applied system innovation

MDPI

Article

Multimodal AI and Large Language Models for Orthopantomography Radiology Report Generation and Q&A

Chirath Dasanayaka [1,†], Kanishka Dandeniya [1,†], Maheshi B. Dissanayake [1,*], Chandira Gunasena [2] and Ruwan Jayasinghe [2]

[1] Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya 20400, Sri Lanka; chirath.dasanayaka@eng.pdn.ac.lk (C.D.); kanishka.d@eng.pdn.ac.lk (K.D.)
[2] Faculty of Dental Sciences, University of Peradeniya, Peradeniya 20400, Sri Lanka; chandirag@dental.pdn.ac.lk (C.G.); ruwanja@dental.pdn.ac.lk (R.J.)
* Correspondence: maheshid@eng.pdn.ac.lk
† These authors contributed equally to this work.

Abstract: Access to high-quality dental healthcare remains a challenge in many countries due to limited resources, lack of trained professionals, and time-consuming report generation tasks. An intelligent clinical decision support system (ICDSS), which can make informed decisions based on past data, is an innovative solution to address these shortcomings while improving continuous patient support in dental healthcare. This study proposes a viable solution with the aid of multimodal artificial intelligence (AI) and large language models (LLMs), focusing on their application for generating orthopantomography radiology reports and answering questions in the dental domain. This work also discusses efficient adaptation methods of LLMs for specific language and application domains. The proposed system primarily consists of a Blip-2-based caption generator tuned on DPT images followed by a Llama 3 8B based LLM for radiology report generation. The performance of the entire system is evaluated in two ways. The diagnostic performance of the system achieved an overall accuracy of 81.3%, with specific detection rates of 87.9% for dental caries, 89.7% for impacted teeth, 88% for bone loss, and 81.8% for periapical lesions. Subjective evaluation of AI-generated radiology reports by certified dental professionals demonstrates an overall accuracy score of 7.5 out of 10. In addition, the proposed solution includes a question-answering platform in the native Sinhala language, alongside the English language, designed to function as a chatbot for dental-related queries. We hope that this platform will eventually bridge the gap between dental services and patients, created due to a lack of human resources. Overall, our proposed solution creates new opportunities for LLMs in healthcare by introducing a robust end-to-end system for the automated generation of dental radiology reports and enhancing patient interaction and awareness.

Keywords: dental healthcare; large language models; vision language models; Blip-2; Llama-3; multimodal AI; instruction tuning

check for updates

Academic Editors: Alberto Comesaña-Campos, Jorge Cerqueiro Pequeño, María Luísa Torres-Durán, Alberto Fernandez-Villar and Manuel Casal-Guisande

Received: 19 November 2024
Revised: 9 March 2025
Accepted: 11 March 2025
Published: 17 March 2025

Citation: Dasanayaka, C.; Dandeniya, K.; Dissanayake, M.B.; Gunasena, C.; Jayasinghe, R. Multimodal AI and Large Language Models for Orthopantomography Radiology Report Generation and Q&A. *Appl. Syst. Innov.* **2025**, *8*, 39. https://doi.org/10.3390/asi8020039

Copyright: © 2025 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The convergence of multimodal artificial intelligence (AI) and large language models (LLMs) is poised to revolutionize healthcare [1]. LLMs, with their capacity to process and generate human-like text, offer exciting possibilities for enhancing clinical documentation, patient communication, and decision support [2]. While the application of these technologies in dentistry is still in its nascent stages, the potential is significant, particularly for tasks requiring the integration of visual and textual information [3]. Orthopantomography,

a widely used panoramic dental imaging technique, plays a crucial role in dental and maxillofacial diagnostics [4]. These images offer a comprehensive visualization of the teeth, jaws, and adjacent anatomical structures, enabling the detection of a wide range of conditions, including dental caries, impacted teeth, bone fractures, cysts, and tumors. While orthopantomograms provide invaluable diagnostic information, their interpretation requires specialized training and expertise [4]. The increasing volume of dental imaging data in clinical practice presents a significant challenge, potentially impacting the timeliness and accuracy of interpretations by radiologists and other dental professionals. This necessitates the exploration of automated solutions to assist in the efficient and accurate analysis of these critical diagnostic images. Automated analysis has the potential to improve diagnostic accuracy, reduce clinician workload, and ultimately enhance patient care.

In dentistry, the automatic generation of radiology reports from orthopantomograms and the ability to respond to patient inquiries regarding radiographic data and general dental health present a compelling use case for vision–language models and LLMs. A system that integrates visual analysis with natural language generation could streamline report creation, reduce diagnostic delays, and support practitioners in detecting abnormalities with greater accuracy. Additionally, such a system could serve as an intelligent virtual assistant, enhancing clinical decision-making and improving patient education. It is worthwhile to mention that numerous studies have been conducted on applying convolutional neural network (CNN)-based AI architectures for the dental healthcare sector, such as segmentation of mental foramen in dental panoramic tomography (DPT) [5] and CNN-based multiclass image classifications [6]. However, LLM-based radiology report generation represents a fundamentally different approach. Unlike traditional classification [7] and segmentation [8,9] tasks, which focus on categorizing images into predefined classes, LLM-based systems aim to synthesize visual and textual data to generate comprehensive, human-like reports.

Multimodal techniques are being used more and more in current AI research for medical diagnosis systems as they are capable of processing, integrating, and reasoning across multiple types of data inputs or modalities simultaneously [1,10]. These modalities may include text, images, audio, video, and structured data, each representing information in fundamentally different ways. Unlike single-modality AI models, multimodal AI has the ability to capture complementary information across these diverse data types and their complex interrelationships. Through this integration, patient outcomes are optimized, individualized treatment is made possible, and diagnostic accuracy is increased. In the research presented, this strategy is demonstrated by the combination of medical imaging with clinical findings, which improves the accuracy of the radiology report generation.

1.1. Large Language Models

LLMs are a subset of deep learning architectures designed to understand, generate, and manipulate natural language by training on vast corpora of text data. Unlike traditional natural language processing approaches that rely on rule-based systems or statistical methods with limited parameters, LLMs utilize transformer-based neural network architectures with billions of parameters to capture complex linguistic patterns and contextual relationships [11,12]. These models learn language representations through self-supervised learning on diverse textual data, enabling them to perform a wide range of language tasks without task-specific training. LLMs have demonstrated remarkable proficiency in tasks such as text summarization, machine translation, question-answering, and conversational AI, making them highly versatile across various domains, including healthcare [13,14].

Recent years have introduced vision–language models (VLMs), which enable AI to understand both images and text, crucial for tasks like image captioning intelligence [15].

They comprehend the connections between images and descriptions as they were trained on image–text pairs. VLMs combine computer vision and natural language processing techniques to bridge the gap between visual and textual understanding. Hence, they are able to carry out tasks like image captioning and visual question-answering, improving AI's capacity to perceive multimodal data.

Pre-training and fine-tuning are the two stages of the training process for these models. Pre-training entails self-supervised learning on large text datasets, where LLMs gain a thorough grasp of linguistic structures by predicting masked words or following tokens, thereby acquiring broad language patterns and world knowledge. Then, using smaller, task-specific datasets along with model fine-tuning techniques such as low-rank adaptation (LoRA) [16] and instruct fine-tuning [17], these previously trained models are adapted to particular tasks. LLM answers can be improved by methods like supervised fine-tuning, parameter-efficient fine-tuning, and reinforcement learning from human feedback (RLHF) [18], bringing them into line with human preferences and enhancing safety.

These training approaches enable LLMs to acquire a broad understanding of linguistics, facilitating their application in diverse fields, such as research, education, and healthcare. While LLMs have automated numerous linguistic tasks previously requiring human expertise, challenges such as bias, hallucinations, and ethical dilemmas remain significant concerns for LLM users. This work explores the application of LLMs, specifically the VLMs concept in medical applications, investigating their ability to automate the process of radiology report generation and user interaction in the Sinhala Language. By leveraging the power of LLMs and VLMs, we aim to demonstrate their potential to enhance patient care in dental health care, while acknowledging and addressing the inherent limitations outlined in contemporary research [19].

1.2. Related Work

A recent study by Park et al. [20] assesses the potential of AI-generated radiology reports in terms of summary quality, patient friendliness, and recommendations, highlighting the unique challenges and opportunities in evaluating report accuracy and utility. In this study, two radiologists independently performed qualitative and quantitative evaluations of the AI-generated reports, using the original reports as the reference standard. Following this, two non-physician raters independently assessed the AI-generated reports for content quality and patient-friendliness. The ChatGPT-3.5-turbo model was used to generate patient-friendly summaries and recommendations, with the original MRI reports serving as prompts. Their results indicate a 1.12% of artificial hallucinations and a 7.4% of potentially harmful translations, in the generated report. Furthermore, ref. [21] highlights the capabilities of LLMs in language generation and understanding, relevant to the healthcare domain. Their review encompasses major LLM architectures, including GPT, Llama, and Bloom, noting their scale, often comprising billions of parameters. The authors also analyze trends in medical datasets used for training these models, categorizing them by subject matter, source, and size. LLMs demonstrate the potential to enhance patient care, accelerate medical research, and improve healthcare efficiency. The paper also addresses the limitations of these models and their practical application in the medical field.

Kassner et al. [22] deeply discussed the challenges that need to be addressed before LLMs can be safely and effectively used in healthcare, such as data privacy, the accuracy of the information, and potential bias. This work investigated the multilingual pre-trained language model mBERT using translated benchmarks named TREx and GoogleRE in 53 languages. The study explored mBERT's capability as a multilingual knowledge base, its language-dependent performance, and the impact of additional training text. Results indicated reasonable performance for 21 languages but instability across 42 languages.

Ref. [23] discussed the instruction tuning of LLMs for biomedical natural language processing. The paper proposed BioInstruct, a novel method of instruction-tuning LLMs with natural language instructions and their corresponding inputs and outputs. The Section 3 further discussed self-instruction, low-rank adaptation (LoRA) [16], instruction tuning, and multi-task learning technologies that can be used in the fine-tuning process. The BioInstruct model has been able to demonstrate performance gains of 17.3% in question-answering, 5.7% in information extraction, and 96% in text generation tasks.

Numerous studies have demonstrated the efficacy of vision–language models in medical imaging analysis. Thawkar et al. proposed XrayGPT [24], a system that leverages medical vision–language models to summarize chest radiographs. Their work demonstrates the potential of vision–language models for understanding and summarizing medical images, specifically in the context of chest radiographs. By combining visual understanding with language processing, XrayGPT provides concise and informative summaries that can assist radiologists in interpreting complex chest radiographs. On the other hand, Wang et al. introduced MedClip [25], a contrastive learning framework that integrates medical images and text to enhance clinical decision support. Their study showcases the effectiveness of vision–language models in capturing meaningful medical information from unpaired image–text data. MedClip offers valuable insights for improving diagnostic accuracy and aiding in treatment planning, demonstrating the potential of vision–language models in clinical decision-making. Eslami et al. explored the benefits of CLIP (contrastive language–image pre-training) in the medical domain, specifically in visual question-answering (VQA). Their study, PubMedCLIP, investigates how CLIP can enhance VQA tasks by utilizing medical domain-specific data from PubMed [26]. By integrating medical knowledge with visual understanding, PubMedCLIP improves the accuracy of answering medical questions based on image inputs. This research highlights the potential of vision–language models to facilitate effective patient communication by providing accurate and informative responses to medical queries. Chaoyi et al. introduced MedKLIP, a method to enhance self-supervised visual-language pre-training in the medical domain by utilizing paired image–text reports [27]. The method includes a report filter for extracting medical entities, an entity embedding module utilizing external knowledge bases, and a transformer-based fusion model. Similarly, Yakoub et al. presented a transformer encoder–decoder architecture for a medical visual question-answering system, combining image features from a vision transformer model and textual encoding for questions, achieving promising results [28]. A study by Chen et al. investigated the use of vision–language models to generate interactive and informative radiology reports [29]. The model analyzed medical images, extracted key findings, and generated natural language summaries, enabling patients to better comprehend their radiology reports and engage in discussions with their healthcare providers.

