XIMIN HU

Senior Data Scientist ◇ Seattle ◇ +1(443) 214-9707 ◇ summer07.nanjolno@gmail.com

SUMMARY

Experienced Data Scientist & Researcher: 6 years of research & working experience with 10 papers, 2 Python packages, and 1 patent published. Expertise in natural language processing, especially LLM fine-tuning & implementation, prompt engineering, and RAG pipeline development. Strong critical thinking, problem-solving and communication skills.

SKILLS

Programming Language	Python, Java, C#, R
Machine Learning	Natural Language Processing (LLMs), Deep Learning, OpenCV, Scikit-learn
Engineering	Azure & GCP, Git, Spark, SQL, Databricks, Ray, Apache Airflow

EDUCATION

University of Washington Sep 2019 - Aug 2023
Ph.D. in Civil Engineering, Data Science

Johns Hopkins University Sep 2017 - Dec 2018
M.S.E in Environmental Engineering

Tongji University Sep 2013 - Jun 2017
B.S. in Water Supply and Wastewater Engineering

EXPERIENCE

Senior Data Scientist Jan 2025 - Present
AstrumU *Bellevue, WA*

- Designed and developed an automated machine learning system for generic knowledge graph generation and update with topic models (BERTopic) and LLMs (e.g., Claude 3.5, DeepSeek) for HR related dataset. Implemented RLFH strategy to improve the pipeline performance, and filed a patent for innovation in automated graph processing.
- Led the design for a knowledge graph-based RAG system with LlamaIndex for efficient data retrieval and analysis.
- Designed and implemented a robust pipeline to evaluate and improve LLM-based models regarding knowledge graph related tasks. Filed one associated patents.

Data Scientist Feb 2024 - Dec 2024
AstrumU *Bellevue, WA*

- Led the development of a scalable machine learning pipeline integrating transformer models for text understanding and analysis regarding professional skills. Performed fine-tuning for text classification and NER tasks, achieving 90% accuracy for skill extraction and classification. Released two major services for downstream team.
- Developed and deployed a graph database and an interactive visualization dashboard using Neo4j to support skill taxonomy management and visualization. Developed demo with Streamlit for customer communication.

Postdoctoral Researcher Aug 2023 - Jan 2024
University of Washington *Seattle, WA*

- Conducted data mining on sophisticated instrumental data to evaluate tire rubbers samples. Using clustering algorithms and regression models to discover the potential toxicants, and facilitate environmental risk assessments.
- Developed a machine learning-based workflow to quantify the source of pollution in water samples with a complex data set. Achieved prediction accuracy rates that exceed 99% for identification and $R^2=0.95$ for quantification with an optimized algorithm on real-world data from different sites. Released a python package for the pipeline.

Machine Learning Scientist Jun 2023 - Aug 2023
Wayfair / Internship *Boston, MA*

- Directed a pivotal project to develop and implement macroeconomic, time-series, and NLP features, which significantly enhanced the user behavior prediction model's performance by over 13%.
- Demonstrated expertise in utilizing the GCP on a daily basis, including dataset processing on BigQuery, conducting feature engineering (including text embedding generation (TF-IDF)) to uncover underlying data patterns. Designed and implemented Airflow DAG for daily data updates and feature engineering workflows.
- Proactively collaborated with cross-functional teams to apply business applications for developed features.