

Elena Ajayi

elenaajayi@outlook.com | [linkedin.com/in/elenaajayi](https://www.linkedin.com/in/elenaajayi) | github.com/elenaajayi | <https://elenaajayi.github.io/>

EDUCATION

St. John's University, New York City, NY

M.S. in Data Science (GPA: 3.71); B.S. in Biomedical Sciences, Minor: French (GPA: 3.77)

TECHNICAL SKILLS

Languages: Python, R, SQL, MATLAB **ML/Data:** PyTorch, Hugging Face Transformers, NumPy, pandas, scikit-learn
Infrastructure: Linux/UNIX, bash, Git/GitHub, Docker, PostgreSQL, Redis, pytest

PUBLICATIONS

- Ajayi, E.**, Chowdhury, A., & Lazar, S. (Jun 2026). **Incoherent Values? Probing LLM Preferences Through Parametric Variation.** *arXiv* (Preprint). [Link](#).
- Gilda, P., Dungarwal, P., Thongkham, A., **Ajayi, E. T.**, et al. (Jan 2026). **Measuring and Fostering Peace through Machine Learning and Artificial Intelligence.** *arXiv* (Preprint). [Link](#).
- Ajayi, E.** & Victor, J. (Aug 2023). **Statistical characterization of medical images of bone.** *Journal of Vision* (Abstract). [Link](#).

PROFESSIONAL EXPERIENCE

Black in AI Safety and Ethics Fellowship

Remote

Research Engineer Fellow (PI: Anagha Late)

April 2026 – Present

- Extracted residual-stream activations from **Llama 3.1 8B** with TransformerLens and trained linear probes under stratified and leave-scenario-out cross-validation to test whether linear deception-detection methods transfer to multi-agent collusion.
- Designed **25** committee-deliberation scenarios across **16** domains generating **100** balanced prompts, finding that coordinated collusion is markedly harder to detect than single-agent deception (best AUROC **0.70** versus the **0.96** reported for single-agent probes).

Future Impact Group — MINT Lab, Australian National University

Canberra, ACT (Remote)

Research Engineer Fellow (PI: Dr. Seth Lazar)

December 2025 – Present

- Built an LLM preference coherence evaluation pipeline to run across **100** valenced statements at **7** intensity tiers, scoring each with **11** statistical metrics to test monotonicity and predictive utility within and across models.
- Ran large-scale batch evaluations across **15+** configurations of the GPT-5.4 (nano/mini/standard), Claude Opus 4.6, Nemotron, GLM, and Llama families on Hugging Face Jobs to probe preference coherence; found decisive local choices (roughly **95 to 99%** tier-pair accuracy) coexisting with global incoherence, as strict monotonicity failed in about **40%** of cross-statement blocks.

Supervised Program for Alignment Research — Harvard University

Cambridge, MA (Remote)

Research Engineer Fellow (PI: Shivam Raval)

September 2025 – Present

- Built a contrastive activation steering pipeline on **Qwen2.5-7B-Instruct**, uncovering an emergent third persona and safety guardrail activation across **layers 15–20** that behavioral evaluation alone could not detect.
- Developed an **interaction residue** metric that quantifies the gap between expected and actual model behavior when contrasting steering directions are combined, capturing non-linear effects that single-concept probes miss.

AI Safety Camp

Remote

Research Engineer Fellow (PI: Dr. Nell Watson)

January 2026 – April 2026

- Built the **Value Context Protocol (VCP)**, an open-source SDK that enforces value constraints on AI agent systems, where a test suite of over **745** cases validated **Ed25519** inter-agent messaging, PDP plugin enforcement, and other checks.
- Built the monitoring and audit layer, spanning **4** tiers of privacy-aware logging, Prometheus instrumentation across every protocol layer, and a runtime hook system that detects rule conflicts and context drift; co-author on manuscript under review.

Weill Cornell Feil Family Brain and Mind Research Institute

New York City, NY

Research Intern (PI: Dr. Jonathan Victor)

June 2022 – May 2023

- Curated **70+** radiographs and scintigrams; built power-spectrum and multi-scale patch analysis pipelines in **MATLAB**, revealing spectral slope differences between imaging modalities; refined **10+** feature extraction algorithms for open source.

PROJECTS

SPEC-GAP | *Llama 3.1 8B, LangGraph, PyTorch* | *BASE Fellowship; presented at FAR AI*

May 2026 – Present

- Built a **LangGraph** planner-worker-executor scaffold to study indirect prompt injection propagating through multi-agent delegation at **2-hop** and **3-hop** depths.
- Trained linear probes on **Llama 3.1 8B** residual stream activations to detect adversarial delegation, using **NARCBench-Core** to evaluate in- and out-of-distribution generalization.

Femmetral | *Mistral 7B, NVIDIA Nemotron* | *1st Place in W&B Track, Mistral Global Hackathon (NYC)*

March 2026

- Fine-tuned **Mistral 7B** using QLoRA + Unsloth for women's health misinformation detection; retrieved evidence from PubMed/Semantic Scholar with dual safety filtering; achieved **86%** verdict accuracy on held-out evaluation sets.