1.3. Contributions

It should be noted that both vision language models and LLMs have the potential to play a substantial role in dental healthcare. Vision language models contribute to interpreting diagnostic images for radiology report generation in precise treatment planning, while LLMs improve communication between healthcare providers and patients, offering educational resources, and enhancing information exchange. As such, combining these models can significantly improve the quality of care, making dental healthcare more accessible, efficient, and effective. The proposed study investigates the potential of large language models (LLMs) in dentistry, focusing on tasks that require the integration of visual and textual data. Specifically, we explore two key applications: automated radiology report generation from DPT, i.e., orthopantomography images, and a question-answering (Q&A) system capable of interpreting radiographic findings and addressing general dental

healthcare inquiries. A significant novelty of the proposed Q&A system is its operation in the Sinhala language. This research aims to evaluate the feasibility and effectiveness of LLMs in these critical areas, thereby facilitating their broader adoption in dental practice.

The main contributions of this research are as follows:

1. Radiology report generation system: A dental radiology report refers to a document generated by a radiologist or dentist after reviewing DPT images. These reports include key information such as findings, impressions, and recommendations, aiding in the diagnosis and monitoring of various dental health conditions. Currently, dental professionals conduct manual reviews of DPT images for radiology report generation. This research introduces an automated radiology report generation methodology aimed at improving diagnostic efficiency for dental professionals.
2. Sinhala language question-answering platform: The second component of this research addresses the critical need for clear and accurate communication between healthcare providers and patients, essential for effective diagnosis and treatment. The World Health Organization (WHO) has reported significant challenges in accessing adequate healthcare information, particularly in developing countries, often due to language barriers, high costs, and limited availability of dental services [30]. Facilitating access to professional medical opinions in local languages is therefore crucial. To address this, we present a robust end-to-end system for dental radiology report generation and a Sinhala language question-answering system focused on the dental domain.
3. Introduction of four new datasets: This research introduces several novel datasets to support future research in this area. Namely, the DPT Image and Caption Dataset [31], The Radiology Reports Dataset [32], the Sinhala Language Corpus Dataset [33], and the Sinhala Language Question-Answering Dataset for the Dental Domain [34].
4. Utilization of multimodal AI: In our work, we utilized dental panoramic tomograms (DPTs) and generated radiology reports to assess patient health. We then combined these findings with past patient queries to formulate answers and to address user/patient inquiries effectively. In summary, we developed a system that uses multimodal AI to generate clinical findings from images, integrating medical imaging and patient history to conduct meaningful Q&A sessions.

2. Datasets

2.1. Ethical Clearance and Data Gathering

The research involved the collection of DPT image data and sample radiology reports from patients. Hence, ethical clearance was sought from the relevant authorities, namely Teaching Hospital Peradeniya, Sri Lanka, and the Postgraduate Institute of Science University of Peradeniya. The patient data were collected at Teaching Hospital Peradeniya, Sri Lanka. Ethical clearance was obtained under the guidelines and regulations set by the respective hospital ethics committees and international ethical standards. Additionally, compliance with the Personal Data Protection Act No. 9 of 2022 (PDPA Sri Lanka) was ensured to further strengthen personal data protection measures [35]. It was important to emphasize that patient privacy and confidentiality were strictly maintained throughout the research. To protect patient identities, all personally identifiable information (PII) was removed from the collected data, and the data were anonymized.

2.2. Datasets

DPT images are valuable tools in dentistry for diagnosing various dental and oral conditions, such as impacted teeth and jaw disorders, and assessing overall dental health. Their panoramic perspective helps dental professionals in treatment planning by assisting

them in obtaining a clear picture of the patient's oral anatomy. Hence, this study uses DPT images as the input information and automatically generates radiology reports along with a user support service in the form of a Q&A platform in the Sinhala language. With this aspiration, our study generates and utilizes four distinct datasets, each focused on a distinct task, caption generation for DPT images, radiology report generation, Sinhala language understanding, and fine-tuning for question-answering in the Sinhala language.

2.3. The DPT Image and Caption Dataset

One of the four datasets, named The DPT Image and Caption Dataset, comprises 1000 image–caption pairs of DPT images. Each image in the dataset was descriptively annotated by a panel of experts specializing in dental and maxillofacial radiology, each with over 10 years of experience. These annotations provide detailed and clinically relevant captions for each image. Figure 1 shows the distribution of oral complexities in the DPT image dataset. The disparity in the frequency of occurrence of each oral complexity in the general population causes the class imbalance in this dataset. Adhering to the ethical committee suggestions, a subset of this dataset has been made available publicly under the Apache License 2.0, allowing for its utilization in future research within the domain. This dataset can be accessed using [31]. Table 1 shows samples of the DPT Image and Caption Dataset [31], i.e., image–caption pairs.

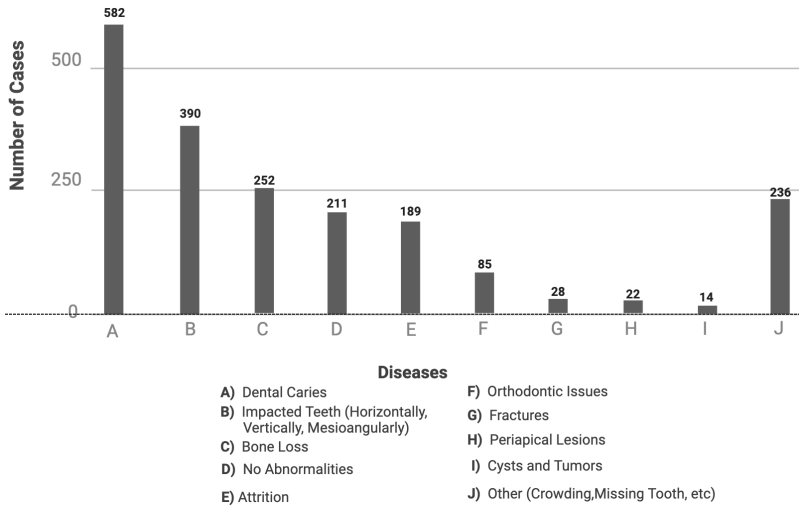


Figure 1. Case distribution for the DPT image dataset.

Table 1. Samples from the DPT image–caption dataset.

DPT Image	Captions (Findings)
	1. Grossly carious 2. Pulp calcification 3. Mild level generalized bone loss in the maxilla and mandible
	1. Bilaterally vertically impacted mandibular 3rd molar 2. Fracture on the left mandibular parasymphyseal region 3. Dental caries are present
	1. Vertically impacted mandibular 3rd molar 2. Mesially tilted mandibular 3rd molar 3. Radiopaque lesion in the right submandibular region, most likely a calculus

2.4. The Radiology Reports Dataset

Another dataset prepared for this research is the radiology reports dataset. It consists of a combination of approximately 1200 radiology reports that were manually curated by medical professionals. The dataset consists of sample reports of a wide variety of dental health-related diseases, including some rare diseases such as lesions and jaw bone fractures. Before feeding the data into the model, a manual verification was conducted by medical experts to ensure the accuracy and reliability of the reports. This dataset is also made publicly accessible under the Apache License 2.0 to permit its use in subsequent future research endeavors within the domain [32]. The radiology reports dataset is arranged in markdown format, and an example data point is shown in Table 2. The full radiology report is not shown due to space limitations and the readers are directed to [32] for further details on the contents of the radiology reports dataset.

Table 2. Extracted sample from the radiology reports dataset.

DPT Image–Caption	Radiology Report
1. ** Bilaterally impacted third molars: ** Bilaterally impacted third molars (wisdom teeth) are noted, with the left third molar showing mesial angulation and the right third molar showing distal angulation. 2. ** No Evidence of dental caries: ** No radiographic evidence of dental caries is observed in the visualized teeth.	**Patient Information:** Patient Name: Date of Birth: Referring Physician: Date of Examination: **Clinical History:** Not Present **Imaging Study:** Dental Panoramic Tomography (Pantomogram) **Findings:** The dental panoramic tomography image reveals the following findings: 1. ** Bilateral impacted third molars: ** Bilateral impacted third molars (wisdom teeth) are noted, with the left third molar showing mesial angulation and the right third molar showing distal angulation.

2.5. The Sinhala Language Corpus Dataset

The next dataset compiled for this research is a Sinhala corpus dataset. This was built using web scraping of Sinhala websites, blogs, PDF files, movie subtitles, and Wikipedia articles written in the Sinhala language. This corpus comprises 12.3 billion tokens, making it the largest Sinhala language dataset currently available on Hugging Face. Similar to previous datasets, this corpus is publicly accessible under the Apache License 2.0, facilitating its application in future research within the domain [33]. This dataset is utilized for the continued pre-training of the Llama 3 8B model [36].

2.6. The Sinhala Language Question-Answering Dataset for the Dental Domain

As the final dataset of this research, a Sinhala language question-answering dataset for the dental domain was curated by the authors. Figure 2 shows the sample entries from this particular dataset.

This dataset, available to the public at [34], includes general questions and answers of the dental domain, addressing concerns within the Sinhala-speaking community. Our objective at this stage is to build an interactive environment for patients to gather information related to the DPT report generated and the follow-up actions. To achieve this objective, we manually extracted 2500 dental-related questions and answers from Facebook community

groups. During the selection of answers provided for community-raised questions, we strictly chose responses solely from group administrators and moderators, who we verified as dental doctors and consultants. The process involved the thorough filtering out of non-relevant questions, and no PII was extracted in adherence to privacy standards.



Question string · lengths 25 899	Answer string · lengths 4 1.25k	text string · lengths 82 1.31k
Doctor මගේ පුතාට bruxism තියෙනවා. මටත් තියෙනවා. පුතාට වයස 13. මොකක්ද කරන්න පුලුවන් ?	Mouth guard එකක් වගේ හොඳ appliance එකක් පාවිච්චි කරන්න පුළුවන්.හානිය වලක්වගන්න සහ පුරුද්ද…	### Instruction: Doctor මගේ පුතාට bruxism තියෙනවා. මටත් තියෙනවා. පුතාට වයස 13. මොකක්ද කරන්…
Docter,මගෙ පුතාගෙ කිරි දත් වැටිල නැහැ.යටින් ස්ථිර දත් දෙක එනවා.කිරි දත් ගලවන්න වෙයිද.	කිරි දත් දෙක ගලවගන්න.	### Instruction: Docter,මගෙ පුතාගෙ කිරි දත් වැටිල නැහැ.යටින් ස්ථිර දත් දෙක එනවා.කිරි දත් ගලවන්න වෙයිද.…
මට මේකට පිළිතුරක් දෙන්න.මන් දැන් දවස් පහකට විතර කලින් කම්බි මාරු කරා එතකොට ඒ මාරු කරද්දි අර චූටි…	උතුර කන නාසය සම්බන්ධ (ENT) වෛද්‍යවරයෙකුට පොඩ්ඩක් පෙන්නන්න.	### Instruction: මට මේකට පිළිතුරක් දෙන්න.මන් දැන් දවස් පහකට විතර කලින් කම්බි මාරු කරා එතකොට ඒ මාරු…

Figure 2. Samples from the Sinhala language question-answering dental dataset. English Translation see Table 3.

Table 3. English Translation of the Sample Sinhala Language Dataset Illustrated in Figure 2.

