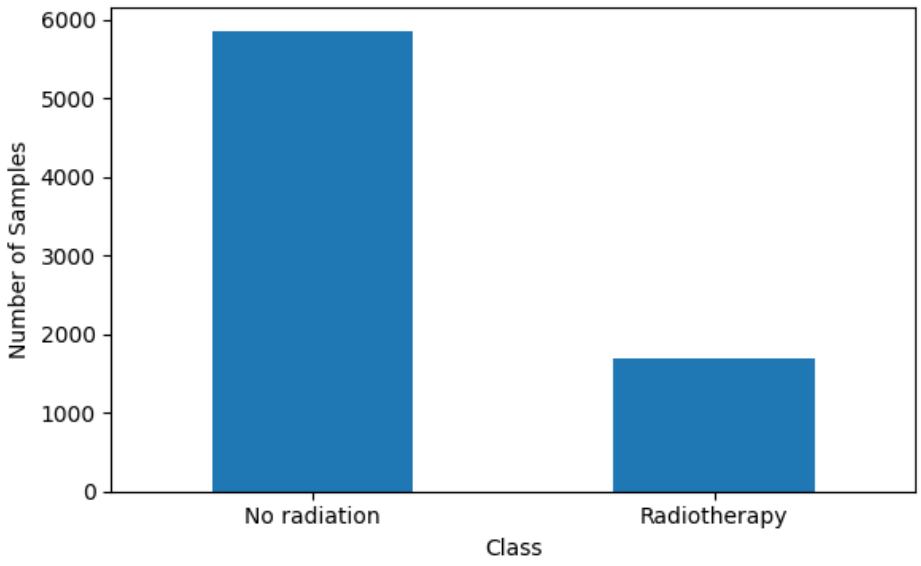

Figure 3: Class imbalance in the radiotherapy dataset

Figure 3 presents the class distribution in the training dataset regarding radiotherapy treatment labels. Thus, 5,860 patients received no radiotherapy treatment, while 1,691 patients received it. The above means that class imbalance occurs since the number of

samples belonging to “No Radiotherapy” is significantly higher compared to Radiotherapy.

Class imbalance affects model evaluation because it can produce high accuracy if the model predicts the majority class. It means that additional metrics should be taken into account such as precision, recall, F1-score, and confusion matrix.

Step	Training Loss
25	0.717400
50	0.442100
75	0.427200
100	0.392100
125	0.381200
150	0.368000
175	0.344300
200	0.380000
225	0.343800
250	0.363100
275	0.318600
300	0.358000
325	0.358600
350	0.364700
375	0.338400
400	0.335500
425	0.311700
450	0.344700

Figure 4: Training loss progress

Figure 4 shows the values of training losses in the process of the BioGPT model fine-tuning based on the QLoRA method of adaptation. The loss decreases significantly from 0.7174 at step 25 to around 0.3117 and step 425. It demonstrates that the model able to capture domain specific contextual relationships that are present within the data. There are minor fluctuations at certain steps 200, 300 and 450. These variations are considered normal, especially in case of transformer-based LLMs. These variations exist due to batch level optimisation along with data set complexity. To simplify testing, validation and result interpretation, 15 records out of 944 were used.

4.2 Fine-tuning BioGPT results

The proposed evaluation starts by observing training loss as given in following figure 5.

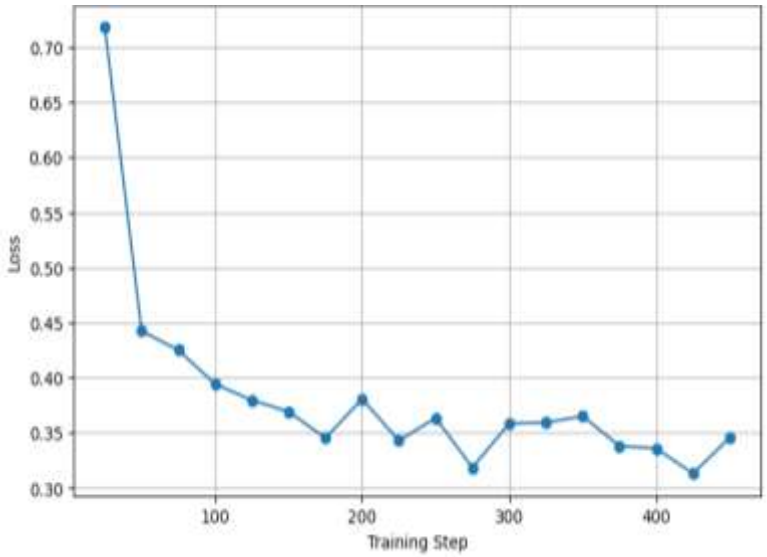

Figure 5: Training loss during BioGPT fine-tuning

It shows the training loss curve, which decreases steadily during BioGPT fine-tuning using QLoRA method. It drops sharply from about 0.72 to 0.44. After that it stabilises around 0.31 to 0.34. This trend indicates shows stable convergence when QLoRA based fine tuning is applied across epochs.

```
Device: cuda
Selected rows: 15
Record 21: No radiation
Record 22: No radiation
Record 23: No radiation
Record 24: No radiation
Record 25: Radiotherapy
Record 26: No radiation
Record 27: Radiotherapy
Record 28: No radiation
Record 29: No radiation
Record 30: Radiotherapy
Record 31: No radiation
Record 32: No radiation
Record 33: No radiation
Record 34: No radiation
Record 35: Radiotherapy
Saved: /content/drive/MyDrive/Thesis/FineTunning/BIOGPT_only2_21_35.csv
```

Figure 6: BioGPT model prediction output for test records

Figure 6 shows the results of prediction using the fine-tuned BioGPT model on unknown patient records in the test dataset. A prediction generation was done on a GPU device (CUDA), which allowed the inference process to run faster. A group of 15 patient records

were picked from the test dataset and evaluated. The model analysed features in each prompt. It produced treatment recommendations corresponding to two predefined categories using constraint decoding. The predictions were subsequently stored in a CSV file where the performance is again to be evaluated, and the accuracy is to be checked.

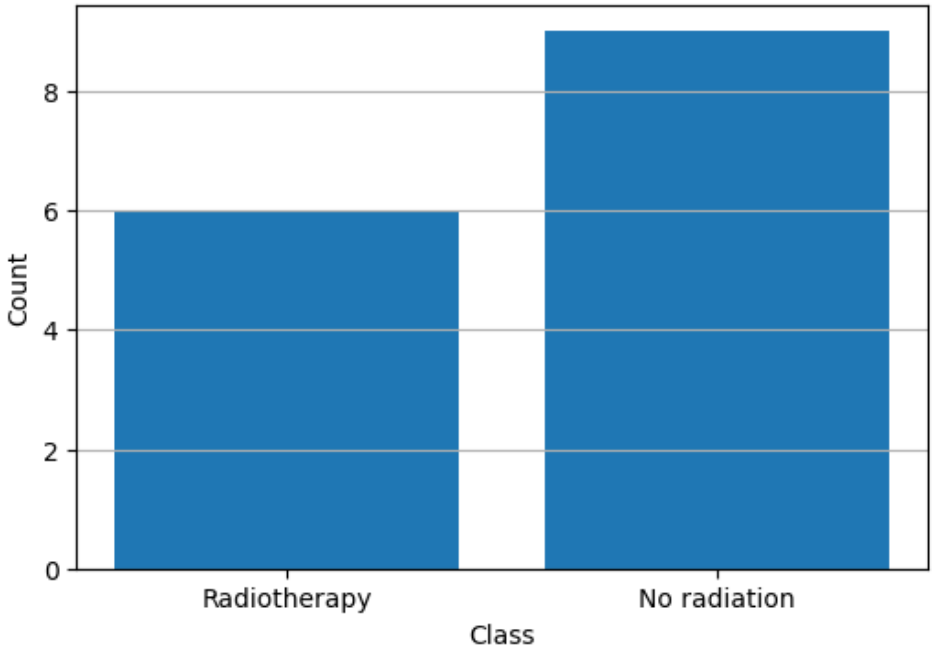

Figure 7: Distribution of true labels in the test dataset

Figure 7 shows the distribution of the true treatment labels in the selected test data of 15 records. There are 6 radiotherapy and 9 no radiation cases as shown in bar plot. Such a distribution implies that the assessment sample is a little more represented by the no radiation category. The performance of the model can be understood by looking at true label distribution. The figure offers the background to the further assessment outcomes, such as the confusion matrix and performance metrics that were obtained while testing.

