# Hongjun Liu

Zhejiang University
My Homepage: https://holayan.github.io/
janne150.lhj@gmail.com | +86 13926516916

## EDUCATION

**Zhejiang University**                                                                                             **August 2020 - June 2024**

• CHU KOCHEN HONORS COLLEGE

• **Dual Bachelor's Degree: Computer Science & Environmental Resource Management,**

•**Relevant coursework:** Machine Learning (online); Elements of the theory of computation; Advanced Data Structure and Algorithm Analysis; Operating Systems and Systems Programming; Discrete Mathematics

## SKILLS

• **Machine Learning Toolkits:** model development (Scikit-learn, TensorFlow, PyTorch, Keras); model evaluation (ROC, AUC, confusion matrix); model deployment (Docker)

• **Programming:** data analysis (pandas, scipy), data visualization (matplotlib, seaborn), data engineering (SQL, MongoDB)

• **Toefl: 107  (Speaking: 24*; Writing:27)**

## LEADERSHIP EXPERIENCE

• **Organizer & Student Union Manager** (3 years) Zhejiang University Guangdong Students Union

• **Sports Team Leader** (2 years) Zhejiang University CHU KOCHEN HONORS COLLEGE Women's Basketball Team

## PUBLICATIONS

• Zhao, Yilun*, **Hongjun Liu***, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. "KnowledgeMath: Knowledge-Intensive Math Word Problem Solving in Finance Domains." (*Equal Contribution)

•Zhao, Yilun*, Yitao Long*, **Hongjun Liu**, Linyong Nan, Lyuhao Chen, Ryo Kamoi, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. "DocMath-Eval: Evaluating Numerical Reasoning Capabilities of LLMs in Understanding Long Documents with Tabular Data."  (*Equal Contribution)

## RESEARCH EXPERIENCES

**New York University**

*Research Assistant, Advisors: Chen Zhao*                                                                       *Nov 2023 - Present*

Project: Developing and Enhancing Language Models for Improved Counterfactual Claim Processing and Evidence Retrieval

• Utilize a large language model (LLM) like GPT-4 to transform counterfactual claims into retriever-friendly versions through decomposition, rephrasing, or similar methods, and subsequently address these reformulated claims.

• Accumulate pairs of original and counterfactual claims to train an editing model, focusing on refining models such as LLAMA2 for this purpose.

• Enhance the capabilities of a dense retriever by fine-tuning it, thereby facilitating more efficient evidence discovery using claims edited by the LLM.

**Yale University & New York University**

*Research Assistant, Advisors: Arman Cohan & Chen Zhao*                                              *July 2023 - Present*

Project: Extending Capabilities of Large Language Models for Knowledge-Intensive Financial Exam QA

• Construct a dataset about Complex Financial Exam QA with a hybrid of textual and tabular content and require college-level knowledge in the finance domain for effective resolution.

• Provide expert-annotated, detailed solution references in Python program format for each QA in the dataset, ensuring a high-quality benchmark for LLM assessment.

• Evaluate a wide spectrum of LLMs on our newly constructed dataset with different prompting strategies like Chain-of-Thoughts

and Program-of-Thoughts. The current best-performing system (i.e., GPT-4 with Program-of-Thoughts) achieves only 45.4% accuracy, leaving substantial room for improvement.

• Use the question as the retrieval query, acting as a knowledge retrieval module to retrieve the top-n knowledge terms with the highest similarities from our constructed knowledge bank and enhance the capabilities of LLMs for solving knowledge-intensive hybrid QA.

• Enhance the proficiency of language models such as LLAMA2 to augment the performance of large language models (LLMs) in addressing complex, knowledge-intensive hybrid question-answering tasks.

**Alibaba DAMO Academy (Research Apartment)**

*Research Assistant, Advisor: Jieping Ye*                                    *Oct 2022 - June 2023*

Project: Use Machine Learning to Establish a Mapping Relationship from Genetic Variation to Phenotypic Diversity

• Spearheaded the annotation of 3D scan data for the skulls of bird species, identifying key points and lines to delineate structural features. Conducted preprocessing of the point cloud data, involving detailed annotation of specific features and conversion of data formats for enhanced usability and analysis.

• Implemented a systematic approach for correlating genomic data with phenotypic traits, focusing on beak development genes and their association with beak width, length, and body mass.

• Applied statistical and machine learning methods commonly used in Bio career, including PCA for dimensionality reduction, ANOVA for inter-group differences, and various regression models (simple linear, polynomial, Lasso & Group Lasso) to uncover complex relationships.

• Conducted gene annotation and feature extraction from protein sequences, leveraging an array of trained models such as ProteinBert-Tape and MSA Transformer to extract significant information regarding protein structure and function.

• Collaborated in a multidisciplinary team to rank genes based on their correlation with phenotype data, integrating ecological measurement data for comprehensive analysis.

• Establish a phenotypic prediction model, which linked relationships between multiple genes and phenotype traits and achieved 84% accuracy.

**Zhejiang University Student Research Program**

*Research Assistant, Advisor: Xili Zhang*                                    *Sept 2021 - Feb 2022*

Project: Financial Statement Fraud and Valuation Misestimation Issues Detection

• Constructed a dataset based on the financial reports of 100 companies over the past ten years.

• By performing a chi-square test, five indicators of accounts receivable turnover, inventory turnover, cash ratio, intellectual property ratio, and current ratio were identified as significant factors for detecting companies that engage in financial statement fraud. These factors were then used to establish data classification labels.

• Based on restructured data, evaluate various kinds of recognition models for value estimation, such as linear regression, logistic regression, decision trees, random forests, and neural networks.

**MAIN REFERENCE**

**Professor Armon Cohan**
**Assistant Professor**
**Computer Science Department**
**Yale University & Allen Institute for AI (AI2)**
**arman.cohan@yale.edu**

**Professor Chen Zhao**
**Assistant Professor**
**The Center for Data Science**
**NYU Shanghai & New York University**
**cz1285@nyu.edu**