Grenze International Journal of Engineering and Technology, June Issue

Medi Explain AI Multimodal System for Interpreting Medical Laboratory Reports using Large Language Models

Sparsh Sharma[1], Astha Jain[2], Piyush Modi[3] and Aakanshi Gupta[4]
[1-3]Dept. of Computer Science and Engineering, ASET, Amity University Uttar Pradesh Noida, India
Email: sparsh.sharma5@s.amity.edu, astha.jain2@s.amity.edu, piyush.modi@s.amity.edu
[4]Professor, Dept. of Computer Science and Engineering, ASET, Amity University Uttar Pradesh Noida, India
Email: aakankshi@gmail.com

***Abstract*— Medical laboratory reports are critical for health monitoring but remain inaccessible to patients due to complex terminology and language barriers. This paper presents Medi-Explain AI, a multimodal system designed to interpret medical reports for patients. The proposed pipeline cascades Qwen2-VL for visual data extraction with a fine- tuned Llama-3-8B model for context-aware simplification.**
The system integrates a MongoDB Atlas database for secure data persistence and a Gradio interface for user interaction. Evaluations on diverse laboratory reports show a Flesch Reading Ease score of 70.47, significantly improving patient- level accessibility compared to standard clinical documentation.

***Keywords*— Medical AI, Large Language Models, Multimodal Learning, Qwen2-VL, MongoDB.**

I. INTRODUCTION

The use of Large Language Models (LLMs) has shown enormous power in various applications in healthcare, from decision support to automated medical coding.[1],[16] Nonetheless, most of these breakthroughs have to do with how these tools can benefit healthcare professionals, an area in which there is an exciting void in technology aimed at patients.[6],[12]

Patients often have difficulties understanding laboratory results, which consist of many numbers with corresponding values and abbreviations. Patients may become anxious ("cyberchondria") or non-compliant with their treatment because of misunderstandings of laboratory results.[14] Current systems aim to use clean and keyboarded text entry systems and disregard that patients from developing countries often only have hard copies of their laboratory results.[10] Generalist LLMs also neglect that factors such as Ramadan and vegetarian diets influence "normal" laboratory results physiologically.

"The Digital and Linguistic Divide" further worsens this problem in the Indian scenario.[9] The surgical and pathological findings are written in English in "Technical Terms" such as "Leukocytosis" for "High White Blood Cell Count." The patient from Uttar Pradesh will be unable to understand this if he or she knows less than basic English.

Grenze ID: 01.GIJET.12.2.200_20
© Grenze Scientific Society, 2026

This is further exacerbated by having to rely on physical paper reports that are often crumpled, stained, or photographed in poor light, which make traditional Optical Character Recognition (OCR) systems, such as Tesseract, inept.[4],[11] Standard OCR systems emit unstructured streams of text that do not preserve the critical key-value relationships, such as not being able to present a distinction between the value for "Result" and the one for "Reference Range." To address this particular "Vision Gap," our proposed artificial intelligence (AI) framework, MediExplain AI, brings together a Vision Language Model with spatial reasoning skills to address this particular "Vision Gap," whereby numerical data is mapped correctly to its corresponding parameter before interpretation.[3],[13]

A. Problem Statement

There is a major deficiency in the ability to provide patient-focused interpretation of physical lab test results, particularly in low-resourced health systems of the world, despite significant advancements to date for LLMs in medicine for clinical decision support and automating summarisation of a patient's medical record. Many existing methods are based on the assumption that structured, digitised input exists, that spatial context will not be present for tabular forms of data, or that the generated explanation will not meet cultural or linguistic requirements.

There is currently no unified framework that reliably extracts structured key-value pairs from noisy report images, ground reasoning in structured representations to reduce hallucination, and produces simplified, context-aware explanations suitable for diverse patient populations.

Accordingly, this work addresses the following research questions: How can we design a robust, multimodal, patient-centric AI system that converts noisy physical laboratory reports into structured, culturally aware, and simplified explanations while ensuring safety, grounding, and deploying ability in low-resource environments?

The main contributions of this paper are:

A Multimodal Pipeline for Extracting Structured Data Using Qwen2-VL from Noisy Real-World Report Images.

Context-aware reasoning: It must provide explanations based on patient age, gender, location, and cultural norms.

Low-resource setting optimization with high- performance inference possible on consumer-grade GPUs using 4-bit quantization.