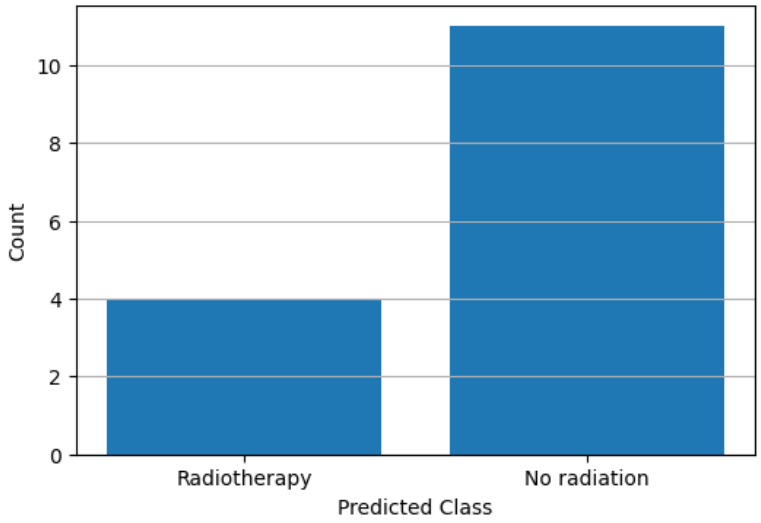

Figure 8: Distribution of predicted labels in the test dataset

Figure 8 demonstrates how the BioGPT model would label most of the predicted treatments on the selected test set of 15 records. As indicated in the chart, the figure of 4 cases was predicted to be radiotherapy, and 11 cases were predicted to be no radiation. Most of the predictions fall within no radiation category as predicted by model. To determine the success of the model in identifying each category of treatment in the prediction process, the figure can be compared with the actual distribution of labels.

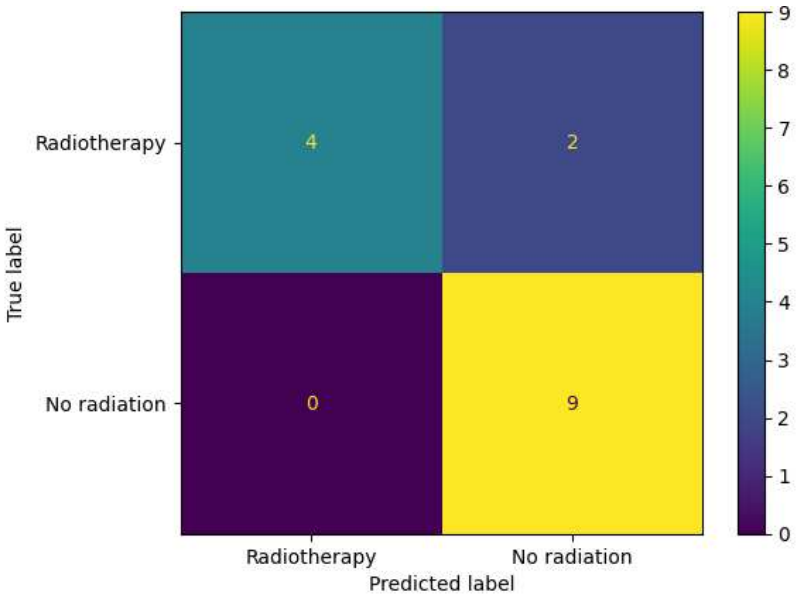

Figure 9: Confusion matrix for radiotherapy prediction

Figure 9 shows the confusion matrix. It shows strong classification performance. The model correctly identified 4 radiotherapy and 9 no radiation cases. The degree of misclassification is only 2 and no false positives occurred for radiation. Overall, high specificity and reliable detection of non-radiation cases have been predicted with the model.

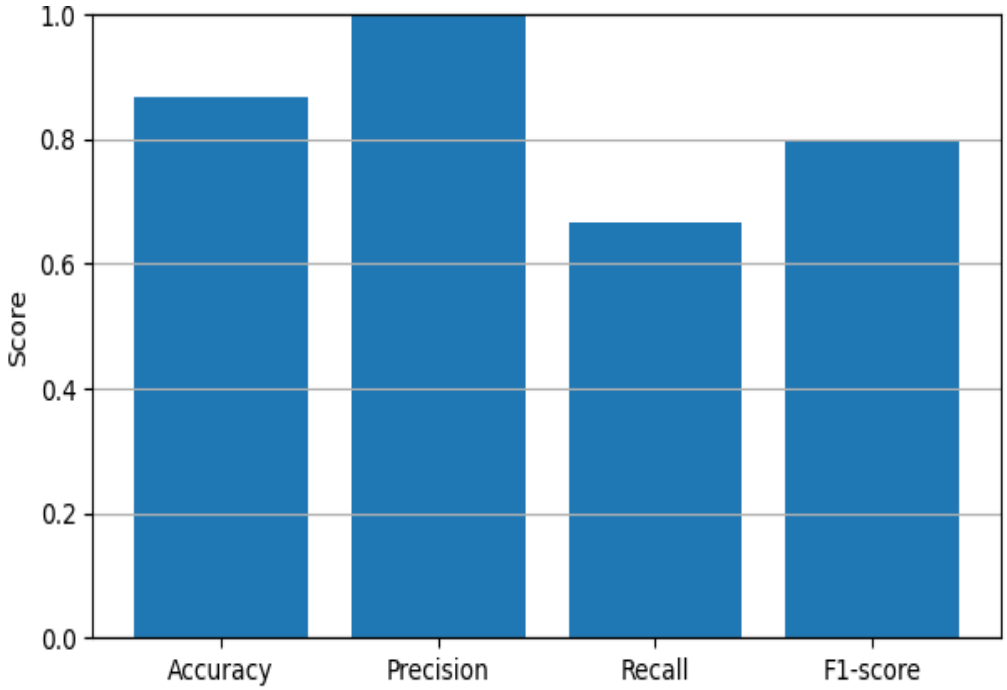

Figure 10: Performance metrics of the BioGPT model

The performance evaluation metrics of the BioGPT model are provided in Figure 10. The model achieved an accuracy of approximately 0.87. This metric shows total correct predictions generated by model. Precision value of 1.00 suggest it does not predict any false positive cases of radiotherapy. The recall value of 0.67 shows that the model correctly identifies positive cases. The balance between the precision and recall is represented by the resulting F1-score of about 0.80. These results indicate model has strong predictive power.

4.3 Fine-tuning Llama results

When the predictions were compared to the original labels of treatment in the dataset, the model yielded a high level of accuracy of about 93.34 percent on the sample that was evaluated. The Llama training loss curve is used during the training process to indicate how prediction error decreases across training steps.

