Statement of Purpose

Motivation

My interest in machine learning arose when I worked under Prof. Nick Hawes on reinforcement learning [1]. From the project, I realized the power of machine learning from an application perspective. I am interested in developing machine learning techniques that can solve real-world problems in natural language processing (NLP) and computer vision (CV) domains. I worked with Prof. Anastasia Borovykh to study the intrinsic structure of the model in the latent space [2]. I also found the application of machine learning in the biomedical domain very intriguing. In recent years, many general large language models (LLMs), such as BERT-series and GPT-series, have been presented. A lot of biomedical LLMs are developed on the foundation of general LLMs and can assist medical professionals in clinical applications. Still, many problems remain:

1. **Hallucination.** Hallucination of LLMs refers to the phenomenon where the generated output contains inaccurate or nonfactual information. It can be categorized into intrinsic and extrinsic hallucinations. Intrinsic hallucination refers to generating outputs logically contradicting factual information while extrinsic hallucination happens when the output generated cannot be verified. When integrating LLMs into the medical domain, fluent but nonfactual LLM hallucinations can lead to the dissemination of incorrect medical information, which can cause misdiagnoses, inappropriate treatments, and harmful patient education.
2. **Domain Data Limitations.** Current datasets in the medical domain remain relatively small compared to datasets used to train general-purpose LLMs. The medical knowledge domain is vast; existing datasets are limited and do not cover the entire space. Current benchmarks fail to evaluate important LLM-specific metrics such as trustworthiness, faithfulness, helpfulness and explainability.

I worked with students at Oxford to come up with a survey about LLMs in the biomedical domain [3] and found the following area interesting for further studies:

Reliable biomedical LLMs: Current solutions to mitigate LLM hallucination can be categorized into training-time correction, generation-time correction, and retrieval-augmented correction. The first solution, training-time correction, aims to mitigate hallucination by adjusting model weights and thus reducing the probability of generating hallucinated outputs. Examples of training-time correction include factually consistent reinforcement learning and contrastive learning. Another solution to reduce hallucination is to add a 'reasoning' process to the LLM inference to ensure reliability. Methods include drawing multiple samples or using a confidence score to identify hallucinations before the final generation. A third approach is the retrieval-augmented correction method, which utilizes external resources to help mitigate hallucination. For example, using factual documents as prompts or chain-of-retrieval prompting technique. Benchmarks such as TruthfulQA and HaluEval evaluate more LLM-specific metrics, such as truthfulness, but fail to cover the medical domain. Future research is necessary to develop more medical and LLM-specific benchmarks and metrics.

Multimodal in biomedical LLMs: Multimodal LLMs (MLLMs) are LLM-based models designed to perform multimodal tasks. While LLMs primarily address NLP tasks, MLLMs support a broader range of tasks, such as comprehending the underlying meaning of a meme and generating website codes from images. This versatility suggests promising applications of MLLMs in healthcare. For example, recent works have introduced MLLM-based frameworks

that integrates vision, audio, and language inputs for automated diagnosis in dentistry and cardiology. However, there are only very few medical LLMs that can process time series data, such as electrocardiograms (ECGs) and sphygmomanometers (PPGs). These time series data are important for medical diagnosis and monitoring. The multimodal nature of MLLM also introduces unique issues, including limited perception capabilities, fragile reasoning chains, sub-optimal instruction-following ability, and object hallucination. Therefore, more research is needed to address these issues.

Machine Learning in the preclinical area: Recent research focuses more on the question answering ability of LLMs. However, medical LLMs can also serve as a tool for preclinical analysis. PICO is a framework for creating specific clinical questions in Evidence-Based Medicine (EBM). There are only limited LLMs that evaluate their performance in generating PICO reports. Current metrics for PICO are adopted from entity extraction and fail to capture the intended meaning or the relevance of the relationships between them, so the need for other evaluation metrics arises. Another underdeveloped area is the translation of animal preclinical models to humans. Four categories of validity are considered in preclinical animal studies. Various validity issue, such as whether the results can be generalised to experiments conducted in other population and time points, affects the performance of medical models for further downstream tasks. More studies are needed to analyze the bias in preclinical studies and merge it into the potential future medical NLP tasks.

Obejectives

My objective in the long term is to build reliable machine learning tools in the biomedical domain that can significantly save medical researchers' time and reform the current diagnosis paradigm. Current studies concentrate more on the experimental aspect rather than the theoretical aspect. To achieve reliable models, I believe we need to understand the intrinsic structure and reveal the black box in an overcomplicated model. As a formal mathematics and a current computer science master, I believe I can contribute productively to the research in NLP at XXX both experimental and theoretical aspects. Due to the financial-consuming nature of current NLP studies, such research is impossible for an individual to carry out, and that's why I want to work in the XXX lab. I also believe that as a well-reputed university, XXX's interdisciplinary departments and dynamic academic atmosphere make it my top choice to continue my research.

References

[1] Hongjian Zhou, Boyang Gu, and Chenghao Jin. Reinforcement learning approach for multi-agent flexible scheduling problems. In *Journal of Physics: Conference Series*, volume 2580, page 012053. IOP Publishing, 2023.

[2] Boyang Gu and Anastasia Borovykh. On original and latent space connectivity in deep neural networks. *arXiv preprint arXiv:2311.06816*, 2023.

[3] Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*, 2023.