Shukai Gong

(+86) 15721022905 | shukai_gong@ruc.edu.cn

Education

Renmin University of China Beijing, China

Double Major in Data Science and Economics; ***GPA: 3.90/4.00; Rank: 1*** *2022.09 – 2026.07*

Relevant Courses: Data Structure and Algorithms(4.0), Probability Theory(4.0), Mathematical Analysis(4.0), Advanced Linear Algebra(4.0), Optimization Theory(4.0), Stochastic Process(4.0), Parallel Computing(4.0)

Honor: The Premium Academic Excellence Scholarship **(2024 Jing Dong Scholarship)**

University of California, Berkeley Berkeley, United States

2024 Spring Exchange Student; ***GPA: 4.00/4.00*** *2024.01 – 2024.05*

Relevant Courses: STAT 135: Concepts of Statistics(A+), ECON 141 Econometrics (Math Intensive) (A+)

Publication

Shen Yuan*, **Shukai Gong***, Hongteng Xu. USPTO-LLM: A Large Language Model-Assisted Information-enriched Chemical Reaction Dataset, **WWW 2025** Resource Track. *(Co-first author)* (*Zenodo Link*)

Research Experience

GSAI, Renmin University of China, Research Intern *2024.03.08 – 2024.12.13*

- **Supervisor**: Hongteng Xu
- **Topic**: Constructed the first chemical reaction dataset (USPTO-LLM, 247K entries) containing abundant reaction condition information by using LLM APIs to extract data from the USPTO patent database. Experiments showed that: (1) Pretraining sequence-based retrosynthesis models on USPTO-LLM significantly improves their performance. (2) Incorporating reaction condition information enhances retrosynthesis accuracy for both graph-based and sequence-based models.

Internship

Intuitive Fosun, R&D Intern *2025.01.15 – 2025.02.15*

- Leveraged OCR to extract high-frequency operating parameters from the transducer screen of the daVinci surgical robot. Processed video frames using OpenCV to minimize noise from parameter fluctuations, achieving an OCR accuracy of 99%.
- Trained a single-layer Transformer Encoder to extract and classify key information from 2,100 after-sales feedbacks of daVinci surgical robots into 11 categories, which greatly helped after-sales engineers identify issues and provide solutions.

Projects

Micro-blockchain System Development and Data Analysis | *GitHub* *2023.10.08 – 2023.12.01*

- Designed the **data structures of accounts, transactions and transaction graphs in a micro-blockchain system** with $O(1)$ **reading and operation speed**. Read 2129 block information and 1048575 transaction records in 15 seconds.
- Developed a **query system** that retrieves account transfer records, current balances based on user input within 5 seconds.
- Implemented loop-detection in the transaction network by **topology sorting**. Used an **priority-queue-optimized Dijkstra algorithm** to find the shortest path between any two accounts in the trading network within 3 seconds.

Image-to-text and Text-to Image Generation | *GitHub* *2025.01.03 – 2025.01.12*

- Implemented an image-to-text model using CLIP as the image encoder, TinyLlama as the text decoder, and Qformer for cross-modal alignment. Achieved BLEU = 0.26 on the Flickr8k dataset, significantly outperforming our baseline model that uses CLIP and Transformer Decoder aligned by cross-attention module.
- Trained a Latent Diffusion model with **self-implemented** VQVAE and U-Net on Flickr8k dataset as a text-to-image baseline and fine-tuned pre-trained SDXL and VAR models to achieve significantly improved performance in text-to-image tasks.

Contests

2023 COMAP's Mathematical Contest in Modeling, Meritorious Prize

2023 Chinese University Math Contest in Modeling, Beijing First Prize

Technical Skills

Programming Language: C/C++, Python, MATLAB, R, STATA

English Proficiency: CET4 650; IELTS 7.5 (Listening: 8.0, Reading: 8.5, Writing: 7.0)