Question	Answer	Text
My son has bruxism. I have it too. My son is 13 years old. What can I do?	You can use a small appliance like a mouth guard to prevent damage and break the habit.	### Instruction: Doctor, my son has bruxism. I have it too. My son is 13. What can I do? ### Response: You can use a small appliance like a mouth guard to prevent damage and break the habit.<eos>
Doctor, my son's baby teeth haven't fallen out. Two permanent teeth are coming in underneath. Do I have to remove the baby teeth?	Remove the two baby teeth.	### Instruction: Doctor, my son's baby teeth haven't fallen out. Two permanent teeth are coming in underneath. Do I have to remove the baby teeth? ### Response: Remove the two baby teeth.<eos>
Please answer this. I just had the wire changed about five days ago, and when I was changing it, the doctor didn't remove that little wire from my mouth for a while.	Please show it to an ENT doctor.	### Instruction: Please answer this. I just had the wire changed about five days ago, and when I was changing it, the doctor didn't remove that little wire from my mouth for a while. ### Response: Please show it to an ENT doctor.<eos>

2.7. Summary of Datasets

Table 4 summarizes the four datasets used in this study. Data quality was ensured through a verification process applied to random samples from each dataset. For the radiology report generation model, this verification was performed manually. For the language model datasets, a combined manual and algorithmic approach was used. The algorithms facilitated tasks such as language validation, duplicate removal, and tag cleaning (e.g., HTML, XML).

Table 4. Summary of datasets.

Dataset Name	Training Data Points	Testing Data Points	Dataset Validation Method
DPT Image and Caption Dataset	900	100	Manual
Radiology Reports Dataset	1000	200	Manual
The Sinhala Corpus Dataset	11.1 Billion Tokens	1.2 Billion Tokens	Manual & Algorithmic
Sinhala Language Question-Answering Dataset	2500	1000	Manual & Algorithmic

3. Methodology

3.1. Proposed Complete System Architecture

The high-level architecture diagram of the proposed system is presented in Figure 3. The generation of radiology reports and enabling user engagement in the form of discussions when needed are the main responsibilities of the proposed system. The radiology report generation is carried out in two stages. First, the DPT image was passed through a Blip-2-based image captioning API, and then its output was fed into the fine-tuned Llama 3 8B model for radiology report generation. Sinhala language queries were routed to the same Llama 3 8B model that was enhanced for understanding the Sinhala language to facilitate user interaction. All user feedback was stored in the Google Firestore in order to enable refinement and enhancement of the model in the future using reinforcement learning with human feedback (RLHF) [37]. This process aimed to improve the model's responsiveness and effectiveness in handling Sinhala queries. The user interface of the proposed system was developed using the Chainlit Python framework. It was containerized using Docker and deployed in Google Cloud using Cloud Run. The backend APIs were deployed using Hugging Face inference endpoints and AWS SageMaker.

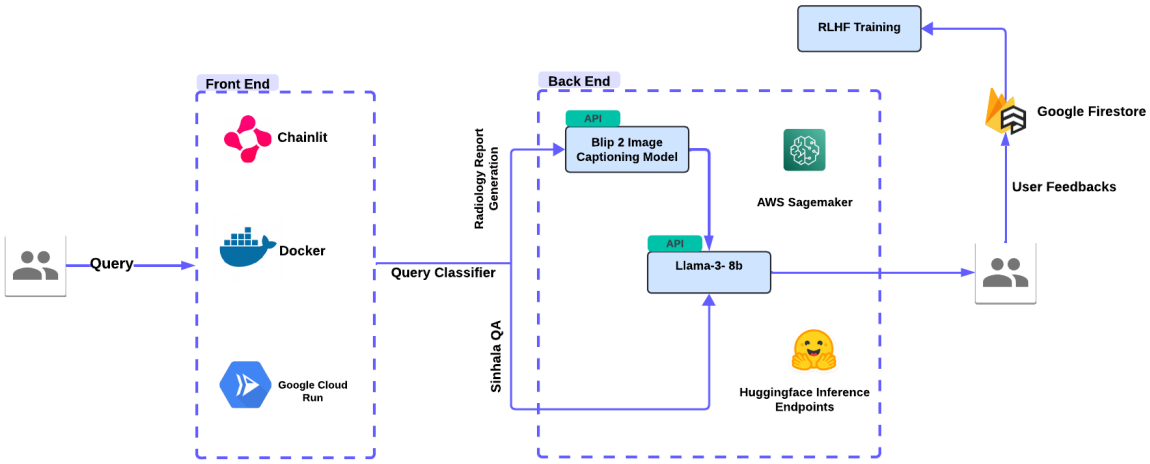


Figure 3. High-level application architecture of the proposed system.

3.2. Blip-2 Architecture for Caption Generation

The first step in radiology report generation is caption generation for input, DPT images. The Blip-2 [38] model was adopted for the image–text retrieval task. These retrievals, recorded as captions, cover a range of conditions and complexities that can be extracted from the DPT image, such as dental caries, periodontal diseases, cysts, tumors, and other oral complexities. Additionally, data augmentation techniques, particularly grid distortion and elastic transformation [39], were applied to improve the model training process and to achieve enhanced diagnostic accuracy. Blip-2 architecture bridges the gap

between vision and language by using frozen pre-trained image encoders and frozen LLMs. Figure 4 presents the architecture of the Blip-2 model used in this research.

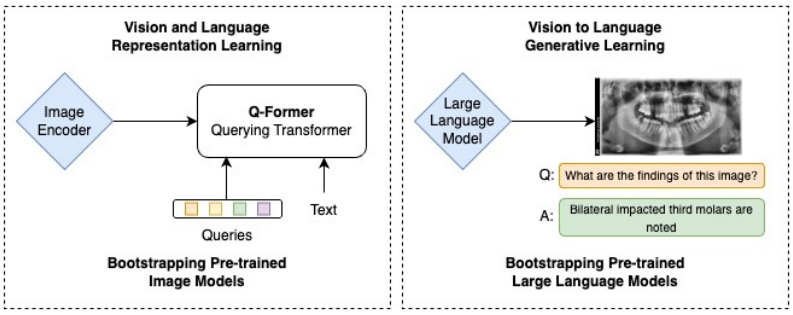


Figure 4. Proposed Blip-2 architecture.

As presented, the Blip-2 model consists of three main modules, which are a frozen image encoder, a frozen LLM, and a querying transformer (Q-Former). The image encoder processes the image input and extracts image embeddings, while the LLM processes the text input and generates text embeddings. Vision transformer [40] and Flan T-5 [41] models were utilized to generate the image embeddings and text embeddings. The Q-Former is a lightweight querying transformer that bridges the modality gap between the image encoder and the LLM. It is pre-trained in two stages: the first stage bootstraps vision–language representation learning from a frozen image encoder, and the second stage bootstraps vision-to-language generative learning from a frozen language model. While this same architecture could have been utilized to generate the radiology report for the input DPT images, the limitations of the 512-token context window and the language capabilities of the Blip-2 model led us to seek a different model for the report generation task.

3.3. LLM Model Selection for Report Generation and Domain Adaptation for LLM Training

LLaMA 3 8B model was selected for radiology report generation and question-answering tasks based on its performance across multiple benchmark evaluations, including MMLU (73.0), IFEval (80.4), GPQA (32.8), and MGSM (68.9), as shown in Table 5. These benchmarks assess general knowledge, reasoning, and complex problem-solving abilities, which are crucial for generating clinically accurate and contextually coherent radiology reports. In comparison, models like Mistral 7B and Gemma 2 9B showed notably lower scores, reflecting limitations in complex text generation tasks. Additionally, encoder-only models such as BERT and RoBERTa are primarily designed for natural language understanding (NLU) tasks like classification and named entity recognition, and encoder–decoder models like T5 are better suited for sequence-to-sequence tasks. Neither architecture excels in long-form, free-text generation needed for radiology reports, further justifying our preference for decoder-only LLMs like LLaMA 3 8B in this domain. However, while large language models like LLaMA 3 8B have demonstrated remarkable capabilities in various natural language processing (NLP) tasks, their performance significantly degrades when applied to domains that differ from the training data. This is because LLMs are trained on massive datasets of text and code, which may not adequately represent the specific terminology, nuances, and context of a particular domain. To address this challenge, domain adaptation techniques are employed. In domain adaptation, an LLM already trained on a different domain is fine-tuned on a target domain using target domain-specific data while incorporating domain-specific constraints into the models' architecture, with the scope of improving accuracy, robustness, and generalization. LLM domain adaptation is particularly crucial in domains where data are scarce or sensitive, especially in the case of the medical domain. Through the adaptation of LLM to the medical domain, organizations

can utilize its capabilities for tasks such as medical diagnosis assistance. Most importantly, this adaptation prioritized the privacy and confidentiality of sensitive medical information by allowing users to run a privately hosted version of an LLM model instead of relying on commercially available LLM models like ChatGPT, Gemini, etc. Hence, with domain adaptation, it is possible to adapt data-hungry LLMs to tasks in the data-scarce medical domain, enhancing their performance and applicability in real-world scenarios.

Table 5. Benchmark Scores for Llama 3 8B, Gemma 2 9B, and Mistral 7B [42].

Model	MMLU	IFEval	GPQA	MGSM
Llama 3 8B	73.0	80.4	32.8	68.9
Gemma 2 9B	72.3	73.6	-	53.2
Mistral 7B	60.5	57.6	28.8	29.9

Our study explored domain adaptation in LLMs through instruct fine-tuning and continued pre-training approaches. Although these approaches share some similarities, they also exhibit notable differences. The process of continued pre-training involves further training a pre-trained model using a large dataset, maintaining the same objective function and training setup as the initial pre-training phase. This methodology ensures the model's understanding of the domain's linguistic nuances and patterns while reinforcing its foundational language understanding. On the other hand, instruct fine-tuning generally involves adding a task-specific layer to the pre-trained model, utilizing a smaller, task-specific dataset for training. This process aims to align the model with a task-specific objective function to optimize the model for a particular task. In summary, these two fine-tuning approaches offer an effective mechanism for training a new LLM from the ground up, without demanding intensive resources using the domain adaptation concepts. Furthermore, we adopted domain adaptation in LLMs in three distinct instances within the scope of the study, as discussed below.

3.3.1. Instruct Fine-Tuning of Llama-3-8B Model for Radiology Report Generation

As pointed out earlier, the proposed system generates a radiology report for input DPT images using the medical captions generated at Blip-2. A fine-tuned Llama-3-8B LLM model was adopted for the report generation. We selected the Llama model family due to its noticeable performance against state-of-the-art benchmarks [43].

Instruct fine-tuning is a specialized technique adopted in LLMs to tailor the model to perform specific tasks based on explicit instructions. Hence, it plays a significant part in aligning the pre-trained model's behavior to effectively cater to a specified task. Although this process introduces some task-specific patterns to the model, alignment is the main emphasis of model fine-tuning, rather than direct knowledge injection.

Llama-3 is a family of pre-trained LLMs, with the capacity to fine-tune further. From the Llama-3 family, the Llama-3-8B parameter model was selected, considering several aspects, such as the computational efficiency, cost associated with training and hosting, and capacity to generate accurate radiology reports after fine-tuning. This selection followed a trade-off between the computational resources and the model's ability to produce precise and relevant radiology reports. The fine-tuning process was carried out primarily using AWS SageMaker training instances. Potential hardware constraints were effectively managed using quantized low-rank adaptation (QLoRA) [44] which is a parameter-efficient fine-tuning (PEFT) technique [45]. QLoRa employs the following three innovative methods to reduce the memory footprint: the 4-bit NormalFloat quantization technique, double quantization, and the page optimizer.