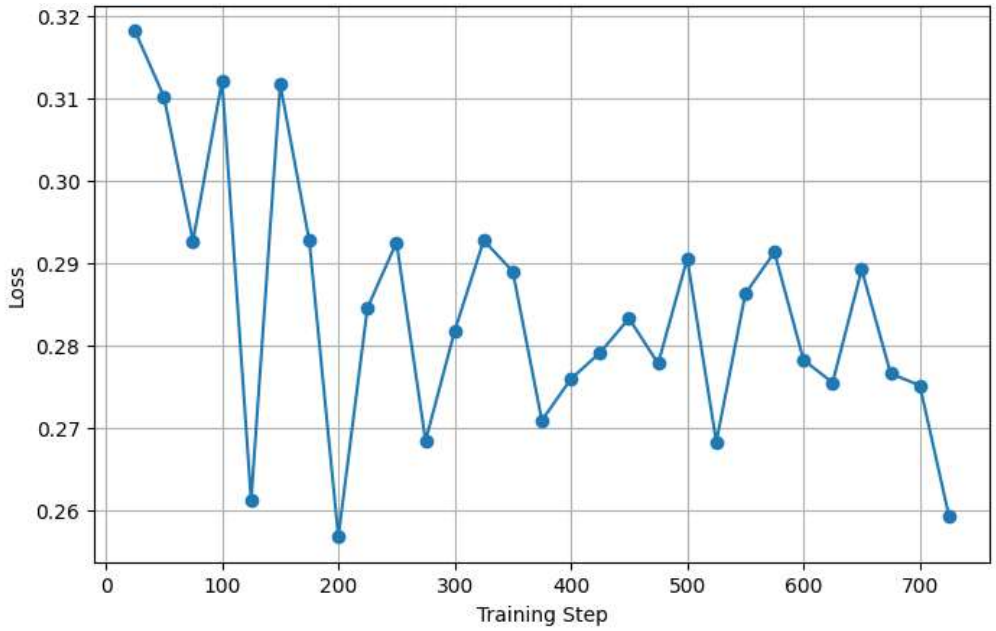

Figure 11: Llama fine-tuning training loss

Figure 11 shows the training loss curve of the Llama model over several training steps of the fine-tuning. This loss first oscillates at 0.31 but decreases steadily as training advances to about 0.26 at later training steps. These variations show that the model is learning patterns based on the data. The general decreasing tendency proves that convergence is stable, and the efficiency of the model learning is enhanced, indicating that the fine-tuning procedure managed to optimise the Llama model on the radiotherapy prediction task.

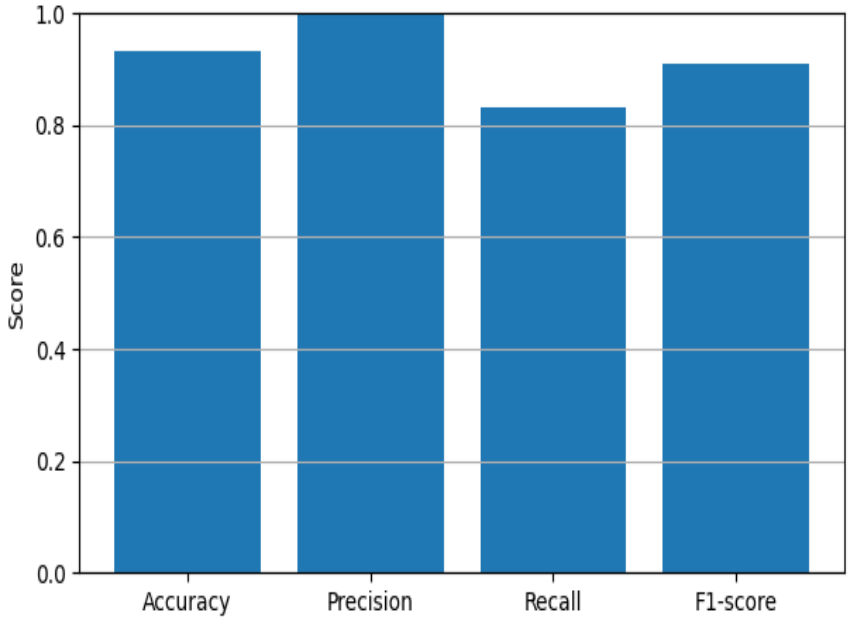

Figure 12: Performance of Llama model

The performance of Llama model indicates that the accuracy is more than 90% indicating correctly classified cases. The precision is 100%. This metric is crucial as it indicates positive cases predicted through the model were correct. Recall as well as F1-score are also significantly higher and are above 80%.

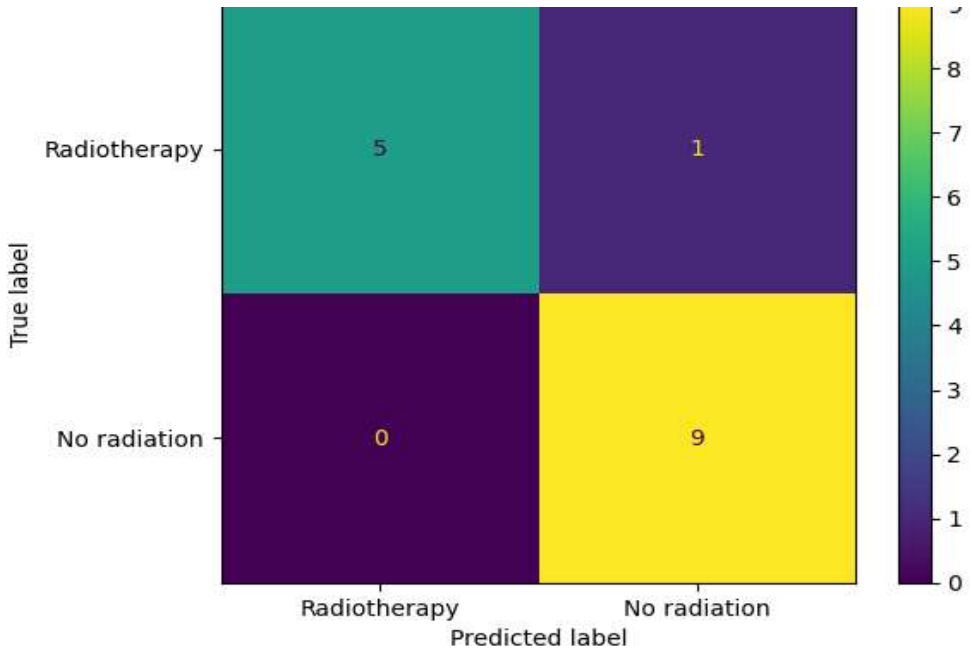

Figure 13. Confusion matrix for Llama

Figure 13 shows the confusion matrix. Strong classification performance has been predicted. The model correctly identified 5 radiotherapy and 9 no radiation cases. The degree of misclassification is only 1 and non-radiation treatment cases were correctly predicted, and there were no false positives for radiotherapy. The findings clearly indicate that the model successfully predicts non-radiation cases with high specificity.

4.4 Fine-tuning Meditron results

The training loss curve result is given in Figure 14, indicating that training loss is gradually decreasing.

