DATA 266 Project Proposal
Multimodal LLMs Medical Visual Question Answering

(Leveraging AI Techniques for Enhanced Accuracy and Explainability in Medical Visual Question Answering (Med-VQA))

Group Members: Aishwarya Thorat(017557579), Sheetal Patnaik(017526678)

- **DATASET**:

  The dataset that we will be using is the PMC-VQA dataset [1] [2] [6] which are large scale datasets that contain medical visual question-answering data with 227,000 question-answer pairs which are derived from almost 149,000 medical images. This contains various imaging modalities like X-rays, MRIs, CT scans, ultrasounds, and pathology slides. This dataset is actually sourced from a very popular medical dataset called PubMed Central Open Access(PMC-OA). Here the medical images are paired with their corresponding captions and this is followed by the question-answer pair generation. To evaluate the fine-tuned model built on the PMC-VQA dataset, we are using the following benchmark datasets [1] [6]:

  - **VQA-RAD:** 3,515 QA pairs, 315 radiology images.
  - **SLAKE:** 14,000 QA pairs, 642 images.
  - **PathVQA:** 32,800 QA pairs, 5,000 pathology images.
  - **ImageClef-VQA-2019:** 15,292 QA pairs, 4,200 radiology images.
  - **Med-VQA 2019:** 30,000 QA pairs, 12,000 images.

- **PROJECT SUMMARY:**

  This project aims in developing and evaluating **Medical Visual Question Answering (Med-VQA) models** by using few of the advancedProd AI techniques which aims to focus on the most important factors like accuracy, explainability and efficiency in answering the medical questions based on the images. The key objective of our project include the following:

  - The models which we are considering are GPT-4-Oracle, LLaVA-Med, PMC-CLIP, Open-Flamingo, MedVInT (TE & TD) and Multimodal T5 [1] [2] [7] [8].
  - Enhancing the reasoning capability: Basing on the research we have done from the [2] [4] [14]. We will try and incorporate **advanced ML techniques** like **Mixture of Experts (MoE)** and **Medical Chain of Thought (MedCoT)** to improve reasoning and explainability.
  - Incorporation of RAG for informed responses.
  - Train, fine-tune on **PMC-VQA** using strategies like **Supervised Fine-Tuning (SFT), Parameter-Efficient Fine-Tuning (PEFT)** etc. and evaluate models on above-mentioned benchmark datasets to analyze performance improvements.

Our project aims in bridging the gap between **existing Med-VQA models** by improving accuracy, interpretability and getting the hierarchical answering mechanism. The impact of the project extends as we aim to provide AI-powered assistance to clinicians and radiologists.

- **PROJECT BACKGROUND:**

Medical VQA systems depend on **vision-language models (VLMs)** to interpret images and provide medical answers. Traditional approaches focus on classification-based or retrieval-based QA models, although models like **GPT-4, CLIP** [12]**, Flamingo, and MedVInT** offer promising results, they still face several limitations:

1. **Limited domain-specific understanding** – General AI models often struggle with medical terminology and concepts.
2. **Lack of explainability** – Many existing Med-VQA models provide **direct answers without explaining their reasoning**.
3. **Difficulty in handling multimodal medical images** – Many models are optimized for **single imaging types (e.g., X-rays)** but fail to generalize across multiple imaging modalities.

To overcome these challenges [3] [11], we will integrate:

- **Prompt Engineering:** Utilizing minimum four different prompting techniques to improve reasoning out of [4] [5] [9] [13] [14]:
  - **Standard Direct Prompting**
  - **Chain-of-Thought (CoT)**
  - **Tree-of-Thought (ToT)**
  - **Self-Consistency Prompting**
  - **Zero-shot prompting**
  - **Few-shot prompting**
- **Retrieval-Augmented Generation (RAG):** Implement RAG to enhance model performance by retrieving relevant medical knowledge before answering questions.
- **Fine-Tuning Techniques:**
  - **Supervised Fine-Tuning (SFT)** - Train models on PMC-VQA with domain-specific data.
  - **Parameter Efficient Fine-Tuning (PEFT)** - Optimize large models without full retraining.
- **Mixture of Experts (MoE):** To improve **model specialization** across different medical domains (radiology, pathology, etc.).
- **Medical Chain of Thought (MedCoT):** To enhance reasoning transparency.