By applying these techniques, a single linear layer in the quantized base model with a single LoRA adapter can be defined as follows:

$$\mathbf{Y}^{\text{BF16}} = \mathbf{X}^{\text{BF16}} \cdot \text{doubleDequant}(c_1^{\text{FP32}}, c_2^{k\text{-bit}}, \mathbf{W}^{\text{NF4}}) + \mathbf{X}^{\text{BF16}}\mathbf{L}_1^{\text{BF16}}\mathbf{L}_2^{\text{BF16}}. \quad (1)$$

Here doubleDequant(·) is defined as,

$$\text{doubleDequant}(c_1^{\text{FP32}}, c_2^{k\text{-bit}}, \mathbf{W}^{k\text{-bit}}) = \text{dequant}(\text{dequant}(c_1^{\text{FP32}}, c_2^{k\text{-bit}}), \mathbf{W}^{4\text{bit}}) = \mathbf{W}^{\text{BF16}} \quad (2)$$

where $\mathbf{Y}^{\text{BF16}}$ is the output matrix after applying the QLoRA adaptation to the input matrix, and $\mathbf{X}^{\text{BF16}}$ is the input matrix in the brain floating point 16 (BF16) format. $\mathbf{W}^{\text{NF4}}$ stands for the weight matrix in a specific 4-bit quantized format (NF4), $c_2^{k\text{-bit}}$ is the scaling factor in k-bit precision used in the second step of dequantization, and L_i^{BF16} denotes the low-rank adaptation matrices in BF16 format used in the ith linear layer. The training continued until the validation loss was reduced to 0.04. The Llama 3 8B base model was fine-tuned on the radiology reports dataset. Following the fine-tuning phase, the model was transformed into an API endpoint and subsequently hosted using Hugging Face's inference endpoints. The instruct prompt used in the fine-tuning process is shown in Table 6 below.

Table 6. Prompt template used to fine-tune the Llama 3 8B model.

Prompt Template Used for Fine-Tuning Radiology Report Generation
<BOS> **###Context:** [Patient history if available] **###Instruction:** [Image–caption that was generated from the Blip-2 model] **###Response:** [Radiology report] **<EOS>**

3.3.2. Continued Pre-Training of Llama 3 8B Model

With emphasis placed on giving patients easy and direct access to dental healthcare information, we also explored the possibility of creating an AI-powered discussion platform for patient engagement in the native Sinhala language.

Sinhala is the primary language of Sri Lanka, spoken only by the country's approximately 21 million people. The adoption of LLMs for Sinhala is challenging due to resource scarcity, language complexity, and linguistic diversity. LLMs are primarily trained on English textual data, and the limited availability of high-quality digitized Sinhala text datasets, along with the rich inflectional morphology of Sinhala, creates challenges for tokenization and language modeling. In this section, we present an efficient and cost-effective method to enhance the language comprehension of LLM from English to Sinhala. The method presented can be adapted to train LLM for any other language as well.

First of all, several open-source LLM models such as Llama 3 8B, Mistral-7b [46], Falcon-7b [47], and Gemma-7b [48] were evaluated to understand their effectiveness in handling the Sinhala language. The observations revealed that most of them failed to provide acceptable responses to the Sinhala language prompts, except Llama 3 8B. The outcome of the Llama 3 8B model for the prompt *"Good morning, how are you?"* in Sinhala is shown in Figure 5. The outcome shows that while the Llama 3 8B model was able to produce some responses in Sinhala, the results lacked meaningful context.

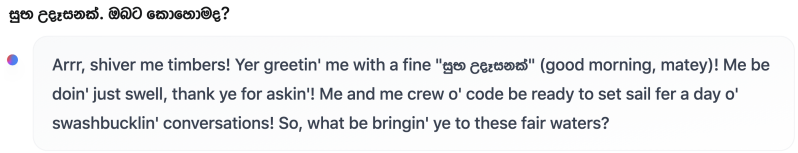


Figure 5. Response of the Llama 3 8B model for a general Sinhala language prompt.

Upon investigating further on the Llama 3 8B model tokenizer, it was understood that the tokenization process was inefficient and often broke words down to the character level, which resulted in an undesirable output. This could be mainly due to the scarcity of the Sinhala language tokens in the pre-trained Llama tokenizer. The model's limited exposure to Sinhala language content during its initial training phase is a key factor contributing to its inability to handle the Sinhala language effectively. To further enhance Llama 3 8B's Sinhala language understanding, we continued its pre-training on a large-scale Sinhala corpus. By exposing the model to a large amount of Sinhala text data, we enabled it to capture Sinhala linguistic patterns and nuances, resulting in improved language understanding. Figure 6 illustrates this process.

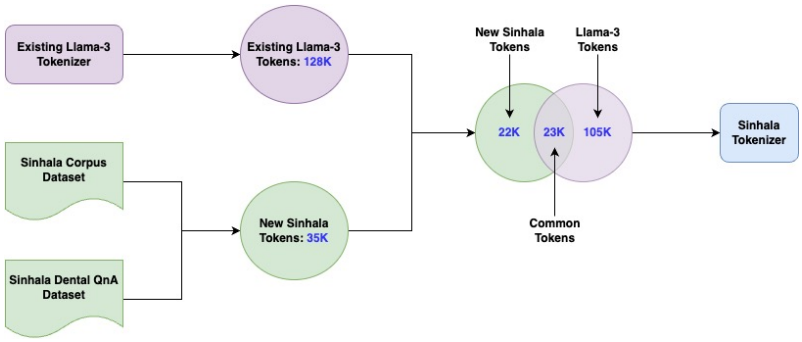


Figure 6. The updating process of the existing Llama 3 8B tokenizer with Sinhala tokens.

For the continued pre-training, we used an A100 GPU equipped with 40 GB of VRAM, which was crucial for handling the large corpus. Additionally, the system was equipped with a 78.2 GB SSD and 83.5 GB of RAM, essential for efficiently managing larger datasets and saving model checkpoints, thereby enhancing overall training efficiency. As the first step in the continued pre-training process, we trained a new tokenizer using both the Sinhala corpus dataset and the Sinhala language question-answering dataset. This resulted in generating 22k tokens for the Sinhala language. Following this, a new vocabulary was created by combining the Sinhala token set with the Llama 3 8B's token set, removing any common tokens. The final vocabulary resulted in 150k tokens. It should be noted that the average token count for the Sinhala Language question-answering dataset was approximately 800 tokens per question-answering pair before fine-tuning the tokenizer for Sinhala. After optimization, the token count was significantly reduced to an average of around 200 tokens, indicating a percentage drop of 75% in the generated tokens. Figure 7 shows the tokenizer comparison for a sample input from the Sinhala question-answering dataset. The Sinhala Corpus dataset was used to refine Llama 3 8B for Sinhala language understanding. The dataset was first split into chunks, each with 4096 tokens. These chunks were then utilized for training the model. As with the instruct fine-tuning previously discussed, domain adaptation strategies like PEFT and QLoRA [49] were also used in this instance to overcome hardware constraints.

Sinhala text : උඩ ඇන්දේ දත් ගැලෙව්වාම.මොලේ ස්නායු වලට බලපානවා කියන කතව ඇත්තද?

Tokenizer	First Few Tokens	Token Length
Llama-3-8b Tokenizer	'▁', '<0xE0>', '<0xB6>', '<0x8B>', '<0xE0>', '<0xB6>', '<0xA9>', '▁', '<0xE0>', '<0xB6>', '<0x87>', 'න', 'ේ', '<0xE0>'	119
Llama-3-8b Sinhala Tokenizer	'▁උඩ', '▁ඇ', 'න්දේ', '▁දත්▁', 'ගැ', 'ලෙ', 'ව්වා', 'ම.', 'මෙ', 'ලේ', '▁ස්නා', 'යු', '▁වලට', '▁බලපාන'	20

Figure 7. Tokenizer comparison for a sample input from the Sinhala question-answering dataset.

3.3.3. Instruct Fine-Tuning of the Continued Pre-Training Llama 3 8B Model for Sinhala Question-Answering

After continued pre-training of the Llama 3 8B model as in Section 3.3.2, it was fine-tuned on the Sinhala language question-answering dataset to align the model output with more human-like responses to Sinhala queries in the domain of dental sciences. The process is similar to the instruct fine-tuning method that was discussed previously in Section 3.3.1.

4. Results

4.1. Results Evaluation Metrics

Evaluating the performance of an LLM, especially in a domain-specific task, demands a combination of qualitative and quantitative metrics. Furthermore, to obtain a better understanding of the system's performance, it is necessary to assess the system's capability under different tasks at different intermediary stages. The generation of the radiology report and creating an interactive dialog with the patient in the Sinhala language as a decision support system are the two most important tasks within the purview of our study. With this scope, we have selected the metrics presented in Table 7 for the task of evaluating the system's performance.

The existing objective evaluation metrics for LLMs, such as perplexity [50], and the bilingual evaluation understudy (BLEU) [51] provide insights into the quantitative performance of LLMs, while mainly focusing on language-related aspects. Furthermore, reference-based overlap utility for gisting evaluation (ROUGE) [52] measures the overlap of n-grams, word sequences, and word pairs between the generated and original text, ensuring that the generated report accurately reflects the essential content of the reference report. The BLEU metric primarily focuses on machine translation and may not effectively capture the specific content requirements of a radiology report. Perplexity measures the probability of a sequence of words and reflects the model's fluency in language generation. However, it does not directly measure content accuracy and relevance, which are vital for radiology reports. Considering the above-described reasons and the usefulness of the ROUGE metric in capturing the correct medical terminology, findings, and conclusions, we employed ROUGE to quantitatively evaluate the radiology reports.

Table 7. Evaluation methods for each model.

Scenario	Qualitative Evaluation	Quantitative Evaluation
Radiology report generation	Weighted average score given by dental professionals	ROUGE
Sinhala language Q&A in the dental domain	Weighted average score given by dental professionals	DDSLU benchmark

However, in the medical domain, which is rich in complex nuances and contextual understanding, relying fully on objective metrics may result in a misleading understanding

of the model's performance. Subjective evaluation, i.e., human evaluation, is crucial in this domain. Human factors, such as domain expertise, ethical considerations, subjective judgment, and adaptability to new information as well as background information, play a pivotal role in decision-making in the medical domain. Hence, in this study, we prioritize a human-centered evaluation approach over objective metrics. To conduct this qualitative assessment, feedback was taken from three categories of dental professionals, both before and after the domain adaptation of the LLM. Each evaluator gave a score out of 10 for the output at report generation and Q&A in Sinhala. The rating process was iterated 10 times per professional to avoid biasing and random noise. The final score was determined by the weighted average of their ratings. The evaluators consisted of (1) a senior doctor with over 10 years of experience, who has completed a four-year residency training program and is a specialist in dental and maxillofacial radiology; (2) a doctor with 5 years of clinical experience; and (3) a graduate doctor—a recently qualified BDS degree holder with 1 year of experience. The weights for each category, tabulated in Table 8, were allocated by considering their professional experience.

Table 8. The weights assigned to each dental professional category.

Category of Dental Professionals	Experience	Given Weight
Senior Doctor	Over 10 years	5
Doctor	5 years	3
Graduate Doctor	1 year	2

One of the bottlenecks faced when performing the quantitative evaluation of the Sinhala language discussion platform is the unavailability of an established benchmark dataset. Hence, we took the initiative to create a benchmark dataset for Sinhala, comprising 1000 multiple-choice questions (MCQs) related to dentistry in the Sinhala language. The benchmark dataset, named the Dental Domain and Sinhala Language Understanding Dataset (DDSLU) [53], was created by authors with the input and responses of dental professionals, and it consists of a wide range of general dental-related inquiries. Our approach was inspired by the Massive Multitask Language Understanding (MMLU) benchmark dataset [54]. The performance of the Llama 3 8B model on the DDSLU dataset was evaluated before and after the domain adaptation.

