YUANKAI LI

(+86) 156-0196-1787◇ yuankaili21@m.fudan.edu.cn◇ Website

EDUCATION

Fudan University   *Sept. 2021 - present*
B.Eng. in Artificial Intelligence (Honor Class, Data Science Track)
GPA: 3.63/4.00 Rank: 4/26
Relevant coursework Method of Optimization (A), Stochastic Processes (A), Data Structure and Algorithm Design (A), Artificial Intelligence (Honor, A), Mathematical Statistics (Honor, A), Pattern Recognition and Machine Learning (A)

RESEARCH INTERESTS

My research interest lies broadly in **natural language processing** and **machine learning**. Currently, I've been concentrating on understanding and extending the ability of **Large Language Models (LLMs)**, with a specific focus on:

- Enhancing retrieval-augmented language models and developing fair evaluation methods for LLMs

- Improving the reasoning ability of LLMs while ensuring reliability and robustness

- Building helpful AI agents and computer systems that communicate through natural language and continuously improve through interaction

PUBLICATION

BRIEF: Bridging Retrieval and Inference for Multi-hop Reasoning via Compression
Yuankai Li*, Jia-Chen Gu*, Di Wu, Kai-Wei Chang, Nanyun Peng
Submitted to NAACL 2025, under review

SCP: Scalable and Customizable Generation of Planning-specific Corner Cases in Autonomous Driving
Lingfeng Zhou, Jin Gao, Mohan Jiang, Yufeng Liu, **Yuankai Li**, Dequan Wang
Submitted to ICRA 2025, under review

Dissecting Dissonance: Benchmarking Large Multimodal Models Against Self-contradictory Instructions
Jin Gao, Lei Gan*, **Yuankai Li***, Yixin Ye, Dequan Wang
ECCV 2024

RESEARCH EXPERIENCE

University of California, Los Angeles   *May 2024 - present*
Research Intern, PLUS Lab

- **Query-aware multi-hop compression for RAG systems**
  Advisor: Prof. Nanyun (Violet) Peng and Dr. Jia-Chen Gu   *May 2024 - Oct. 2024*

  – Proposed a synthetic data pipeline by extracting atomic propositions from the source documents. The pipeline has an automatic validation mechanism to ensure high-quality summaries for multi-hop context compression. It involves solely open-source LLMs and has the potential to scale up, offering a data-centric approach to constructing data for context compression.

- Trained a T5-based compression model (BRIEF) that performs query-aware multi-hop reasoning by compressing retrieved documents into highly dense textual summaries, which can integrate into any form of downstream tasks. BRIEF significantly reduces the computation cost in an RAG system.

- BRIEF constantly improves the compression rate and EM/F1 scores by more than 3 percent compared to SOTA baselines across different multi-hop QA datasets. The improvement persists when we change to reader models of varying sizes.

- This work results in a first-authored paper submitted to NAACL 2025.

Shanghai Jiao Tong University *June 2023 - Feb. 2024*
Research Intern, Qing Yuan Research Institute

- **Demonstration Selection for In-context Learning**
  Advisor: Prof. Dequan Wang and Prof. Zhijie Deng *Dec. 2023 - Feb. 2024*

  - Proposed a training-free few-shot demonstration selection method for LLMs on knowledge-intensive QA tasks using sparse retrieval methods.

  - Introduced a novel prompting paradigm that achieved 5% accuracy improvement on QA tasks including medicine and college mathematics together with the selected demonstrations.

  - Explored in depth what makes good demonstrations for domain-knowledge-intensive tasks.

- **Benchmarking LMMs against Self-contradictory Instruction**
  Advisor: Prof. Dequan Wang *Sept. 2023 - Nov. 2023*

  - Introduced the idea of self-contradictory instructions in Large Multimodal Models (LMMs), emphasized its potential harm, and sought to benchmark and address this problem.

  - Created a diverse benchmark dataset that aims to assess the capability of LMMs to handle self-contradictory instructions, covering both the language and the vision modality.

  - Conducted thorough experiments on various LMMs and proved that current SOTA LMMs perform badly even when equipped with in-context learning.

  - Contributed to a second-authored paper published at ECCV 2024.

- **Building LLMs Agent in Autonomous Driving Simulation**
  Advisor: Prof. Dequan Wang *June 2023 - Sept. 2023*

  - Developed an autonomous driving simulation system using LLMs in the decision-making stage and devised a method to translate a series of LLM decisions into simulation trajectories.

  - Proposed the idea that descriptive natural language can be used to generate rare corner case driving simulations, *e.g.* the vehicle executes an evasive maneuver when detecting an accident.

  - Through massive experiments, proved that LLMs can handle such decision-making tasks and respond accordingly to natural language.

  - Contributed to a paper submitted to ICRA 2025.

AWARDS AND SCHOLARSHIPS

2nd *Prize*, Excellent Students Scholarship, Fudan University *Sept. 2024*
1st *Prize*, East China Mathematical Contest in Modeling (Rank 4th in East China) *Nov. 2023*
2nd *Prize*, Contemporary Undergraduate Mathematical Contest in Modeling (Shanghai) *Nov. 2023*

LANGUAGE & SKILLS

Python, C, Matlab, R, Pytorch