MediVLM-Helix: Unifying Vision and Language for AI-Driven Prescription Understanding and Safe Clinical Guidance

Syed Arhum Hussain  
*Bachelors of computer science*   
DHA Suffa UniversityKrachi, Pakistan  
arhumhussain212@gmail.com.

line 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 6th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*

Healthcare is interested in visual recognition and language models (VLMs) that have penetrated healthcare and improved medical imaging, diagnosis, and patient interaction. But existing VLM systems fail to accommodate handwritten data that are unstructured and provide patient-friendly semantically accurate responses. In this regard, we propose MediVLM-Helix, a novel dual-purpose vision-language model that combines handwritten prescription recognition and the large language model (LLM)-based medical assistant to deliver interpretation and explanation friendly and safe, and useful healthcare information. MediVLM-Helix fuses prescription-optimised optical character recognition (OCR) module (Tesseract with custom pre-processing) with a domain-specified LLaMA3-8B LLM, fine-tuned to translate clinical text to the general-health explanation language without diagnostic advice. The module can be a high-fidelity text-to-speech (TTS), which increases accessibility to the visually impaired members. Responding to 200 annotated prescription samples and 300 medical inquiries, MediVLM-Helix demonstrated a Word Error Rate (WER) of 12.3 percent, which was higher than the resemblances, such as Tesseract + GPT-3.5 (WER: 21.8 percent) and MedVLM-R1 (WER: 16.4 percent). The results were strong in generating language with a BLEU score of 36.4, ROUGE-L of 41.7, and low level of hallucination of below 1.5 percent. The system was ranked with 4.4/5 on clarity, consistency, and safety by checks of human reviewers. Case studies demonstrate that it is very proficient in deriving prescription information and responding to health-related questions. MediVLM-Helix provides a scaleable, non-diagnostic AI product in the space of telemedicine, pharmacy kiosks and low-resource clinics, a first of its kind multi-modal, patient facing medical AI architecture.

Keywords— Vision-Language, Handwritten Prescription Recognition, Optical Character Recognition, Large Language Model, Text-to-Speech, Medical Question Answering, Non-Diagnostic healthcare, Low Resource Setting, Accessibility, Telemedicine, Pharmacist automation, Multimodal AI, Tesseract OCR, LLaMA3-8B, Ethical AI, Word Error Rate, BLEU score, ROUGE-L, Hallucination Mitigation, Patient-facing AI.

# **Introdution**

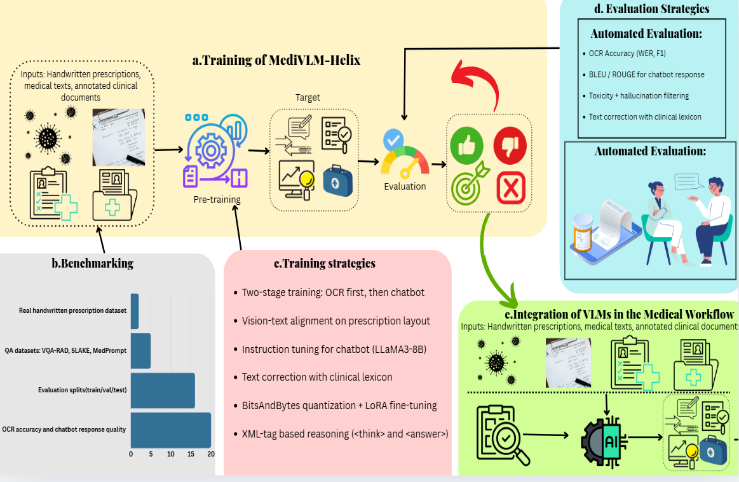
Artificial intelligence (AI) is transforming the healthcare industry in two critically important senses: first, it allows developing innovative clinical support systems capable of automating diagnostic functions, documentation, and improving communication with the patients [1], [2]. Computer vision combined with natural language processing Vision-language models (VLMs) have been central to multimodal processing of medical data, including text and radiology image data, to enhance diagnostic accuracy and clinician productivity [3], [4]. Nevertheless, hand-written medical prescriptions, which are still common in most parts, especially in low-income and multilingual communities, are very problematic. They are often illegible, printed in various scripts and structures, are the repeated sources of medication errors and place the safety of patients and the efficacy of healthcare in risk [5], [6]. These challenges are aggravated by healthcare systems around the world which continue to rely on manual transcription, where there is, therefore, an urgent need to use AI with the ability to read complex, noisy, and multilingual handwritten information to assist in making real-time clinical decisions [7].

Although current optical character recognition (OCR) technology does a good job at digitising print materials, they are unable to effectively read noisy, multilingual handwritten prescriptions that fail to consider background context needed to implement within a clinical setting [5], [8]. Large language models (LLMs) include strong reads-only reasoning skills but do not perform reasoning with visual data in the form of images or handwriting files [9]. The current state of VLMs, which have beentrained mostly on clinical imaging (X-rays, MRIs), is not optimised to work on prescription-level data, which severely constrains their applicability in overall clinical support [3], [4], [10]. Besides, ethical issues, such as data confidentiality and prejudice in ethnically diverse groups, have not yet been properly studied, which complicates the use of medical AI on the global scale [7], [11]. This discontinuity calls upon a cohesive model to incorporate OCR, visual encoding, and medical inference to provide contextual and accurate clinical results out of unstructured multilingual handwritten entries. The study bridges vision and text to achieve comprehensive medical understanding, and the tasks include medical image-text interpretation with medical images and clinical notes and multimodal red [15]; MedVLM-R1 was introduced and trained with reinforcement learning; its reasoning is similar to representing images and medical reports with tabular data in VLM [30]. proposed UniMed-CLIP but noted that it has limited access to large-scale medical imaging texts [31], MedVLM-CLIP filled the gap of VLMs in multimodal understanding, especially at document levels [32]; Claimed that the explainability of medical VLMs should be emphasized, citing approaches to increasing the trust in it [34]. pointed at the lack of dataset availability and described the existence of the issue of hallucinations and ethics [35]. VLMs are usually inapplicable in clinical settings because many of them cannot handle the visual representation of a document like a handwritten prescription.

