XIAOZHE LI

Github: https://github.com/OliverLeeXZ/
4800 Cao'an Road, Jiading district, Shanghai, 201804
(+86) 13572467181 ⋄ Lxxzzz@tongji.edu.cn

EDUCATION

Tongji University, Shanghai (985, 211, Double First Class) *2022.06 - 2028.03*
Ph.D. in Computer Science and Technology
Supervisor: Prof. Qingwen Liu

Donghua University, Shanghai (211, Double First Class) *2018.09 - 2022.06*
Bachelor of Computer Science and Technology GPA: 88.5/100 Rank3/34

PUBLICATIONS

Long Range Barcode Scan Using Resonant Beam: **Xiaozhe Li**, Shuaifan Xia, Qingwen Liu*, Yunfeng Bai, SPAWC 2023, Shanghai

PROJECT EXPERIENCE

Educational Large Language Models for Teaching Services: System Construction and Demonstrative Applications
· Ministry of Science and Technology's Changjiang Science and Technology Innovation Project, collaborating with USTC and iFLYTEK, with a vertical fund of 10 million RMB
· The project on large language models in the education field can be divided into three sub-projects: data construction and evaluation, human-computer interaction, and cognitive models
· Data construction and evaluation: material integration, supervised cleaning, and effectiveness evaluation
· Multimodal data input and output: multimodal unified encoding, feature alignment, and intent recognition
· Enhanced logical professionalism: domain pre-training, retrieval prompts, and logical enhancement

Research on Large Language Models with Integrated External Knowledge
· Ministry of Education's University-Industry-Research Fund, with a vertical fund of 100,000 RMB.
· The large language model system in the research field can be divided into four sub-problems: LLM model fine-tuning, external knowledge base construction, knowledge retrieval recall, and model capability assessment.
·LLM model retraining: 1) Using LLM to parse and extract domain keywords from a large number of papers; 2) Collecting public datasets; 3) Using regular matching to filter the dataset based on domain keywords; 4) Manually reviewing filtered data and constructing a domain dataset; 5) Training LLM with domain data.
· External knowledge base construction: 1) Collecting domain paper documents; 2) Using the Nougat model to convert all PDF papers to LaTeX format; 3) Constructing LaTeX papers into a document tree format; 4) Extracting question-answer pairs from all papers; 5) Building an external knowledge base.
· Knowledge retrieval recall to achieve user intent recognition and extraction of key question terms: 1) Designing and testing intent recognition thinking chain templates; 2) Testing the accuracy of keyword similarity recall algorithms.
·Building a research knowledge test set for LLM capability assessment: 1) Creating domain problems as multiple-choice questions, true/false questions, and open-ended questions; 2) Testing the scores of research LLM and other mainstream LLMs on the test set; 3) Providing feedback to improve system

shortcomings.

Large-Scale Graph Reachability Queries
· Read multiple papers from top conferences in computer theory and data management, including FOCS, SIGMOD, VLDB, and ICDE
· Reproduce algorithms proposed in top conference papers FOCS, SIGMOD'14, and SIGMOD'17, and conduct experimental analysis
· Improved the bottom-up closure size estimation algorithm and proposed a top-down verification algorithm
· Proposed a new way of generating trees to obtain more accurate values of node closure size
· Introduced two optimization strategies to improve accuracy while maintaining faster speed
· Improved closure size estimation algorithm in linear time and space complexity

Intelligent Speech Recognition and Input for Structured Electronic Medical Records
· Project leader for a national undergraduate innovation project
· Integrated speech recognition technology into traditional medical record entry, achieving real-time voice input and keyword input
· Implemented voice/manual modification for incorrectly entered information and enabled mobile recording during ward rounds
· Achieved a 97% accuracy rate in recognizing colorectal cancer text with the core speech recognition model
· Applied for a software copyright: 2020SR1605226

INTERNSHIP EXPERIENCE

Shanghai Guan'an Information Technology Co., Ltd. Jiading District, Shanghai
Algorithm Intern *November 2021 - January 2022*
· Design and implementation of deep learning-based network intrusion detection algorithms
· Reading and reproducing multiple papers from EMNLP and NIPS
· Proposing a network intrusion detection model based on Text (Char) CNN and self-attention mechanism
· Conducted ablation experiments to demonstrate that attention mechanisms can effectively improve the performance of intrusion detection models. Compared with traditional machine learning models and existing deep learning models, the proposed model showed superior performance

COMPETITION AWARDS

Mathematical Contest In Modeling (MCM) Honorable Mention
China Collegiate Programming Contest— Team Programming Ladder Competition National Bronze Medal
National College Student Computer Design Contest National Third Prize
Shanghai Computer Application Skills Competition Shanghai Third Prize

SCHOLARSHIPS AND HONORS

Outstanding Graduate of Donghua University
Donghua University Academic Excellence Scholarship

SKILLS

Programming: Python, C++, Java, Linux, Matlab
Tools: Git, Latex
Languages: English (CET-6: 520)