Junzhi Chen

+1-347-431-8891 | jc13140@nyu.edu | google scholar

Education

- **New York University** *Sep 2024 - May 2026*
  Master of Science in Computer Engineering New York, United States
  - GPA: 3.78/4.00

- **The Chinese University of Hong Kong, Shenzhen** *Sep 2020 - May 2024*
  Bachelor of Engineering in Computer Science and Engineering Shenzhen, China

Publications

[1] *Smurfs: Multi-Agent System using Context-Efficient DFSDT for Tool Planning*
Junzhi Chen, Juhao Liang, Benyou Wang
NAACL 2025

Research Experience

- **Shenzhen Research Institute of Big Data, CUHKSZ** *Jun 2023 - Aug 2024*
  Research Assistant Shenzhen, China
  - Conducted research on context-efficient multi-agent systems, focusing on tool-using large language models (LLMs) for complex problem solving.
  - Identified limitations of the Deep-First-Search-Decision-Tree (DFSDT) algorithm and proposed "Smurfs", a novel multi-agent system (MAS) that enhances DFSDT with a modular, context-efficient, and training-free design. Detailed design can be seen at project page
  - Reduced token usage by 60.9% compared to DFSDT and enabled Mistral-7b to perform on par with GPT-4-DFSDT on StableToolBench.

- **Machine Learning for Language (ML²) Lab, NYU** *Nov 2024 - Present*
  Research Assistant New York, United States
  - Building a benchmark to evaluate LLM agents' ability to collaborate with users in task environments where instructions are either unsolvable or underspecified.

Work Experience

- **ModelBest, Shenzhen Intermediate People's Court** *Mar 2024 - Jul 2024*
  Machine Learning Engineer Intern Shenzhen, China
  - Collaborated with two colleagues to clean and curate 15TB of pre-training data from a vast amount of legal document.
  - Collaborated with judges and a team of five colleagues to develop and refine a downstream task pipeline by extracting instruction fine-tuning data from the court database, instruction tunning the pre-train model, prompt engineering, evaluating performance with judicial feedback, and iterating on data refinement/training to continuously improve model performance.
  - Achieved the first Chinese LLM for Judicial Trials, which covers 85 legal processes, including case filing, document reviewing, trials, and document drafting. The system has assisted in filing 291,000 cases and generated 11,600 draft documents, significantly improving the quality and efficiency of legal proceedings since trail operation.

Course Project

- **New York University** *Jan 2025 - May 2025*
  DS-GA 1012: Natural Language Understanding New York, United States
  - Designed and implemented an end-to-end pipeline to enable pretrained LLaMA-3 models to use explicit memory (sparse attention key-value pairs) for math and code reasoning tasks.
  - Designed and implemented the codebase for knowledge base construction, memory writing/reading/sparsification, model inference, training, and evaluation.
  - Conducted profiling-based optimization for training code, achieving 20× speed-up via batched memory encoding and FAISS IVFPQ acceleration.
  - Detailed design can be seen at project page and project report.