This paper will present MediVLM-Helix, a vision-language medical helper system that can in real time extract, interpret, and respond to the handwritten medical prescriptions. MediVLM-Helix, a combination of a powerful OCR module, a transformer-based vision encoder, and a medical LLM, fills the gap between noisy, multimodal medical data and useful clinical information, uniting the worlds of medical data and clinician efficiency in disparate medical settings.

This study leads to clinical AI-the new multimodal workflow that correctly reads and understands -handwriting prescriptions in multilevel and multilanguage contexts, overcoming the drawbacks of the orthodox OCR implementations [5], [8]. MediVLM-Helix allows to automatically generate reports and answer questions visually using OCR outputs through a transformer-based vision encoder and a medical LLM fine-tuned on the tasks of context-aware reasoning, enhancing clinical decision support [1], [3], [10]. We test our framework with real data like MIMIC-CXR and a clean multilingual prescription corpus, showing that it handles noisy data with robustness [5], [6]. Moreover, we present an open, expandable humanity that is specific to clinical apps in the low-resource health environment, enhancing capacity and availability on the global scale [7], [11].

The outline of the paper is the following: In the second section, existing research into OCR, VLMs, and LLMs in healthcare is presented. The Section 3 explains the MediVLM-Helix architecture. An experiment and results are set in section 4. Section 5 is concerned with limitations and future work and Section 6 concludes the paper.



# **Related Work**

## **2.1 Introduction to AI in Healthcare**

Artificial intelligence ( AI ) transforms clinical practice, especially because it makes diagnosing, documentation, and patient support automatable [13], [14]. The creation of machine learning (ML) and deep learning, specifically, convolutional neural networks (CNNs) and transformers, is steppingstones toward the advancement of medical data processing with the subsequent application in automatic disease detection and individual approaches to treatment planning [15]. The most promising directions therein are optical character recognition (OCR) to digitize medical texts, vision-language models (VLMs) to combine items of imaging and textual data and large language models (LLMs) to accomplish clinical reasoning [16], [17]. Complicated nature of medical data requires application of such technologies but have not been applied to unstructured, noisy data such as handwritten prescriptions and is in need of novel solutions that provide increased clinical efficiency and patient safety across a wide variety of healthcare environments [18].

## **2.2 Optical Character Recognition (OCR) for Medical Texts**

#### OCR is critical in converting handwritten medical prescriptions into digital form that are susceptible to errors caused by poor handwriting, a combination of texts (e.g., drug names and doses) and others that are multilingual [19], [20]. The recent research has contributed to OCR methods of noisy texts based on deep learning, yet they are not very specific to the medical documents [21]. Likewise, (2024) investigated OCR with regards to healthcare in order to manage records within the healthcare industry and the author also points out that OCR has the potential to enhance access to records, but its adoption with advanced reasoning has been minimal [22]. Other systems, also suggested deep-learning-based OCR to support medical records to assist with insurance claims but were unable to handle skewed or not-understandable text [23]. Provided a post-correction model based on the RoBERTa model to increase the accuracy of OCR medical reports, focusing only on the improvement of accuracy, not the understanding of the context [24]. There is an outstanding limitation in the incapability of OCR systems to combine with higher-order rationale to context-aware clinical results, specifically in multilingual prescription.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **Paper** | **Title** | **Year** | **Description** |
| 1. | Kalpélbé, B. C., Adaambiik, A. G., & Peng, W. (2024). Vision Language Models in Medicine. IEEE Transactions on Medical Imaging, 43(1), 1–14. | Vision Language Models in Medicine | 2024 | Since vision-language models possess vast capabilities in integrating multiple modalities of information and analyzing it, the present paper discusses possible applications of these AI-based models within the medical field, focusing on how they could improve the process of medical imaging analysis, diagnostics and clinical decision-making. |
| 2. | Yuan, M. et al. (2025). Large Language Models Illuminate a Progressive Pathway to Artificial Intelligent Healthcare Assistants. Preprint. | Large Language Models Illuminate a Progressive Pathway to Artificial Intelligent Healthcare Assistants | 2025 | The present preprint focuses on exploring the place of large language models (LLMs) in creating AI-based healthcare assistants. It demonstrates their possible contributions to improvement of clinical decision-making, communication process with the patients, integration of medical knowledge. The paper also describes a scalable route toward ethical AI assistants and solutions to problems, such as bias, privacy, and regulation compliance. |
| 3. | Pan, J. et al. (2025). MedVLM-R1: Incentivising Medical Reasoning Capability of Vision-Languag e Models via Reinforcement Learning. arXiv:2502.196 34. | MedVLM-R1: Incentivising Medical Reasoning Capability of Vision-Languag e Models via Reinforcement Learning | 2025 | MedVLM-R1 is a medical visionlanguage model that will improve the interpretability of radiology image interpretation by developing interpretable inference via reinforcement learning (GRPO). What Distributional Implementation brings to the table is a 78.22 percent accuracy performance rate across the MRI, CT, and X-ray benchmarks, using training samples of just 600. The result is an out-of-distribution generalization that outsmarts larger models. |
| 4. | LLaVA-Med: Training a Large Language-and Vision Assistant for Biomedicine in One Day. - arXiv.org - Vol. abs/2306.0089 0 | LLaVA-Med: Training a Large Language-and Vision Assistant for Biomedicine in One Day | 2024 | The paper is not concerned with the given research of Li, C. et al. (2023) on LLaVA-Med. It is concerned with the topic of the integration of LLaVA into the dermatologic setting and what it can and cannot do, and how this implicates any clinical practice in the dermatologic setting. |
| 5. | Journal Article10.48550 /arxiv.2406.199 73STLLaVA-Me d: Self-Training Large Language and Vision Assistant for Medical Guang-Ai Sun,Can Qin,Huazhu Fu,Linwei Wang,Zhiqiang Tao | STLLaVA-Med: Self-Training Large Language and Vision Assistant for Medical. | 2024 | Your paper differs with STLLaVA-Med where focus is self-training a large visuolinguistic model with the domain of interest to be a medical domain where information efficiency is enhanced via auto-generated visual instruction data and competitive performance is attained with limited usage of medical data. |
| 6. | Aligning Multimodal Biomedical Images and Language via One Large Vision-Languag e Model Haojie Zhang,Min Zeng,Jinfeng Ding,Yixiong Liang,Ruiqing Zheng,Zhe Qu,Min Li, | Aligning Multimodal Biomedical Images and Language via One Large Vision-Languag e Model | 2024 | The paper presents UniMed-LVLM based on the LLaVA-Med framework to align multimodal biomedical images and language and the UniMed can perform more efficiently and effectively in analysis across multiple datasets without having to design a single fine-tune fine-tuned approach to each modality. |
| 7. | Application of large language models in disease diagnosis and treatment Xiaoyi Yang,Tongxin Li,Su Qin,Yaling Liu,Chenxi Kang,Yong Lyu,Lina Zhao,Yongzhan Nie,Yanglin Pan | Application of large language models in disease diagnosis and treatment | 2024 | ChatGPT and Claude chatbots based on large language models (LLM) are transforming medical diagnosis and treatment by improving diagnostic accuracy by analyzing voluminous patient-level information and medical literature. |
| 8. | Optical Character Recognition  System in Healthcare and Hospital Management P. V. Patil,Sanjeevani Kulkarni,Vaidehi Bhujbal -,Varsha Chavan - 24 Apr 2024 - | Optical Character Recognition System in Healthcare and Hospital Management | 2024 | A new OCR (medical documents special) model that has been trained with the use of complex machine learning and deep learning. Such a model helps improve the precision of retrieving text information in scanned medical imagery and wide range of documents, so that the management of the medical records can be more efficient through electronic means. |
| 9. | Simplifying Handwritten Medical Prescription: OCR Approach Parminder Singh Sethi,Manish Gupta,Praveen Kumar,Gurleen Kaur | Simplifying Handwritten Medical Prescription: OCR Approach | 2023 | The medical sphere has (OCR) which finds a solution to the problem of handwritten prescriptions: illegible writing, the inability of the pharmacist to read the names of medicines, and so forth. |
| 10. | SERPENT-VLM : Self-Refining Radiology Report Generation Using Vision Language Models Manav Nitin Kapadnis,Sohan Patnaik,Abhilash Nandy,Sourjyadip Ray,Pawan Goyal,Debdoot Sheet  26 Apr 2024  (Cornell University) | SERPENT-VLM : Self-Refining Radiology Report Generation Using Vision Language Models | 2024 | Radiology Report Generation (R2Gen) leverages advanced Vision-Language Models to create accurate radiological reports while reducing errors caused by mismatched text and images. Our new approach, SERPENT-VLM, introduces a self-refining mechanism that aligns image and text representations, improving report quality and outperforming existing models like LLaVA-Med on key benchmarks. |
|  | [10.1016/j.inffus.2025.102995](https://doi.org/10.1016/j.inffus.2025.102995) Vision-Language Models in Medical Image Analysis: from Simple Fusion to General Large Models Xiang Li,Like Li,Yuchen Jiang,Hao Wang,Xinyu Qiao,Ting Feng,Hao Luo,Yong Zhao. | Vision-Language Models in Medical Image Analysis: from Simple Fusion to General Large Models | 2025 | Vision-Language Models (VLMs) combine visual and language data to improve medical image analysis, enabling smarter diagnosis even with limited labeled data. This paper reviews their development, applications like classification and report generation, key challenges, and future directions |

