YU YIN

Email: yinyu201906@gmail.com | Homepage

EDUCATION

Imperial College London UK
MSc in Applied Computational Science & Engineering | Distinction Degree 09/2023 - 09/2024

University of Liverpool UK
BSc in Computer Science | First Class Degree 09/2021 - 07/2023

Xi'an Jiaotong-Liverpool University China
BSc in Information Computer Science 09/2019 - 07/2021

Relevant Coursework

- **Undergraduate:** Linear Algebra, Multivariable Calculus, Discrete Mathematics and Statistics, Complexity of Algorithms, Computer Systems, Computer Programming, Programming in Java, Advanced Object Oriented C Languages, Database Development, Data Structures, Advanced Artificial Intelligence, Software Engineering, Decision, Computation & Language, Principles of C and Memory Management and Game Theory
- **Master:** Computational Mathematics, Data Science and Machine Learning, Deep Learning, Advanced Programming, Patterns for Parallel Programming, Applying Computational/Data Science and Modern Programming Methods

RESEARCH INTERESTS

- Natural Language Processing and LLMs, Medical Imaging and multimodal data analysis, Biomedical Text Mining and Information Retrieval, Artificial Intelligence-assisted Biomedical Applications

PUBLICATION

- ***Y. Yin****, H. Kim*, X. Xiao, C. Wei, Z. Lu, H. Xu, M. Fang, Q. Chen. Augmenting Biomedical Named Entity Recognition with General-domain Resources. (Accepted in Journal of Biomedical Informatics 2024)
- Z. Qin*, ***Y. Yin****, X. Wu, N. Liu, and X. Zhang, Q. Chen. LMOD: A Large Multimodal Ophthalmology Dataset and Benchmark for Vision–Language Models. (Accepted in NAACL Findings 2025)
- W. Han, M. Fang, ***Y. Yin***, Z. Song, Q. Chen, M. Pechenizkiy. MedINST: Meta Dataset of Biomedical Instructions. (Accepted in EMNLP Findings 2024)
- ***Y. Yin***, J. Zhao, Q. Zhang, M. Pechenizkiy, and M. Fang. CovidNews: Answering Questions on Multilingual Covid News. (Under Edition)

RESEARCH

Augmenting Biomedical Named Entity Recognition with General-domain Resources
Researcher, Yale 02/2024 - 06/2024
Developed a novel and effective multi-task learning method that incorporates general-domain named entity recognition (NER) datasets to improve biomedical named entity recognition (BioNER)

- Revealed significant limitations associated with multi-task learning when applied to various BioNER datasets, particularly concerning label ambiguity and convergence challenges
- Conducted multi-task learning to train a pre-trained biomedical language model utilizing both the target BioNER dataset and the general-domain NER dataset. Subsequently, Fine-tuned the models specifically for the BioNER dataset
- Utilized widely recognized general-domain NER datasets such as CoNLL2003 and GUM to demonstrate an average enhancement of 0.9% average improvement over baseline models across eight distinct entity types. Notably, Our model exhibited significant improvements in F1 scores for datasets characterized by a limited number of training instances

- Performed an ablation study to compare the performance of models trained on individual entity types against those trained on all entity types. Examined and identified the factors contributing to the suboptimal performance observed in specific instances

LMOD: A Large Multimodal Ophthalmology Dataset and Benchmark for Vision–Language Models

Researcher, Yale 02/2024 - 06/2024

Developed a comprehensive dataset and benchmark to evaluate large vision-language models (LVLMs) using ophthalmology images, with an emphasis on anatomical understanding, diagnostic analysis, and demographic extraction

- Curated 21,993 images derived from five distinct ophthalmic imaging modalities: optical coherence tomography (OCT), scanning laser ophthalmoscopy (SLO), ocular photographs, surgical scenes, and color fundus photographs (CFP) sourced from existing datasets. The dataset offers multi-granular annotations, including bounding boxes, region annotations, and image-level annotations. The pipeline is readily applicable to new datasets
- Benchmarked 13 state-of-the-art (SOTA) LVLMs across three key areas (1) anatomical recognition, which quantifies the effectiveness of LVLMs at the regional level; (2) disease diagnosis and classification, which assesses the capability of LVLMs in diagnosing and classifying ophthalmic diseases; and (3) demographic information extraction, aimed at determining whether LVLMs can infer demographic characteristics such as age and gender from ophthalmic images
- Generated bounding boxes and corresponding labels, along with the coordinates of the regions of interest, derived from the original open-source datasets for the purpose of evaluating anatomical recognition
- Formulated a comprehensive classification including both the staging of macular holes into four distinct stages and the classification of glaucoma, with a focus on evaluating the performance of models in predicting the appropriate stage of macular holes and determining the presence of glaucoma based on the visual characteristics presented in ophthalmic images
- Evaluated the capability of various models to accurately infer these biomarkers from ophthalmological images. Approached sex classification as a binary task, whereas age estimation employed a detailed categorization scheme, segmenting patients into distinct age groups
- Performed extensive experiments with 13 SOTA LVLMs to demonstrate their limitations in comprehensively understanding the complexities associated with ophthalmological images, particularly evident in their capabilities related to spatial reasoning, diagnostic analysis, and the management of out-of-domain queries

