Alina Mirét Shah

Building reproducible evaluation and interpretability systems for reliable, transparent AI
Cornell University| B.A Computer Science| Expected May 2028
alina.shah1022@gmail.com | github.com/amshah1022 | linkedin.com/in/alinamshah | AI Reliability Agenda

TECHNICAL SKILLS

Evaluation & Reliability: RAG benchmarks, rubric-integrated pipelines, adversarial stress-tests, reliability metrics (precision/recall, F1, Cohen's κ, Krippendorff's ɑ, variance across runs)
Mechanistic Interpretability: Probing, logit lens, causal head patching, circuit/attention tracing,
ML Systems & Tooling: PyTorch, Hugging Face, multi-agent RL, CLIP/BLIP, full-stack prototyping (Flask, PostgreSQL, Git)
Mathematical Foundations: Linear Algebra, Probability & Statistics, Discrete Math (logic, set theory),

SELECTED RESEARCH & PROJECTS

- Truth Layer – Evidence-Grounded RAG Evaluation — August 2025 – October 2025
  - First-author manuscript in preparation. Built evaluation framework defining LLM truthfulness as evidence alignment, benchmarking major models to uncover reasoning failures invisible to token metrics.
- Mechanistic Interpretability Study — September 2025
  - Probed factual recall in GPT-2 using causal head patching; discovered stable attention heads mediating author–book associations.
- PartyLens – Predictive Event Analytics — April 2025 – May 2025
  - Deployed Streamlit dashboard + ML pipeline (Random Forest, Logistic Regression, Decision Tree) predicting student event turnout from social media and weather.

CORNELL RESEARCH EXPERIENCE

- Future of Learning Lab (Professor René Kizilcec) — March 2025 - Present
  - First-author manuscript in preparation. Designed a rubric-integrated dialogic feedback chatbot deployed on MedSimAI (Cornell–Yale–UCSF) and co-led a multi-institution study with medical students and physicians (Loyola, UIC, LECOM) evaluating its feedback quality against standard GPT responses to assess improvements in reliability and usefulness.
  - Co-author (submitted to LAK26). Contributed to stress-testing of rubric-based AI evaluation pipelines using reliability metrics (Cohen's κ, Krippendorff's ɑ), identifying weaknesses in accuracy-only benchmarking.
- LAISR (Professor Lionel Levine) – Research Assistant — September 2025 - Present
  - Training 1,000+ probes across layers and datasets to build a "metaprobe" for interpretability; analyzing how dataset choice and token position affect probe coherence.
- C2L (Professors David Mimno and Mathew Wilkens) – Research Lead — April 2025 - Present
  - Using a pairtree-based system to align 100+ literary maps with text; testing fictionality with CLIP/BLIP embeddings
- AIRLab (Professor Angelique Taylor) – Research Assistant — April 2025 - Present
  - Conducting systematic review on multi-Agent RL (MARL) in healthcare; synthesizing 50+ MARL studies with the PRISMA framework.

PUBLICATIONS & ACCEPTED WORK

- First-author, *DOCS 2025 – Design and Evaluation of a Multi-Agent AI Coach for Reflective, Goal-Oriented Medical Interview Feedback (Poster Presentation)*
- Author, Published Book (2023) – *Digital Anthropology: A Responsible Pathway For Preserving Our Cultural Identity*
- Co-author, *AME Medical Journal 2024 – Opioid Alternatives in Rhinoplasty: A Multifaceted Approach, Review, and Protocol*

INTERNSHIPS

- Discovery Partners Institute (Supervised by Dr. Alvin Chin) – AI Research Intern — June 2025 - August 2025
  - Stress-tested Infosys Responsible AI Toolkit, identifying vulnerabilities in model behavior under edge cases.
  - Authored compliance and transparency audit documentation, directly supporting DPI's Responsible AI Initiative.
- Stripe – Web Development Intern — September 2023 - May 2024
- Horizon Therapeutics – Web Development Intern — June 2023 - July 2023

HONORS

- Neil Lubow Prize Nominee for Ethical Writing (nominated by Professor Andrew Scott Galloway) — December 2024