# Evan Ryan Gunter

evgunter@gmail.com website github 510 812 7851

## Professional Experience

**ML Alignment & Theory Scholars**

*Research Scholar – David Lindner* *collab: Evžen Wybitul, Mikhail Seleznyov* 01/24—03/24

To assess the capacity of vision-language models to be process supervisors for RL, developed a benchmark for fundamental capabilities and compositions of capabilities in complex tasks [x]

*Long-Term Future Fund Grantee* 09/23—01/24

Investigated loss landscape geometry theoretically using results from high-dimensional math and physics; tested related predictions of Singular Learning Theory

*Research Scholar – Victoria Krakovna* *collab: Yevgeny Liokumovich* 06/23—09/23

Proved theorems on stability of non-power-seeking for Markov decision process (MDP) policies with bounded gradient [x]; developed improved formalism of power as optionality for MDPs [x]

**Granica**

*Research Engineer* 02/22—07/23

Data compression research; used sketch algorithms, statistical modeling, black box optimization, Bayesian estimation, spectral clustering, integer linear programming, automated hyperparameter tuning, gradient descent, data imputation, singular value thresholding

*Software Engineer* 02/21—02/22

Streamlined the cloud infrastructure deployment process and developed new CLI functionality for a data compression product; improved internal tools for reducing costs and building packages

**Berkeley Existential Risk Initiative** *Research Assistant to Anders Sandberg* 04/21—12/21

For a book draft, checked physics derivations and consulted on philosophical, scientific, and other content in weekly one-on-one meetings

**Theorem LP** *Engineering Intern* 07/18—09/18

Detected duplicate applications using nearest neighbor search in SQL; created custom Protocol Buffer serialization scheme to improve performance; configured ETL pipelines with Terraform

**California Institute of Technology**

*Head Deans' Tutor, Calculus of One and Several Variables & Linear Algebra* 09/18—06/19

*Head Deans' Tutor, Classical Mechanics and Electromagnetism* 04/18—06/18

*Teaching Assistant: Fundamentals of Computer Programming, Introduction to Discrete Mathematics, Principles of Biology* 04/18—03/19

*Summer Undergraduate Research Fellowship* *advisor: Erik Winfree* 06/17—08/17

Studied the implementation of randomized algorithms with stochastic chemical reaction networks; developed example algorithms and proved performance bounds for them

*Deans' Tutor, misc. math, physics, and computer science courses* 09/16—06/19

*Peer Tutor, Hixon Writing Center* 04/16—06/19

**JPL Science Data Modeling and Computing Group** *Intern* 07/16—09/16

Improved Climate Model Diagnostic Analyzer data processing pipeline generality and reliability

## Education

**California Institute of Technology**

*BS Mathematics, BS Computer Science, BS Philosophy* *3.6 GPA* 09/15—06/19

Only 2019 triple major; research in computer science and philosophy; thesis in philosophy of physics; eight A+'s; 14 physics courses (3 graduate-level); peer tutoring and teaching assistance in computer science, math, biology, writing, and physics

## Projects

**Independent**

Personalized ML chording keymap optimization, Rust macros to control expansion order of other macros, Telegram bot for GPT-4 API access, improved repositories for project dependencies [x][x], found a high-severity security bug in Android

**Mentorship for Alignment Research Students (MARS)** *Mentor* 01/24—05/24

Mentored students in projects on loss landscape geometry with different optimizers [x], extending results on deep linear net minima to nets with ReLUs, and how whether model training Markov chains "mix" quickly enough for the training process to resemble MCMC

**California Institute of Technology**

*Undergraduate Projects in C.S.* *advisor: Mike Vanier* *collab: Aidan Swope* 04/19—06/19

Investigated AlphaZero-inspired efficient tree search for automated theorem proving

*Philosophy thesis* *advisor: Chip Sebens* *Anthropic reasoning in infinite worlds* 09/18—06/19

Argued that the self-indication assumption for anthropic reasoning is less arbitrary, more predictive, and has fewer counterintuitive consequences than Bostrom's self-selection assumption; addressed mathematical issues in infinite worlds; applied findings to spacetime dimensionality

*Reading in Philosophy* *advisor: Frederick Eberhardt* *collab: Alex Denko* 01/19—04/19

Wrote two papers: one against Tegmark's mathematical universe hypothesis with arguments from Russellian monism; one on implications of panpsychism and personal identity for Parfit's repugnant conclusion in population ethics

**University of California, Berkeley**

*Linguistics Research Apprentice Practicum (Ling. 197)* 01/15—05/15

Prepared phonetics data for analysis; wrote code to do some preparation programatically

*Linguistic Typology (Ling. 222)* 01/14—05/14

Synthesized linguistic data into original analyses; wrote 40-page research paper on syntactic phenomena in the language Kolyma Yukaghir

*Introduction to Phonetics and Phonology (Ling. 110)* 08/13—12/13

Collected and analyzed phonetic data; wrote 20-page research paper on Mandarin phonetics

## Papers

*Quantifying stability of non-power-seeking in artificial agents* 01/24

*collab: Yevgeny Liokumovich, Victoria Krakovna*

*NGD converges to less degenerate solutions than SGD* 09/24

*collab: Moosa Saghir, Raghavendra Narayan Rao, Zihe Liu*