4.2. Radiology Report Generation: Blip-2 Results

The performance of the fine-tuned Blip-2 model was evaluated, both qualitatively and quantitatively, on 100 DPT image–caption pairs. The evaluation focused on the accuracy of the captions generated for the corresponding DPT images. Figure 8 illustrates the total number of cases against the correctly identified cases for each dental condition detected in DPT images. It should be noted that some DPT images indicate more than one dental condition, hence, there are more cases than the number of DPT images in the test set.

The findings indicate that the model can accurately identify 87.9% of dental caries, 89.7% of impacted teeth, 88% of bone loss, and 81.8% of periapical lesions. These four cases were the dental conditions that produced a detection accuracy above 80%. Bone fractures and orthodontic issues had the lowest detection accuracy, with 60% and 62.5% accuracy values, respectively. The decreased number of cases in the training dataset is the primary cause of the lower ratings for these two conditions. Therefore, model performance can be enhanced over time by increasing the size of the dataset and the density of cases for each dental condition.

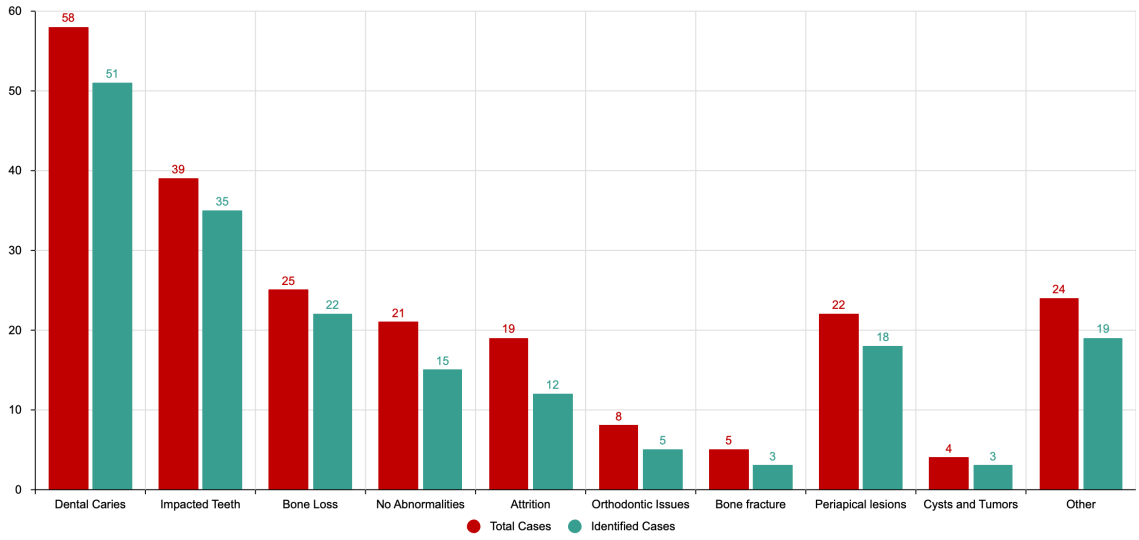


Figure 8. Total number of cases and correctly identified cases for the test set.

4.3. Radiology Report and Generation: LLM Results

We conducted a qualitative evaluation of the generated radiology reports to understand the models' accuracy compared to reports generated by dental professionals. The ROUGE-1 value for this evaluation stood at 72.7%. In the subjective quality evaluation of the radiology reporting, a group of dental professionals evaluated the generated results, individually using the guidelines mentioned in Table 9, and assigned a rating within 0–10 with 10 indicating best performance and 0 indicating worst performance. This assessment utilized 100 radiology reports.

The inter-observer variability among each pair of evaluators (senior doctors vs. doctors, senior doctors vs. graduate doctors, and doctors vs. graduate doctors) was analyzed using Bland–Altman [55] plots as illustrated in Figures 9–11. The mean difference observed between the senior doctor and doctor, as well as between the senior doctor and the graduate doctor, was approximately −1, indicating that the senior doctor's evaluations tended to be slightly more stringent. Owing to this observation, a higher weight was assigned to the senior doctor's grading, as presented in Table 10. Moreover, the mean difference between the doctor and the graduate doctor was around 0.2, suggesting closer alignment in their evaluations. This observation led to assigning weights of 3 and 2, respectively, to doctors and graduate doctors. Furthermore, only a few data points were outside the acceptance regions for each plot, indicating general agreement among the evaluators within acceptable limits.

Table 9. Sample guidelines for rating the generated reports.

Evaluation Criteria	Score
Identification of Findings	7.0
Detail of Descriptions	7.9
Comprehensive Reporting	8.0
Diagnostic Accuracy	8.3
Recommendations Clarity and Conciseness	8.2
Medical Terminology	8.4
Ethical Considerations	8.7
Average Score	8.1

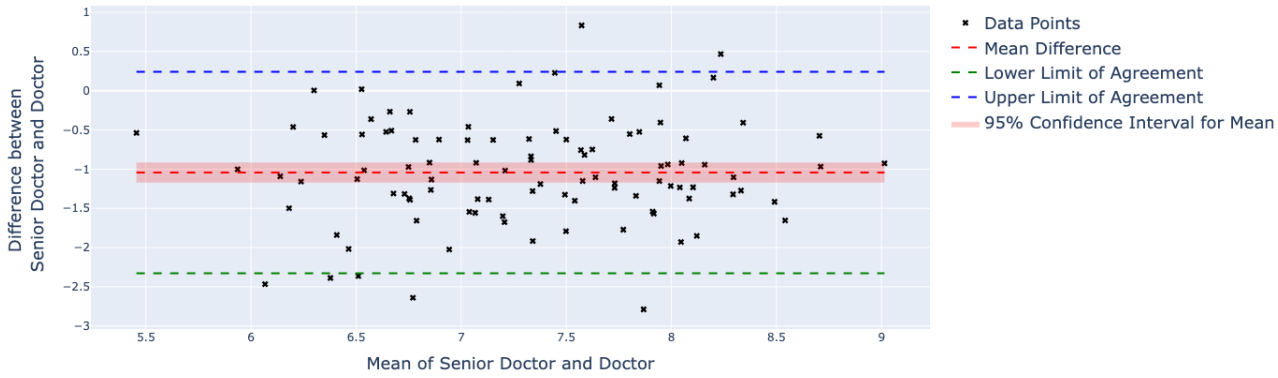


Figure 9. Inter-observer variability (senior doctor vs. doctor).

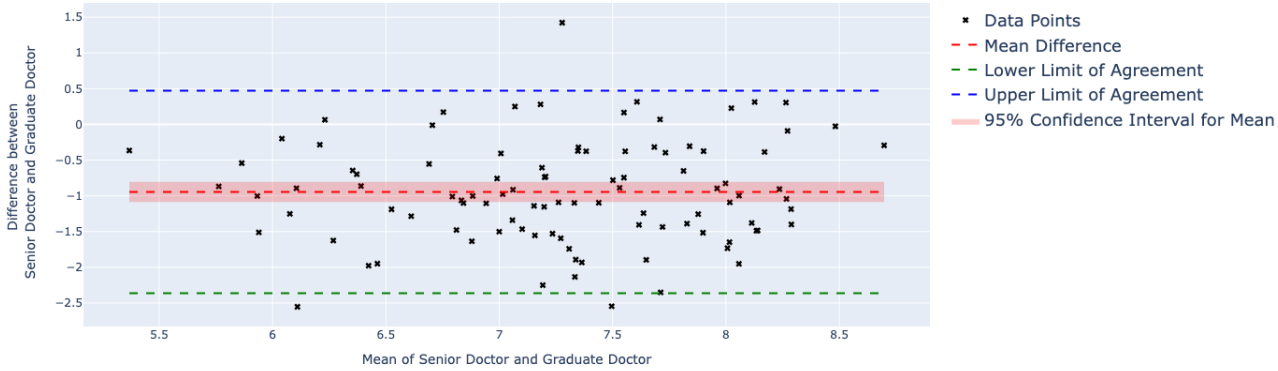


Figure 10. Inter-observer variability (senior doctor vs. graduate doctor).

The subjective assessment of the proposed model's caption generation, i.e., dental condition detection, resulted in an overall score of 8.1/10, while report generation capacity resulted in an overall weighted average rating of 7.5/10 as presented in Table 10. The average rating results for the defined evaluation criteria, as assessed by dental professionals, are illustrated in Figure 12.

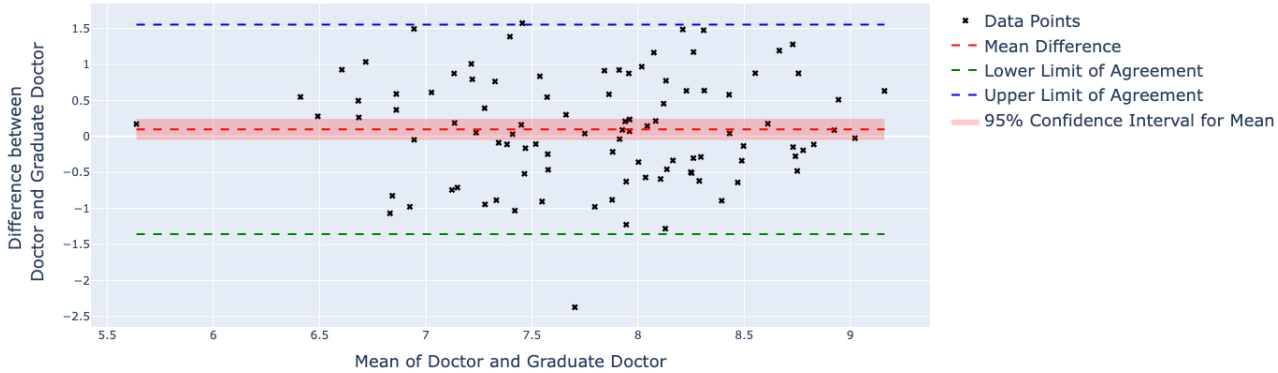


Figure 11. Inter-observer variability (doctor vs. graduate doctor).

Figures A4 and A6 present two AI-generated radiology reports with relatively high and low objective scores, respectively. It is evident that the model accurately identifies key findings, such as grossly carious teeth, caries, pulp calcification, and mild generalized bone loss, demonstrating strong diagnostic accuracy. The detailed descriptions and recommendations are well-aligned with the identified findings, thus offering clear actionable steps for further evaluation by dental professionals. Building upon this, radiologists and dental professionals can utilize this system-generated report as a draft and improve on it to pro-

duce a detailed report if required. Hence, the proposed intelligent clinical decision support system (ICDSS) will act as a supportive tool and significantly contribute to reducing the workload of dental professionals.

Table 10. Qualitative evaluation results for radiology report evaluation.

