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 [3]. 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 [4].

I worked with students at Oxford to come up with a survey about LLMs in the biomedical domain [5] and found the following area interesting for further studies:

Reliable biomedical LLMs There are three main ways to reduce hallucinations. The first solution is to adjust model weights and reduce the probability of hallucinations. Examples of training-time correction include factually consistent reinforcement learning [6] and contrastive learning[7]. Another solution is to add a 'reasoning' process to the LLM inference to ensure reliability. Methods include drawing multiple samples [8] or using a confidence score to identify hallucination 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 [9] or chain-of-retrieval prompting technique [10]. 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 While LLMs primarily address NLP tasks, Multimodal LLMs (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 integrate vision, audio, and language inputs for automated diagnosis in dentistry and cardiology [11]. However, there are only very few medical LLMs that can process time series data, such as electrocardiograms (ECGs) and sphygmomanometers (PPGs), which are important for medical diagnosis and monitoring. The multimodal nature of MLLM also introduces unique issues, including limited perception capabilities [11][12], fragile reasoning chains [13], sub-optimal instruction-following ability [13], and object hallucination [12].

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 [14]. 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 [15]. More studies are needed to analyze the bias in preclinical studies and merge it into the potential future medical NLP tasks.

Objectives 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 CambridgeLTL in 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 CambridgeLTL lab. I also believe that as a well-reputed university, Cambridge'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] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[4] Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. Faithful ai in medicine: A systematic review with large language models and beyond. *medRxiv*.

[5] 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.

[6] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*, 2023.

[7] I-Chun Chern, Zhiruo Wang, Sanjan Das, Bhavuk Sharma, Pengfei Liu, Graham Neubig, et al. Improving factuality of abstractive summarization via contrastive reward learning. *arXiv preprint arXiv:2307.04507*, 2023.

[8] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[9] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

[10] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

[11] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29, 2023.

[12] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[14] Qianying Wang, Jing Liao, Mirella Lapata, and Malcolm Macleod. Pico entity extraction for preclinical animal literature. *Systematic Reviews*, 11(1):1–12, 2022.

[15] Qianying Wang. Preclinical risk of bias assessment and pico extraction using natural language processing. 2022.