Project Proposal: A Reimplementation of Integrating ChatGPT into Secure Hospital Networks: A Case Study on Improving Radiology Report Analysis

Lokanath Das, Jacob Fuehne, Jared Backofen

Department of Computer Science, University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
{ldas2, jfuehne2, jaredb3}@illinois.edu

Abstract

This proposal outlines a plan to reproduce a healthcare NLP study that distills a powerful cloud LLM into an on-premise model for radiology report sentence classification under strict privacy constraints. We summarize the problem, approach, novelty, data access, feasibility, and implementation choices.

Introduction

Hospitals face challenges when it comes to utilizing state-of-the-art LLMs while also preserving the privacy of patients. Compliance requirements such as HIPAA and GDPR restrict hospitals from sending health information to some 3rd-party cloud-hosted LLM services where they do not have strict privacy guarantees of the data. The target paper by Kim *et al.* (1) proposes getting around this limitation by transferring labels generated by a high performance insecure LLM to a smaller, locally-hosted and secure LLM model through knowledge distillation.

Problem Statement

The paper addresses how to retain the analytical performance of a powerful cloud-hosted LLM while meeting hospital privacy requirements that prohibit sending sensitive information off-premise. The core challenge is to distill the knowledge of the cloud model into a smaller model that can run within a hospital's secure network, improving the performance of the secure model while maintaining the patient data security.

Methodology

Specific Approach

The authors use knowledge distillation from the, at the time, SOTA cloud based ChatGPT 3.5 model into smaller, on-premise student models for sentence-level classification of radiology reports. Their method emphasizes Sentence-level Knowledge Distillation (S-KD), introducing a ternary label space (normal / abnormal / uncertain) to explicitly capture ambiguity. Training combines cross-entropy loss with a supervised contrastive loss to better separate classes in representation space. Student backbones include medically oriented BERT-family models such as RadBERT-RoBERTa, BioMed-RoBERTa, BioBERT, and BlueBERT. Performance is evaluated using standard clinical classification metrics: Accuracy, Sensitivity, Specificity, and AUC.

Novelty / Relevance / Hypotheses to be Tested

Novelty/Relevance. The approach operationalizes privacy-preserving deployment by transferring capability from a cloud LLM to an on-premise model, aligning with real hospital constraints. Key innovations include sentence-level distillation (finer granularity than document-level) and an explicit uncertain class to handle ambiguous clinical language. The supervised contrastive component is relevant for sharpening class boundaries in subtle radiology statements.

Why better than baselines. Compared to document-level KD or binary labeling, S-KD with a ternary scheme and contrastive learning is expected to yield better discrimination and safer abstention behavior, improving clinical reliability.

Hypotheses. For the paper hypotheses, they claim that, sentence-level KD improves downstream classification over document-level KD, that adding an "uncertain" class reduces harmful misclassifications without materially degrading overall performance, and that contrastive loss enhances separability and metrics such as AUC.

Ablations / Extensions Planned

We plan to run a few small ablations to see how specific design choices affect performance. For example, we will test training without the contrastive loss and try collapsing the ternary labels into binary to compare results. Additionally, we will experiment with alternative backbones such as BioClinical-BERT (4) and DeBERTaV3 (3). These will help us understand which components contribute most to model accuracy and clinical reliability.

Data Access and Implementation Details

Description of How You Will Access the Data/Model

We will use the MIMIC-CXR corpus (de-identified public chest radiographs with reports). Access is granted via a data use agreement, PhysioNet onboarding. Following the paper, we will reproduce ”teacher” labeling using a high quality cloud based LLM to derive sentence-level ternary labels, then train local ”student” models such as RadBERT-RoBERTa. We were able to find the github code referenced in the paper through one of the author’s personal githubs, and we also found a model checkpoint reference to the trained RadBERT-RoBERTa model that they created that we can use to compare results against.

Discussion of the Feasibility of the Computation

Our team owns an Nvidia 4070, a 5080, and a MacBook, so we’ll be using a combination of Nvidia GPUs and possibly Google Colab to do our training of student models. As for the teacher aspect of the architecture, we will be replacing the ChatGPT 3.5 model (which is no longer offered by OpenAI) with Llama 3.3 70B versatile, hosted by Groq cloud. Groq offers an unlimited free trial, as long as you remain below a rate limit, so this should reduce our API budget to $0. Overall, we expect to face no issues running the project, and to pay little to no cost. After doing calculations, we also determined that it may also be affordable to run ChatGPT 5 as the teacher.

Statement of Whether You Will Use the Existing Code or Not

We will use the available GitHub code that the paper published, as well as the MIMIC data from Physio (2).

References

[1] Kim, K., Park, J., Langarica, S., Alkhadrawi, A. M., & Do, S. (2024). Integrating ChatGPT into Secure Hospital Networks: A Case Study on Improving Radiology Report Analysis. *CHIL 2024 / arXiv:2402.09358 [cs.AI].* https://doi.org/10.48550/arXiv.2402.09358

[2] Johnson, A. E. W., et al. (2019). MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv:1901.07042*.

[3] He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543*.

[4] Sounack, T., Davis, J., Durieux, B., Chaffin, A., Pollard, T. J., Lehman, E., Johnson, A. E. W., McDermott, M., Naumann, T., & Lindvall, C. (2025). BioClinical ModernBERT: A State-of-the-Art Long-Context Encoder for Biomedical and Clinical NLP. *arXiv:2506.10896*. https://arxiv.org/abs/2506.10896