Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences (SD TIB PAN)

SUBJECT:
Verifiable Biomedical Question Answering: Retrieval, Source Attribution and Hallucination Control in Large Language Models

SUPERVISOR:
Prof. Svitlana Vakulenko (main supervisor);
WU (Vienna University of Economics and Business), Vienna, Austria
svitlana.vakulenko@wu.ac.at

Wojciech Kusa, PhD (auxiliary supervisor).
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB), ul. Kolska 12, Warszawa
wojciech.kusa@nask.pl

DESCRIPTION:
Large language models (LLMs), including general models such as GPT-4o and domain-adapted models such as Med-PaLM 2, Meditron and PMC-LLaMA, now match specialised systems on many biomedical question answering benchmarks. Reliability, however, lags behind fluency. Models still fabricate references, attach citations to claims the cited paper does not support, and omit important studies. Shared-task results confirm that reference attribution remains unsolved: even strong systems in the TREC Biomedical Generative Retrieval track fail to support a substantial share of generated sentences. Such errors are particularly damaging in medicine, where an unsupported statement can mislead clinicians, researchers and patients. This project will develop and evaluate methods that make biomedical answer generation verifiable, so that every claim can be traced to a specific passage in the published literature.
The project will compare sparse, dense and hybrid retrieval over PubMed and related corpora, including biomedical retrievers such as MedCPT, and will study how the choice of corpus, retriever and re-ranker affects the factual quality of the final answer rather than ranking quality alone. It will also examine evidence selection: which passages should be shown to the model when studies disagree, when evidence is outdated, or when study designs differ in strength.
The project will develop a citation planner that decomposes an answer into individual claims and links each claim to supporting passages before or during generation. Candidate citations will be checked with a verification model, for example a natural language inference classifier trained on biomedical text, so that unsupported links can be revised or removed.
Performance will be measured with citation precision and recall, following the evaluation protocols established by ALCE and the TREC BioGen track.
The project will fine-tune models on verified, attributed answers and then apply preference optimisation, such as Direct Preference Optimization, using preference pairs derived from automatic factuality signals including FActScore and entailment-based checks. A related question is when a model should abstain or express uncertainty because the retrieved evidence is insufficient, rather than producing a confident answer. Evaluation will combine

established biomedical benchmarks (MIRAGE, MedHallu, TREC BioGen) with expert assessment, since automatic factuality metrics are themselves imperfect and their agreement with clinical judgement needs to be quantified.

REQUIREMENTS:

- MSc degree in computer science, computational linguistics, bioinformatics, artificial intelligence or a related field
- Strong programming skills in Python, with experience in PyTorch and the Hugging Face ecosystem
- Familiarity with information retrieval and evaluation methodology
- Knowledge of current research on LLMs, retrieval-augmented generation and model fine-tuning
- Interest in biomedical text and willingness to work with clinical and life-science experts
- Advanced level of English (spoken and written)

BIBLIOGRAPHY:

- Asai A. et al. (2024) Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflection
- Bohnet B. et al. (2022) Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models
- Gao T. et al. (2023) Enabling Large Language Models to Generate Text with Citations
- Gupta D. et al. (2024) Overview of the TREC 2024 Biomedical Generative Retrieval (BioGen) Track
- Gupta D. et al. (2025) Overview of the TREC 2024 Biomedical Generative Retrieval (BioGen) Track
- Jin Q. et al. (2023) MedCPT: Contrastive Pre-trained Transformers with Large-Scale PubMed Search Logs for Zero-Shot Biomedical Information Retrieval
- Min S. et al. (2023) FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation
- Pal A. et al. (2023) Med-HALT: Medical Domain Hallucination Test for Large Language Models
- Pandit S. et al. (2025) MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models
- Rafailov R. et al. (2023) Direct Preference Optimization: Your Language Model is Secretly a Reward Model
- Singhal K. et al. (2023) Towards Expert-Level Medical Question Answering with Large Language Models
- Xiong G. et al. (2024) Benchmarking Retrieval-Augmented Generation for Medicine