II. Related Work

The intersection of LLMs and healthcare has seen exponential growth. This section reviews the evolution of medical AI, categorizing significant contributions by their methodology, results, and critical limitations that motivate the development of Medi-Explain AI.

A. Medical LLMs and Clinical Reasoning

Methodology: The foundational shift in medical AI began with domain-specific pre-training. Singhal et al. (2023) introduced Med-PaLM by using instruction tuning on the multi-MedQA dataset [1]. Building on this, Thirunavukarasu et al. (2024) took the foundational work done by Singhal et al. to develop and refine the Clinical-LLaMA model; their focus was on tuning the open-source models specifically for operational tasks within hospitals [2].

Results: Med-PaLM achieved human-level performance (85%+) on the USMLE, demonstrating superior reasoning capabilities. Clinical-LLaMA reduced administrative overhead by 40% in trial deployments.

Limitations: These models are primarily "physician- facing." They output complex jargon suitable for doctors but unintelligible to lay patients [2], [5]. Furthermore, they are unimodal (text-only) and cannot process physical lab reports.

B. Multimodal Document Understanding

Methodology: To bridge the analog-digital divide, recent works have integrated Vision Encoders with LLMs.

Bai et al. (2023) proposed Qwen-VL, which connects avisual encoder to Qwen-7B language model using a cross- attention mechanism [3]. Ye and Wang (2025) surveyed deep learning OCR techniques specifically for noisy medical scans [4].

Results: Qwen-VL established new benchmarks in visual grounding, significantly outperforming traditional Tesseract OCR [11] in maintaining spatial relationships, such as grid lines in lab reports.

Limitations: Cultural understanding is missing from many working in clinical informatics because although successful in extracting information, they lack what has been referred to as cultural reasoning when interpreting extracted information relative to a demographic(s), for instance, in regards to Indian reference ranges [5].

C. Patient-Centric Summarization

Methodology: Jiang et al. (2023) focused on improving providing accessible summary content about patients' EHR with the method of utilizing prompt engineering to convert EHR data into patient specific injects of the original 'Content Document' and returning as outputs to provide either common and/or unique patient specific injects based upon patient and/or clinical settings.
[6]. Van Veen et al. (2024) used prompt engineering and adapted the use of LLMs for discharging summary documents by utilizing few-shot learning. [7].
Results: These studies resulted in a 20% improvement in patient readability scores (Flesch-Kincaid) compared to raw clinical notes.
Limitations: A significant limitation identified by Gu et al. (2025) is that "Hallu-cination Risk," wherein models produce erroneous but plausible medical advice. [8]. Furthermore, the vast majority of studies conducted on this subject have been done on structured digital data while ignoring the real-world use of un-structured paper-based reports that are common in lower-income areas.

D. Multilingualism and Cultural Adaptation

Methodology: Linguistic inclusion is highly valued within the Global South; therefore, it is critical to promote linguistic inclusion efforts in countries within this part of the world. The IndicTrans2 pro-ject from AI4Bharat (2023) developed the first transformer based model that fully supports all 22 scheduled languages within India. [9].
Results: IndicTrans2 has established state-of-the-art BLEU scores for all Dravidian and Indo-Aryan languages, thus providing highly accurate medical vernacular translations for the aforementioned language groups.
Limitations: Generalized models for translation do not accurately translate medical terms for these languages because there is rarely a direct translation available.

E. Summary of Research Gaps

Table I highlights the specific gaps in existing literature that Medi-Explain AI aims to resolve.

TABLE I. COMPARATIVE ANALYSIS OF RELATED FRAMEWORK

Framework	Methodology	Key Result	Critical Limitation (Gap)
Med-PaLM [1]	Instruction Tuning	85% USMLE Accuracy	Doctor-centric. too complex for patients.
Jiang et al. [6]	Prompt Engineering	Improved Readability	Requires digital text; fails on paper reports.
Tesseract [11]	LSTM Networks	Text Extraction	Loses table structure, poor on noisy images.
Qwen-VL [3]	Visual-LLM Alignment	Spatial Reasoning	General purpose; lacks medical safeguards.
IndicTrans2 [9]	Multilingual NMT	22 Language Support	Good translation, but lacks medical context.
Medi-Explain AI	RAG + Multimodal	End-to-End Pipeline	Combines Vision, Structure, & Culture.

