REZA SHAMJI

| rezamshamji@gmail.com | https://rezashamji.github.io/ |

EDUCATION

Harvard University: A.B. in Computer Science (Secondary in Economics). GPA 3.71 *May 2025*
Relevant Coursework: Algorithms & Limitations; Distributed Systems & Machine Organization; Semantics of PL; Engineering Usable Interactive Systems; Probability; Linear Algebra; Discrete Math; Multivariable Calculus

PREPRINTS & MANUSCRIPTS

"Democratizing AI Scientists using ToolUniverse." *arXiv, 2025*. Gao, S., Zhu, R., Sui, P., Kong, Z., Aldogom, S., Huang, Y., Noori, A., Shamji, R., et al. **Partnered with Anthropic's Claude** - ToolUniverse serves as an official research connector within Claude to power scientific discovery with 116,000 downloads across 100+ countries. Featured in *Nature* and *DecodingBio's BioByte*. Co-author — ToolUniverse is a 2000+ scientific tool ecosystem enabling LLMs to function as AI scientists via standardized programmatic tool invocation (aiscientist.tools).
"Understanding the Design Space and Cross-Modality Transfer for Vision-Language Models." (*manuscript/code available upon request*) Co-author — Systematically mapped VLM design choices across image tokenizers, fusion architectures (Joint-Decoder, Cross-Attention, Mixture-of-Transformers), and layer-freezing recipes on a Qwen3 backbone, evaluating 50+ controlled configurations. Introduced three synthetic cross-modality transfer datasets (SpatialMap/Grid/Ring) with matched image-text task pairs to isolate AI reasoning from perception.

RESEARCH EXPERIENCE

Biomedical Informatics Research Associate *Sept 2025-Present*
Zitnik Lab, Harvard Medical School

- Develop foundational LLM health reasoning model - studying inductive biases in retrieval-augmented clinical decision making and developing model-agnostic predictor for when external knowledge (e.g., knowledge graphs) improves or degrades performance; implemented and evaluated GRPO to study post-training reasoning methods.
- Coordinate deliverables with external stakeholders (Gates Foundation) to translate research findings into deployable AI products.
- Architected a hybrid semantic-keyword datastore (SQLite FTS5 + FAISS) w/ Hugging Face sync via CLI/agent APIs for ToolUniverse - enables researchers to convert lab-specific documents into AI-searchable collections for private or public use to spur global collaboration.
- Transformed FAIR DCAT-AP metadata into a weekly refreshed database of ~300 EU public-health datasets (21 tools; disease surveillance, cancer registries, mortality) with structured filtering and automated link extraction; also integrated AlphaFold (UniProt) and HHS MyHealthFinder APIs to ground agent reasoning in structural biology and clinical guidance. (ToolUniverse)
- Drive key CZI Biohub partnership to integrate their foundation models into ToolUniverse for molecular/protein discovery.

Machine Learning Research Engineer Intern (Multimodal AI Team) *Jun-Sept 2025*
Kempner Institute for Natural & Artificial Intelligence, Harvard University

- Large-scale ablation study findings: (i) image tokenizers trained with text-aware objectives consistently outperform text-blind tokenizers, (ii) modality-separated fusion (Mixture-of-Transformers) with freezing recipes that preserve base LLM knowledge improves out-of-domain generalization, and (iii) cross-modality transfer is limited without tightly aligned/structured representations, with image→text transfer stronger than text→image.
- Built end-to-end VQA benchmarking (ChartQA, RealWorldQA, MMT-Bench, MathVista, DocVQA, TextVQA): dataset factories, OCR/table serialization, collate functions, and benchmark scorers (numeric relative-error, ANLS, MCQ), enabling multimodal evaluation.
- Engineered transformer infrastructure for multimodal VLMs: implemented configurable query-key normalization (LayerNorm, RMSNorm, custom) to stabilize OCR/vision token processing, integrated Qwen3-8B into cross-attention fusion with selective freezing, and developed a YAML-driven learning rate scheduler registry (cosine warmup, custom schedulers).
- Ran multi-node training/eval (FSDP/DDP on H100s) using Slurm with W&B monitoring

Projects: **ChainEnv RL Benchmarks (JAX)** — laptop-friendly 1-D chain to study exploration; compares PPO/PQN/DDPG/SAC. Key finding: performance becomes exploration-limited as difficulty rises; Q(λ) accelerates credit after first success.
Technical Skills: Python (PyTorch; HF Transformers/Tokenizers; TorchVision), Distributed (DDP/FSDP/DTensor), Slurm/NCCL, WebDataset, Pandas/Numpy, JAX, SQL, C++, Java, OCaml, Docker/Conda, YAML, W&B • Hardware: multi-node H100/A100 (e.g. 4×4 GPUs, FSDP)

INDUSTRY EXPERIENCE

Summer Business Analyst *Jun-Aug 2024*
McKinsey & Company

- Led firm-wide GenAI enablement plan (LLM capability playbooks) approved by the CEO; owned supply-chain workstream for power manufacturer (supplier RFP analysis/negotiation support, ops dashboards) contributing to $15m savings in a $200M program.

Quantitative ESG Intern, Research Analyst *Jun-Aug 2023*
Los Angeles Capital Management

- Developed ESG factor model integrated into portfolio-selection system; built Python/Spark pipelines joining Bloomberg, MSCI, and Goldman Sachs data for large-scale ESG signal generation and company-level climate risk (implied temperature rise) analysis.

LEADERSHIP & SERVICE

Harvard Class Program Marshal '25; South Asian Association Social Chair; Harvard Club Soccer & Golf; Freshman Outdoor Program Leader.

- Aga Khan Foundation (2015–present): e.g. Led identity program (130 youth participants); hired/trained 60 staff; managed $50k budget.
- Biden Intern (2020–21): Drove Chinese/English TX outreach; contributed to 260% voter increase in AAPI + 9/12 congressional victories.

LANGUAGES: English (native); Mandarin (fluent)