ELLIOT TOWER

elliot@elliottower.ai | elliottower.ai | Boston, MA

EDUCATION

University of Massachusetts Amherst — Amherst, MA

M.S. in Computer Science — Concentration in Data Science (Machine Learning) — *Sep. 2020 – May 2022*

B.S. in Mathematics — Concentration in Computing, *Second major in Philosophy* — *Sep. 2016 – May 2020*

EXPERIENCE

Visiting Researcher (Machine Learning) — University of Edinburgh, School of Informatics Jun. 2026 – Present

- Mechanistic Interpretability: developed *Subspace Attribution Patching* (SAP) for localizing causally important subspace communication channels per circuit edge. Performs causal variable and circuit localization jointly (MIB).

Machine Learning Engineer — Eluve Inc Jul. 2024 – Mar. 2026

- Designed and validated multi-level evaluation frameworks for summary quality, hallucination detection. Benchmarked frontier models, multi-step pipelines. Worked with clinicians and designed automated preference alignment system.
- Developed tooling for longitudinal patient state and insights. Combined information extraction, retrieval, and medical knowledge graphs to produce alerts for drug-drug interaction, documentation gaps, and clinical decision support

Machine Learning Engineer — Swarm Labs May. 2023 – Jun. 2024

- Developed novel LLM evaluation benchmarks for adversarial robustness and red-teaming, AI safety and AI security.
- Benchmarked frontier and open-source LLMs, discovered novel failure modes, and created model scorecard reports.

Open-Source Project Manager — Farama Foundation Mar. 2023 – Jun. 2024

- Project manager of PettingZoo, the API standard for multi-agent RL. Developed ML training integrations and tooling.

INDEPENDENT RESEARCH

Reproducible Science [Website] [Paper] - NeurIPS 2026, AI4Science Jul. 2026 – Present

- Developed an API standard and engine for declarative, continuous, machine-checkable provenance layer for research claims. Enables work to be reproducible from the start, verifying every claim. Staleness checks, drift, hallucination detection. Claude Code/Codex plugins. React UI for auditing. Used in preregistered study (*Frontiers in Pharmacology)*

Mechanistic Validity Framework [Website] [Preprint] May 2026 – Aug. 2026

- Developed domain-general validity framework for the triangulation and evidence formalization of mechanistic claims.
- Unified validation methodology from philosophy of science, psychometrics, neuroscience, pharmacology, genetics.
- 36-criteria instrument and scoring engine: analyzed 16 published results (IOI, SAE, probing), identified systemic gaps.

Mechanistic Views Framework [Website] [Preprint] May 2026 – Aug. 2026

- Created framework for formalizing mechanism disagreements across epistemic views (e.g., neuron vs SAEs, J-lens).
- Surveyed 17 open problems in mechanistic interpretability: resolved 8 as ill-posed due to view-evidence mismatch, 2 reframed, 7 experimentally unblocked by adopting a different view. Experimental results with empirical validation.

Taxonomy of LLM Reasoning Failures Mar. 2026 – May. 2026

- Developed suite for novel reasoning failures in frontier models: use-mention, compositional reasoning, self-evaluation
- Developed unique solutions for each failure mode, benchmarking scaffolds for mid-generation detection & mitigation.

PUBLICATIONS

- Das R, Godbole A, Naik A, **Tower E**, Zaheer M, Hajishirzi H, Jia R, McCallum A. **Knowledge Base Question Answering by Case-based Reasoning over Subgraphs.** ICML, 2022. [Proceedings]
- **Tower E. Cross-Design Evidence: Mendelian Randomization Status Alone Explains the Classification of Phase III Drug Outcomes Across Ten Disease Domains.** Frontiers in Pharmacology, 2026. [Link]