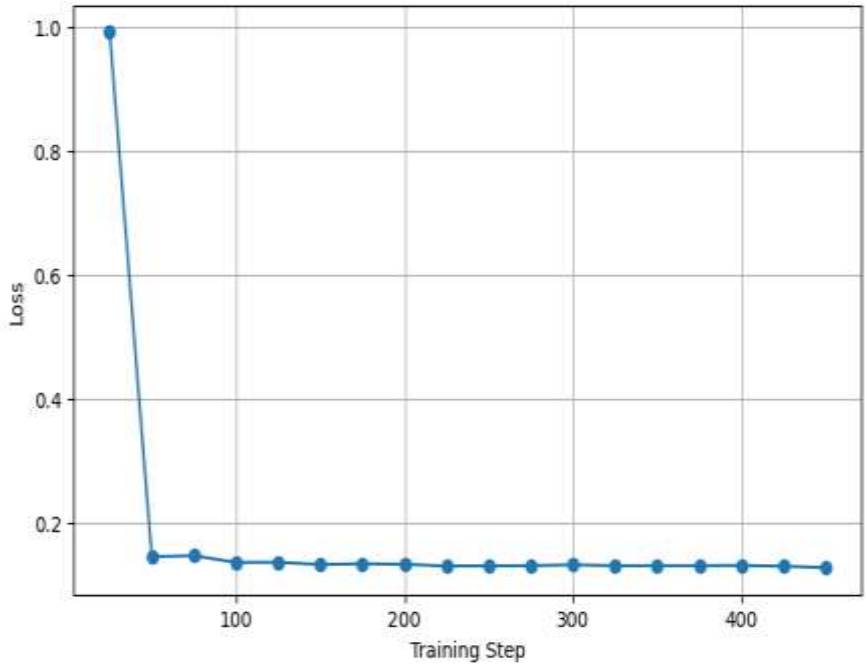

Figure 14: Meditron-7B fine-tuning training loss

There is a steep decline in the loss, i.e., 1.0 at the first stage and almost 0.14 after the first few training steps, which demonstrates that the parameters are learned quickly and adjusted. The loss levels after this first decrease and has only slight variances, indicating that the model has converged. This trend indicates efficient learning and partially validates the success of the fine-tuning process in the case of medical prediction exercises.

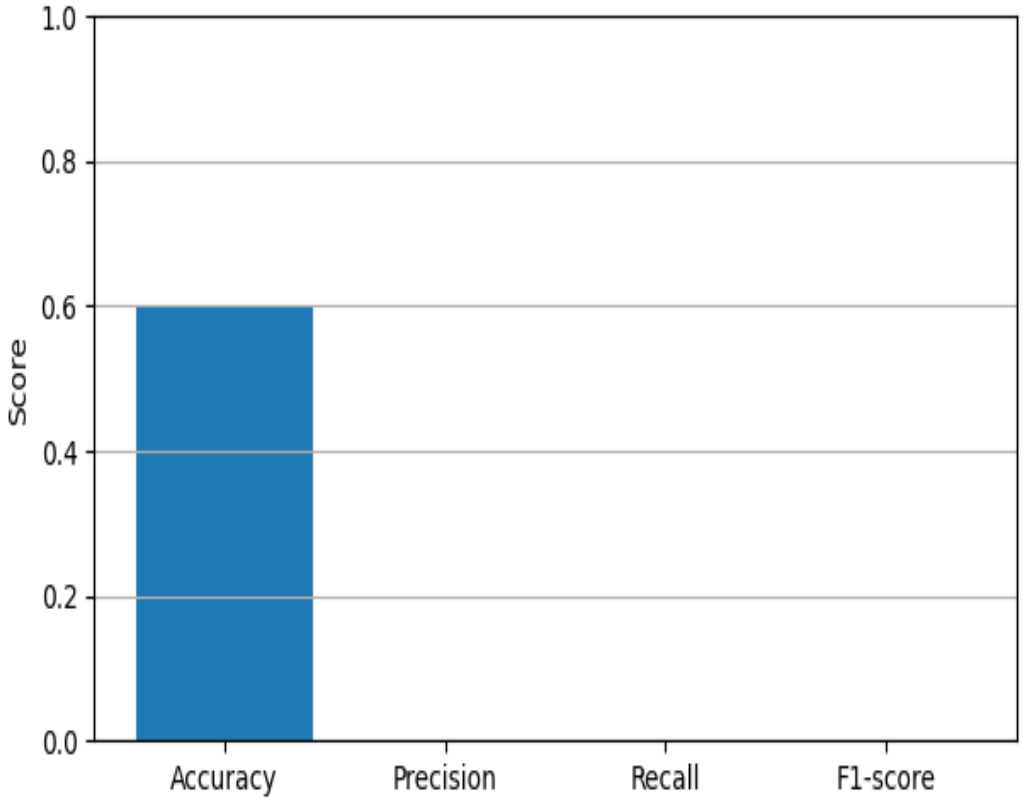

Figure 15: Meditron model performance metrics

Figure 15 shows the metrics used for the Meditron model i such as accuracy, precision, recall, and F1-score. The findings demonstrate moderate accuracy. Precision indicating correctness of positive cases was almost 0. Furthermore, recall values was 0, which means that it is difficult to identify cases of relevant radiotherapy predictions correctly. The reduced precision and recall indicate that the model has false positives or missed predictions. Such outcomes indicate that optimisation of training and refining of datasets or optimisation of model architecture is needed in order to achieve better predictive reliability.

Figure 16 gives the confusion matrix. This model successfully classified nine observations into the “no radiation” category. However, it failed to correctly classify even one observation within the “radiotherapy” category. This indicates that there is a high rate of misclassification in the “radiotherapy” category, with six false negatives and zero false positives.