A Multimodal Biomedical Foundation Model Pretrained on Image-text Pairs on PMC

Researcher, Yale 07/2024 - 11/2024

Developed a comprehensive pipeline for the curation of PMC image-text pair in the context of medical content, consisting of five stages: (1) detection of medical figures; (2) detection of compound figures; (3) separation of compound figures; (4) alignment of subfigures with their corresponding subcaptions; (5)classification of subfigures based on medical criteria. This pipeline enabled the generation of a high-quality image-text dataset subsequently used for vision language model (VLM) fine-tuning. Our methodology exhibited superior performance relative to previous approaches, notably exceeding the capabilities of the SOTA GPT-4v model

- Conducted an extensive literature review to identify existing gaps in the training of biomedical VLMs, emphasizing the urgent need for high-quality medical image-text datasets. This review underscored data quality issues present in open-source projects such as LLaVA-Med and PMC-CLIP, as well as the unavailability of the dataset from the BiomedCLIP project, despite it notable performance
- Developed a systematic curation process for the management of image-text pairs obtained from PMC-OA, ensuring efficient processing while meticulously excluding non-medical, compound images, and misaligned subfigure subcaptions
- Led the initial screening phase by fine-tuning multiple language models, including PubMedBERT, BioLinkBERT, and Bioformer (8L and 16L), utilizing synthetic data generated by GPT-4 to classify images as medical or non-medical images based on their captions. PubMedBERT was ultimately chosen due to its optimal balance of performance and throughput
- Led the task of aligning subcaptions with subfigures, which comprised two primary components: (1) training the LLaVA-LLaMA-3-8B VLM using synthetic data alongside the ImageCLEF dataset for the purpose subcaption-subfigure alignment,; (2) establishing baseline comparisons with PMC-CLIP and BioMedCLIP
- Implemented a hybrid approach for caption separation by integrating regex pattern matching with POS tagging, significantly enhancing alignment accuracy. Augmented this solution across various baseline VLM
- Conducted the fine-tuning of LLaVA-LLaMA-3-8B utilizing synthetic data and in conjunction with the ImageCLEF dataset. Performed a comprehensive comparative analysis against SOTA models, including PMC-CLIP, BioMedCLIP and GPT-4v. Our model demonstrated exceptional performance, surpassing even GPT-4v in tasks related to medical image-text alignments

WORK

Research Assistant| Yale, New Haven, USA 05/2024 - present

Development of a Multi-agent Medical Support Framework

- Identified a critical gap in current medical agent systems through a comprehensive literature review: multi-round role-playing conversations in medical QA tasks often lead large language models (LLMs) to over-rely on their intrinsic knowledge, potentially generating unreliable responses
- Integrated a retrieval-augmented generation (RAG) system that leverages external reliable information sources, significantly enhancing the accuracy and reliability of medical responses
- Developed a sophisticated multi-agent system using the AutoGen framework, integrating both proprietary models (GPT-3.5, GPT-4o) and open-source LLMs (LLaMA-3-8B) with RAG techniques to enable dynamic and precise information retrieval
- Implemented a novel multi-agent pipeline enabling collaborative debate among LLMs, where agents assume different medical roles and leverage both intrinsic knowledge and external databases—including context-specific evidence from the MedRAG corpus (PubMed, Wikipedia, and medical textbooks)—to generate well-referenced medical responses
- Prompted agents to generate detailed reference lists and accurately distinguish sources of external snippets, reducing hallucinations and increasing trustworthiness
- Conducted rigorous experiments across multiple medical QA tasks (MedQA, MedMCQA, MMLU, PubMedQA, BioASQ) and a specialized ophthalmology use case, demonstrating significant improvements in accuracy, reference validity, and hallucination mitigation

Large Language Model Memorization and Privacy Detection

- Conducted extensive literature review that revealed significant memorization risks in medical LLMs, particularly concerning the potential leakage of sensitive patient data such as Electronic Health Records (EHR)
- Implemented a comprehensive LLM inference pipeline using the VLLM framework, incorporating evaluation metrics including consecutive token count, ROUGE score, BLEU score, and BERT score to rigorously assess models' memorization ability
- Designed a systematic memorization detection protocol for pre-trained LLMs, integrating models like MEDITRON, LLaMA1, and LLaMA2. Sampled and processed test datasets from sources using various prompt instructions, splitting inputs and ground truths to enable detailed comparisons for memorization and privacy assessment
- Curated fine-tuned datasets for medical QA tasks to investigate memorization in fine-tuned LLMs with the pre-trained setting.
- Extensive experiments demonstrated that both pre-trained and fine-tuned LLMs exhibit notable memorization of training data, with fine-tuned models, in particular, indicating memorization rates exceeding 50%
- Manually analyzed model outputs to uncover instances of training data leakage involving medical entities, and explored mitigation strategies such as applying random masked tokens in training datasets to reduce memorization in fine-tuned LLMs

Software Engineer Intern | Chengdu Yingnuo Industrial Co., Ltd., Sichuan, China 07/2021 - 08/2021

- Designed and implemented a set of user interfaces with customer requirements
- Finished the writing of user research reports and feasibility analysis reports by collecting and analysing pertinent information regarding user characteristics, emotions, habits, psychology, and needs
- Implemented the functional aspects of user interfaces, resulting in an enhanced interactive experience characterized by a logical layout and user-friendly interaction modes
- Conducted a series of test cases to debug each function and ensure its proper operation; passed the user acceptance testing prior to delivery
- Integrated the user interfaces with industrial systems to facilitate the identification of shadow points

MISCELLANEOUS

- Programming Languages: C/C++, ,Java, Python, R, MySQL, HTML, CSS, JavaScript
- Libraries: HuggingFace, PyTorch, NumPy, SciPy, Pandas, Scikit-Learn, Keras