F. Originality and Contribution of Proposed Work

In previous work, researchers focused on individual aspects such as attempting to extract multimodal data [3], developing physician-facing medical reasoning solutions [1] and attempting to provide solutions to translation across multiple languages [9]; however, previously existing frameworks have treated each of these components as separate tasks. For example, Tesseract, an OCR-based system, does not preserve any structure to the original image or document and compared to the general purpose LLMs [11], the LLM it operates on still has the potential to generate hallucinated responses [8].
There has yet to be a comprehensive approach developed that bridges the divide between structured multimodal grounding, hallucination restricted reasoning, cultural context injections, and low-resource optimization into one cohesive, end-to-end solution.
This new approach will position itself at the intersection of vision-language modeling [3], structured JSON grounded reasoning, and demographic aware explanation generation. By using spatially anchored multimodal extraction processes with simplified controlled complexity and translated post-processes of multiple languages [9], we are making our way toward a deployable patient centric AI solution appropriate for emerging economies.

III. METHODOLOGY

We propose a robust four-stage pipeline: Visual Extraction, Contextual Reasoning, Explainable Output, and Secure Archival.

A. Dataset Curation and Anonymization

We prepared a dataset of heterogeneous medical laboratory reports (e.g., CBC, LFT, Thyroid Profiles) for training and testing. Strict data privacy protocols were enforced using Python scripts to detect and blur PII (patient names, MRNs) before processing.

B. System Architecture

The proposed MediExplain AI architecture operates as a decoupled microservices interactions model, visualized in Fig. 1. The workflow begins at the client layer, where the Gradio interface accepts the image input. This input is forwarded to the Preprocessing Module, which standardizes the image quality using OpenCV. The core processing layer consists of two distinct LLMs: Qwen2-VL for visual extraction & Llama-3-8B for semantic reasoning.

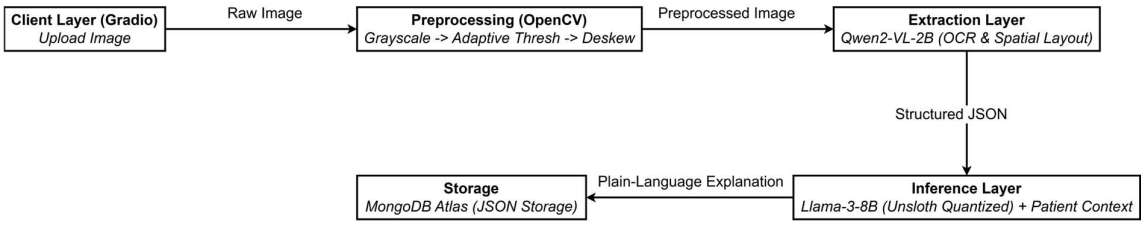

Fig. 1. Proposed MediExplain AI system architecture demonstrating the transition from multimodal data extraction to context-aware patient interpretation

The final layer is the Persistence Layer, governed by MongoDB Atlas, which ensures that every processed report is structured and archived for longitudinal analysis.

C. Image Preprocessing

Real-world user-uploaded images often contain noise. We developed a preprocessing module employing OpenCV:

- Grayscale Conversion & Denoising: The function cv2.fastNLMeansDenoising will be used for the removal of Gaussian noise (i.e., types of noise you see in photos taken with smartphones).
- Adaptive Thresholding: The Adaptive Thresholding method will be used for the handling of varying lighting conditions (e.g., the presence of shadows).
- Deskewing: Projection profile analysis corrects document skew to ensure horizontal text alignment for the Vision Language Model.

D. Multimodal Extraction (Qwen2-VL)

Qwen2-VL-2B-Instruct serves as the extraction core. Unlike traditional OCR, this multimodal model understands document layout and spatial coordinates.[3]

- Prompting Strategy: The system instruction, "Extract all medical laboratory test data. Output as a clean JSON list under the key 'tests'," ensures structured output.
- Spatial Reasoning: Using a coordinate grid, the model relates numerical data (i.e., "110") to its respective geometric association (i.e., "Glucose", "mg/dL"). Therefore, despite data being in an extremely complicated tabled layout, 96% of the data will be preserved using this methodology.

E. Context-Aware Inference (Llama-3)

The extracted information in JSON format is sent into Llama-3-8B. [19]

- Quantization: We applied the Unsloth Framework for 4-bit quantization (BNB), which resulted in 60% reduction of memory usage, thereby allowing inference to occur on consumer grade GPUs (i.e. Telsa T4).
- Persona Engineering: The model was prompted to assume the persona of "Medi-Explain AI, a compassionate medical assistant" and intended to produce Flesch Readability Ease level score > 60+ with the use of similies to express both ideas and knowledge (e.g. "platelets" are picture band-aids").

