John (Chenxi) Song

Email: chs342@pitt.edu GitHub: johncxsong.github.io Locations: Pittsburgh, PA

Summary

1.5+ years research assistant experience in **AI**, **machine learning**, and bioinformatics. Experience in implementing models, including a Java-bases machine learning Bayesian Network for flu diagnosis, and contributing to projects in **large language models** (LLMs) reasoning and pathology images analysis. Passionate about **leveraging AI** and machine learning to drive innovation across industries and create solutions that **benefit society**.

Education

University of Pittsburgh Pittsburgh, PA
M. S. in Information Science, Big Data Analytics Aug, 2022- Dec, 2023
GPA: 3.67/4.0

Geneva College Beaver Falls, PA
B. S. in Computer Science Aug, 2016- Dec, 2018
B. S. in Engineering

North Seattle College Seattle, WA
Associate of Science Degree April, 2015- June, 2016

Fundamental of Engineering Exam (FE) Mechanical Nov. 2018
Credential ID 19-633-53

Publications

Poster

Probabilistic Disease Surveillance Using Large Language Model
Chenxi Song**,** Yuhe Gao, RunXue Bao, Yiming Sun, Julians Tirado Alicia, & Ye Ye. (2024)
AMIA 2025 Informatics Summit

Paper

Transfer Learning with Clinical Concept Embeddings from Large Language Models
Yuhe Gao, Runxue Bao, Yuelyu Ji, Yiming Sun, Chenxi Song, Jeffrey P. Ferraro, & Ye Ye. (2024).
arXiv preprint arXiv:2409.13893.

Online transfer learning for RSV case detection
Yiming Sun, Yuhe Gao, Runxue Bao, Gregory F. Cooper, Jessi Espino, Harry Hochheiser, Marian G. Michaels, John M. Aronis, Chenxi Song, & Ye Ye. (2024, June).
In 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) (pp. 512-521).
IEEE. ***Best Paper Award in Analytics Track***

Research Experience

Bayesian Network Transfer Learning March, 2023- Dec., 2023

Mentor: Dr. Ye Ye (University of Pittsburgh)
Skills: BN Graph, Java, Supervised Learning, Greedy Algorithm
Description: Re-use a source model that learned from electronic medical record (EMR) data to predict diseases, such as influenza, target data set.

- Revamped a moderate-sized, 10-year-old Java codebase by refactoring variables and functions to align with **Object-Oriented (OOP) design** principles, enhancing reusability and maintainability. Compiled files into a JAR package and developed a CLI for users access.
- Verified the **greedy algorithm** for parent nodes search and manually configure the maximum parent nodes for **Bayesian Network structure**.

- Implemented a new feature (advanced multi-source domain data) derived from academic pseudocode for model with **accuracy of 1% improvement** comparing to existing single source model prediction.

Exploring the Integration of Foundational Model — May, 2024- Aug., 2024

Mentor: Dr. Ye Ye (University of Pittsburgh)
Skills: Vision Transformer, OpenSlide, Pillow, PyTorch, Matplotlib, Scikit-Learn, CUDA
Description: Drive insights by combining Prov-GigaPath's computational capabilities with Mesothelioma data Center. Potential replication of the model for other bio-repositories or research focuses.

- Applied Prov-Gigapath model to preprocess and analyze **large-scale pathology slides** (up to 4GB), to distill image features to **1 x 768 embedding representation**.
- Employed **stratified sampling** techniques to create balanced training and testing set, enhancing the reliability and accuracy of model evaluation results.
- Performed **10-fold cross-validation** on image embeddings to conduct **binary classification**, demonstrating robust model validation.
- Established baseline model performance for cancer diagnosis with 0.87 accuracy and 0.70 AUROC, providing a foundation for future collaboration.

Probabilistic Disease Surveillance Using LLM — June, 2024- Sep, 2024

Mentor: Dr. Ye Ye (University of Pittsburgh)
Skills: CoT prompting, RegEx, PyTorch, Transformers, LangChain, Ollama, CUDA
Description: Examine the capability of LLMs to provide probability estimations for detecting infectious disease cases from electronic medical (EHR) data

- **Deployed the LLMs (Llama 3-8B) locally** using Docker to ensure data privacy and secure processing of sensitive healthcare data.
- Converted **text unstructured data** into dense or sparse vectors (embeddings) to set up a **vector database** through **ElasticSearch**, enabling efficient data retrieval.
- Customized few-shot prompting using the **Chain-of-Thought (CoT)** technique, achieving a **15% performance** improvement over zero-shot learning by optimizing prompt structure.

Industrial Experience

Role: Network and System Administrator — Feb, 2024- Present

Employer: Reformed Presbyterian Theological Seminary
Skills: Python, Git, Azure Cloud, Ubuntu, Bash, IoT, ADDS, DNS
Description: Aim to understanding network communication in **large-scale web systems and data exchange**. Gained skills in building systematic frameworks within **domain networks** to enhance visibility and manage a digital presence effectively

- Monitored **150+ network connections** and **30+ systems**, optimized office software for **10% annual cost saving**, and automated workflows to **reduce processing time by 20%**.
- **Established IT support ticketing system** and knowledge base website to centralize asset management and improve accessibility, boosting **problem-solving efficiency by 30%** [link]

Honors & Award

Best Paper in Analytics Track, IEEE 12th ICHI, FL — Aug. 2024
Winner, Geneva College Tower Scholarship, PA — Aug. 2017 -Dec. 2018
Winner, College Hill Church Scholarship, PA — Aug. 2017 -Dec. 2018