Tianjun Zhong

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

EDUCATION

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

- **GPA**: 3.97/4.00
- **Relevant Courses:** Natural Language Processing (A+), Spoken Language Processing (A), Computer Vision (A), AI (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 20%), 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] **SWE-Spot: Building Small Repo-Experts with Repository-Centric Learning**
Jinjun Peng*, Magnus Saebo*, **Tianjun Zhong**, Yi-jie Cheng, Junfeng Yang, Baishakhi Ray, Simin Chen, Yangruibo Ding
ICML 2026 (under review) [arXix]

[2] **From Chains to DAGs: Probing the Graph Structure of Reasoning in LLMs**
Tianjun Zhong, Linyang He, Nima Mesgarani
ACL 2026 (under review) [arXiv]

[3] **Far from the Shallow: Brain-Predictive Reasoning Embedding through Residual Disentanglement**
Linyang He*, **Tianjun Zhong***, Richard Antonello, Gavin Mischler, Micah Goldblum, Nima Mesgarani
NeurIPS 2025 [arXiv] [OpenReview] [GitHub]

[4] **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

Probing the Graph Structure of Reasoning in LLMs
Columbia University Department of Electrical Engineering — New York, NY
Research Assistant, Project Lead (Advisor: Professor Nima Mesgarani) — Nov. 2025 – Present

- Lead the development of Reasoning DAG Probing, a framework for testing whether LLM hidden states encode graph-structured reasoning beyond linear chain-of-thought; Design structural probes to recover reasoning node depth and pairwise distance from hidden states, enabling layer-wise localization and reconstruction of reasoning-relevant structure.
- Conduct large-scale evaluations across model sizes and families, showing that reasoning DAG geometry emerges most strongly in intermediate layers, with deeper reasoning steps becoming maximally recoverable at later layers than shallow ones, scales with model size, and correlates with downstream reasoning correctness; submitted to ACL 2026.

Repository-Centric Learning for Software Engineering Agents
Columbia University Department of Computer Science — New York, NY
Research Assistant (Advisor: Professor Baishakhi Ray, Mentor: Yangruibo (Robin) Ding) — Sep. 2025 – Present

- Develop Repository-Centric Learning (RCL) for training small software engineering agents, enabling them to internalize repository-specific structure through sustained interaction rather than task-centric exposure across unrelated codebases.
- Design a four-unit agent experience pipeline spanning codebase understanding, agentic fill-in-the-middle, evolutionary replay from version history, and semantic-runtime alignment via test generation; train SWE-SPOT-4B, demonstrating strong multi-task performance that outperforms open-weight models up to 8× larger and matches efficiency-focused commercial agents.

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 70B 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); 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.

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

- Developed a residual disentanglement framework to isolate distinct linguistic representations (lexicon, syntax, meaning, reasoning) from LLM activations using minimal-pair datasets; 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*.

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

- Built and deployed a Retrieval-Augmented Generation (RAG) system for academic advising in Columbia CS, reducing faculty workload via automated response handling and escalation of out-of-domain queries.
- Designed an unsupervised out-of-domain detection method using agglomerative knowledge clustering and statistical testing, achieving a 78.8% rejection rate on hallucination-prone queries and improving response reliability.

WORK EXPERIENCES

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

- Designed and implemented three RAG pipeline improvements to improve retrieval quality: enhanced query-document matching, dynamic top-k retrieval, and optimized text splitting.
- Evaluated multiple embedding models and integrated reranking, achieving 98.45% top-1 accuracy on high-difficulty synthetic queries, and built 23 FAISS vector databases from 1.27M+ tokens of scraped data for downstream retrieval.

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

- Developed Python automation tools for monitoring 30+ data centers, automatically generating twice-daily operational reports and submitting operational requests (e.g., on-site inspections), reducing manual workload by ~90%.

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 weekly office hours; delivered detailed, constructive feedback on assignments, projects, and exams, helping students improve both technical skills and conceptual understanding in Python and SQL.

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 language of instruction for 10 years, professional fluency, GRE 335), Mandarin (native speaker), Cantonese (elementary), Japanese (elementary)