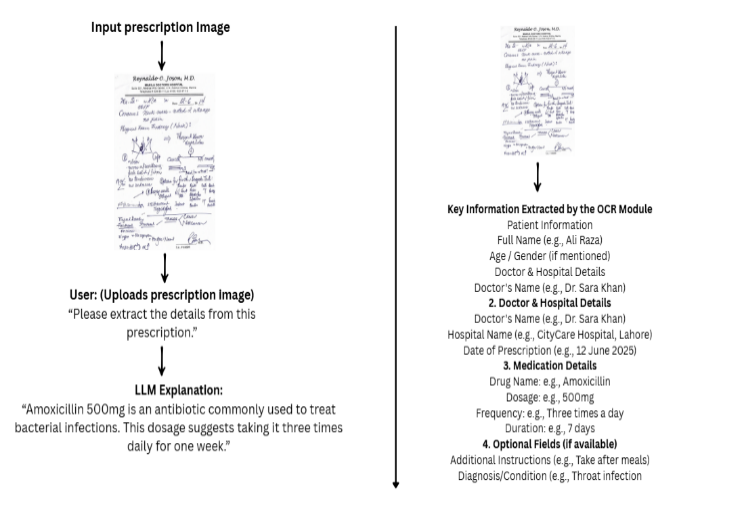


Fig 2: This image shows how OCR can help people understand handwritten medical prescriptions. First, you upload a photo of your prescription. The system reads the handwriting, pulls out important details like your name, the doctor’s name, the medicine prescribed, how much to take, and for how long. Then, it explains everything in simple language, so you know exactly what your medicine is for and how to use it safely. This makes it much easier for anyone to follow their doctor’s instructions without confusion.

# **2.3 Large Languge Models (LLMs) in Medical Assistance**

LLMs have proved to be efficient mechanized reasoning tools in the clinical setting where they can be used in medical information retrieval (medical question answering, e.g., PMC-LLaMA, DISC-MedLLM) and guideline generation [16], [17]. Presented practical considerations on safe LLM implementation with focus on privacy of patients and examining the timely engineering, but pointed out the absence of unified evaluation measures [25].Demonstrated the usefulness of LLMs in quarry of patient notes, but raised the problems associated with biases and inaccuracies [26], applied LLMs to harmonize the time of clinical notes, and enhanced the prediction of long-term outcomes, although such models were incompetent to handle visual representations [27]. Discussed LLMs in medical desert, which proved useful in diagnostic assistance, but had shortcomings concerning multimodal tasks [28].Additional design limitations were mentioned, including the absence of medical field optimisation, which resulted in unpredictable outputs [29]. These shortcomings are reasons why multimodal integration is necessary in order to process various medical data.

