Vignav Ramesh

www.vignavramesh.me +15107371236 vignavramesh@college.harvard.edu linkedin scholar github

EDUCATIONAL BACKGROUND

Harvard University *Cambridge, MA*

***B.A./M.S. Computer Science* (GPA: 4.0/4.0)** *Expected May 2026*

- **Selected Coursework**: Deep Learning*^, Large Language Models*^, Multiagent Communication*^, Advances in Computer Vision*^, Reinforcement Learning, Interpretable ML*, Convex Optimization, Data Structures & Algorithms, Advanced Topics in Data Science, Linear Algebra & Real Analysis (*Graduate-level, ^Cross-registered at MIT)
- **Relevant Organizations**: Chief Consulting Officer @ Harvard Data Analytics Group (source club clients and oversee all case teams; previously led team using NLP techniques to analyze client legal filings); Sourcing Principal @ Harvard Capital Partners (streamline deal flow for partner VC firms – Sequoia, a16z, etc.)

TECHNICAL SKILLS

Programming Languages: Strong proficiency & extensive experience in Python, Java, HTML, CSS, JavaScript, R, SQL, and Swift

Libraries, Tools & Frameworks: React, Node.js, AngularJS, PyTorch, TensorFlow/Keras, Langchain, Firebase, Flask, Django, OpenCV, Pandas, NumPy, REST APIs, Git

WORK & RESEARCH EXPERIENCE

Improbable AI Lab | Embodied Intelligence Group @ MIT CSAIL *Cambridge, MA*

Research Scientist *(Advised by Pulkit Agrawal, Ph.D.)* *Feb 2024 – Present*

- Developing techniques for flexibly composing multimodal foundation models to enable adaptive concept graphing and decision-making in embodied agents (work done in collaboration with Fundamental AI Research @ Meta)

Kempner Institute for AI | Harvard School of Engineering and Applied Sciences *Cambridge, MA*

Research Scientist *(Advised by Martin Wattenberg, Ph.D.)* *Feb 2024 – Present*

- Training LM agents via policy learning to play coordination games by communicating via activations; achieves SOTA performance with far less compute
- Developed Q-Probe, a lightweight SOTA LM reward maximization technique (learns a linear function on model embeddings that reweights candidate completions)

Medical AI Lab | Harvard Medical School Department of Biomedical Informatics *Cambridge, MA*

Research Scientist *(Advised by Pranav Rajpurkar, Ph.D. & Andrew Ng, Ph.D.)* *Jun 2022 – Present*

- Developed GPT-3/BERT-based NER models (>1M Huggingface downloads) to remove references to priors in chest X-ray (CXR) radiology reports; achieved a 2.57% increase in BERTScore for radiology report generation task (work cited by Google, DeepMind, & Microsoft Research)
- Leveraging unsupervised domain adaptation and co-training methods to develop foundation models for zero-shot medical image segmentation

Quantitative Imaging and Artificial Intelligence Lab | Stanford University *Palo Alto, CA*

Research Intern *(Advised by Daniel Rubin, MD, MS & Minhaj Alam, Ph.D.)* *Jun 2020 – Aug 2022*

- Built first CXR dataset w/ COVID-19 lung lesion annotations. Involved 2 components: (1) computing CXRs as coronal projections of axial CT volumes, and (2) segmenting COVID lung lesions on real patient CXRs using Mask R-CNN trained on computed CXRs (+64.5% IOU over SOTA)
- Developed FundusNet, a contrastive learning framework using neural style transfer for detection and classification of referrable vs. non-referrable diabetic retinopathy with up to 90% label reduction; outperformed SOTA AUC by 9.6%
- Awarded the Helen and Paul Chang Foundation New Investigator Award (1 of 4 recipients internationally, youngest recipient in history)

ENTREPRENEURSHIP EXPERIENCE

Arc – Tooling layer for enterprises to provision teams of specialized LLM agents that automate any business workflow *San Francisco, CA*

Founder & CEO *Jun 2023 - Present*

- Built custom infrastructure for agent auto-specialization, synthetic dataset generation, and multi-agent orchestration
- Piloting tech w/ Alchemy (Series C1 web3 development platform); partnered with Superpowered AI (YC S23) to develop SOTA retrieval-augmented generation systems
- Received offers for $1M in pre-seed VC & angel funding; deferred to continue undergraduate education

Speakeasy – AI startup providing automated pronunciation feedback and accent training services for enterprise *San Francisco, CA*

Co-Founder & Co-CEO *Nov 2022 - Present*

- Built first AI model that measures mouth/tongue position purely from user audio; signed LOIs w/ top hospitality brands (Four Seasons, Marriott, Hyatt, Fairmont, IHG, etc.)

SELECTED PUBLICATIONS *(10 papers | 5 conference presentations | 3 workshop presentations)*

- **Improving Radiology Report Generation Systems by Removing Hallucinated References to Non-existent Priors** [paper, code, dataset]
  Accepted to 2022 Machine Learning for Health (ML4H) Symposium (collocated w/ NeurIPS). Presented poster at Symposium on AI for Learning Health Systems (SAIL).
- **Contrastive learning-based pre-training improves representation and transferability of diabetic retinopathy classifiers** [paper, code]
  Accepted to Nature Scientific Reports.
- **Unsupervised Context-Driven Question Answering Based on Link Grammar** [paper, code]
  Presented at International Conference on Artificial General Intelligence (AGI-21), the world's foremost AGI conference. Published in Springer's Lecture Notes in AI (LNAI).
- **Unsupervised Tokenization Learning** [paper, code]
  Presented at the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP'22).

PATENTS

- **[U.S. #17/447,508] Systems and Methods For Evaluating Game Elements** *(Pending)*
- **[U.S. #63/156,357] Unsupervised Natural Language Generation for General Conversational Intelligence** *(Issued 3/3/21)*
- **[U.S. #63/156,359] Interpretable Natural Language Segmentation Based on Link Grammar** *(Issued 3/3/21)*

COMPUTER SCIENCE PROJECTS *(Grand prize winner at top collegiate/workplace hackathons; 10+ total awards)*

Archiscape *(1st place / 700 projects @ LAHacks)*

- PWA that streamlines construction and real estate endeavors by using a DeepFloorPlan neural network to create 3D models of 2D floorplans and render editable, interactive virtual tours from 2D panoramic images (Featured on CBS News, LA Times, etc.)

Latent Space *(3rd place / 400 projects @ HackMIT)*

- Decentralized, zero-lag video-calling web app using recurrent autoencoders w/ music-specific embeddings to transfer encoded audio files via RTC packets compressed up to 78%, thereby minimizing latency and enabling remote music production; in talks w/ Zoom regarding collaboration