College of Engineering and Computer Science
Office of the Dean
777 Glades Road, EE96, Room 308
Boca Raton, FL 33431
561.297.3400

Announces the Ph.D. Dissertation Defense of

Apoorv Tripathi

for the degree of Master of Science (MS)

LLM for Clinical Named Entity Recognition: A Study on RAG for PubMed and UMLS

03/24/2026,
3pm to 4pm EE96, 405
777 Glades Road
Boca Raton, FL

DEPARTMENT: Electrical Engineering and Computer Science

ADVISOR: Dr. Xingquan (Hill) Zhu, PhD. PH.D.

SUPERVISORY COMMITTEE: Yufei Tang, PhD, Sareh Taebi, PhD

ABSTRACT OF DISSERTATION

THE FIRST STEP OF BIOMEDICAL NLP IS RECOGNIZING CLINICAL NAMED ENTITIES, WHICH CONSIST OF IDENTIFYING AND CATEGORIZING A VARIETY OF CLINICAL ENTITIES SUCH AS DISEASES, SYMPTOMS, GENETICS, DIAGNOSTIC TESTS, PROCEDURES, ETC. FROM A BODY OF UNSTRUCTURED CLINICAL TEXT. THIS STUDY PRESENTS A PUBMED BASED RETRIEVAL AUGMENTED GENERATION FRAMEWORK WHICH IMPROVES THE PERFORMANCE OF THE LARGE LANGUAGE MODELS TO IDENTIFY CLINICAL ENTITIES BY PROVIDING CONTEXT. IN PARTICULAR, THE FRAMEWORK CONSISTS OF A TWO-STAGE PIPELINE, WHERE CANDIDATE TOKENS ARE IDENTIFIED FROM INITIAL LLM-BASED CLASSIFICATION AND REFINED WITH RETRIEVED CONTEXT FROM EITHER PUBMED OR UMLS. THE PROPOSED FRAMEWORK IS ASSESSED ACROSS TWO ESTABLISHED BIOMEDICAL DATASETS, THE NCBI DISEASE CORPUS (BINARY CLASSIFICATION) AND MEDMENTIONS (MULTI-CLASS CLASSIFICATION) AND ASSESSED USING THREE LLMS, LLAMA-70B, QWEN-35B, AND GPT-5. THE RESULTS OF THE EVALUATION INDICATE THAT RETRIEVAL-BASED ON PUBMED-SOURCE CONSISTENTLY IMPROVED OR MAINTAINED F1 SCORES WHILE UMLS-BASED RETRIEVAL DECREASED PERFORMANCE ACROSS NEARLY ALL CONFIGURATIONS. THEREFORE, THE RESULTS INDICATE THAT RETRIEVAL-SOURCE SELECTION IS A CRITICAL ASPECT OF RETRIEVAL-AUGMENTED GENERATION BIOMEDICAL NLP SYSTEMS.