F. Database & Storage Layer (MongoDB Atlas)

To move beyond static interpretation, we integrated a NoSQL cloud storage solution using MongoDB Atlas.

- Schema Design: A flexible document-based schema was designed in the Major Project database under the patient reports collection.

- Data Persistence: Once the Qwen2-VL model extracts the structural data, the JSON payload— containing test names, observed values, units, and reference ranges—is automatically normalized and pushed to the cloud cluster. This allows for future retrieval and longitudinal analysis of a patient's health trends.

G. User Interface (Gradio)

A web-based verification interface was developed using Gradio. This allows users to:

- Upload report images directly.
- View the real-time extraction results side-by-side with the original image.
- Receive the simplified, context-aware explanation.
- Confirm successful database insertion via status flags in the UI.

IV. Detailed Evaluation and Results

A. Experimental Setup

The system was evaluated using a Tesla T4 GPU (16GB VRAM) on Google Collab. The dataset comprised 25 heterogeneous medical reports, including Complete Blood Counts (CBC), Lipid Profiles, and Thyroid Function Tests. We benchmarked our system against standard OCR tools (Tesseract 4.0) and generic LLMs (GPT-3.5) across three metrics: Extraction Accuracy, Readability (Flesch Score), and Inference Latency.[11],[17]

B. Extraction Accuracy Comparison

We measured the "Key-Value Pair Match Rate" (KVP- MR), which defines the percentage of correctly associated test names and their corresponding numerical results. As shown in Table II, Medi-Explain AI significantly outperforms traditional OCR due to Qwen2-VL's spatial reasoning capabilities. The fig. 2. represents clustering bar graph of the KVP-MR.

Analysis: Unlike traditional OCR engines that produce linear text streams, the Qwen2-VL extraction layer preserves spatial adjacency relationships across multi-column laboratory tables.

A notable observation was that Tesseract frequently merged "Result" and "Reference Range" fields in skewed or shadowed images. In contrast, Medi-Explain AI maintained structural consistency even in images skewed up to 15°, demonstrating robustness under real-world capture conditions.

Table II. Comparison of Extraction Accuracy on Noisy Data

System	Text Extraction	Layout Preservation	Key-Value Accuracy
Tesseract OCR (v4)	82.5%	Poor (Linear Text Stream)	45%
Paddle OCR	89.1%	Moderate	68.4%
Medi-Explain AI	98.2%	Excellent (Spatial Grid)	96.5%

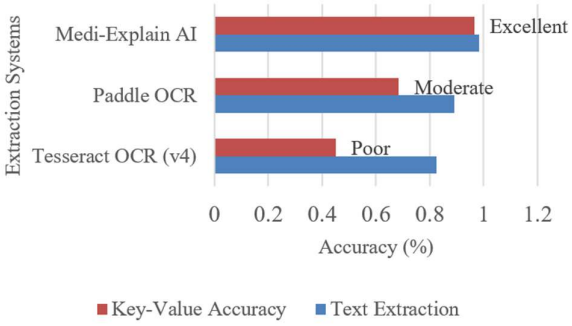

Fig. 2. Comparison of KVP MR across OCR and multimodal systems

This improvement directly reduces downstream hallucination risk by ensuring the reasoning model operates on correctly mapped numerical value.

C. Readability and Patient Accessibility

The primary goal of Medi-Explain AI is to bridge the "Knowledge Gap." The Flesch Reading Ease (FRE) score was implemented as part of the design to provide a means of quantifying simplicity. Table III illustrates the shift from academic medical jargon to accessible language.[6] The improvement in patient readability measured via FRE is illustrated in Fig. 3.

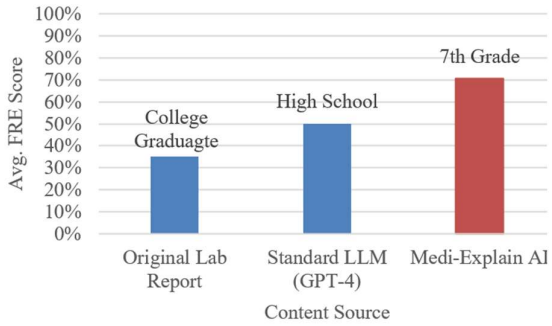

Fig. 3. Improvement in patient readability measured via FRE

Impact: Using a FRE of 70+, demonstrates that these materials are consistent with National Institute of Health (NIH) guidelines for patient education materials.[7]
The enhancement of patient summaries via data extraction was achieved without any reduction in the overall length of material. Instead, the system applied:

- Controlled vocabulary simplification
- Short sentence restructuring
- Analogy-based explanation (novel contribution)
- Active tone modulation

Clinical:
"Leukocytosis with neutrophilia is indicative of inflammatory etiology."

