Artemis Veizi

aveizi@stanford.edu | (703) 462-0980 | linkedin.com/in/artemisveizi

EDUCATION

Stanford University *Expected Dec 2027*
M.S. Computer Science (Artificial Intelligence) GPA: 4.225/4.0
Coursework: AI Measurement Science, Continuous Methods for ML, Algorithmic Fairness
Research Collaborator in the LDR (Language, Data, Reasoning) Lab under Amin Saberi and Shayan Talaei; eliciting higher LLM output diversity via weight-space perturbation.

Princeton University Class of 2023
B.S.E. Electrical and Computer Engineering (Computer Systems) GPA: 3.81/4.0
Senior Thesis: "ML for Network Optimization: Buffer Sizing via Reinforcement Learning"
Advised by Dr. Maria Apostolaki
Honors: G. David Forney Jr. Thesis Prize, Sigma Xi, magna cum laude

SKILLS

Languages & Frameworks: Python, C++, C, PyTorch, Hugging Face Transformers, TensorFlow, scikit-learn, NumPy, pandas, Jupyter

AI/ML: reinforcement learning, constrained optimization, on-device learning, quantitative model evaluation, statistical anomaly detection, LLM benchmarking, probabilistic modeling, psychometrics, measurement theory

RESEARCH & PROJECTS

Language Model Benchmark Contamination Leaves a Person-Fit Signature in the Graded Response Matrix — Under review, AAAI-27. Detected benchmark contamination from the graded response matrix alone via guessing-aware IRT person-fit; validated under controlled fine-tuning injection across 4 base models × 3 benchmarks (AUROC 0.94–0.99).
Person-Fit Statistics as a Black-Box Detector for LLM Sandbagging on Hazardous-Capability Benchmarks — Detected deliberate LLM underperformance on WMDP via Rasch IRT person-fit statistics; validated on AISI model organisms.
Cold-Start Predictive Evaluation on AI Benchmarks — Predicted per-item LLM correctness on unseen questions via LLM-as-judge categorization, online Platt calibration, and hierarchical Bayesian shrinkage (−0.60 log-loss, 0.70 AUC-ROC).

EXPERIENCE

Apple, Inc. Silicon Engineering Group Cupertino, CA
Silicon Validation Engineer August 2023–Present
- Designed and implemented automated evaluation frameworks for silicon validation data, defining metrics for measurement consistency and applying statistical methods to detect anomalous results and optimize test coverage across multiple hardware platforms.
- Developed ML pipeline combining on-device simulated annealing with off-device reinforcement learning for silicon validation; extended optimization tool across multiple test platforms. Presented in an Apple internal research forum.

Silicon Validation Engineering Intern June 2020–May 2021, May–August 2022
- Designed and implemented a lightweight simulated annealing algorithm in C for resource-constrained embedded environments, targeting SoC performance counters to optimize validation workloads.
- Implemented and evaluated warm-start & other improvements for on-device optimization algorithm, reducing memory requirements for resource-constrained embedded deployment.
- Built a Python data processing and anomaly detection pipeline for SoC performance data.

EXTRA–CURRICULAR ACTIVITIES

Co-Founder & Co-Chair, Early Career Professionals at Apple (2023–2025)

Lab Teaching Assistant, ECE302: Robotic and Autonomous Systems Lab (2023)

Princeton Varsity Women's Lightweight Rowing, D1 (2018–2023)
National Champions (2022, 2023), IRA All-Academic Team (2022, 2023)

Writing Center Fellow, Princeton University Writing Center (2019–2022)

INTERESTS

Albanian (first language), French (fluent), billiards, road cycling, trail running, sourdough.