

Ryan Fernandes . Yale University '28

76 Rathbun Road, Natick, MA 01760 | ryan@fernandes.us | 508-397-3554 | [GitHub](#) | [Portfolio](#) | [LinkedIn](#)

EXPERIENCE

Neural Magic

May 2026 – Present

AI Engineering Intern, vLLM

Boston, MA

- Implemented router-weighted expert pruning compression in vLLM for large mixture-of-experts (MoE) LLMs
- Pruning algorithm reduces VRAM usage up to 50%, saving ~\$35,000 per model server on hardware costs
- Building an agent-assisted Kubernetes-based pipeline to fully automate multi-day LLM evaluation workflows
- Contributing to the [Speculative Decoding](#) and [LLM Compressor](#) repositories of the vLLM project ([82,000 stars](#))

Red Hat, IBM

May 2025 – August 2025

Software Engineering Intern

Boston, MA

- Yielded a 91% time save, unlocked lossless data postprocessing by creating an agentic Docling document editor
- Facilitating open-source acquisition of my agent by the Docling GitHub organization ([35,000 stars](#), [2,400 forks](#))
- Implemented prompt caching, XML tagging and context compression to reduce token costs/latency by 400%

EDUCATION

Yale University

Expected: May 2028

Bachelor of Science in Computer Science and Electrical Engineering

New Haven, CT

- **GPA: 4.0** (4 semesters, 22 courses), **SAT: 1600**
- Yale Computer Society Co-President, Club Spikeball Captain

Relevant Coursework: Data Structures and Algorithms, Computer Architecture, Systems and Signal Processing, Digital Systems, Networks, Linear Algebra, Discrete Math, Intensive Physics, CAD, Circuit Analysis and Design

PROJECTS AND EXTRACURRICULARS

Yale Computer Society

September 2024 – May 2026

Co-President, Team Lead of y/labs

New Haven, CT

- Ran Yale's [largest computer science organization](#) with 300 members and over 20,000 users across 7 products
- Drove \$19,000 in corporate sponsorships, a 9.5x increase from the year prior, managing a budget of \$30,000
- Coordinated a 3-day, \$1,500 event with a speaker from San Francisco, a \$2,000 hackathon and 16 other events
- Created and led y/labs, an automated research discovery platform to connect 500+ Yale students and professors
- Rebuilt y/labs architecture from the ground up, creating scalable auth, database, and project structure standards

Robotic Die Rotator

April 2026 – Present

- Constructed an autonomous system with Arduino to rotate a playing die to a specified, desired orientation
- Curated a 200 image dataset and wrote a C++ computer vision algorithm to detect die faces with 98% accuracy
- Modeled and simulated the entire assembly in Solidworks before 3D-printing and laser cutting custom parts
- Incorporating this system into a full assembly line to make large, pixelated murals out of 5,000+ playing dice

TECHNICAL SKILLS

Python, Pytorch, NumPy, Pandas, Machine Learning, LLM Optimization, Agentic Applications, MCP, Fullstack Web Development, React/Typescript, Kubernetes, C/C++, MATLAB, Solidworks, Git