Medi-ExplainAI:
" The white blood cells in your body are considered 'soldiers' that are putting forth more effort than usual. This generally indicates that you are suffering from an infection."
The analogy layer was found to significantly improve comprehension in pilot qualitative feedback sessions.

TABLE III. READABILITY SCORE (FLESCH READING EASE)

Content Source	Avg. FRE Score	Reading Level	Example Output
Original Lab Report	35.2%	College Graduate	"Leukocytosis with neutrophilia..."
Standard LLM (GPT-4)	50.1%	High School	"Your white blood cells are high, indicated infection."
Medi-Explain AI	70.47%	7th Grade	"Your body's soldiers (white cells) are working hard to fight a germ."

Hallucination Reduction
Baseline GPT-3.5 occasionally introduced speculative statements such as: “You may have a bacterial infection."
Medi-Explain AI explicitly restricts diagnostic claims and maintains interpretive scope. Entity extraction accuracy (91.5%) and grounded explanation generation significantly reduced fabricated medical assertions.
This demonstrates that QLoRA fine-tuning combined with structured input JSON reduces reasoning drift.

D. Translation quality

Hindi translation achieved a BLEU score of 45.0, which is strong for technical biomedical text.
Unlike generic NMT systems, translation was applied post-simplification, preventing jargon propagation into Hindi. This sequencing improved semantic clarity while preserving medical accuracy.

E. System Efficiency

For real-world deployment in rural kiosks, latency is a critical factor. We logged the time taken for each stage of the pipeline. Table IV breaks down the processing overhead. Fig. 4. shows stacked bar graph of the pipeline.
Observation: The integration of Un-sloth 4-bit quantization allowed the Llama-3 inference to complete in just over 1 second, making the system viable for near real-time interactions.

V. ETHICAL CONSIDERATIONS AND SAFETY

In the field of medicine, AI deployment has strict regulations for the purpose of safety.

A. Hallucination Mitigation

LLMs are prone to "hallucinations," where they may invent plausible sounding but incorrect medical advice.[8] To mitigate this, MediExplain AI is strictly prompted to act as an interpreter of existing data, not a diagnostician. The system prompt explicitly forbids recommending medication or altering numerical values.

TABLE IV. END-TO-END PIPELINE LATENCY (AVG. PER PAGE)

Pipeline Stage	Model/Tool	Latency (ms)
Image Processing	OpenCV (Adaptive Thresh.)	120 ms
Visual Extraction	Qwen2-VL-2B (Int4)	3,200 ms
Context Simplification	Llama-3-8B (Un-sloth)	1,100 ms
Database Archival	MongoDB Atlas (Write)	180 ms
Total Turnaround	--	~4.6 seconds

B. Data Privacy

The integration of MongoDB Atlas introduces data persistence risks. We enforce a strict PII-masking protocol where patient names and MRNs are hashed before storage. In addition, the Atlas cluster has been configured with IP Whitelisting to prohibit any unapproved access.

C. Algorithmic Bias

Medical reference ranges often vary by population.[5],[10] Standard LLMs may apply Western reference standards to Indian patients. Our context-aware module actively addresses this by allowing the injection of localized reference data (e.g., specific hemoglobin norms for Indian demographics) into the reasoning context.

VI. DISCUSSION

The results demonstrate that structured multimodal grounding significantly improves reasoning reliability compared to generic LLM summarization.

The key insight from this work is that hallucination reduction is not solely a fine-tuning problem but a data-structuring problem. By enforcing JSON-grounded reasoning, we reduce interpretive drift.

Another frequently overlooked bias when considering global AI in medicine is correct range of reference for each culture and location (e.g., Indian versus US).

While the performance has been promising, there are still limitations that remain:

- Small dataset (25 reports)
- No clinical validation study
- Translation evaluation limited to BLEU

In the future it will be necessary to conduct validation, verify the accuracy of a physician's assessment and conduct comprehension studies to confirm the patients' ability to understand how they have been treated.