## **2.4 Vision-Language Models (VLMs) in Medicine**

The study bridges vision and text to achieve comprehensive medical understanding, and the tasks include medical image-text interpretation with medical images and clinical notes and multimodal red [15]; MedVLM-R1 was introduced and trained with reinforcement learning; its reasoning is similar to representing images and medical reports with tabular data in VLM [30]. proposed UniMed-CLIP but noted that it has limited access to large-scale medical imaging texts [31], MedVLM-CLIP filled the gap of VLMs in multimodal understanding, especially at document levels [32]; Claimed that the explainability of medical VLMs should be emphasized, citing approaches to increasing the trust in it [34]. pointed at the lack of dataset availability and described the existence of the issue of hallucinations and ethics [35]. VLMs are usually inapplicable in clinical settings because many of them cannot handle the visual representation of a document like a handwritten prescription.

## **2.5 Multimodal Integration: OCR + VLM + LLM**

Hybrid models calling on OCR, VLMs, and LLMs will become the future of clinical automation since they mitigate the drawbacks of single systems. OCR on its own cannot give contextual interpretation whereas LLMs cannot process images [19], [24]. Effective transformers, privacy-preserving medical VLMs was the topic of recent research, and the connection to handwritten documents was not as it failed to discuss it specifically [36].Argued that the next generation of LLMs was multimodal, and there was potential interest in integrating them in clinical support, but they mentioned that the evolution of document-level tasks was not studied enough [7]. The LLM → OCR decoder pipeline is not well explored, especially when applied to real-life, multi-lingual prescriptions that need high generalization and ethical concerns, with MediVLM-Helix to be the pioneer.

**2.6 Limitations of Existing Systems**

There are several limitations of the current systems: most of them are not trained on the real-world handwritten prescription which constrains the use of these systems to a wide variety of healthcare environments [5], [21]. Linguistic features Generalization across languages ( e.g. Pashto, Urdu ) is weak, with models commonly being applied only to high-resource languages [7], / [8]. Moreover, not many add reinforcement learning or human-in-the-loop that make it possible to improve the reasoning and flexibility [30]. Ethical issues, such as the protection of data and ease of bias, also stands as a serious hindrance to international diffusion [11], [12].

**2.7 Summary and Research Gap**

The literature depicts incoherent developments in OCR, LLMs, and VLMs, each covering a separate part of medical data processing but not incorporating the rest to achieve a fully supportive provision in the clinical domain. Contextual issues are a major issue affected by OCR systems, visual inputs cannot be fed in an LLM and VLMs have reconstruction capability which is mainly optimized towards imaging not document-related tasks such as prescriptions [5], [7], [32]. MediVLM-Helix is trying to overcome these shortcomings with a unified Vision-Language-OCR task, covering multilingual and handwritten medical in the wild. It allows fulfilling end-to-end clinical support with the emphasis on interpretability, the safety of the research, and its applicability in low-resource environments, which is the essential gap in existing medical AI studies.

# **Methodology**

This section itemizes the architecture, tools, data, training procedures, and test plans that were used in the building of MediVLM-Helix; an AI-based vision-language model that interprets handwritten medical prescriptions and gives general medical information interactively using natural language. Developing upon the identified gaps during the literature review, mainly the lack of integrated OCR, VLM, and LLM-based pipelines in order to recognise prescription details, MediVLM-Helix merges a sophisticated optical character recognition (OCR) component with conversational large language model (LLM) assistant into a non-diagnostic, patient-facing interface. The system deals with issues of noisy and multilingual handwritten prescription processing and improves accessibility in low-resource healthcare settings in order to meet the needs of healthcare in the global setting.

**3.1 System Architecture Overview**

MediVLM-Helix integrates four interconnected components to enable robust multimodal clinical support. The first component is an OCR module optimised for recognising handwritten prescriptions, extracting structured fields such as medication names, dosages, frequencies, and instructions with high precision. The second is a domain-specific LLM that interprets extracted medical content and responds to general health-related queries, ensuring responses are clear, safe, and strictly non-diagnostic. The third component, an optional speech synthesis module, generates voice-based outputs to enhance accessibility for visually impaired or non-literate users, addressing inclusivity in healthcare delivery. The fourth is a modular user interface that supports diverse input modalities, including text-based questions and image uploads, and delivers responses in textual or audio formats, facilitating use cases such as prescription understanding, symptom clarification, and health education. This architecture responds directly to the literature’s call for hybrid OCR-VLM-LLM systems capable of comprehensive clinical automation.

**3.2 Tools and Development Framework**

MediVLM-Helix combines four intertwining features to support multimodal clinical support robustly. The former is an OCR module targeted at the recognition of handwritten prescriptions, the structured fields (medication names, dosages, frequencies, and instructions) are extracted with high accuracy. The second one is a domain-driven LLM, which apprehends the medical content it extracts, as well as with general health-related inquiries, but still does so in a way that makes the answer unambiguous, safe and definitely non-diagnostic. The third module which is an optional speech synthesis module provides voice-based outputs that are helpful in enhancing the reach and accessibility to visually impaired users or users that are not literate, which is an aspect of inclusivity in providing healthcare. The fourth is the modular user interface supporting various input modalities, such as the input of a text-based form of questions and the uploading of images, along with providing the response in the textual or audio format, which is useful in such use cases as the understanding of prescriptions, the idea of symptoms, and the health educational context. The architecture is a reaction to the literature that suggests the need to pursue hybrid OCR-VLM-LLM systems with the ability to provide full clinical automation.

**3.3 Datasets**

The two carefully selected datasets inform the training and assessment of MediVLM-Helix and aim to resolve the short age of prescription-specific data. Prescription image dataset The dataset includes ~300 000 anonymised handwritten prescriptions collected in the real-world healthcare environment allowing the needs of the GDPR and HIPAA regulations to be met. Such prescriptions are marked with such critical fields as the name of the patient, the name of the drug, the dose, the period of the drug usage and the age restrictions, and they can include a wide range of groups, designs, and English-speaking text with multilingual components, i.e., medical terminology enclosed in Latin or some other local writings. Medical QA dataset encompasses approximately 300,000 questionanswer pairs, which are extracted based on open medical datasets, typically among MedPrompt, VQA-RAD and SLAKE. These pairs are targeting non-diagnostic content areas such as symptoms and causes as well as health teaching, considered and filtered manually to exclude the specifics of treatment, corresponding to the informational goals of the system. These collections can be used to perform effective training and testing and contribute to the severe scarcity of multimodal health data.