Category of Dental Professionals	Given Weight	Avg Score for Generated Captions	Avg Score for Generated Radiology Reports
Senior Doctor	5	7.9	7.1
Doctor	3	8.4	7.9
Graduate Doctor	2	7.9	7.8
Weighted Averages		8.1	7.5

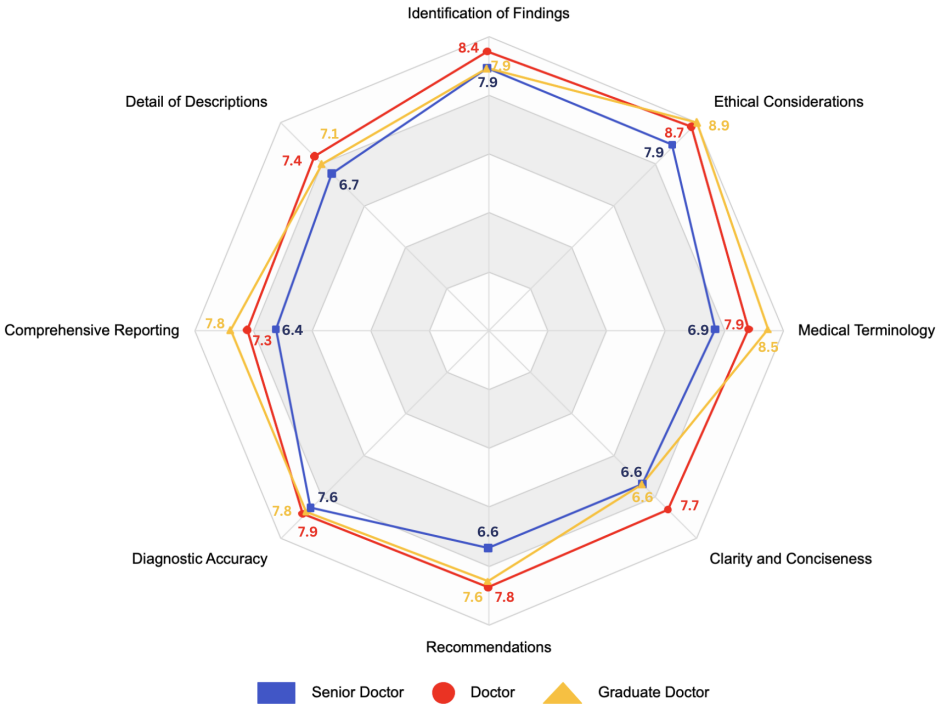


Figure 12. Radial graph showing the average rating from each evaluator.

4.4. Sinhala Q&A in the Dental Domain

Figure 13 exhibits the text prompts generated by Gemma-7b, Mistral-7b, and Llama 3 8B for the dental query in Sinhala, meaning "What are the remedies for tooth sensitivity?". Before the domain adaptation of the Llama 3 8B model, the generated Sinhala response lacked accuracy and context. Moreover, the other popular LLM models tested—namely Gemma-7B, ChatGPT, and Mistral-7B—also failed to provide a comprehensive response to the input query. However, after the domain adaptation, the model produced a precise and context-aware response in Sinhala, translated into English as "Use appropriate toothpaste, minimize consumption of high and very low-temperature foods, and seek dental treatment for tooth decay". This illustrates that the model was effectively fine-tuned through the techniques presented in Section 3.3.

Model Name	Question	Answer
Gemma 7b	දන්ත සංවේදීතාව සඳහා කල යුත්තේ මොනවද?	දන්ත සංවේදීතාව පිළිබඳ ආරම්භ කළ දැනට ජීවිත අධ්‍යාපන හා සමාගම ඉඩ දෙන ආකාර විද්‍යාව තොරතුරු ඇත. දන්ත සංවේදීතාව එකතු පිළිබඳ උපායෝග හා පෞඩ්ඛ විද්‍යාත්මක තාක්ශණික තොරතුරු දෙක වූ අතිරේක ප්‍රදේශ ප්‍රතිපත්ති ලක් දෙයි.
Mistral 7b	දන්ත සංවේදීතාව සඳහා කල යුත්තේ මොනවද?	දන්ත සංවේදීතාව එකට පිරිසක් ඇති ආරක්ෂක තොරතුරු වෙත මොනවද අතර, උපම ඉරැඩ කිරීම, ඉතිhashes, විණෝ ජීවනය, බොද්ධ විද්‍යාත්මක පරිගණන A/B පරීක්ෂණ, දන්ත බොද්ධ ක්‍රම පිළිබඳ විවෘත තොරතුරු එන්නේ.
ChatGPT 3.5	දන්ත සංවේදීතාව සඳහා කල යුත්තේ මොනවද?	ඔබගේ දන්ත සම්පූර්ණ කිරීම සඳහා කෙසේදැයි ඔබට විශ්වාස සම්පූර්ණ කළ යුතු වේවිය යැයි මොනවද? කෙසේ අදාළ සංවේදී මොනවද? මේ පිළිබඳ නිතර සැකසුම් මෙහි මාසිකව නැතිවන බවත්, කලුන් එක්ස්පෝර් හෝ බිල්පත් අඩවියේ දිස්වෙනුත් පුරා මෙහි සකස් කරන්න.
Llama-3-8b before training	දන්ත සංවේදීතාව සඳහා කල යුත්තේ මොනවද?	පාරිභෝගිකයන්ගේ අපේක්ෂා ඉටු කිරීම පාරිභෝගික සේවාව උසස් කිරීම පාරිභෝගිකයන්ගේ ප්‍රතිචාර වැඩි කිරීම පාරිභෝගිකයන්ට අවශ්‍ය තොරතුරු සපයා දීම මෙම ක්‍රියාමාර්ග අනුගමනය කිරීමෙන් පාරිභෝගික සංවේදීතාව වැඩි කළ හැක.
Llama-3-8b after training	දන්ත සංවේදීතාව සඳහා කල යුත්තේ මොනවද?	**සුදුසු දන්නාලේප භාවිත කිරීම, අධික හා ඉතා අඩු උෂ්ණත්ව සහිත ආහාර භාවිතය අවම කිරීම සහ දත් දිරාපත්වීම සඳහා දන්ත ප්‍රතිකාර ලබා ගැනීම.**

Figure 13. Generated results for each LLM for a given Sinhala query. English Translation see Table 11.

Considering that one of our research goals is to explore effective methods of applying automation to dental healthcare for the non-English-speaking community, we focused on evaluating the model's comprehension of dental domain inquiries in the Sinhala language. We evaluated the fine-tuned Llama-3-8B model's question-answering capabilities in the dental application domain by using the DDSLU benchmark dataset. Before domain adaptation, Llama 3 8B accurately answered 624 questions out of 1000, resulting in a correct answering rate of 62.4%. Figures 14 and 15 present the training loss curves of the continued pre-training and the instruct fine-tuning of the Llama 3 8B model, respectively. After the domain adaptation, the model accurately answered 741 questions, resulting in an accuracy improvement of 11.7%. Upon qualitative evaluation by dental professionals, the weighted average rating for the Sinhala language Q&A session stood at a value of 6.1/10. Table 12 presents the qualitative evaluation results for the fine-tuned model for question-answering in the Sinhala language.

Moreover, Figure A1 showcases the final output of the user interface designed for the proposed application in a mobile environment.

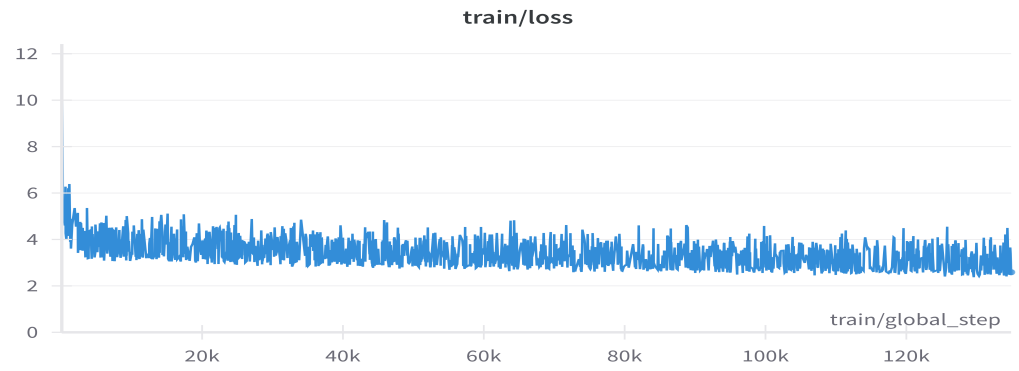


Figure 14. Training loss of the continued pre-training of Llama 3 8B for Sinhala language adoption.

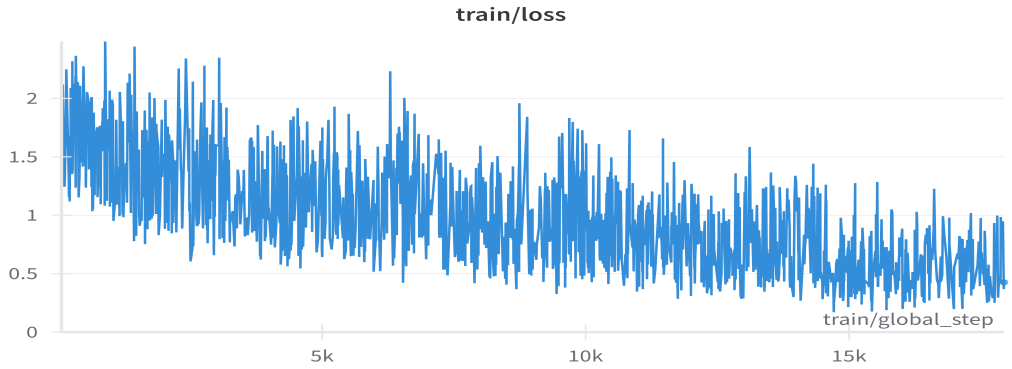


Figure 15. Training loss of instruct fine-tuning of Llama 3 8B for Sinhala language adoption.

Table 11. English Translation of the Sinhala Language Q&A Illustrated in Figure 13.

Model Name	Question	Answer
Gemma 7b	What should be done for tooth sensitivity?	There is a wealth of information about life sciences and the company that allows for basic knowledge of tooth sensitivity. Additional areas of interest include the application of dental sensitivity and the application of scientific technology to dental sensitivity.
Mistral 7b	What should be done for tooth sensitivity?	What are the security information available for sensitive teeth, and open information on image erasure, hashes, Chinese life, scientific A/B testing, and data mining methods.
ChatGPT 3.5	What should be done for tooth sensitivity?	How do you know if you need to complete your data? What are the relevant sensitive data? To avoid this, please set this up here, even if it appears on the Kalun Espor or Billp site.
Llama 3 8b before training	What should be done for tooth sensitivity?	Meeting customer expectations, improving customer service, increasing customer responsiveness, and providing the customer with the information they need can increase customer awareness by adopting these measures.
Llama 3 8b After training	What should be done for tooth sensitivity?	Use appropriate toothpaste, minimize the consumption of foods with high and very low temperatures, and seek dental treatment for tooth decay.

Table 12. Qualitative evaluation results of the domain-adapted Llama 3 8B model for question-answering in the Sinhala language.

Dental Professional Category	Given Weight	Average Rating	Average Weighted Rating
Senior Doctor	5	5.6	28.0
Doctor	3	6.4	19.2
Graduate Doctor	2	6.7	13.4
Weighted Avg	-	-	6.1

5. Discussion

The primary objective of this research was to develop a robust end-to-end system for dental radiology report generation and a Sinhala language question-and-answer platform with discussion capabilities. This research specifically focuses on addressing the needs of non-English-speaking communities in developing countries and presents a complete prototype of a dentistry application that can be used in their native language. Within our scope, we have explored the domain adaptability of LLMs for a specific task. While our immediate focus was on the Sinhala language, the methodology introduced can be extended

to any other language with a similar application scenario. The size and quality of the dataset emerged as the critical factor influencing the model's capability of language understanding. Our findings suggest that a corpus of at least a few billion tokens is required for optimal language comprehension. Moreover, it is important to emphasize that the LLaMA 3 8B model is a critical component of our system, and the reported performance is specific to this model. While we have demonstrated its effectiveness in our framework, results may vary with other LLMs. Larger models or those fine-tuned with more extensive datasets may achieve better performance, particularly in handling complex tasks or challenging languages with limited resources.