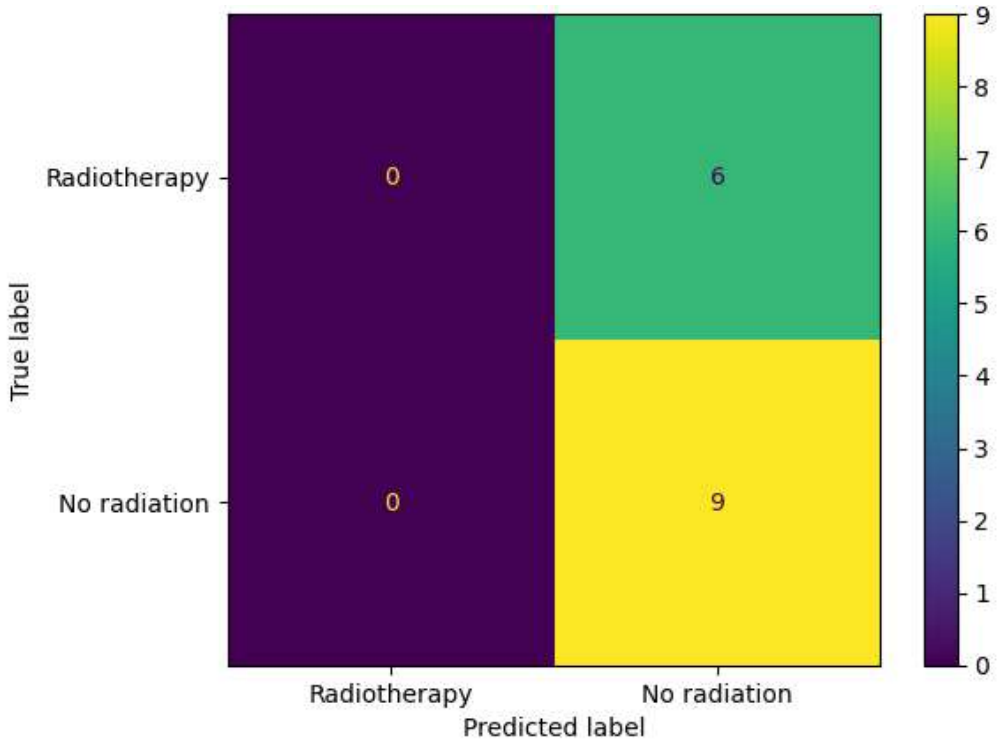

Figure 16: Meditron confusion matrix

4.5 Fine-tuning GEMMA results

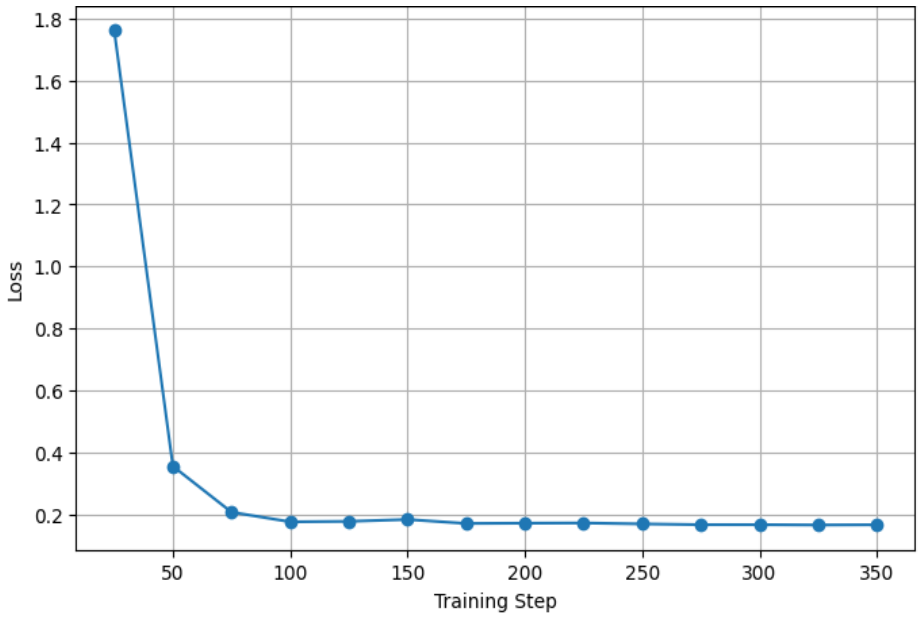

Figure 17: GEMMA fine-tuning training loss

This curve shows how the GEMMA model loses training on fine-tuning. The loss begins at a higher value of around 1.8 and declines at a high rate in the initial training steps, which represents better learning. Following this, loss levels reached at 0.17, implying convergence and the optimisation of the parameters. The gradual loss trend justifies that the fine-tuning mechanism is effective in enhancing the learning ability of the GEMMA model and makes it ready to predict, although reliability varied (Figure 17).

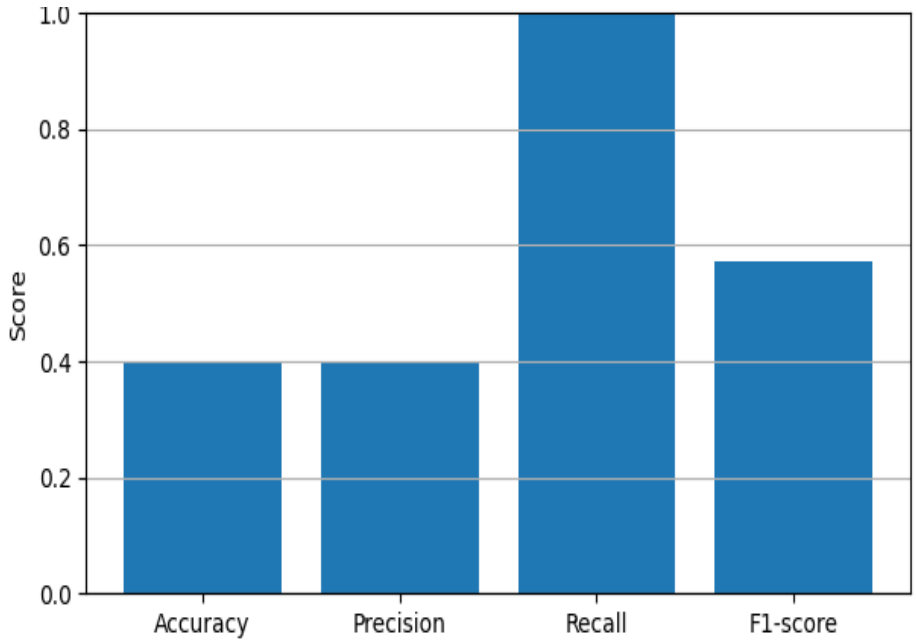

Figure 18: GEMMA model performance metrics

This figure 18 shows the performance of the GEMMA model in terms of accuracy, precision, recall, and F1-score. The findings show excellent recall rates, which is the reason why the model can identify most positive cases in the data. The precision and accuracy were moderate. This indicates that there exist false positive predictions. The balanced F1-score is an indicator of the precision and recall trade-off.

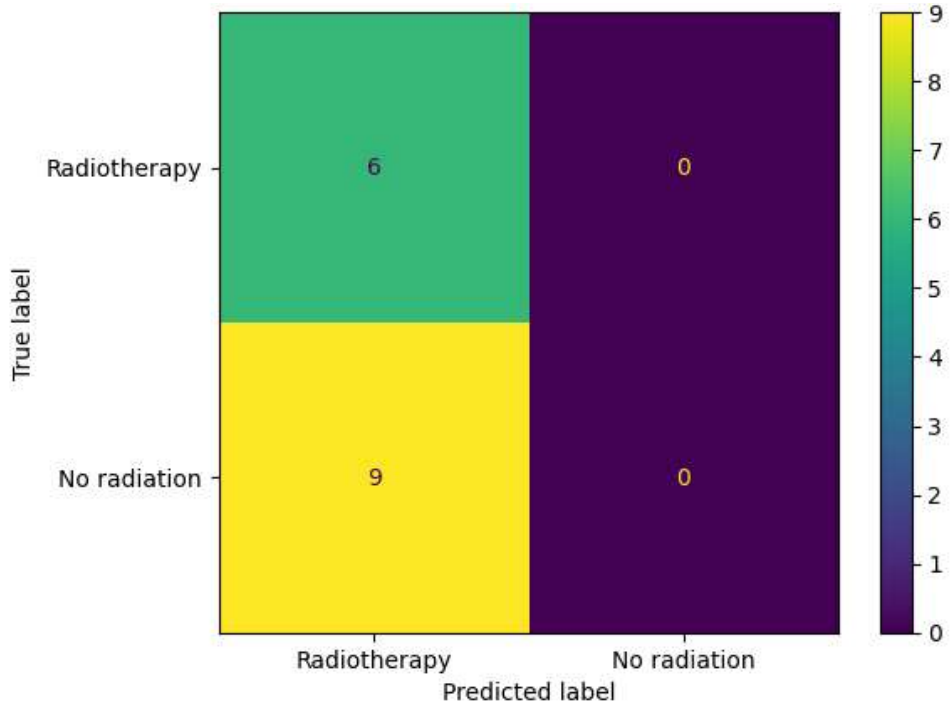

Figure 19: Confusion matrix of Gemma

The confusion matrix of Gemma model is shown in Figure 19. All six observations, which were associated with radiotherapy, have been predicted accurately by the model. Model however not able to predict non-radiation cases accurately leading to nine false positive errors, while none of the true negatives could be identified.

