




ZHAOYI HOU (JOEY)

 joeyhou.com  joeyhou@seas.upenn.edu  (858)729-8929

RESEARCH INTEREST

Artificial Intelligence, Machine Learning, Natural Language Processing
Commonsense Reasoning, Knowledge Representation, Intelligent System

EDUCATION

University of Pennsylvania Aug 2021 - Present
Data Science, Master of Science in Engineering

University of California, San Diego Sep 2017 - Jun 2021
Data Science, Bachelor of Science (Cum Laude)

PUBLICATION

- Xiaochen Kev Gao, **Zhaoyi Hou**, Yifei Ning, Jingbo Shang, Vish Krishnan. *Towards Comprehensive Patent Approval Predictions: Beyond Traditional Document Classification*. [ACL 2022]
- Caitlin A. Stamatis, Jonah Meyerhoff, Tingting Liu, **Zhaoyi Hou**, Garrick Sherman, Brenda L. Curtis, Lyle H. Ungar, David C. Mohr. *The association of language style matching in text messages with symptoms of affective psychopathologies*. [Procedia Computer Science]
- Artemis Panagopoulou, Manni Arora, ...(6 more), **Zhaoyi Hou**, Alyssa Hwang, Lara Martin, Sherry Shi, Chris Callison-Burch, Mark Yatskar. *QuakerBot: A Household Dialog System Powered by Large Language Models*. [Alexa Prize TaskBot Challenge Proceedings]
- Emily Nicole Corbett Manoogian, Adena Zadourian, ...(8 more) **Zhaoyi Hou**, Jason G. Fleischer, Shah Golshan, Pam R Taub, Satchidananda Panda. *Feasibility of Time-Restricted Eating and Impacts on Cardiometabolic Health in 24-Hour Shift Workers: The Healthy Heroes Randomized Clinical Trial*. [Cell Metabolism, In Press]

RESEARCH EXPERIENCE

Computer and Information Science (CIS) at Penn Engineering Sep 2022 - Present
NLP Researcher (advised by Prof. Chris Callison-Burch)

- Building a **schema learning** model to generate event-based schema by distilling knowledge from pre-trained large language models (e.g. GPT-3).

Computer and Information Science (CIS) at Penn Engineering May 2022 - Present
NLP Researcher (advised by Li Zhang and Prof. Chris Callison-Burch)

- Building a **natural language generation** model to generate "branching" options in a daily script;
- Building an **explanation generation** module to generate rationales for each option based on common sense.

Computer and Information Science (CIS) at Penn Engineering Jan 2022 - Aug 2022
NLP & Psychology Researcher (advised by Prof. Lyle Ungar)

- Defined and built a pipeline to extract **semantic mirroring** from mobile text conversations;
- Analyzed the correlation between **psychological conditions** and semantic mirroring in text messages.

Shang Data Lab at UCSD (SDLab) Jun 2020 - Apr 2021
NLP Researcher (advised by Prof. Jingbo Shang)

- Built a **text data ETL and classification** pipeline to handle **600,000** patent documents;
- Implemented a customized **BERT-based** text classification model and improved true negative rate (specificity) from **60% to 86%** (heavily unbalanced data with **84%** positive instance).

Salk Institute for Biological Studies Jul 2019 - Jun 2021
Data Science Researcher (advised by Prof. Satchidananda Panda)

- Built a **data ETL and analysis** pipeline for user behavior analysis (more than **500,000** records).

PROJECT

Amazon Alexa TaskBot Competition

Nov 2021 - Apr 2022

Information Retrieval

- Implemented the **document retrieval** module for the Alexa TaskBot competition;
- Improved the retrieval success rate by **25%** and **advanced to the final list**.

Music Re-listen Prediction

Nov 2021 - Dec 2021

Recommendation System

- Built a **click-through-rate(CTR)** prediction model based on historical user behavior;
- Built a feature engineering pipeline and **deep factorization machine** model to achieve **67% AUC** in test data.

Stack Overflow Question Quality Classification

May 2021 - Jul 2021

Text Mining & Cloud Computing

- Built a **text data ETL and analysis** pipeline for **60,000** Stack Overflow question texts;
- Built an **XGBoost** classification model with AWS SageMaker and deployed it as an **AWS SageMaker Endpoint**;
- Achieved **87% accuracy** in the "High-Quality Question" classification task.

Food Text Parser

Jul 2020 - Apr 2021

Natural Language Processing

- Built an open source **text parsing pipeline** to extract food content from a health monitor app;
- Automatically correct typos and extract food-related phrases from the user's input text with **85% parsing accuracy**.

Machine Learning for Ophthalmological Diagnosis

Nov 2019 - Dec 2020

Computer Vision & Healthcare

- Built an **image classification pipeline** to pre-diagnose common eye diseases with a convolutional neural network;
- Achieved **75% accuracy** for classifying involutional ptosis, thyroid eye disease, and normal eyes.

TEACHING / TUTORING

Head Teaching Assistant in Machine Learning

Jul 2021 - Aug 2021

X Academy 2021

- Designed the machine learning capstone project for the teenager-oriented summer camp in China

Machine Learning Tutor

Mar 2021 - Jun 2021

Hahcioğlu Data Science Institute, UC San Diego

- Tutored CSE151A (Intro to Machine Learning) with Prof. Jingbo Shang

Data Structure Tutor

Mar 2018 - Aug 2019

Hahcioğlu Data Science Institute, UC San Diego

- Tutored DSC20 (Intro to Data Structure) with Prof. Marina Langlois

COMMUNITY INVOLVEMENT

Chinese Computer Community (Triple C) at UCSD

May 2020 - Apr 2021

President

- Project-based student community with more than **100 members and 15 student projects**;
- Designed the **onboarding technical training** for data science members.

Big Data Digest - AI Scholar

Jul 2019 - Dec 2020

Voluntary Content Contributor

- Translated the latest **AI papers and research news** from English to Chinese;
- Promoted the understanding of AI technologies on Chinese social media.

TECHNICAL SKILL

Programming:

Python, Java, C, C++, Data Structure, Object Oriented Programming, SQL

Machine Learning:

LSTM, Transformer, CNN, Boosting, Logistic Regression, SVM, Decision Tree, K-Means

Software & Tools:

PyTorch, Hugging Face, AWS SageMaker, Pandas, Scikit-learn, MySQL

Data Analysis:

Bootstrap, ETL, Hypothesis Testing, Research Design