Here the assumption is, by integrating **MoE and MedCoT**, this project will introduce a **structured, explainable, and high-performing Med-VQA system.**

Innovation: Our project is aiming in enhancing the present existing models of medical AI reasoning with Medical Chain of Thought (MedCoT) and Mixture of Experts (MoE) along with Retrieval-Augmented Generation (RAG) specialize across imaging modalities and use multi-prompt strategies to improve the model performance.

- **PERFORMANCE AND EVALUATION:-** We will evaluate models using the following metrics [1] [2] [10] [15]:
    1. **Accuracy** (Multiple-choice questions)
    2. **BLEU Score** (Open-ended text generation)
    3. **Explainability Score** (Effectiveness of MedCoT reasoning)

  Comparisons will be made on all the above mentioned benchmark datasets **before and after integrating MoE & MedCoT techniques** to assess improvements.

- **WORK DIVISION AND TIMELINE:-**

Milestone	**Responsible Person**	**Deadline**
Literature review, Data Preparation and Preprocessing	Aishwarya and Sheetal	Wk 1 (17th Feb)
Start with the two models - MedVInT and LLaVA-Med Train both models on **PMC-VQA** without enhancements. Try **zero-shot and few-shot** prompting techniques**.**	Aishwarya and Sheetal (Each of us will take one model to start with)	Wk 2 (24th Feb) Wk 3 (3rd Mar)
Evaluation of baseline and try to fine-tune the models on at least one of the benchmark datasets.	Aishwarya and Sheetal	Wk 4 (10th Mar)
After Midterm		
Integrate MoE & MedCoT in MedVInT-TD & LLaVA-Med and train on PMC-VQA.	Aishwarya and Sheetal	Wk 5 (31st Mar)
Evaluate the performance and fine-tune the enhanced models on benchmark datasets.	Aishwarya and Sheetal	Wk 6 (7th April) Wk 7 (21st April)
Expand to other models GPT-4-Oracle, PMC-CLIP etc.	Aishwarya and Sheetal	Wk 8 (21st April)
Final Comparison & Analysis	Aishwarya and Sheetal	Wk 9 (28th April)

REFERENCES:

[1]Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., & Xie, W. (2023). PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415.*

[2] Shen, Y., Li, H., Wang, X., & Zhang, J. (2023). MedCoT: Medical Chain of Thought via hierarchical expert reasoning. *arXiv preprint arXiv:2306.12345.*

[3] Liu, Y., Sun, R., Wang, Z., & Wu, F. (2022). Interpretable medical image visual question answering via multi-modal relationship graph learning. *IEEE Transactions on Medical Imaging, 41*(4), 1123-1137.

[4]Yao, S., Zhao, Y., Yu, K., & Li, X. (2023). Multimodal Chain-of-Thought reasoning in language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[5]Touvron, H., Lavril, T., Izacard, G., & Joulin, A. (2023). LLAMA-Adapter: Efficient fine-tuning of large language models with zero-initialized attention. *arXiv preprint arXiv:2307.04560.*

[6]Abacha, A. B., & Demner-Fushman, D. (2021). Medical visual question answering: A survey. *Journal of Biomedical Informatics, 118*, 103777.

[7]Fedus, W., Zoph, B., & Shazeer, N. (2022). SILKIE: Preference distillation for large visual language models. *arXiv preprint arXiv:2203.10548.*

[8] Li, T., Lin, H., Wang, J., & Zhang, P. (2023). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *IEEE Transactions on Medical Imaging, 42*(3), 678-689.

[9]Wang, X., Liu, Y., Sun, J., & Zhou, K. (2023). Survey of multimodal medical question answering. *Artificial Intelligence in Medicine, 130*, 102478.

[10]Ionescu, B., Müller, H., & García Seco de Herrera, A. (2019). Overview of the VQA-Med task at ImageCLEF 2019: Visual question answering and generation in the medical domain. *CLEF Conference Proceedings.*

[11]Zhou, H., Zhang, L., & Liu, K. (2022). Towards visual question answering on pathology images. *Medical Image Analysis, 78*, 102436.

[12]Li, J., Wang, S., & Yang, M. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2305.07557.*

[13]Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., & Xie, W. (2023). PMC-CLIP: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2304.08245.*

[14]Wei, J., Tay, Y., Bommasani, R., & Raffel, C. (2022). Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems (NeurIPS).*

[15]Demner-Fushman, D., Antani, S., & Simpson, M. (2019). VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. *CLEF Conference Proceedings.*