4.6 Comparative analysis

The comparison chart evaluates BioGPT, Llama-3.1, Gemma, and Meditron for radiotherapy prediction. Llama-3.1 demonstrates the strongest overall performance. It achieves high accuracy (0.93). The perfect precision (1.00), strong recall (0.86), along with the highest F1-score (0.91), is achieved by Llama-3.1. This indicates balanced classification capability. BioGPT also performs well with perfect precision. However, it has a lower recall (0.67). This indicates missed radiotherapy cases. Gemma achieves perfect recall (1.00). However, it has weaker precision and accuracy. This indicates false positives are on the higher side. Meditron shows moderate accuracy (0.60). However, it shows limited reliability overall. Based on comparative analysis, it is shown that Llama-

3.1 is the most consistent and effective model for radiotherapy prediction tasks (Figure 20).

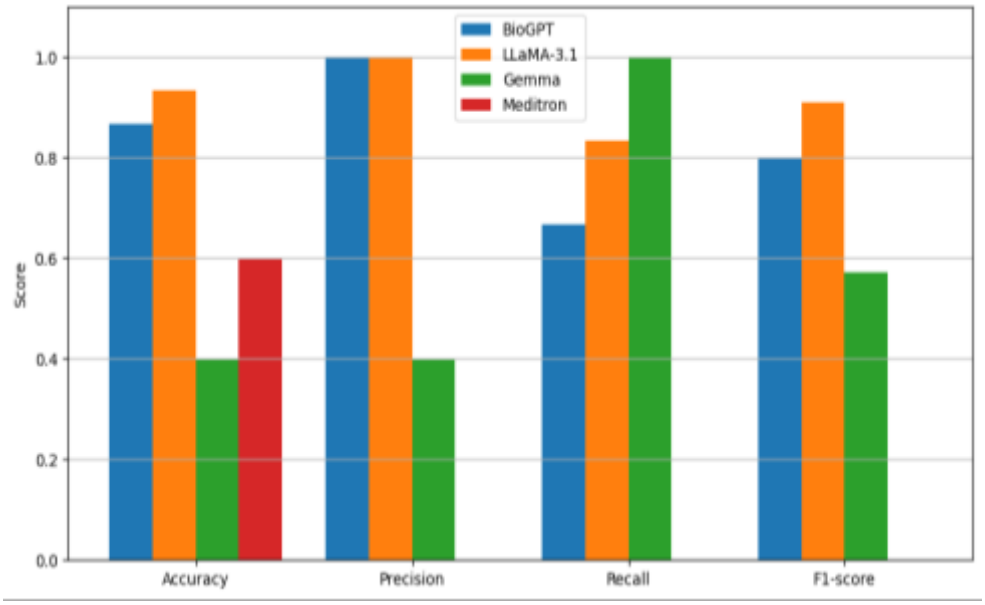

Figure 20: Performance comparison of different language models for radiotherapy prediction

Although BioGPT showed strong biomedical knowledge representation due to its training on biomedical literature, Llama-3.1 achieved superior overall model performance, indicating that it is the most suitable architecture for experimental decision-support applications in this study.

4.7 Testing with a smaller training and test size

LLMs required heavy computation and in case large dataset, they will consume large amount of time to training as well as testing. To handle this situation, size of the dataset both for training and testing was limited to 7551 samples and 15 samples. This strategy enabled the model to be trained faster and yet sufficient data to test the performance of the model. The reduced dataset experiment was aimed at studying the performance of the LLM when it is trained with small amount of data. These experiments can be applied in real life, where, in some cases, large clinical records are not always accessible. Through testing the model in these conditions, it is possible to test the strength and flexibility of the suggested method.

The link to web based prognostic tool is given as follows

https://radiotherapyclassifer.netlify.app/

5 Discussion

5.1 Evaluation

This thesis explored the use of four main LLMs (BioGPT, Llama, GEMMA, and Meditron) for radiotherapy treatment recommendations in patients with OSCC. Additionally, the LLMs were compared and the highest-performing LLM was integrated as a web-based prognostic tool for treatment recommendation. This thesis showed that LLM fine-tuning is a productive method of specialising pretrained biomedical models to specific tasks. This is because QLoRA only changes very few parameters. Therefore, it makes the training process efficient. Additionally, the use of prompt engineering ensured that the fine-tuning process was successful, such that the model became capable of being incorporated into domain-specific problems, like in medical decision-making. Llama achieved experimental accuracy of 0.93 (93%). BioGPT showed good performance with 0.86 (86%) accuracy score in limited evaluation environment. However, Gemma and Meditron showed weaker performance in terms of accuracy with 0.4 and 0.6 for patients with OSCC. This suggests significance of prompt engineering and LLM fine-tuning in improving decision-making for clinicians.

Limiting the decoding strategy led to a high degree of prediction. This decoding strategy helps in reducing the number of outcome possibilities. This causes models to focus on relevant patterns only and hence degree of misclassification reduces significantly. The output of generative models can be erratic and cannot be easily assessed without this mechanism. The limitation of the space of outputs of the model provided consistency of the predictions with the selected categories of treatment. It is possible that LLMs can assist oncologists in their analysis of patient data and coming up with initial treatment recommendations. Nevertheless, these systems must be considered as decision-support, but not alternatives to clinical experience. Human supervision is still necessary, especially in complicated medical cases where various modes of treatment have to be taken into account. To facilitate testing of these models by researchers and clinicians, the highest performing model was integrated as a web-based tool. This can facilitate the introduction of such tools into a workable clinical decision-support environment.

According to the results of the comparative evaluation, the most efficient model, which was Llama, was developed into a prototype web-based tool of prognostics used to help clinicians analyse patient data and provide treatment options. The system enables a structured entry of clinical parameters by clinicians and thus the model is able to manipulate patient specifications and offer guideline-congruent treatment recommendations.

5.2 Limitations

In spite of the encouraging outcomes, this thesis has some limitations that should be properly considered in the interpretation of the results presented in this thesis. First, the test data is relatively small. A future study that will explore the use of large test data is warranted. Second, the model used basic clinicopathological characteristics of the patients. It did not consider other important characteristics such as imaging parameters, pathological reports, clinical notes, sociodemographic parameters, or genome-based information. Incorporating these other data sources would enhance the quality of prediction and further produce a model that is representative of the patients' characteristics. Third, LLMs are probabilistic in nature. This implies that the treatment recommended by these models can change slightly with prompt format or decoding hyperparameters. Even though constrained decoding minimises this variability, additional appraisal is necessary to guarantee the similarity of performance under various clinical settings. Additionally, there was a class imbalance in the dataset that had an implication, probably causing a biased model. Future studies should consider an approach to handling data imbalance to achieve a model that is generalisable.

6 Conclusion

This thesis examined the possibility of clinical decision-support systems based on LLMs to enhance treatment decision-making in the management of OSCC. The main aim was to determine whether LLMs could analyse structured patient data for radiotherapy treatment recommendations. The thesis examined four different LLMs (BioGPT, Llama, GEMMA, and Meditron) and concluded that these LLMs have the potential for treatment recommendations. From the experiment results, it is evident that the LLMs under consideration can predict credible radiotherapy treatment suggestions and interpret structured clinical features. The experiments revealed that both Llama and BioGPT had high prediction performances in their respective classifications in radiotherapy. Nonetheless, Llama-3.1 had the best overall prediction performance compared to other models, thus making it a strong candidate for clinical decision support systems. The training loss curve declines throughout the fine-tuning process, indicating convergence and learning of clinical patterns.

