Juntai Cao

+1 778-723-7270 | jtcao7@cs.ubc.ca | juntaic7.github.io

LinkedIn | Github | Google Scholar | Semantic Scholar

Vancouver, BC, Canada

Profile

I am a passionate researcher exploring the frontiers of large language models (LLMs) and agents.

Experience

- **University of British Columbia** *06/2024 - 06/2025*
  Research Assistant Vancouver, Canada
  - Research topic: Multi-Document Summarization (MDS).
  - Introduced test-time scaling via repeated sampling and aggregation to improve MDS performance.
  - Proposed Consistency-Aware Preference (CAP) Score to mitigate positional bias when using LLM-as-Judge.
- **University of British Columbia** *06/2023 - current*
  Teaching Assistant Vancouver, Canada
  - CPSC 121 Models of Computation (2025S2)
  - CPSC 221 Basic Algorithms and Data Structures (2025S1)
  - CPSC 340 Machine Learning and Data Mining (2023W2)
  - CPSC 320 Intermediate Algorithm Design and Analysis (2023W1)
  - CPSC 320 Intermediate Algorithm Design and Analysis (2023S)

Education

- **University of British Columbia** *09/2023 - current*
  Master of Science in Computer Science Vancouver, Canada
  - Advisor: Jiarui Ding
- **University of California, Berkeley** *06/2021-08/2021*
  Visiting Student (Remote) Vancouver, Canada
- **University of British Columbia** *09/2020 - 05/2023*
  Bachelor of Computer Science Vancouver, Canada
- **University of British Columbia** *09/2016 - 05/2020*
  Bachelor of Applied Science in Materials Engineering with Distinction Vancouver, Canada

Projects

- **Retrieval Augmented Generation (RAG) in Commonsense Question Answering** *01/2024 - 03/2024*
  Commonsense, Question Answering, RAG, Agent [GitHub]
  - We leverage retrieval-augmented generation (RAG) techniques to enhance the commonsense reasoning capabilities of small language models. By dynamically retrieving relevant external knowledge, our approach allows smaller models to effectively incorporate contextual information, thereby improving their performance on commonsense reasoning tasks.

Publications & Preprints

[ACL'25 Main] **Why Prompt Design Matters and Works: A Complexity Analysis of Prompt Search Space in LLMs**
Xiang Zhang, **Juntai Cao***, Jiaqi Wei, Chenyu You, Dujian Ding. (2025).*

[In Submission] Tokenization Constraints in LLMs: A Study of Symbolic and Arithmetic Reasoning Limits
Xiang Zhang, **Juntai Cao***, Jiaqi Wei, Yiwei Xu, Chenyu You. (2025).*

[Preprint] Multi^2: Multi-Agent Test-Time Scalable Framework for Multi-Document Processing
Juntai Cao, Xiang Zhang*, Raymond Li, Chuyuan Li, Chenyu You, Shafiq Joty, Giuseppe Carenini. (2025).*

Skills

- **Programming Languages:** Python, C++, C, Julia
- **Machine Learning & Data Science:** PyTorch, PyTorch Lightning
- **Agent:** LangGraph
- **GPU Programming:** Triton
- **LLM Serving:** VLLM, SGLang
- **HPC Development Tools:** Docker, Slurm

Honors& Awards

- **Undergraduate Student Research Awards** *06/2022*
  University of British Columbia [🌐]
  - Awards are offered to undergraduate students to consider graduate studies and/or a research career by providing research work experience that complements their studies in an academic setting.
- **Trek Excellence Scholarship for Continuing Students** *01/2022*
  University of British Columbia [🌐]
  - Scholarships are offered to the top 5% of international undergraduate students.

Service

- **Reviewer**
  - 2025: ARR May

Additional Information

Languages: English (Proficient), Chinese (Native)
Interests: Board Games, Detective Fiction, Hiking