Mingxia Li

Personal Website: cat-mia.github.io
Google Scholar

Email: limingxia1999@gmail.com
Mobile: +86-152-8474-0305

Research Interests

My research interests in CS:#financial AI agents utilizing a multi-agent system, agent simulation and optimization, knowledge graph generation.
My research interests in Business:#business intelligence systems, the intersection of LLMs in IS and OM.

Education

- **University of Electronic Science and Technology of China (UESTC)** — Bachelor of Engineering
  Computer Science and Technology; GPA: 3.95 — *2016.9 - 2020.6*
- **University of Science and Technology of China (USTC)** — Master of Engineering
  Lab for Intelligent Networking and Knowledge Engineering; GPA: 3.65 — *2020.9 - 2023.6*

Professional Experience

- **Microsoft Research Asia (MSRA)** — Research Intern (full-time)
  Innovation Engineering Group — *2021.2 - 2022.3*
- **AntGroup (Fintech Affiliate of Alibaba Inc.)** — AI Applicaiton Engineer
  International Financial Risk Management Department — *2023.8 - 2025.3*
- **The Chinese University of Hong Kong, Business School** — Research Assistant
  Operations Management — *2024.9 - Present*

Publications

- **Dynamic Resource Allocation for Deep Learning Clusters with Separated Compute and Storage (CORE-A*)**: Accepted by ***IEEE International Conference on Computer Communications*** *2023. Authors:* ***Mingxia Li****, Zhenhua Han, Chi Zhang, Yuanchi Liu, Haisheng Tan*
- **SiloD: A Co-design of Caching and Scheduling for Deep Learning Clusters (CORE-A)**: Accepted by ***ACM EuroSys*** *2023. Authors: Hanyu Zhao, Zhenhua Han, Zhi Yang, Quanlu Zhang,* ***Mingxia Li****, Fan Yang, Qianxi Zhang, Binyang Li, Yuqing Yang, Lintao Zhang, Lidong Zhou, Yafei Dai*
- **Asymptotically Optimal Online Caching on Multiple Caches with Relaying and Bypassing (CORE-A*)**: Published in ***IEEE/ACM Transactions on Networking*** *2021. Authors: Haisheng Tan, Shaofeng H.-C. Jiang, Zhenhua Han,* ***Mingxia Li***
- **Online Learning-Based Co-task Dispatching with Function Configuration in Edge Computing (CORE-C)**: Accepted by ***Parallel and Distributed Computing, Applications and Technologies*** *with* ***Best Paper Award****. Authors: Wanli Cao, Haisheng Tan, Zhenhua Han, Shuokang Han,* ***Mingxia Li****, Xiang-Yang Li*
- **Deep Density-based Image Clustering (CORE-B)**: Published in ***Knowledge-Based Systems*** *2019. Authors: Yazhou Ren, Ni Wang,* ***Mingxia Li****, Zenglin Xu*
- **Cross-Model Operator Batching for Neural Network Architecture Search (CCF-C)**: Accepted by ***International Conference on Wireless Algorithms, Systems, and Applications*** *2022. Authors: Lingling Ye, Chi Zhang,* ***Mingxia Li****, Zhenhua Han, Haisheng Tan*

Projects

- **Text2SQL - Interactive & Intelligent Supply Chain Database Query System**
  Integerate Panjiva global supply chain data — *2024.10 - Present*
  - **Contents**: We implemented an advanced interactive global supply chain data query system powered by cutting-edge text-to-SQL AI technology. Seamlessly integrating Panjiva's extensive database with a user-friendly graphical interface, this system allows users to input natural language queries, which are then interpreted by an LLM that expertly generates SQL statements to access the Panjiva database. The system delivers insightful, precise responses to user questions, complete with dynamic visualizations that enhance the clarity and impact of query results.
  - **Personal Contributions**: transform Panjiva data into database tables, build the text2SQL user query system driven by GPT-4, build the knowledge base of data description, test and analyze various types of queries sets.
- **Management Science Paper Replication Using Multi-Agent System**
  Successfully replicated 2 papers in MS replication project — *2024.9 - 2024.10*
  - **Contents**: In MS replication project comprising several studies conducted with human subjects, I used a multi-agent system to replicate two specific experiments: *Designing Pricing Contracts for Boundedly Rational Customers: Does the Framing of the Fixed Fee Matter?* and *Inferring Quality from Wait Time.* The replication results aligned with the original findings. To enhance the simulation, I assigned different personas to the LLM agents and prompted them to generate reasoning outputs. Analysis suggests that multi-agent systems can effectively model human interactions in complex environments.
  - **Personal Contributions**: primary role, build up the multi-agent system, transform original instructions into prompts objectively, assigning different personas to LLM agents, fix LLM mathematical errors and result analysis.