Moreover, the similarities and differences between multiple LLMs have been shown by the comparison of the language model's architectures. This happens as training of models is done iteratively. The loss that occurred in the previous iteration was reduced through the learning mechanism. Factors such as model scale, transformer depth and attention mechanisms affect how the model captures the complex insights within clinical data. One possible explanation for Llama model's better performance is its strong architecture and generalisation to a wide range of input forms, regardless of the fact that this is general-purpose LLM. The primary reason for its better performance includes strong contextual understanding and improved fine-tuning efficiency. Furthermore, self-supervised learning used in Llama helps in extracting useful patterns from dataset for accurate predictions. In contrast, Meditron, which is trained on biomedical literature, showed strong domain-specific knowledge representation but was comparatively limited in adapting to structured classification for clinical prediction tasks.

Additionally, domain-specific training helps to enhance contextual cognition by allowing models to acquire medical-specific vocabulary and clinical reasoning style. Nonetheless, the findings of this research indicate that domain training does not suffice to ensure the best performance. Rather, it is more down to the combination of a strong architecture and fine-tuning techniques. This underscores the fact that with a combination of fine-tuning and directing general-purpose models, they can perform as well as or better than domain-specific models. On the whole, these results highlight the value of striking a balance between the architectural capabilities and domain adaptation in the creation of AI systems used in healthcare. The choice of an appropriate model must also, however, take into account its training domain as well as its structural ability to learn and generalise on clinical data. In conclusion, LLMs have great potential to enhance clinical decisions in the management of OSCC. Given adequate fine-tuning and guided by intentional prompts, LLMs can analyse complex data about patients and generate treatment suggestions that align with clinical reasoning. Nevertheless, these tools are supposed to be considered as supportive tools instead of substitutes for medical experts. As more validation is achieved, with increased balanced datasets and better interpretability systems, LLM-based decision support systems could become an integral part of future intelligent healthcare frameworks.

References

Ahmadsei, M., Christ, S. M., Seiler, A., Vlaskou Badra, E., Willmann, J., Hertler, C., & Guckenberger, M. (2022). Quality-of-life and toxicity in cancer patients treated with multiple courses of radiation therapy. *Clinical and Translational Radiation Oncology, 34,* 23-29. https://doi.org/10.1016/J.CTRO.2022.03.006

Ahn, J., Kang, B. G., Chang, M., & Yoon, S. (2025). Applications and future perspectives of large language models in otolaryngology-head and neck surgery: A comprehensive survey. *Clinical and Experimental Otorhinolaryngology, 18*(4), 283–295. https://doi.org/10.21053/ceo.2025-00121

Ah-thiane, L., Heudel, P.-E., Campone, M., Robert, M., Brillaud-Meflah, V., Rousseau, C., Le Blanc-Onfroy, M., Tomaszewski, F., Supiot, S., Perennec, T., Mervoyer, A., & Frenel, J.-S. (2025). Large language models as decision-making tools in oncology: Comparing artificial intelligence suggestions and expert recommendations. *JCO Clinical Cancer Informatics, 9, e2400230.* https://doi.org/10.1200/cci-24-00230

Alabi, R. O., Almangush, A., Elmusrati, M., & Mäkitie, A. A. (2021a). Deep machine learning for oral cancer: From precise diagnosis to precision medicine. *Frontiers in Oral Health, 2,* 794248. https://doi.org/10.3389/froh.2021.794248

Alabi, R. O., Bello, I. O., Youssef, O., Elmusrati, M., Mäkitie, A. A., & Almangush, A. (2021c). Utilizing deep machine learning for prognostication of oral squamous cell carcinoma: A systematic review. *Frontiers in Oral Health, 2,* 686863. https://doi.org/10.3389/froh.2021.686863

Alabi, R. O., Mäkitie, A. A., Pirinen, M., Elmusrati, M., Leivo, I., & Almangush, A. (2021b). Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer. *International Journal of Medical Informatics, 145,* 104313. https://doi.org/10.1016/j.ijmedinf.2020.104313

Anil, A. K. P., & Singh, U. K. (2023). An optimal solution to the overfitting and underfitting problem of healthcare machine learning models. *Journal of Systems Engineering and Information Technology (JOSEIT), 2*(2), 77–84. https://www.researchgate.net/publication/374421747_An_Optimal_Solution_t

o_the_Overfitting_and_Underfitting_Problem_of_Healthcare_Machine_Learning_Models

Anisuzzaman, D. M., Malins, J. G., Friedman, P. A., & Attia, Z. I. (2025). Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health, 3*(1), 100184. https://doi.org/10.1016/J.MCPDIG.2024.11.005

Benary, M., Wang, X. D., Schmidt, M., Soll, D., Hilfenhaus, G., Nassir, M., Sigler, C., Knödler, M., Keller, U., Beule, D., Keilholz, U., Leser, U., & Rieke, D. T. (2023). Leveraging large language models for decision support in personalized oncology. *JAMA Network Open, 6*(11), e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689

Buhr, C. R., Ernst, B. P., Blaikie, A., Smith, H., Kelsey, T., Matthias, C., Fleischmann, M., Jungmann, F., Alt, J., Brandts, C., Kämmerer, P. W., Foersch, S., Kuhn, S., & Eckrich, J. (2025). Assessment of decision-making with locally run and web-based large language models versus human board recommendations in otorhinolaryngology, head and neck surgery. *European Archives of Oto-Rhino-Laryngology, 282*(5), 1593–1607. https://doi.org/10.1007/s00405-024-09153-3

Chen, D., Avison, K., Alnassar, S., Huang, R. S., & Raman, S. (2025a). Medical accuracy of artificial intelligence chatbots in oncology: A scoping review. *The Oncologist, 30*(4), oyaf038. https://doi.org/10.1093/oncolo/oyaf038

Chen, D., Parsa, R., Swanson, K., Nunez, J.-J., Critch, A., Bitterman, D. S., Liu, F.-F., & Raman, S. (2025b). Large language models in oncology: A review. *BMJ Oncology, 4*(1), e000759. https://doi.org/10.1136/bmjonc-2025-000759

Chen, Z., Romanou, A., Bonnet, A., Hernández-Cano, A., Alkhamissi, B., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Swamy, V., Sakhaeirad, A., Krawczuk, I., Bayazit, D., Marmet, A., Mi, L., Boillat-Blanco, N., ... Bosselut, A. (2024). MEDITRON: Open medical foundation models adapted for clinical practice [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-4139743/v1

Coletta, R. D., Yeudall, W. A., & Salo, T. (2024). Current trends on prevalence, risk factors and prevention of oral cancer. *Frontiers in Oral Health, 5,* Article 1505833. https://doi.org/10.3389/froh.2024.1505833

Deng, M., Lin, Y., Yan, L., Fei, Z., Chen, C., & Ding, J. (2025). Global burden of head and neck cancer from 1990 to 2021: A comprehensive analysis and projections to 2030 based on the Global Burden of Disease Study 2021. *PLOS ONE, 20*(9), e0330805. https://doi.org/10.1371/journal.pone.0330805

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, Article 441, 10088–10115. https://dl.acm.org/doi/10.5555/3666122.3666563

Dunn, L. A., Ho, A. L., & Pfister, D. G. (2026). Head and neck cancer: A review. JAMA, 335(6), 531–541. https://doi.org/10.1001/jama.2025.21733

Dusin, J., Melanson, A., & Mische-Lawson, L. (2023). Evidence-based practice models and frameworks in the healthcare setting: A scoping review. *BMJ Open, 13*(5), e071188. https://doi.org/10.1136/BMJOPEN-2022-071188

