Zachary Burton

617-417-1070 | zacnburton@gmail.com | linkedin.com/in/zac-burton | github.com/zacn04

Education

Massachusetts Institute of Technology (MIT) — Cambridge, MA
B.S. in Mathematics (GPA: 4.6/5.0) — *Sep 2022 – Feb 2026*

Publications

Inference-Time Diversity in RL-Trained Lean Theorem Provers: A Diagnostic Study | Under Review | arXiv:2601.16172v2

QEDBench: Quantifying the Alignment Gap in Automated Evaluation of University-Level Mathematical Proofs | ICML 2026 (Poster) | PhD-level expert evaluator (probability, discrete mathematics, algorithms) | icml.cc/virtual/2026/poster/63067

Poissonization-based collision threshold derivation for random walks on lattices | arXiv:2505.02973

Experience

Meta — Feb 2026 – Present
Machine Learning Engineer — *New York, NY*

- Shipped a revenue optimization launch boosting billing-based revenue by **+0.251%** across Facebook Reels surfaces; fastest ramp-to-launch on team
- Designed and drove a cohort-aware calibration system for a listwise ranking model, segmenting users into decile-based revenue cohorts and iterating through 7 experiment configurations with a **+0.22%** revenue lift in early signal
- Co-authored two launch proposals for position-based video length tuning in a late-stage reranker, yielding **+0.185%** revenue-per-view (head-load) and **+0.525%** tail-load watch time globally

Research

Inference-Time Diversity in RL-Trained Lean Theorem Provers — Jan 2026 – Present
Solo project; arXiv:2601.16172v2; submitted to AI4Math Workshop @ ICML 2026

- Diagnostic study of inference-time mode collapse in RL-trained Lean provers; built a 4-model experimental matrix (DeepSeek-Prover-V1.5-RL/Base, V2-7B, Goedel-Prover-SFT) on miniF2F-test, with multi-seed (n=3) variance reporting and a leave-one-out empirical formalization-difficulty stratification methodology
- Showed RL-vs-SFT asymmetry at $\mathbf{3.7}\sigma$: V1.5-RL gains $\mathbf{+12.3 \pm 4.2}$ theorems under a fixed 15-skeleton schedule across 3 seeds, while Goedel-Prover-SFT loses -10.0 ± 4.4 under the same intervention; V1.5-Base (no RL) proves zero theorems with or without skeletons, identifying RL as the locus of both proof capability and inference-time collapse
- Direct distributional measurement of mode collapse: across **13,159** stochastic samples, the median V1.5-RL theorem receives only **2** distinct first-tactic heads from 64 i.i.d. samples (median Shannon entropy 0.27 bits / 6.0 max), with **49%** of miniF2F receiving a deterministic single strategic opening
- Built multi-pod vLLM orchestration for 4 open-source provers across A100 80GB/40GB tiers; diagnosed and fixed a reasoning-mode output-extractor bug (unfenced `import Mathlib` + nested theorem pattern emitted by Kimina) via stderr instrumentation; 8-way sharded experiments with per-attempt JSON logging supporting downstream entropy analysis
- Pinned Lean v4.9.0-rc1 / Mathlib commit `7fa489a5` to V1.5-RL's training environment to rule out the newer-tactic-solver confound that invalidated v1 of this work

PIEFACE: Live RLHF Platform for Interactive Policy Personalization — 2025 – Present
Solo full-stack project; deployed at pieface.ai

- Deployed full-stack ML web application: Flask + SQLAlchemy backend serving 3 production Maskable PPO models (sb3-contrib), Cytoscape.js graph-visualization frontend, session-based state management with rate-limited API endpoints (200/day, 50/hour, 2/sec) and CORS-locked production deployment
- Top-K action inference with action masking over a **32**-dim observation / **217**-dim discrete action space; per-action confidence scores surfaced to the UI for human selection
- Personalization pipeline: SQLite-backed preference table (state vector, action, accept/reject label, timestamp), custom PyTorch `RewardMLP` trained on collected preferences (SiLU + dropout), action re-ranking by live reward-model scores, per-episode telemetry (success, used_rlhf flag, reward-model version) for offline analysis (1/2)

– End-to-end RLHF loop in production: user steps through agent trajectories, accepts or rejects individual moves, reward model updates on collected preferences, and the agent's subsequent action distribution shifts toward the user's revealed preferences

Representational Limits of Tabular and Neural RL on Symbolic Algebra Fall 2025

MIT 6.7920 (Reinforcement Learning) final project

– Built custom Gym environment for solving single-variable linear equations as sequential AST rewrites; tested PPO (GRU encoder), tabular Q-Learning, SARSA, and a Neural Q-Learning baseline across equation depths 1–4 with 20K–50K episodes per agent

– PPO solve rate degrades from **38%** (depth-1) to **1%** (depth-4); tabular Q-Learning and SARSA are statistically indistinguishable from a random baseline (∼**5%**) at every depth—a representational, not sample-complexity, failure

– Confirmed the diagnosis with two ablations: a state-normalization scheme that compresses **99%** of the state space ($341 \rightarrow 2$ unique states for depth-1) with no solve-rate improvement, and a GRU-encoded neural baseline that delivers 2–4× tabular performance but still fails depth-2 (**4%**)

RL for Automated Search over Motion-Planning Gadget Structures Nov 2024 – Aug 2025

MIT CSAIL UROP | Advised by Jayson Lynch & Erik Demaine

– Built the OOP infrastructure for the project: gadget definitions (Door, Toggle2, AntiParallel2Toggle, Crossing2Toggle, locking variants), Hopcroft DFA minimization, equivalence-verification test suite, and DFS/BFS/random baseline searches over gadget combination operations

– Formulated gadget-simulation construction as a Maskable PPO problem over $\sim 10^4$ discrete actions per state and identified a generalization failure: trained agents memorized task-specific action traces rather than learning compositional construction logic, with catastrophic failure under planarity constraints and held-out gadget classes