- **Knowledge Graph-Enhanced Business Strategy Recommendation System**
  for Lazada (e-commerce platform in Southeast Asia under Alibaba Inc.) *2024.5 - 2024.8*
  - **Contents**: Lazada's business strategies heavily rely on expert experience and code linked to payment event features, lacking readability and textual identifiers. To overcome this, we generate a knowledge graph from these strategies. We first use LLMs to decompose tasks, annotate unlabelled data and code in natural language, then combine these annotations to construct the knowledge graph. By applying graph RAG and integrating it into our agent-driven AI assistant, we develop a recommendation assistant that engages in dialogue with user and suggests the most relevant strategies for queried risk events.
  - **Personal Contributions**: primary role, lead the entire project, data preparation, task decomposition and merging, construct the knowledge graph with neo4j, and implemented the agent-driven assistant.
- **Reflection-Based Self-Revolutionary Domain Knowledge Generation**
  for the cross-border payments risk control of Alipay+ *2024.5 - 2024.8*
  - **Contents**: In global remittance products, risks like money laundering and fraud exist. Alipay+ uses an internal SOP(Standard Operation Procedure) for risk control, but as fraudulent tactics evolve, the SOP becomes insufficient. We create a multi-agent system where a reflection agent reviews error cases and updates the SOP daily. This ensures that as fraudulent tactics evolve, our defense strategies are also updated.
  - **Personal Contributions**: propose the idea of multi-agent system with a reflection agent, implement the whole system, construct the evaluation framework, and suggest incorporating a fraudulent agent for simulation.
- **Multi-Round LLM-Driven AI Assistant for Payment Risk Decision-Making**
  for B2B international transactions of Alibaba's settled merchants *2023.11 - 2024.2*
  - **Contents**: For cross-border financial risk events, many cases like forged contracts and doctored ID documents can't be decided by existing strategies or algorithms and require human review, which is inefficient and costly. We developed an AI assistant using a fine-tuned LLM to interact with human reviewers. First, we fine-tuned the LLM using internal company materials. Then it generates questions to assess risk, and based on human responses, it produces the next round of questions. After several rounds, the AI assists in making a risk control decision.
  - **Personal Contributions**: prompt engineering and participate in building the system.

- **Dynamic Resource Allocation for Deep Learning Clusers with Separated Compute and Storage**
  IEEE International Conference on Computer Communications *2022.7 - 2023.1*
  - **Contents**: With the goal of minimizing the compute and storage financial costs of training unpredictable deep learning workloads in cloud platforms, we do this trade-off by identifying the huge cost difference and heterogeneous preference among different workloads to derive the dynamic allocation.
  - **Personal Contributions**: formulate the problem of trade-off between compute and storage costs; design an $O(n)$ algorithm for resource allocation and prove its optimal; achieve significant combined cost reduction on Microsoft Philly production.
- **Catur: Balancing Demand and Supply of Hardware Resources in Azure Cloud Services**
  Co-work with Microsoft Azure(Raymond) during MSRA internship *2021.9 - 2022.3*
  - **Contents**: The overbooking phenomenon in Microsoft Azure cloud services is a supply and demand issue where Azure sells more service packages than available hardware, assuming not all users will use the service simultaneously. This strategy maximizes revenue while minimizing hardware costs. However, if all users access the service at once, performance degradation may occur due to unpredictable user demand and mismatched resource allocation. Catur uses RL-based methods (DQN) to optimize resource allocation and minimize users' perception of performance decline.
  - **Personal Contributions**: Build the pipeline for preprocess and cross-validate the raw data; create env for training the Q-network and implement baseline policies for comparison; use supervised learning methods to extract and analysis the correlation between RL state, action and reward to help verify and improve the RL design - the correlation coefficient was increased from 0.42 to 0.96; define and implement the evaluation metrics

Skills Summary

- **Technical Proficiency**:
  - **System Implementation**: Java, Python, C++, Docker, Linux, MySQL, Neo4j
  - **LLM-Tools**: Llama-Index, PyTorch, Pandas, OpenAI API
  - **LLM-Application**:Solid background in industrial intelligence systems: LLM-driven multi-agent systems, knowledge systems with RAG, natural language query systems, automated knowledge graph construction
- **Academic Skills**: LaTeX, build website for lab/project, host reading groups, Git(for open source research and backup)
- **Soft Skills**: collaboration and communication, innovative problem solving
- **English level**: **IELTS** band 7.5, **GMAT Focus Edition** 95th Percentile

Honors and Awards

- Mathematical Contest in Modeling - **Meritorious Winner** *2018.2*
- National English Competition for College Students - **the Second Prize** *2019.6*
- **Outstanding Graduate**, **Excellent Thesis Award (Top 1%)** of UESTC *2020.6*
- Microsoft Research Asia Stars of Tomorrow Internship Program - **Award of Excellence** *2022.5*
- The First-Class Student Scholarship of UESTC *2016-2020*
- The First-Class Student Scholarship of USTC *2020-2023*