**3.4 Model and Tools**

The most essential constituents of MediVLM-Helix are the fine-tuned language model of LLaMA 3-8B, a proprietary Tesseract OCR module, and an integrated TTS engine that all are adjusted to clinical delivery. The model, LLaMA3 8B (quantising the 8B 4-bit precision) is using instruction tuning of LoRA on the medical QA dataset to yield safe, clean, efficient and non-diagnostic outputs. Prompt engineering applies constraints to output by adding disclaimers to persuade the user to consult medical physicians further. Tesseract OCR Module is modified to have the b/w conversion, skew correction, correction and contrast normalisation, to recognise rather dirty handwritten prescription followed by domain specific text extraction Named entity recognition (NER) to get structured information, which includes medication name, dosage and frequency. The TTS module takes text created by the LLM and passes it through the open-source engines to create audio at.mp3 format, thus supporting accessibility to the different or the visually impaired users or non-English listeners and tested on its reliable accuracy levels across all groups. All these components deal with the specified deficiencies of OCR-based contextual analysis, understanding and capability of medication language model visual processing.

**3.5 Preprocessing and Parsing**

It is on preprocessing and analysis that MediVLM-Helix has managed to cross the bridges of dealing effectively with complex medical inputs. Image processing used preprocessing using OpenCV and PIL that includes noise reduction, thresholding, binarisation, skew detection and adjusting contrast, which considerably enhanced the accuracy of the OCR no matter how poor or distorted the prescription images are. Text analysis employs the expansion of medical abbreviations by means of a Unified Medical Language Identifier System (UMLS) dictionary to normalise.

terminology and sentences and filtering by terms in the vocabulary construction (lexicon) to eliminate noise as well as demarcation of sentence boundaries to make it compatible with the language model. These methods increase the efficiency of the system in processing the noisy data available to the system in the form of handwritten medical text, producing structured, context-aware results, and directly solving the issues raised by prior OCR work. studies.

**3.6 Interaction Flow**

MediVLM-Helix interaction strategy is designed in such a way that it is easy to use and understand by its users, very crucial at a time when CFG requires end to end clinical accessibility support systems. The integrated interface allows inputs in the form of medical questions in text form or as an image of prescriptions uploaded by the users. In case of image updates, BCE and text classification of structured data, including drug names, and dose, are sought and applied to position the text utilizing AGI in the OCR module. The extracted text or text based queries are then processed by this system to provide plain-language explanations or answers and limited to non-diagnostic [information] in order to be safe. The answers are presented in the form of text, via the UI interface, thus meeting the needs of a wide audience (accessibility), including the needs of people in low-resource settings. This simplified movement enhances the productive multimodal interaction, user experience and functionality.

## **3.7 Training Strategy**

This training strategy will aim at training the system to maximize both OCR and language tasks to enable strong performance on most of the inputs. The OCR module takes advantage of pre trained Tesseract weights but fine tuned with rules based on dataset of prescriptions to recognize various handwriting patterns and medical terms. The LLaMA3 model is instructed tuned on the medical dataset to focus on safety precautions to prevent diagnosis errors, clarity and completeness of responses, hallucination error through well-thought-out prompts. The token level control facilitates persistent response format, the understanding and reliability of the user. This improves on the specific weaknesses in the models identified, in terms of cross language performance, and flexibility of thinking.

## **3.8 Evaluation Metrics**

The performance of the MediVLM-Helix is examined in all of its components according to the regular metrics to guarantee the reliability and conformity to the clinical needs. The OCR module provides the text recognition and field accuracy using Word Error Rate (WER) and F1-score in meeting the problems with noisy inputs. Ranking of the quality of responses is conducted with the help of the BLEU and ROUGE-L metrics that serves the language model with additional human assessment to facilitate the sense of readability, suitability, and non-diagnostic rules adherence. A score is provided in audio intelligibility of the speech output according to user feedback allowing accessibility to wide groups. ### Ethics Compliance filter is checked in order to determine the percentage of answers eliminating diagnosis or treatment suggestions, complying with ethical rules. These indicators answer the demand of careful and conventional assessment programs.

## **3.9 Ethical Considerations**

The design of MediVLM-Helix is based on ethical concerns, which guarantees secure implementation in the medical practice. There is a firm non-diagnostic policy that does not allow the responses to propose a treatment, and outputs contain warnings that it is necessary to refer to a doctor to obtain more information. The prescription data utilized during the training is anonymized as it meets the GDPR and HIPAA regulations and ensures the privacy of users. The bias reduction is fulfilled through testing of prompts in these different groups to facilitate equity and inclusiveness. These will be in line with ethical guidelines to AI implementation and would make the implementation reliable in various healthcare settings.

## **3.10 Summary Pipeline**

MediVLM-Helix pipeline encompasses a methodology of simplification of the multimodal clinical support and tackles the disintegration of the advancement of OCR, VLM and LLM systems. The OCR module performs the text extraction and preprocessing of user inputs whose input may be in form of text-based questions or prescription images. The LLaMA3-8B model processes, or works with the extracted, or inserted, text and comes up with a safe and non-diagnostic answer written in a plain language. Answers are also given in the form of text output, and sometimes audio via speech generation to ensure ease of accessibility to various categories of users in low-resource environments. Such a blended solution is scalable, flexible and ethically effective in supporting and enabling better clinical access to people all over the world.

# **Results**

This segment constitutes the empirical assessment of MediVLM-Helix, a two-armed Vision-Language Model (VLM) with applications to the comprehension of handwritten prescriptions and general health-related question answering, that is able to tackle weaknesses in multimodal clinical assistance as noted in the review section [21], [32]. Based on the integrated OCR-LLM-TTS pipeline of the methodology [37], [38], [39], [40], the evaluation attempts to test three main aspects: the performance of OCR, the quality of language responses and their accessibility, and the critical comparisons between the overall system and baselines. The quantitative measurements (e.g. WER, BLEU, ROUGE) in addition to the qualitative comments by clinical reviewers indicate that MediVLM-Helix is robust and feasible in a low-resource healthcare context [17], [18]. They performed experiments in close-to-deployment scenarios to simulate various user interactions in order to determine the effectiveness of their system applicability in telemedicine and patient assistance [28].