Erdat, E. C., & Kavak, E. E. (2025). Benchmarking LLM chatbots' oncological knowledge with the Turkish Society of Medical Oncology's annual board examination questions. *BMC Cancer, 25,* Article 197. https://doi.org/10.1186/s12885-025-13596-0

González-Moles, M. Á., Aguilar-Ruiz, M., & Ramos-García, P. (2022). Challenges in the early diagnosis of oral cancer, evidence gaps, and strategies for improvement: A scoping review of systematic reviews. *Cancers, 14*(19), 4967. https://doi.org/10.3390/cancers14194967

Hao, Y., Qiu, Z., Holmes, J., Lockenhoff, C. E., Liu, W., Ghassemi, M., & Kalantari, S. (2025). Large language model integrations in cancer decision-making: A systematic review and meta-analysis. *npj Digital Medicine, 8,* Article 450. https://doi.org/10.1038/s41746-025-01824-7

Homer, J., & Winter, S. (Eds.). (2024). Head and neck cancer: United Kingdom National Multidisciplinary Guidelines. *The Journal of Laryngology and Otology,* 138(S1), S1–S224. https://doi.org/10.1017/s0022215123001615

Huang, S. H., & O'Sullivan, B. (2013). Oral cancer: Current role of radiotherapy and chemotherapy. *Medicina Oral, Patologia Oral y Cirugia Bucal, 18*(2), e233–e240. https://doi.org/10.4317/medoral.18772

Jiang, C., Wang, S., & Zhu, L. (2025). Efficacy and safety of immunotherapy for head and neck squamous cell carcinoma: A meta-analysis of randomized clinical trials. Frontiers in Oncology, 14, Article 1489451. https://doi.org/10.3389/fonc.2024.1489451

Johnson, D. E., Burtness, B., Leemans, C. R., Lui, V. W. Y., Bauman, J. E., & Grandis, J. R. (2020). Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers, 6,* 92. https://doi.org/10.1038/s41572-020-00224-3

Kang, Y. J., Lee, Y. G., Chung, M. J., Kim, J., & Choi, N. (2025). Deep learning-based artificial intelligence models predict survival in patients with oral cavity squamous cell carcinoma. *Scientific Reports, 15,* Article 43537. https://doi.org/10.1038/s41598-025-27428-5

Karuppan Perumal, M. K., Rajan Renuka, R., Kumar Subbiah, S., & Manickam Natarajan, P. (2025). Artificial intelligence-driven clinical decision support systems for early detection and precision therapy in oral cancer: A mini review. *Frontiers in Oral Health, 6,* Article 1592428. https://doi.org/10.3389/froh.2025.1592428

Kasa, K., Yoshimaru, D., Ohta, H., Ohki, T., & Okano, H. J. (2026). Performance of large language models on the Japanese cardiovascular surgery board examination: A comparative analysis of eight contemporary AI models with educational implications. *General Thoracic and Cardiovascular Surgery,* 1–13. https://doi.org/10.1007/S11748-026-02304-9

Kijowska, J., Grzegorczyk, J., Gliwa, K., Jedras, A., & Sitarz, M. (2024). Epidemiology, diagnostics, and therapy of oral cancer: Update review. *Cancers, 16*(18), Article 3156. https://doi.org/10.3390/cancers16183156

Li, H.-X., Gong, Y.-W., Yan, P.-J., Xu, Y., Qin, G., Wen, W.-P., & Teng, F.-Y. (2024). Revolutionizing head and neck squamous cell carcinoma treatment with nanomedicine in the era of immunotherapy. *Frontiers in Immunology, 15,* 1453753. https://doi.org/10.3389/fimmu.2024.1453753

Lotfian, G., Parekh, K., & Suthar, P. P. (2025). Performance review of Meta LLaMA 3.1 in thoracic imaging and diagnostics. *IRADIOLOGY, 3*(4), 279–288. https://doi.org/10.1002/ird3.70013

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6), bbac409. https://doi.org/10.1093/bib/bbac409

Mäkitie, A. A., Alabi, R. O., Ng, S. P., Takes, R. P., Robbins, K. T., Ronen, O., Shaha, A. R., Bradley, P. J., Saba, N. F., Nuyts, S., Triantafyllou, A., Piazza, C., Rinaldo, A., & Ferlito, A. (2023). Artificial intelligence in head and neck cancer: A systematic review of systematic reviews. *Advances in Therapy, 40*(8), 3360–3380. https://doi.org/10.1007/s12325-023-02527-9

Mathew, A., Davis, S., Boby, J. M., R I, A., Suryavanshi, M., Dawood, S. S., Panda, P. K., Nag, S. M., Das, A., Rohatgi, N., Popat, S., Shah, R. N. H., Thampy, C., Parikh, A. R., Yadav, S., Mehta, P., Singh, R., Mukherji, D., Shilpakar, R., … Sirohi, B. (2024). Discordance in recommendation between next-generation sequencing test reports and molecular tumour boards in India. *JCO Global Oncology, 10*(10), e2300330. https://doi.org/10.1200/GO.23.00330

Miserocchi, G., Spadazzi, C., Calpona, S., De Rosa, F., Usai, A., De Vita, A., Liverani, C., Cocchi, C., Vanni, S., Calabrese, C., Bassi, M., De Luca, G., Meccariello, G., Ibrahim, T., Schiavone, M., & Mercatali, L. (2022). Precision medicine in head and neck cancers: Genomic and preclinical approaches. *Journal of Personalized Medicine, 12*(6), 854. https://doi.org/10.3390/jpm12060854

National Cancer Institute (NCI). (2025). SEER cancer stat facts: Oral cavity and pharynx cancer. Retrieved April 25, 2026, from https://seer.cancer.gov/statfacts/html/oralcav.html

Phadtare, A. S., Kulkarni, N., & Devare, M. (2026). A dual transformer-based LLaMA framework for efficient disease prediction and drug recommendation in healthcare. Network Modeling Analysis in Health Informatics and Bioinformatics, 15, Article 54. https://doi.org/10.1007/s13721-025-00687-y

Saha, H. N., Bhattacharya, D. C., Dutta, S., Bera, A., Basuray, S., Changdar, S., Banerjee, S., & Turdiev, J. (2026). Transforming healthcare with state-of-the-art medical-LLMs: A comprehensive evaluation of current advances using benchmarking framework. *Computers, Materials and Continua, 86*(2), 1–56. https://doi.org/10.32604/CMC.2025.070507

Sarker, I. H. (2024). LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling. Discover Artificial Intelligence, 4, Article 40. https://doi.org/10.1007/s44163-024-00129-0

Suri, S., Boora, G. S., Kaur, R., Chauhan, A., Ghoshal, S., & Pal, A. (2024). Recent advances in minimally invasive biomarkers of OSCC: From generalized to personalized approach. *Frontiers in Oral Health, 5,* Article 1426507. https://doi.org/10.3389/froh.2024.1426507

Tan, Y., Wang, Z., Xu, M., Li, B., Huang, Z., Qin, S., Nice, E. C., Tang, J., & Huang, C. (2023). Oral squamous cell carcinomas: State of the field and emerging directions. *International Journal of Oral Science, 15,* Article 44. https://doi.org/10.1038/s41368-023-00249-w

Tolley, A., Hassan, R., Sanghera, R., Grewal, K., Kong, R., Sodhi, B., & Basu, S. (2023). Interventions to promote medication adherence for chronic diseases in India: A systematic review. *Frontiers in Public Health, 11,* 1194919. https://doi.org/10.3389/fpubh.2023.1194919

Appendices

Appendix-A

GitHub Link: https://github.com/MuhammadUsmanManzoor/oral-cancer-llm-treatment-recommendations