Curriculum Vitae - Tevin Wang

Info

Google Scholar: Tevin Wang
Github: TevinWang
Email: tevin(at)cmu.edu
Website: tevinwang.com

Research Interests

I am interested in improving the capabilities of LLMs, specifically through tightening the generation-verification gap and raising the verification ceiling.

Education

Carnegie Mellon University, Pittsburgh, PA

B.S. Computer Science, Expected Graduation: May 2026

- *Concentration in Machine Learning*
- *GPA*: 4.00/4.00
- *Relevant Coursework*: Introduction to Deep Learning, Generative AI, Machine Learning with Large Datasets, Introduction to Machine Learning, Mathmatical Foundations for Computer Science, Matrices and Linear Transformations, Concepts of Artificial Intelligence

Publications

AutoRule: Reasoning Chain-of-thought Extracted Rule-based Rewards Improve Preference Learning
Tevin Wang, Chenyan Xiong
arXiv preprint

Deep Research Comparator: A Platform For Fine-grained Human Annotations of Deep Research Agents
Prahaladh Chandrahasan*, Jiahe Jin*, Zhihan Zhang*, **Tevin Wang**, Andy Tang, Lucy Mo, Morteza Ziyadi, Leonardo F.R. Ribeiro, Zimeng Qiu, Markus Dreyer, Akari Asai, Chenyan Xiong
arXiv preprint

Understand User Opinions of Large Language Models via LLM-Powered In-the-Moment User Experience Interviews
Mengqiao Liu*, **Tevin Wang***, Cassandra A. Cohen, Sarah Li, Chenyan Xiong
Findings at ACL 2025

Interpret and Control Dense Retrieval with Sparse Latent Features
Hao Kang, **Tevin Wang**, Chenyan Xiong
NAACL 2025

RAGViz: Diagnose and Visualize Retrieval-Augmented Generation
Tevin Wang, Jingyuan He, Chenyan Xiong
EMNLP 2024 System Demonstrations

Research Experience

Language Technologies Institute (LTI), CMU
Undergraduate Researcher, Advisor: Chenyan Xiong — Jan 2024 - Present

- Conducted research on mitigating LLM reward hacking, with hands-on experience in post-training (SFT, RL) and benchmarking methods
- Adapted mechanistic interpretability techniques for LLMs to dense retrieval
- Contributed to the modeling team in an Amazon AGI-LTI collaboration on RAG, co-organizing a competition at NeurIPS 2025 for multimodal RAG

Industry Experience

OpenAI
Member of Technical Staff (Intern) — Summer 2025

- Team: Integrity

Pinecone
Software Engineering Intern Summer 2024

- Worked on the Assistants team to build a managed service for performant retrieval augmented generation (RAG)
- Pinecone's RAG product outperforms OpenAI Assistants in RAG tasks related to text-heavy technical data
- Deployed a shared cache layer for blob storage of the Assistants product to improve p95 time to first token latencies
- Built a version migrations tool for the document ingest pipeline that maintains full document availability, utilizing distributed systems concepts like shadowing, pointer swinging, and atomic transactions

LanceDB (YC W22)
Software Engineering Intern Summer 2023

- Developed optimized product quantization feature flag and Rust-based linear algebra implementations for in-house Lance format
- Introduced critique-based contexting: reducing hallucinations from AI agents using long-term memory
- Created AI use cases with LLMs and agents that utilize serverless open source vector database LanceDB

Chess.com
Software Engineering Intern Summer 2021

- Worked in Frontend Stats team to create, organize, and display data visualization pages in Vue and Typescript for user chess puzzle stats
- Cold-contacted CEO for the internship during junior year of high school, interviewed CEO for management/entrepreneurial advice

Leadership Experience

ScottyLabs
Director of Outreach Mar 2023 - Mar 2025

- Managed recruitment and retention, content/social media, and event outreach for CMU's premier technology club

Early Projects

Reducing Hallucinations in LLM-powered Agents using Long-Term Memory
Tevin Wang July 2023

- Created concept of critique-based contexting, building long-term memory system using embeddings + vector DBs and designing a self-critique feedback loop for continual agent improvement

Analyzing the Reddit Subreddit Network using the PageRank and HITS Algorithms
Kevin He, Tevin Wang December 2022

- Utilized analytical and numerical approaches in Julia to apply the PageRank/HITS algorithms to Stanford SNAP datasets on Reddit hyperlinks

Awards

Dean's List, High Honors (5 semesters)
Hudson River Trading Challenge Winner - HackCMU

Skills

RL for LLMs, RLHF, Alignment, Supervised Fine-Tuning, RAG, Mechanistic Interpretability, Deep Research, Python, Deep Learning Systems, PyTorch, Rust, Cloud Infrastructure