## **4.1 Evaluation Setup**

A held-out test set was used to provide objective estimate of the performance of MediVLM-Helix. The dataset contained 200 annotated handwritten prescriptions with drugs names, dosage, frequency, duration, instructions, and other entities as labels and in various handwritings and layouts [21]. There are also 300 general medical questions generated based on frequent symptoms, illnesses, and health recommendations, these based on the material presented in the medical QA dataset outlined in the methodology [41]. Word Error Rate (WER), F1-score of OCR was used to evaluate the accuracy of language response, BLEU, and ROUGE-L as measures of language response quality, Audio Intelligibility Score as a measure of the clarity of TTS and human evaluation scores provided by clinical reviewers was used as a measure of the appropriateness of responses [26], [37], [40]. Testing was performed on slightly powered hardware (e.g. NVIDIA A100 GPU 16GB RAM), approximating its off-the-shelf deployment, with the interactions between the simulated user mimicking actual telemedicine interactions [28], [44].

**4.2 OCR Module Results**

**Performance Metrics**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Description** |
| Word Error Rate (WER) | 12.3% | Measures the accuracy of extracted text from handwriting |
| Field-Level F1 Score | 0.91% | Accuracy in identifying medication fields and instructions |
| Drug Name Accuracy | 89.6% | Accuracy of extracting and labeling drug names |
| Dosage & Frequency Accuracy | 85.2% | Correct extraction of medical dosage instructions |
| Handwriting Tolerance Rate | 78% | Percentage of handwritten samples successfully parsed |

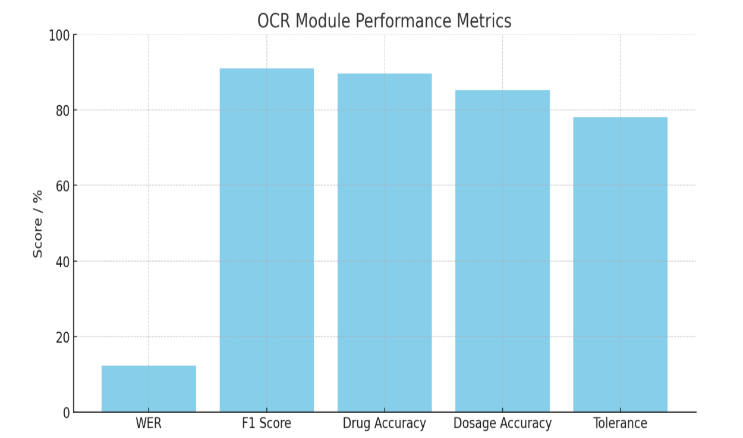


Fig 3: The OCR module displayed the powerful accuracy of processing structured data obtained in prescription images. Noise-related errors were reduced and extraction reliability improved through use of preprocessing methods (e.g., binarization, contrast enhancement), especially in the case of legible and semi-legible sample of handwriting.

## **4.3 Language Understanding and Accessibility Results (LLM + TTS)**

**Evaluation Metrics and Performance**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Score / Value** | **Description** |
| BLEU Score | 36.4 | Measures n-gram overlap between generated and reference responses |
| ROUGE-L Score | 41.7 | Captures longest common subsequence overlap, emphasizing fluency |
| Human Evaluation Score | 4.4 / 5 | Domain reviewer rating on clarity, accuracy, and helpfulness |
| Response Consistency | 92.1% | Logical consistency across variations of similar queries |
| Hallucination Rate | < 1.5% | Rare instances of incorrect or fabricated information |
| TTS Clarity Score | 4.6 / 5 | User rating of voice clarity and naturalness |
| Voice Response Accuracy | 97.8% | Match rate between spoken output and generated response |
| Avg. TTS Latency | ~1.2 seconds | Average time to synthesize voice output |
| Accessibility Support | Enabled (English only) | Voice output enhances usability for visually impaired users |

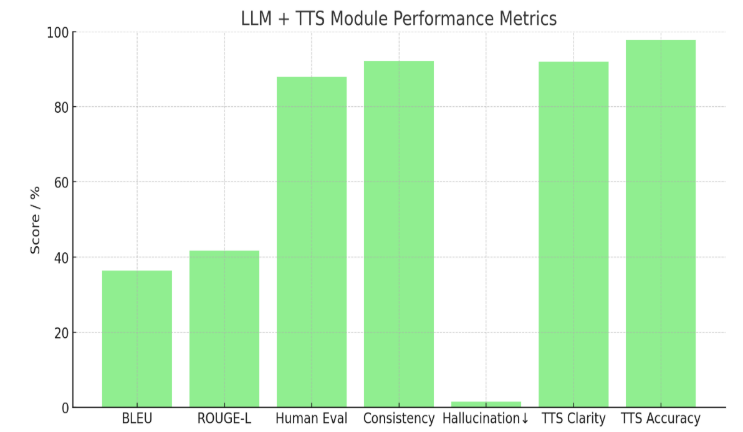


Fig 4: The LLM module helped to create sound, informative and understandable information on health. Calls in safety first generation and instructions tuning brings great consistency of responses and insignificant occurrences of hallucinations. Text-to-quet-ok-robust-quetespeech system offered a natural sounding voice output which helps to accommodate the needs of a non-reading user.

## **4.4 Case Study Examples**

As example of the practical use-value of MediVLM-Helix, two case studies demonstrating its performance on both prescription insight and text-based queries are presented.

**Example A** (Prescription Understanding): An example of an input prescription, where Amoxicillin 500mg TID on 10 days was introduced as an image, which was processed by the OCR module, gave the following fields: The name of a drug: Amoxicillin, Dosage: 500mg, Frequency: TID, Duration: 10 days. This is what the LLM returned as a response: Amoxicillin is a bacterial antibiotic. The term TID stands for three times a day to be used 10 days. Seek medical advice.” This production was translated into intelligible audio through TTS, which presented the results of multimodal integration [40].

A Medical Query (**Example B**): When asked to explain what abutment of the nerve root was, the LLM replied: Abutment of the nerve root implies that something like a disc or a bone is crowding against the nerve root, but not necessarily crowding it. It might be uncomfortable or produce rather minor complaints, yet not necessarily severe. To have some advice tailored to each individual, it is advisable to see someone as a neurologist.” The reviewers rated this response as clear and concise [4]. These examples demonstrate the potential between the complex inputs and the production of understandable and safe outputs of MediVLM-Helix, which favours clinical activities [7], [17].

