Tianjun Zhong

tianjun.zhong@columbia.edu | (470) 855-3939 | LinkedIn | GitHub

EDUCATION

Columbia University, Fu Foundation School of Engineering and Applied Science **New York, NY**
M.S. Computer Science, Track: Machine Learning (Advanced Master's Research) Sep. 2024 - May 2026

- **GPA**: 3.96/4.00
- **Relevant Courses:** Natural Language Processing (A+), Spoken Language Processing (A), Computer Vision (A), Artificial Intelligence (A-)

Emory University, Emory College of Arts and Sciences, Goizueta Business School **Atlanta, GA**
Double Major: B.A. Computer Science, B.B.A. Business Administration Aug. 2020 - May. 2024

- **Overall GPA**: 3.88/4.00, **Computer Science GPA**: 3.97/4.00
- **Awards**: Graduation with Distinction (top 15%), Dean's List Fall 2020, Fall 2021, Spring 2022, Fall 2023, Spring 2024
- **Relevant Courses:** Machine Learning (A), Analysis of Algorithm (A), Data Mining (A)

PUBLICATIONS

[1] **Brain-Predictive Reasoning Embedding through Residual Disentanglement**
Linyang He*, **Tianjun Zhong***, Richard Antonello, Gavin Mischler, Micah Goldblum, Nima Mesgarani
NeurIPS 2025 [OpenReview]

[2] **K2-Think: A Parameter-Efficient Reasoning System**
Zhoujun Cheng*, Richard Fan*, Shibo Hao*, Taylor W. Killian*, Haonan Li*, Suqi Sun*, Hector Ren, Alexander Moreno, Daqian Zhang, **Tianjun Zhong**, Yuxin Xiong, Yuanzhe Hu, Yutao Xie, Xudong Han, Yuqi Wang, Varad Pimpalkhute, Yonghao Zhuang, Aaryamonvikram Singh, Xuezhi Liang, Anze Xie, Jianshu She, Desai Fan, Chengqian Gao, Liqun Ma, Mikhail Yurochkin, John Maggs, Xuezhe Ma, Guowei He, Zhiting Hu, Zhengzhong Liu*, Eric P. Xing
arXiv preprint [arXiv] [Hugging Face] [GitHub]

* Equal Contribution

RESEARCH EXPERIENCES

Efficient Multi-Agent Systems for Collaborative Software Engineering
Columbia University Department of Computer Science New York, NY
Research Assistant (Advisor: Professor Baishakhi Ray) Sep. 2025 – Present

- Designing collaborative LLM agents (<32B) to assist developers with complex software engineering tasks, targeting a paper submission to ICML 2026; the goal is to achieve GPT/Claude/Gemini-level performance in agentic software engineering with 10–100 times lower computational cost and latency, and more details will be updated as the work advances.

Brain-Predictive Reasoning Embedding through Residual Disentanglement
Columbia University Department of Electrical Engineering New York, NY
Research Assistant (Advisor: Professor Nima Mesgarani) Jan. 2025 – Present

- Developed a residual disentanglement framework to isolate distinct linguistic representations (lexicon, syntax, meaning, reasoning) from LLM activations using minimal-pair diagnostic datasets (BLiMP, COMPS, ProntoQA, WinoGrande); employed layer-wise probing and iterative ridge regression to extract residual embeddings that are linearly unpredictable from lower-level features, achieving effective feature separation.
- Validated the framework by showing that each residual embedding selectively captures its intended feature through targeted diagnostic tasks, and mapped reasoning embeddings to human ECoG recordings, revealing stronger correlation with high-level reasoning cortical regions than classical language areas and later neural activation peaks.
- Provided the first evidence of brain-relevant reasoning representations in LLMs, advancing interpretability and enabling more precise analysis of model cognition; framework offers a generalizable methodology for probing and improving reasoning ability in large models, with paper accepted to *NeurIPS 2025*.

