PROPOSED RESEARCH PROJECT

Title of proposed research project	**Student**
A comprehensive exploration and evaluation of foundational and subspecialized Large Language Models (LLMs) in healthcare and medicine	Felicia Liu 1006950042

Outline of proposed research project

Background Context and Motivation

The recent and rapid development of large language models (LLMs) has created many opportunities for improved healthcare; LLMs are assisting clinical decision support systems [1] and patient triaging [2], addressing primary care concerns [3], and summarizing key findings [4], as well as automating clinical note creation [5][6] and assisting with patient counseling and diagnostic recommendations [6][7]. Current research of LLMs in healthcare involves designing LLMs for natural language and text-based tasks, and assumes the use of large-scale datasets [8]. As a result, there remains a lack of comprehensive overviews evaluating accuracy, robustness, and the overall utility of medical LLMs when available data is limited, and beyond text-based problems. The motivation for this research is to cultivate a more informed understanding of LLMs and their capabilities and effectiveness in healthcare beyond the often-explored areas involving large-scale text-based datasets. Assessing when specialized models are necessary compared to when generalizable foundational models suffice, as well as exploring LLM capabilities in non-text based tasks, addresses a key gap in existing LLM research.

Research Objectives and Hypothesis

There are two research objectives: 1) Determine how general-purpose LLMs compare with specialized models in their effectiveness across various healthcare tasks. Specifically, we aim to explore whether general LLMs, such as Clinical Camel [6], are sufficient for certain tasks "out of the box" or if adding specialized training on small/niche datasets significantly improves performance. We hypothesize that fine-tuned models may demonstrate improved accuracy and robustness in specialized cases where available data is limited. 2) Assess the utility of LLMs for vision-based tasks, such as classifying brain tumor pathology [9]-[13] and MGMT promoter methylation status in images, and tumor segmentation [10]-[14]. We hypothesize that LLMs may struggle on image tasks unless adapted or fine-tuned.

Methods and Procedures

To compare general-purpose LLMs and specialized models in task effectiveness, we will test general LLM models, starting with Clinical Camel [6], an open-source foundational medical LLM, on the MIMIC-CXR dataset [15], the largest open source dataset of chest X-rays with radiology reports. Then, we will fine tune these LLMs to assess the extent to which specialized training improves performance, leveraging EleutherAI's evaluation framework [16]. To explore LLMs in non-traditional vision-based tasks, we will modify the base architecture to accept different image inputs, and output classification predictions. These models will be evaluated using the Brain Tumor Segmentation dataset [9]-[14], the most frequently benchmarked open source dataset of brain magnetic resonance images. For image classification, performance will be measured using AUC, sensitivity, specificity, and F1 score, while Dice coefficient and 95% Hausdorff distance will be used for image segmentation. These metrics provide a comprehensive assessment of model performance for both image and text domains.

Significance

This research aims to provide a clearer understanding of when and where LLMs are most effective in healthcare. By comparing general and specialized models, we will establish a framework for selecting the optimal models and training approaches for various medical tasks, highlighting when fine-tuning is most beneficial. Extending LLMs to image classification and understanding when they perform well could revolutionize fields such as radiology, leading to more accurate and personalized healthcare.

References

[1] K. J. Prabhod, “Integrating Large Language Models for Enhanced Clinical Decision Support Systems in Modern Healthcare,” *J. Mach. Learn. Healthc. Decis. Support*, vol. 3, no. 1, Art. no. 1, Jun. 2023.

[2] L. Masanneck *et al.*, “Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study,” *J. Med. Internet Res.*, vol. 26, no. 1, p. e53297, Jun. 2024, doi: 10.2196/53297.

[3] H. Mondal, R. De, S. Mondal, and A. Juhi, “A large language model in solving primary healthcare issues: A potential implication for remote healthcare and medical education,” *J. Educ. Health Promot.*, vol. 13, no. 1, p. 362, Sep. 2024, doi: 10.4103/jehp.jehp_688_23.

[4] K. Singhal *et al.*, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.

[5] D. Yuan *et al.*, “A Continued Pretrained LLM Approach for Automatic Medical Note Generation,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 565–571. doi: 10.18653/v1/2024.naacl-short.47.

[6] A. Toma, P. Lawler, J. Ba, R. Krishnan, B. Rubin, and B. Wang, *Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding*. 2023. doi: 10.48550/arXiv.2305.12031.

[7] E. C. Stade *et al.*, “Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation,” *Npj Ment. Health Res.*, vol. 3, no. 1, pp. 1–12, Apr. 2024, doi: 10.1038/s44184-024-00056-z.

[8] Z. A. Nazi and W. Peng, “Large Language Models in Healthcare and Medical Domain: A Review,” *Informatics*, vol. 11, no. 3, Art. no. 3, Sep. 2024, doi: 10.3390/informatics11030057.

[9] S. Bakas *et al.*, “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge,” Apr. 23, 2019, *arXiv*: arXiv:1811.02629. doi: 10.48550/arXiv.1811.02629.

[10] S. Bakas *et al.*, “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Sci. Data*, vol. 4, p. 170117, Sep. 2017, doi: 10.1038/sdata.2017.117.

[11] B. H. Menze *et al.*, “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[12] S. Bakas, H. Akbari *et al.*, "Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection", The Cancer Imaging Archive, 2017. DOI: 10.7937/K9/TCIA.2017.KLXWJJ1Q

[13] S. Bakas, H. Akbari *et al.*, "Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection", The Cancer Imaging Archive, 2017. DOI: 10.7937/K9/TCIA.2017.GJQ7R0EF

[14] U. Baid *et al.*, “The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification,” Sep. 12, 2021, *arXiv*: arXiv:2107.02314. doi: 10.48550/arXiv.2107.02314.

[15] A. E. W. Johnson *et al.*, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019, doi: 10.1038/s41597-019-0322-0.

[16] L. Sutawika *et al.*, *EleutherAI/lm-evaluation-harness: v0.4.4*. (Sep. 05, 2024). Zenodo. doi: 10.5281/zenodo.13694023.