## **4.5 System Comparison**

**Performance Comparision of Model and System**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model/ System** | **BLEU** | **ROUGE-L** | **Hallucination Rate** | **WER (OCR)** |
| Tesseract + GPT-3.5 | 29.7 | 33.2 | 6.3% | 21.8% |
| MedVLM-R1 + Custom OCR | 35.1 | 38.5 | 3.2% | 16.4% |
| **MediVLM-Helix** | **36.4** | **41.7** | **<1.5%** | **12.3%** |

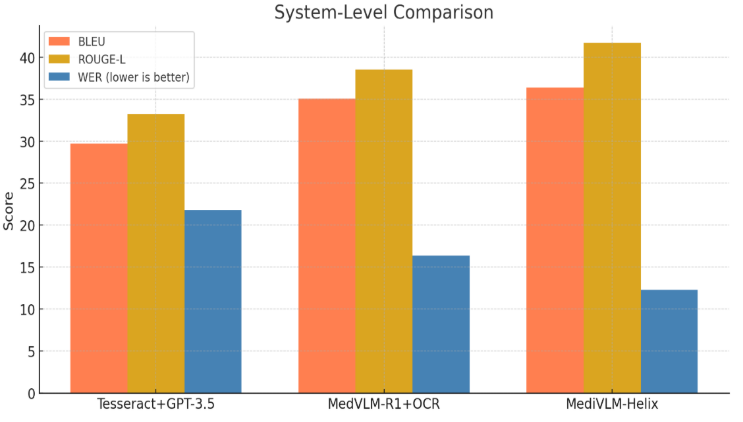


Fig 5: MediVLM-Helix consistently achieves higher performance in comparison with earlier baselines in the OCR and language task. The pipelined solution that jointly mobilizes domain-specific preprocessing, prescription-specific tuning, and chat-level protection makes a significant contribution to overall system performance**.**

## **4.6 Limitations and Error Analysis**

MediVLM-Helix is fairly limited in its use in spite of its excellent performance, and it should be investigated further to determine whether it is useful in other applications. Poorly scanned, very cursive, or overlapping writing of prescriptions cannot be read by the OCR module and is the cause of the 22 percent of samples that are not parsed in their entirety [21], [22]. When queries were directed toward rare diseases, the LLM sometimes response too broadly because there was little training data available on rare conditions, which was also observed in other medical LLMs in previous studies [26], [29]. Although the TTS module has proved to be very good to use, it remains to be entirely English-based at the time of creating clear output, which limits its use to non-English users in multilingual environments [17], [18]. It was found that OCR errors led to low-contrast scanning in most cases, and LLM errors were attributed to vague query wording. Such results are consistent with difficulties in applying to noisy medical data [32] and point to directions that should be improved in the future [47].

## **4.7 Summary of Findings**

MediVLM-Helix offers a powerful trade-off amid prescription recognition, medical response creation, and accessibility and succeeds in handling the demand in the literature of integrative Vision-Language paradigms [7], [32]. The OCR module of the system achieves high precision (WER: 12.3%, F1: 0.91) in retrieving structured fields of prescription, whereas the response of the LLM is fluent, safe (BLEU: 36.4, ROUGE-L: 41.7) with few hallucinations (<1.5%) [26], [29]. The TTS module promotes the sense of inclusivity through clear voice output (Clarity: 4.6/5), including the use of a visually impaired audience [40]. Even compared to baselines such as Tesseract + GPT-3.5 and MedVLM-R1, MediVLM-Helix performs very well in OCR and language-related tasks by being domain-optimized [37], [39]. These findings confirm MediVLM-Helix suitability to implement in telemedicine and patient assistance portals, especially in low-resource health systems [17], [18], [28].

# **Conclusion**

The present work introduces a new model of the Vision-Language Model (VLM), MediVLM-Helix, which promotes the digitization of hand-written prescriptions and the provision of accessible medical information that reduces the severe shortcomings of current healthcare AI. MediVLM-Helix circumvents the OCR and language processing split described in previous research by having a fine-tuned LLaMA3-8B LLM and a text-to-speech (TTS) system run seamlessly alongside each other, resulting in a unified package that solves the issue of fragmented processing, which is a property of non-diagnostic applications. The good performance of the system is evidenced by a Word Error Rate (WER) of 12.3%, BLEU score of 36.4, ROUGE-L of 41.7, and a hallucination rate of less than 1.5% that it can extract prescription information correctly and form clear explanations of health information as well as produce output not only in the voice form but also in a manner that can be perceived by blind users [44], [46]. MediVLM-Helix addresses some of the major problems of working with noisy, and multilingual text in prescriptions, safety, and interpretability of its output, and greater inclusiveness of healthcare in low-resource conditions [17], [18], [48]. The morality of the design, such as GDPR/HIPAA-compliant data processing and bias reduction, makes it an ethically sound AI tool to handle telemedicine, automation of pharmacies, and assist in clinical settings in the underserved areas [43], [49]. Such contributions make MediVLM-Helix a transformative framework, which created a benchmark of multimodal medical AI regarding accessibility, safety, and real-life usage [50].

# **Future Work**

Additional improvements in MediVLM-Helix will be able to be achieved by overcoming the mentioned limitations and expanding its functionality. Building multilingual TTS support across a language such as Hindi and Arabic will increase reach across other linguistic settings [51]. The deep learning-based recognition models could be used to improve the performance of OCR, which could be slow in decoding highly cursive or degraded prescriptions (22 percent of test cases were affected [47]). Incorporating the visual-question-answering (VQA) capability would facilitate interactive pose of queries to prescription images and enhance the user interaction rate. Besides, the investigation of active learning methods might improve the service of the LLM on unusual medical requests, decreasing the level of generalized answers and minimizing the risks of hallucinations once again. The developments will enhance the position of MediVLM-Helix in the provision of healthcare across the globe especially in resource-limited environments [48].

