Carter Teplica

c.teplica@gmail.com ◆ +1 (203) 969-5048 ◆ `github.com/crtep`

Education

New York University — *Sept. 2023 – expected May 2025*

Master of Science in Computer Science
GPA: 3.70

Columbia University — *Sept. 2019 – May 2023*

Bachelor of Arts in Mathematics, concentration (minor) in Physics
GPA: 3.79; Math Dept. GPA: 4.03

Budapest Semesters in Mathematics — *June – August 2022*

GPA: 4.0

Research Experience

Tim G. J. Rudner, New York University, and Arman Cohan, Yale University — *Feb. 2024 – present*

Paper: Carter Teplica, Yixin Liu, Arman Cohan, and Tim G. J. Rudner. SCIURus: Shared Circuits for Interpretable Uncertainty Representations in Language Models. 2024.

Paper and code at `github.com/crtep/sciurus`

First author. Planned research directions and carried out experimental work independently; wrote code and majority of paper.

Studied mechanistic processes by which uncertainty estimates arise in LLMs by applying causal tracing and other interpretability techniques to probing-based uncertainty quantification in novel ways.

Formulated, refined, and statistically tested a hypothesis ("shared circuits") about localization of uncertainty quantification circuitry in LLMs.

Presented at MINT, ATTRIB, SFLLM, and SafeGenAI workshops at NeurIPS in December 2024. Under review for NAACL.

Tal Linzen, New York University: Comp. Linguistics and Cog. Sci. (class final project) — *Sept. – Dec. 2024*

Hongxin Song and Carter Teplica. Sociolinguistic Simulacra: Interactions Between Language and Attitudes in Finetuned Language Models. 2024.

Paper at `tinyurl.com/song-teplica-simulacra`

Formulated research question, designed and curated datasets, and finetuned language models using direct preference optimization. Will submit for publication in early 2025.

Studied influence of low-level linguistic bias in preference datasets on high-level personality and self-reported demographic features in tuned models.

Joan Bruna, New York University: Mathematics of Deep Learning (class final project) — *Jan. – Apr. 2024*
Matus Telgarsky, New York Univ.: Conceptual Gaps in Modern ML (class final project) — *Jan. – Apr. 2024*

Carter Teplica. Singularities and the Edge of Stability. 2024.

Post at `tinyurl.com/teplica-singularities`

Studied the "edge of stability" phenomenon from the perspective of singular learning theory. Wrote a blog post describing experimental results. Project for both classes.

Yibo Jiang and Victor Veitch, XLab, University of Chicago *June – August 2023*
Blog post: Carter Teplica. A Mechanistic Analysis of Counting in Distil-GPT2. 2023.
Post at `tinyurl.com/teplica-distil`

Carried out a mechanistic interpretability study of counting and number representations in a large language model.

István Miklós, Rényi Institute, Budapest Semesters in Mathematics *June – August 2022*

Proved the four-reversal conjecture for the infinite site model, a combinatorics problem with applications to genomics.

Poster presented at Joint Mathematics Meetings, January 2024.

Marcel Agüeros, Columbia University *Jan. 2021 – May 2022*

Completed a project using unsupervised learning to construct a new membership list for the Alpha Persei stellar cluster.

Poster accepted to American Astronomical Society meeting, 2022.

Honors and Awards

Scholarship Grant, Long Term Future Fund *Sept. 2023 – May 2025*

Awarded based on promise as an early career AI safety researcher. Covers tuition, housing, and living expenses.

Bruce Fishkin Scholarship *Sept. 2019 – May 2023*

Merit-based scholarship covering most of my tuition.

Columbia Science Research Fellow *Sept. 2019 – May 2023*

Merit-based funding for summer research.

Van Amringe Mathematics Prize *April 2022*

Best score in Columbia College graduating class on a Putnam-like exam.

Standardized test scores:

GRE: 170 verbal / 169 quantitative / 890 math subject test *Sept. 2022 and May 2023*
ACT: 36 math / 36 reading / 36 science *Fall 2018*

Leadership, Academic, and Professional Experience

Research Mentor, Existential Risk Laboratory (XLab), University of Chicago *June – August 2024*

Mentored an undergraduate student in a mechanistic interpretability research project.
Gave advice on techniques, project management and developing research taste.

Summer Research Fellow, Existential Risk Laboratory (XLab), University of Chicago *June – August 2023*

Completed a research project in mechanistic interpretability (see above). Developed research skills. Attended talks by researchers in AI safety and governance and other existential risk areas. In-person; received a stipend.

Facilitator, AI Safety Fellowship, Columbia AI Alignment Club — *Sept. 2022 – May 2024*

Facilitated three semesters of a technical AI safety reading and discussion group, primarily for graduate students.

Served as discussion group leader, organized weekly meetings, and redesigned the fellowship curriculum.

ML Safety Scholars Fellow, Center for AI Safety — *Sept. – Dec. 2022*

AI Safety Fellowship, Columbia AI Alignment Club — *Jan. 2022 – May 2022*

Tutor, Top Hat Tutors — *Jan. 2018 – August 2019*

Tutored high, middle and elementary school students in mathematics, Latin, and standardized test prep. Designed a mathematics enrichment curriculum with lessons in topology and geometry.

Selected conferences and workshops attended:

NeurIPS, Vancouver, BC — *December 2024*

Presented research in mechanistic interpretability and uncertainty quantification at four workshops. Received travel stipend.

AISST/MAIA AI Safety Workshop, Essex, MA — *March 2024*

AI Risks Workshop, Berkeley, CA — *December 2022*

Workshops on technical AI safety and governance. Attended talks by researchers in interpretability, alignment, and technical governance. Received travel stipends.

Skills and Coursework

Programming skills:

Experience building, training and interpreting deep neural networks. Experience managing complex, compute-intensive research projects on large academic clusters.

Proficiency in Python, C, C++, incl. CUDA; experience with Haskell, Rust, R, JS.

Selected coursework:

Machine learning: deep learning; RL; ML; causal inference; GPUs; math of DL; computational linguistics and cognitive science.

Other computer science: operating systems; cryptography; heuristic problem solving.

Mathematics and physics: graph theory; combinatorics; point-set, algebraic, and differential topology; analysis; algebra; PDEs; quantum mechanics.

Other skills and experience:

Languages: Spanish (conversational, reading-proficient); Latin (reading-proficient); Mandarin Chinese (intermediate).

Layout editor, Columbia Undergraduate Science Journal.

Took AP Calculus BC in seventh grade (age 11).

Wrote and organized an extensive puzzle hunt.

Singer and jazz a cappella arranger.