The computational and operational hardware costs are a significant bottleneck we experienced during this research. The cost breakdown, as detailed in Table 13, highlights the resource allocation for both the LLaMA-3 8B and BLIP-2 models during different stages of training and fine-tuning [56]. The cost of training and fine-tuning the models was approximately $1000, a one-time expense. For deployment, the system requires two ml.g5.xlarge AWS instances, costing approximately $2.424 per hour if hosted in the cloud. However, these costs can be significantly reduced through serverless architectures, which dynamically allocate compute resources based on actual demand, eliminating the need for pre-allocated server capacity. While serverless deployment may introduce slight latency for initial requests, it offers substantial cost savings in return. Additionally, on-premise deployment remains a viable alternative, ensuring data privacy, regulatory compliance, and long-term cost savings. However, both the training and deployment costs are well justified by the clinical and operational advantages the system offers to dental healthcare professionals and patients. Additionally, our work resulted in an emission of 16.8 kg of CO_2. To minimize these expenses as well as to overcome hardware limitations while maintaining optimum results, we incorporated techniques such as quantization, QLoRA, and the usage of freely available GPU resources for specific tasks.

Table 13. Cost breakdown.

Stage	Resource	Model	Cost per Hour	Time Duration	Estimated Cost
Continued pre-training	ml.g5.12xlarge	Llama 3 8B	$5.672	~168 h (7 Days)	~$953
Instruct fine-tuning	ml.g5.2xlarge	Llama 3 8B	$1.212	~48 h (2 Days)	~$58
Fine-tuning	ml.g5.xlarge	Blip-2	$1.006	~12 h	~$12
Deployment (Llama 3 8B)	ml.g5.xlarge	Llama 3 8B	$1.006	-	-
Deployment (Blip-2)	ml.g5.xlarge	Blip-2	$1.006	-	-

Limited resources and the lack of availability of qualified professionals in developing countries for healthcare sector-related tasks often slow down the diagnostic process. We believe that our proposed system can be practically implemented and deployed to be used in the real world, providing valuable assistance to medical professionals as well as to patients. The loss curves of the Sinhala-adopted Llama 3 8B model, as presented in Figures 14 and 15, have shown that the proposed system can be further improved by training for longer iterations with a larger dataset. Additionally, the study presents a cost-effective approach for domain adaptation of these models. However, it is worthwhile to investigate the impact of reinforcement learning with human feedback, specifically with direct policy optimization (DPO), for further improving the efficacy of LLMs.

6. Limitations

In our research, while demonstrating the potential of multimodal AI and large language models in generating radiology reports from dental panoramic tomography (DPT) images, we noted several shortcomings. Lower model performance was observed for com-

plex pathologies such as cysts, tumors, fibro-osseous lesions, and their exact anatomical sites. This was due to the limited availability of data for these rare and complex conditions, as well as the level of granularity and complexity of the annotations.

It should be emphasized that the primary goal of the current study is to demonstrate a prototype that has a comparatively high level of granularity in identifying general dental problems. In this effort, we also look at the prospect of automating medical reporting, a routine yet time-consuming process. The suggested model and the present training methodology are adequate to support the potential breadth of this study's outcomes with this goal. However, we firmly believe that in order to further enhance the suggested model's performance, it must be trained on a more robust database.

Also, the proposed model was purposefully excluded from identifying complex annotations, such as the numbering and positioning of teeth based on standard dental notation systems like FDI (ISO) notation, marking of roots and canals, or detailed mapping of bone structures. This is because achieving this level of detail would require an extensive dataset with highly specific and accurate annotations, which was beyond the scope of our current research. Our model currently focuses on the initial stages of radiology report generation, primarily aiding in the identification and categorization of broad conditions. Moreover, the system's report generation accuracy of approximately 81.3% is promising, yet it is necessary to have professional oversight. Thus far, the AI-generated outputs are intended as support tools for medical professionals.

While the LLaMA 3 8B model was selected for this prototype due to its strong performance across multiple benchmarks, relatively low resource consumption, acceptable output quality, and suitability for radiology report generation, a comprehensive evaluation across multiple large language models was not conducted due to the substantial computational costs and resource limitations associated with training and testing such models. Evaluating other models represents a potential direction for future research.

7. Conclusions

In conclusion, this study highlights the potential of multimodal artificial intelligence and large language models to improve access to high-quality dental healthcare in developing countries. We conceptualized and developed an automated radiology report generation system and a dental question-answering (Q&A) platform in the Sinhala language, benefiting both dental professionals and patients. The LLM language adaption procedure described in this study for Sinhala can be readily extended to other languages. The prototype system proposed utilizes a Blip-2 model for token generation, followed by a continued pre-trained and instructed fine-tuned Llama 3 8B model for report generation and content generation for the Sinhala Q&A session. Our approach highlights a cost-effective and efficient training methodology for the domain adaptation of LLMs. Data augmentation techniques, particularly grid distortion and elastic transformation, were also employed to address the shortcomings related to the scarcity of dental data. All things considered, these techniques enhanced the model training and raised the diagnostic accuracy. For instance, the overall system architecture resulted in an accuracy of 81.3% for radiology report generation.

Moreover, the models can be further improved by using DPO after collecting human feedback over time. By incorporating feedback from medical and dental professionals, the model can be refined to enhance its accuracy, relevance, and usability, ultimately leading to better support for healthcare providers and patients. Furthermore, as a standalone system, our approach also ensures the protection of patients' privacy, which is a critical concern in healthcare, as the proposed model is able to deploy independently at standalone devices.

Finally, the authors hope to improve this work's generalizability in subsequent research. This will ultimately improve patient care and clinical outcomes by strengthening the system's resilience and applicability across a larger range of medical settings and diseases.

Author Contributions: Conceptualization: C.D. and K.D.; methodology design: C.D., K.D. and M.B.D.; data curation: C.G., R.J., C.D. and K.D.; Investigation, C.D., K.D. and M.B.D.; Resources, C.D., K.D., C.G. and R.J.; Visualization, C.D. and K.D.; validation: C.D., K.D., M.B.D., C.G. and R.J.; writing—original draft preparation: C.D., K.D. and M.B.D.; writing—review and editing: C.D., K.D., M.B.D. and R.J.; supervision and project administration: M.B.D., Funding Acquisition, M.B.D. and R.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical clearance was obtained under the guidelines and regulations set by the respective hospital ethics committees and international ethical standards.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The complete codebase used for this research is available at https://github.com/ChirathD/Radiology-Report-Generation (accessed on 13 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

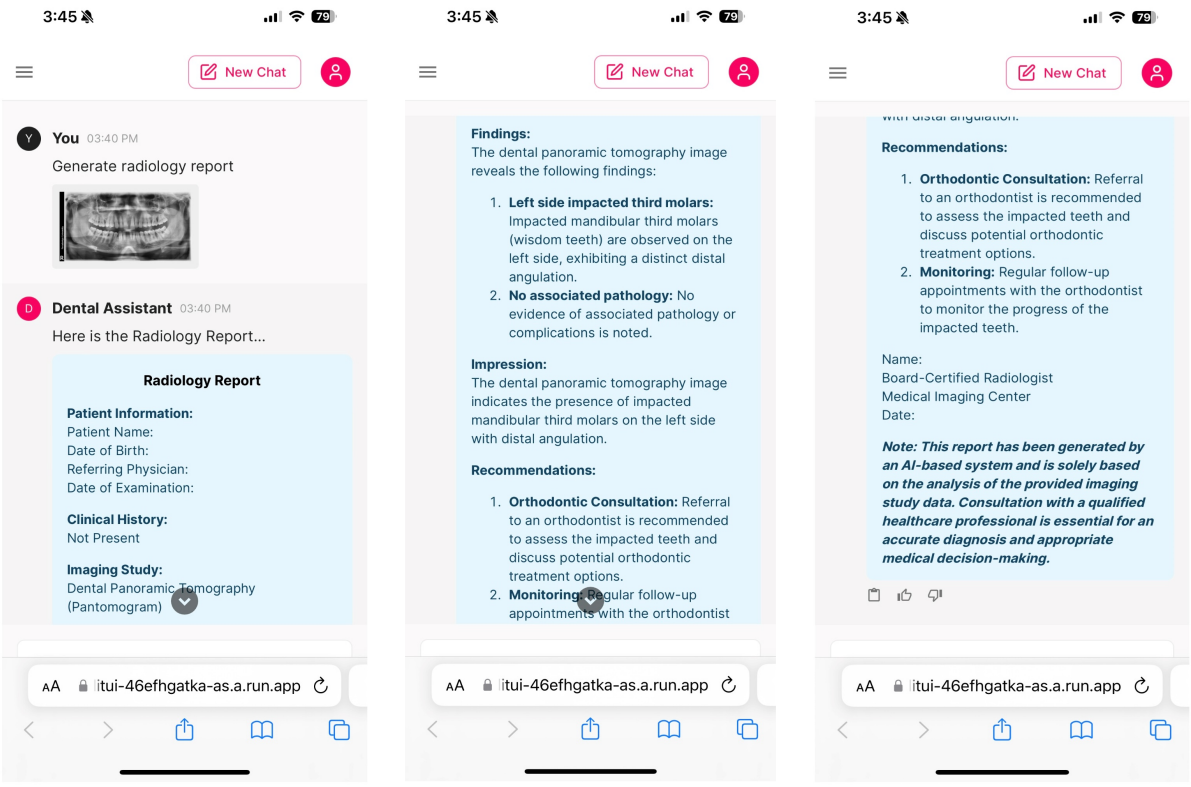


Figure A1. The user interface of the radiology report generation application.

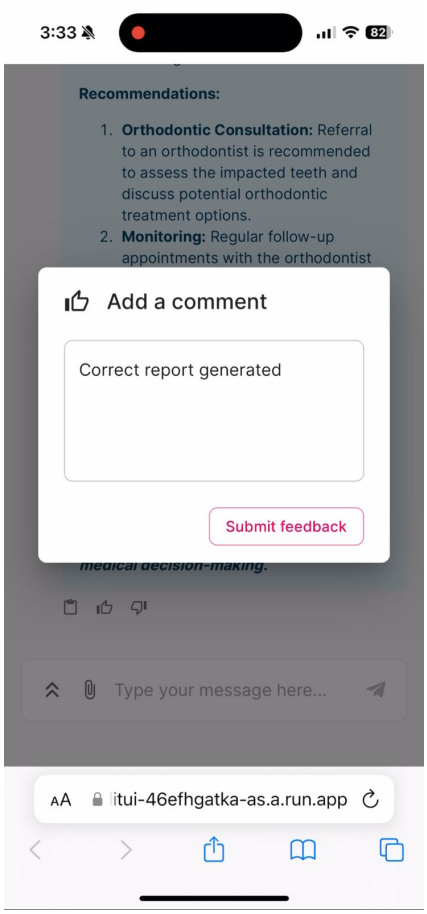


Figure A2. Feedback collection prompt for RLHF.

Appendix B



Figure A3. Sample DPT image 01.

Patient Information: Patient Name:
Date of Birth:
Referring Physician:
Date of Examination:

Clinical History: Not Present

Imaging Study: Dental Panoramic Tomography (Pantomogram)

Findings: The dental panoramic tomography image reveals the following findings:

1. **Grossly Carious Teeth:** Multiple teeth exhibit extensive carious lesions, indicating severe dental decay.
2. **Caries:** Additional carious lesions are present, affecting various teeth.
3. **Pulp Calcification:** Evidence of pulp calcification is observed, which may indicate chronic pulpitis or aging.
4. **Mild Generalized Bone Loss:** There is a mild level of generalized bone loss in both the maxilla and mandible, potentially indicative of early periodontal disease.

Impression: The dental panoramic tomography image indicates the presence of grossly carious teeth, multiple carious lesions, pulp calcification, and mild generalized bone loss, all of which require further dental evaluation and management.