K2-Think: A Parameter-Efficient Reasoning System
UC San Diego Department of Computer Science & Engineering San Diego, CA
Research Assistant (Advisor: Professor Zhiting Hu) May 2025 – Sep. 2025

- Employed large-scale post-training techniques for 32B and 72B parameter LLMs to bridge the performance gap between open-source and proprietary reasoning models; engineered and managed distributed training jobs on

clusters of up to 512 GPUs using SLURM, configuring sequence parallelism, packing, and advanced learning rate scheduling for efficiency.
- Developed automated data curation and evaluation pipelines, authoring tools for model distillation and LLM-powered dataset labeling, leveraging PyTorch, DeepSpeed, and Hugging Face Transformers.
- Achieved state-of-the-art 86.26 on AIME 2024 and a competitive 60.90 on LiveCodeBench via supervised fine-tuning (SFT) and reinforcement learning (RL), outperforming open-source models at comparable scale; the resulting model was released publicly as *K2-Think* on Hugging Face, with the complete recipe—from data to training pipelines—made openly accessible through the project's technical report and GitHub repository.

Retrieval-Augmented Generation with Unsupervised Learning Enabling Hallucination Mitigation
Columbia University Department of Computer Science New York, NY
Research Assistant (Advisor: Professor Vishal Misra) Sep. 2024 - Dec. 2024

- Developed a Retrieval-Augmented Generation (RAG) system to address academic advising bottlenecks, reducing faculty workload and improving student access to timely information; deployed the virtual advising assistant in Columbia's CS Department with automatic escalation of out-of-domain queries to human advisors.
- Designed an unsupervised out-of-domain query detection mechanism by modeling knowledge bases as hyper-ellipsoid clusters in embedding space, implementing agglomerative clustering to group query embeddings and statistical methods to identify queries outside knowledge boundaries.
- Achieved a 78.8% rejection rate for queries prone to LLM hallucinations, significantly improving response factual reliability, and preparing to scale the system to serve all 6,700+ Columbia Engineering students following the successful pilot implementation.

WORK EXPERIENCES

TripleE Group, Department of Product R&D **Shenzhen, China**
AI Engineering Intern May 2024 - Aug 2024

- Designed and implemented three original algorithms to improve RAG pipelines: (1) enhanced query–document matching, (2) dynamic top-*k* retrieval, and (3) optimized text splitting for vector database construction.
- Evaluated three state-of-the-art embedding models for retrieval accuracy, integrated a reranker for relevance ranking, and achieved 98.45% top-1 accuracy on synthetic high-difficulty queries containing ambiguous concepts and expressions.
- Developed efficient web scraping methods to collect over 1.27M tokens of text, constructing 23 FAISS vector databases for downstream retrieval tasks.

Tencent, Department of IDC Platform **Shenzhen, China**
Software Engineering Intern May 2023 - Jun 2023

- Developed Python automation tools to streamline daily operations for the IDC Platform team, reducing manual workload and redundancy by 90%.
- Built programs to monitor 30+ data centers and automatically generate two daily operational reports.
- Integrated solutions with the institution's Git platform and cloud database, creating reusable Python functions for team-wide adoption.

TEACHING & CONTEST

Emory University, Goizueta Business School **Atlanta, GA**
Teaching Assistant (Course: Applied Data Analytics with Coding) Jan 2024 - May 2024

- Mentored 122 students through personalized guidance on Python and SQL during weekly two-hour office hours, explaining complex concepts in clear language to a diverse student body.
- Delivered detailed, constructive feedback on assignments, projects, and exams, helping students improve both technical skills and conceptual understanding.

International Collegiate Programming Contest (ICPC), *Team Leader, Contestant* Sep 2022 - Feb 2024

- Ranked 3rd among 103 teams in 2023 ICPC Southeast USA Regional Contest, 19th among 110 teams in 2022 ICPC
- Exercised Java and Python skills in dynamic programming, data structures, greedy algorithm in weekly two-hour practices

SKILLS

Programming: Python (Transformers, PyTorch, TensorFlow, Scikit-learn, NumPy), Java, SQL, C, R
Language: English (primary instruction language for 10 years, professional fluency, GRE 335), Mandarin (native speaker), Cantonese (intermediate)