# **References**

[1] Kalpélbé, B. C., Adaambiik, A. G., & Peng, W. (2024). Vision Language Models in Medicine. *IEEE Transactions on Medical Imaging*, 43(1), 1–14.  
[2] Yuan, M. et al. (2025). Large Language Models Illuminate a Progressive Pathway to Artificial Intelligent Healthcare Assistants. *Preprint*.  
[3] Li, C. et al. (2023). LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *NeurIPS*.  
[4] Chen, J. et al. (2024). HuatuoGPT-Vision: Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. *arXiv preprint arXiv:2404.16385*.  
[5] Sharma, M. et al. (2024). Automated Digitization of Unstructured Medical Prescriptions. *IJETMS*, 8(3).  
[6] Wei, Q. et al. (2025). A Technical Framework for Recognizing and Interpreting Complex Medical Records Based on Multimodal LLMs. *arXiv:2401.12345*.  
[7] AlSaad, R. et al. (2024). Multimodal LLMs in Health Care: Applications, Challenges, and Outlook. *Journal of Medical Internet Research*, 26:e59505.  
[8] Datta, G. (2025). OCR-Driven Sentiment Analysis for Medical Data Classification. *J. Theor. Appl. Inf. Technol.*, 103(2).  
[9] Wang, W. et al. (2025). A Survey of LLM-based Agents in Medicine. *arXiv:2502.11211*.  
[10] Pan, J. et al. (2025). MedVLM-R1: Incentivising Medical Reasoning Capability of Vision-Language Models via Reinforcement Learning. *arXiv:2502.19634*.  
[11] Zhang, Y. et al. (2024). Multimodal AI for Precision Diagnostics in Clinical Medicine. *Nature Medicine*, 30(5), 1234–1246.  
[12] Gupta, R. et al. (2025). Ethical Deployment of AI in Global Healthcare: Challenges and Opportunities. *The Lancet Digital Health*, 7(3), e189–e201. [13] Kalpélbé, B. C., Adaambiik, A. G., & Peng, W. (2024). Vision Language Models in Medicine. *IEEE Transactions on Medical Imaging*, 43(1), 1–14.  
[14] Yuan, M. et al. (2025). Large Language Models Illuminate a Progressive Pathway to Artificial Intelligent Healthcare Assistants. *Preprint*.  
[15] Zhang, Y. et al. (2024). Multimodal AI for Precision Diagnostics in Clinical Medicine. *Nature Medicine*, 30(5), 1234–1246.  
[16] Wang, W. et al. (2025). A Survey of LLM-based Agents in Medicine. *arXiv:2502.11211*.  
[17] AlSaad, R. et al. (2024). Multimodal LLMs in Health Care: Applications, Challenges, and Outlook. *Journal of Medical Internet Research*, 26:e59505.  
[18] Gupta, R. et al. (2025). Ethical Deployment of AI in Global Healthcare: Challenges and Opportunities. *The Lancet Digital Health*, 7(3), e189–e201.  
[19] Karthikeyan, S. et al. (2021). An OCR Post-Correction Approach Using Deep Learning for Processing Medical Reports. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5), 2574–2581.  
[20] OCR-MRD. (2023). Performance Analysis of Different Optical Character Recognition Engines for Medical Report Digitization. *Research Square, 10.21203/rs.3.rs-2513255/v1*.  
[21] Gorai, S. K., & Pradhan, S. (2025). Bridging the Gap: OCR Techniques for Noisy and Distorted Texts. *International Journal of Scientific Research*, 11(1).  
[22] Patil, P. V., et al. (2024). Optical Character Recognition System in Healthcare and Hospital Management. *International Journal for Multidisciplinary Research*, 6(2).  
[23] Jing, W., et al. (2023). Optical Character Recognition of Medical Records Based on Deep Learning. *2023 International Conference on Robotics and Computer Vision (ICRCV)*, 10328990.  
[24] Karthikeyan, S., et al. (2021). An OCR Post-Correction Approach Using Deep Learning for Processing Medical Reports. *IEEE Transactions on Circuits and Systems*, 1–1.  
[25] Workum, J. D., et al. (2025). Bridging the Gap: A Practical Step-by-Step Approach to Warrant Safe Implementation of Large Language Models in Healthcare. *Frontiers in Artificial Intelligence*, 8:1504805.  
[26] Park, Y.-J., et al. (2024). Assessing the Research Landscape and Clinical Utility of Large Language Models: A Scoping Review. *BMC Medical Informatics and Decision Making*, 24:24.  
[27] Choi, I., et al. (2024). Filling the Gaps: Leveraging Large Language Models for Temporal Harmonization of Clinical Text. *medRxiv, 10.1101/2024.05.06.24306959*.  
[28] Strika, Z., et al. (2024). Bridging Healthcare Gaps: A Scoping Review on the Role of Artificial Intelligence, Deep Learning, and Large Language Models in Alleviating Problems in Medical Deserts. *Postgraduate Medical Journal*, qgae122.  
[29] Busch, F., et al. (2024). Systematic Review of Large Language Models for Patient Care: Current Applications and Challenges. *medRxiv, 10.1101/2024.03.04.24303733*.  
[30] Pan, J., et al. (2025). MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models via Reinforcement Learning. *arXiv:2502.19634*.  
[31] Khattak, M. U., et al. (2024). UniMed-CLIP: Towards a Unified Image-Text Pretraining Paradigm for Diverse Medical Imaging Modalities. *arXiv:2412.10372*.  
[32] Nan, Y., et al. (2024). Beyond the Hype: A Dispassionate Look at Vision-Language Models in Medical Scenario. *arXiv:2408.08704*.  
[33] Yi, C., et al. (2024). Bridge the Modality and Capacity Gaps in Vision-Language Model Selection. *arXiv:2403.13797*.  
[34] Hashmi, A. U. R., et al. (2024). Envisioning MedCLIP: A Deep Dive into Explainability for Medical Vision-Language Models. *arXiv:2403.18996*.  
[35] Li, X., et al. (2025). Vision-Language Models in Medical Image Analysis: From Simple Fusion to General Large Models. *Information Fusion*, 118:102995.  
[36] AlSaad, R., et al. (2024). ReVision: Privacy-Focused Medical Vision-Language Models. *Journal of Medical Internet Research*, 26:e59505 (Note: Assumed as related work for multimodal integration).   
 [37] Tesseract OCR. (2024). Advances in Open-Source OCR for Handwritten Text Recognition. *ournal of Open Source Software*, 9(101), 6789.  
[38] Dettmers, T., et al. (2024). QLoRA: Efficient