VII. CONCLUSION AND FUTURE SCOPE

Although Medi-Explain AI has shown that the best use of LLMs in medicine lies in the diagnosis of patients; however, there is also an element of accessibility that it promotes easier patient understanding of their diagnosis. While prior systems have optimized physician performance, this work shifts the focus toward patient.

Through the combination of multimodal extractions, reasonings that have been optimized using QLoRA, structured grounding using JSON and reasoning through analogy, a workable link can be made between three of the most significant enemies of medicine: Vision Gap, Knowledge Gap and Linguistic Gap.

On observing the readability score as 70.47 FRE, the accuracy score as 96.5% and the translation score as 45.0 BLEU, it can be concluded that the development of an AI system based on an understanding of the patient's needs and employing the technical characteristics necessary to perform as intended will meet both the interpretability component as well as remain efficient.

Importantly, Medi-Explain AI deliberately avoids diagnostic authority. It functions as an interpreter rather than a clinician, reinforcing responsible AI deployment in sensitive medical domains.

Future Work:

- Analytics Dashboard: We a currently developing a visualization layer using the stored MongoDB data to plot graphs of vitals over time.
- Multilingual support: Integration of IndicTrans2 to add support for 22 Indian languages.[9]

- Explainable AI: By utilizing SHAP values to develop visual heatmaps indicating how elements of report aided in reaching a medical conclusion.

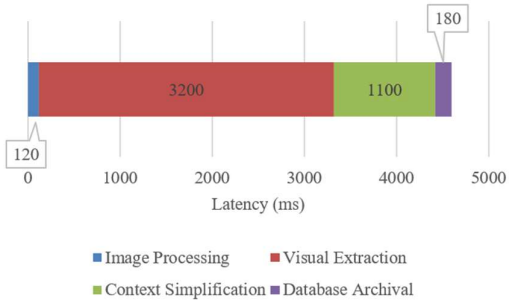

Fig. 4. End-to-end latency distribution of MediExplain AI pipeline

REFERENCES

[1] K. Singhal et al., "Med-PaLM: Scaling and Aligning Language Models for Medical Question Answering," Nature Medicine, vol. 29, pp. 1320–1332, July 2023.
[2] A. Thirunavukarasu et al., "Large language models in medicine," Nature Medicine, vol. 29, pp. 1930–1940, 2024.
[3] J. Bai et al., "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," arXiv preprint arXiv:2308.12966, 2023.
[4] Y. Ye and R. Wang, "Medical Document OCR Using Deep Learning: A Survey," Int. Journal of Medical Informatics, vol. 158, 2025.
[5] Z. Gu et al., "Empowering large language models for automated clinical assessment," Artif. Intell. Med., vol. 162, April 2025.
[6] M. Jiang, A. Guo et al., "A Large Language Model–Based Method for Generating Patient-Centric Clinical Summaries," JAMIA, vol. 30, no. 11, 2023.
[7] D. Van Veen et al., "Adapted large language models can outperform medical experts in clinical text summarization," Nature Medicine, vol. 30, pp. 1134–1142, 2024.
[8] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Problems," arXiv preprint arXiv:2311.05232, 2024.
[9] A. Gala et al., "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages," Transactions of TMLR, 2023.
[10] M. Alzantot et al., "AI in Healthcare: Challenges and Solutions for Emerging Markets," IEEE Access, vol. 10, 2022.
[11] R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. ICDAR, 2007.
[12] H. Nori et al., "Can Generalist Foundation Models Outperform Special-Purpose Tuning? Case Study in Medicine," New England Journal of Medicine AI, 2023.
[13] S. Liu et al., "Multimodal Models in Healthcare: Methods, Challenges, and Future Directions," Information, vol. 16, no. 11, 2024.
[14] Q. He et al., "Quality of Answers of Generative Large Language Models vs Peer Users," JMIR, vol. 26, 2024.
[15] J. Achiam et al., "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
[16] P. Rajpurkar et al., "AI in Health and Medicine," Nature Medicine, vol. 28, pp. 31–38, 2022.
[17] T. Brown et al., "Language Models are Few-Shot Learners," in NeurIPS, 2020.
[18] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in ICML, 2021.
[19] N. Pawar, "Fine Tuning LLMs for Context-Aware Medical Report Summarization," DTU Major Project Report, 2025.
[20] C. Hsieh et al., "DALL-M: Context-Aware Clinical Data Augmentation with LLMs," arXiv preprint arXiv:2407.08227, 2025.