Recommendations:

1. **Comprehensive Dental Examination:** A thorough dental examination is recommended to assess the extent of carious lesions and plan appropriate restorative treatment, including addressing grossly carious teeth.
2. **Endodontic Consultation:** Consider consultation with an endodontist to evaluate the teeth with pulp calcification and determine the need for potential endodontic therapy.
3. **Periodontal Assessment:** Referral to a periodontist for evaluation and management of the observed generalized bone loss.

Name:
Board-Certified Radiologist
Medical Imaging Center
Date:

Note: This report has been generated by an AI-based system and is solely based on the analysis of the provided imaging study data. Consultation with a qualified healthcare professional is essential for an accurate diagnosis and appropriate medical decision-making.

Figure A4. AI-generated report for DPT image 01. This report received a high subjective score for identifying oral conditions in DPT Image 01.



Figure A5. Sample DPT image 02.

Patient Information: Patient Name: Date of Birth: Referring Physician: Date of Examination:

Clinical History: Not Present

Imaging Study: Dental Panoramic Tomography (Pantomogram)

Findings: The dental panoramic tomography image reveals the following findings:

1. **Generalized Bone Loss:** Evidence of generalized bone loss affecting both the maxilla and mandible, with severe involvement in the anterior regions of the upper and lower jaws. This suggests advanced periodontal disease.
2. **Periapical Lesions:** Periapical lesions are noted in several teeth, indicating potential chronic infection or inflammation at the tooth roots, which may be secondary to long-standing dental conditions.
3. **Severe Mesial Bone Loss / Deep Pocket:** A specific region demonstrates severe mesial bone loss, accompanied by a deep periodontal pocket, suggesting advanced localized periodontal disease that may require surgical intervention.

Impression: The panoramic image reveals significant periodontal disease, characterized by generalized bone loss, periapical lesions, and localized severe mesial bone loss with a deep pocket. Immediate dental and periodontal evaluation is recommended.

Recommendations:

1. **Comprehensive Dental Examination:** A full dental evaluation is advised to assess the extent of bone loss and periapical pathology, and to determine appropriate treatment, including endodontic or surgical options.
2. **Endodontic Consultation:** Referral to an endodontist is recommended to evaluate the teeth with periapical lesions for potential root canal therapy or extraction.
3. **Periodontal Treatment:** Referral to a periodontist for further management of the generalized and localized periodontal issues, including consideration of surgical intervention for deep periodontal pockets.

Name: Board-Certified Radiologist Medical Imaging Center Date:

Note: This report has been generated by an AI-based system and is solely based on the analysis of the provided imaging study data. Consultation with a qualified healthcare professional is essential for an accurate diagnosis and appropriate medical decision-making.

Figure A6. AI-generated report for DPT image 02. This report received a low subjective score for identifying oral conditions in DPT image 02.

References

1. Shaban-Nejad, A.; Michalowski, M.; Bianco, S. (Eds.) Multimodal Artificial Intelligence: Next Wave of Innovation in Healthcare and Medicine. In *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*; Springer International Publishing: Cham, Switzerland, 2023; pp. 1–9. [CrossRef]
2. Geantă, M.; Bădescu, D.; Chirca, N.; Nechita, O.C.; Radu, C.G.; Rascu, S.; Rădăvoi, D.; Sima, C.; Toma, C.; Jinga, V. The Potential Impact of Large Language Models on Doctor–Patient Communication: A Case Study in Prostate Cancer. *Healthcare* **2024**, *12*, 1548. [CrossRef] [PubMed]
3. Huang, H.; Zheng, O.; Wang, D.; Yin, J.; Wang, Z.; Ding, S.; Yin, H.; Xu, C.; Yang, R.; Zheng, Q.; et al. ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model. *Int. J. Oral Sci.* **2023**, *15*, 29. [CrossRef] [PubMed]
4. Izzetti, R.; Nisi, M.; Aringhieri, G.; Crocetti, L.; Graziani, F.; Nardi, C. Basic Knowledge and New Advances in Panoramic Radiography Imaging Techniques: A Narrative Review on What Dentists and Radiologists Should Know. *Appl. Sci.* **2021**, *11*, 7858. [CrossRef]
5. Dasanayaka, C.; Dharmasena, B.; Bandara, W.R.; Dissanayake, M.B.; Jayasinghe, R. Segmentation of Mental Foramen in Dental Panoramic Tomography using Deep Learning. In Proceedings of the 2019 14th Conference on Industrial and Information Systems (ICIIS), Kandy, Sri Lanka, 18–20 December 2019; pp. 81–84. [CrossRef]
6. Turosz, N.; Chęcińska, K.; Chęciński, M.; Brzozowska, A.; Nowak, Z.; Sikora, M. Applications of artificial intelligence in the analysis of dental panoramic radiographs: An overview of systematic reviews. *Dentomaxillofac. Radiol.* **2023**, *52*, 20230284. [CrossRef]
7. Wickramasinghe, W.M.S.P.B.; Dissanayake, M.B. Vision transformers for glioma classification using T1 magnetic resonance imaging. *Artif. Intell. Health* **2024**, *2*, 68–80. [CrossRef]

8. Kaldera, H.N.T.K.; Gunasekara, S.R.; Dissanayake, M.B. MRI based Glioma segmentation using Deep Learning algorithms. In Proceedings of the 2019 International Research Conference on Smart Computing and Systems Engineering (SCSE), Colombo, Sri Lanka, 28 March 2019; pp. 51–56. [CrossRef]
9. Gunasekara, S.R.; Kaldera, H.N.T.K.; Dissanayake, M.B. A systematic approach for MRI brain tumor localization and segmentation using deep learning and active contouring. *Dournal Healthc. Eng.* **2021**, *2021*, 6695108. [CrossRef]
10. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef]
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I.; Kaiser, Ł. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
12. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
13. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 24824–24837.
14. Liu, H.; Cheng, M.; Wei, J.; Cao, Y.; Cheng, H.; He, P. Medical LLMs: Capabilities, limitations, and safety guardrails from the perspective of healthcare delivery. *npj Digit. Med.* **2023**, *6*, 209. [CrossRef]
15. Zhang, J.; Huang, J.; Jin, S.; Lu, S. Vision-Language Models for Vision Tasks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5625–5644. [CrossRef]
16. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
17. Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned Language Models are Zero-Shot Learners. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
18. Bonfigli, A.; Bacco, L.; Merone, M.; Dell'Orletta, F. From pre-training to fine-tuning: An in-depth analysis of Large Language Models in the biomedical domain. *Artif. Intell. Med.* **2024**, *157*, 103003. [CrossRef] [PubMed]
19. Ong, J.C.L.; Chang, S.Y.; William, W.; Butte, A.J.; Shah, N.H.; Chew, L.S.T.; Liu, N.; Doshi-Velez, F.; Lu, W.; Savulescu, J.; et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit. Health* **2024**, *6*, e428–e432. [CrossRef] [PubMed]
20. Park, J.; Oh, K.; Han, K.; Lee, Y.H. Patient-centered radiology reports with generative artificial intelligence: Adding value to radiology reporting. *Sci. Rep.* **2024**, *14*, 13218. [CrossRef]
21. Nassiri, K.; Akhloufi, M.A. Recent Advances in Large Language Models for Healthcare. *BioMedInformatics* **2024**, *4*, 1097–1143. [CrossRef]
22. Kassner, N.; Dufter, P.; Schütze, H. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. *arXiv* **2021**, arXiv:2102.00894.
23. Tran, H.; Yang, Z.; Yao, Z.; Yu, H. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *arXiv* **2023**, arXiv:2310.19975. [CrossRef]
24. Thawkar, O.; Shaker, A.; Mullappilly, S.; Cholakkal, H.; Anwer, R.; Khan, S.; Laaksonen, J.; Khan, F. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv* **2023**, arXiv:2306.07971.
25. Wang, Z.; Wu, Z.; Agarwal, D.; Sun, J. Medclip: Contrastive learning from unpaired medical images and text. *arXiv* **2022**, arXiv:2210.10163.
26. Eslami, S.; Meinel, C.; De Melo, G. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 1181–1193.
27. Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; Xie, W. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv* **2023**, 2023–01.
28. Bazi, Y.; Rahhal, M.M.A.; Bashmal, L.; Zuair, M. Vision–language model for visual question answering in medical imagery. *Bioengineering* **2023**, *10*, 380. [CrossRef]
29. Chen, Z.; Shen, Y.; Song, Y.; Wan, X. Cross-modal memory networks for radiology report generation. *arXiv* **2022**, arXiv:2204.13258.
30. Jain, N.; Dutt, U.; Radenkov, I.; Jain, S. WHO's Global Oral Health Status Report 2022: Actions, Discussion and Implementation. *Oral Dis.* **2024**, *30*, 73–79. [CrossRef] [PubMed]
31. Dasanayaka, C. DPT Image Caption Dataset. Available online: https://huggingface.co/datasets/LexiconShiftInnovations/DPT_Image_Caption_Dataset (accessed on 15 May 2024).
32. Dasanayaka, C. DPT Caption Dataset Version 2. Available online: https://huggingface.co/datasets/ChirathD/dpt-caption-dataset-version-2/settings (accessed on 22 July 2023).
33. LexiconShift_Innovations. SinhalaCorpusLarge (Revision 881ce28). Available online: https://huggingface.co/datasets/LexiconShiftInnovations/SinhalaCorpusLarge (accessed on 20 January 2024).

34. LexiconShift_Innovations. Dental QnA Instruct. Available online: https://huggingface.co/datasets/LexiconShiftInnovations/Dental_QnA_Instruct (accessed on 20 February 2024).
35. Parliament Of Shri Lanka. Personal Data Protection Act, No. 9 of 2022. Available online: https://www.parliament.lk/uploads/acts/gbills/english/6242.pdf (accessed on 22 January 2024).
36. AI@Meta. Llama 3 Model Card. Available online: https://huggingface.co/meta-llama/Meta-Llama-3-8B (accessed on 13 March 2025).
37. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 53728–53741.
38. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023), Honolulu, HI, USA, 23–29 July 2023*; Vuppala, A.K., Ed.; Proceedings of Machine Learning Research; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 19730–19742.
39. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. [CrossRef]
40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **2024**, *25*, 1–53.
42. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
43. Hugging Face. Hugging Face Open LLM Leaderboard. Available online: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (accessed on 20 April 2024).
44. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 10088–10115.
45. Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; Raffel, C.A. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1950–1965.
46. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
47. Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Goffinet, É.; Hesslow, D.; Launay, J.; Malartic, Q.; et al. The falcon series of open language models. *arXiv* **2023**, arXiv:2311.16867.
48. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; et al. Gemma: Open models based on gemini research and technology. *arXiv* **2024**, arXiv:2403.08295.
49. Xu, L.; Xie, H.; Qin, S.Z.J.; Tao, X.; Wang, F.L. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv* **2023**, arXiv:2312.12148.
50. Jelinek, F.; Mercer, R.L.; Bahl, L.R.; Baker, J.K. Perplexity—A measure of the difficulty of speech recognition tasks. *J. Acoust. Soc. Am.* **1977**, *62*, S63. [CrossRef]
51. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
52. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
53. Hugging Face. DDSLU. Available online: https://huggingface.co/datasets/LexiconShiftInnovations/DDSLU_Benchmark (accessed on 1 March 2025).
54. Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; Steinhardt, J. Measuring massive multitask language understanding. *arXiv* **2020**, arXiv:2009.03300.
55. Altman, D.G.; Bland, J.M. Measurement in medicine: The analysis of method comparison studies. *J. R. Stat. Soc. Ser. D Stat.* **1983**, *32*, 307–317. [CrossRef]
56. Amazon Web Services. Amazon EC2 G5 Instances. Available online: https://aws.amazon.com/ec2/instance-types/g5/ (accessed on 1 March 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.