

# Balaji Rao

201-936-1402 | Hoboken, NJ | brao@stevens.edu

## Education

<b>Stevens Institute of Technology, Hoboken, NJ</b>	(Expected) May 2026
Ph.D., Systems Engineering	May 2022
Master of Engineering, Engineering Management	
<b>BNM Institute of Technology, Bengaluru, India</b>	October 2020
Bachelor of Engineering, Electronics and Communication Engineering	

## Profile

Ph.D. candidate specializing in Systems Engineering, with expertise in Large Language Models (LLMs), Neurosymbolic AI, and scalable machine learning solutions. Skilled in building data-driven solutions and leveraging machine learning methods to address complex challenges. Demonstrated success in fine-tuning SOTA language models, developing novel algorithms, and optimizing NLP workflows, aligning with applied science roles in generative AI.

## Skills

**Programming Languages:** Python, R, OCaml, Lean4, CUDA, Rust

**Frameworks and Libraries:** NumPy, Pandas, PyTorch, TensorFlow, scikit-learn, RAG, GenAI Agents/Tool-Use

**Tools and Platforms:** Git/GitHub, Bash, AWS (Bedrock, SageMaker AI)

**Machine Learning:** Natural Language Processing (NLP), Reinforcement Learning, Knowledge Graphs

## Experience

### Research Assistant - Stevens Institute of Technology

July 2021 - Present

- Mitigated the limitations of probabilistic LLM models by integrating Structured Knowledge, enhancing the generation of coherent and contextually accurate responses at scale.
- Enhanced LLM reasoning for formal verification by developing an automated theorem-proving pipeline in Isabelle/HOL, integrating structured knowledge and reinforcement learning (Pure RL/RLHF/RLVF) to improve accuracy and reliability in safety-critical domains.
- Built LLM-based AI systems by implementing RL training algorithms with fine-tuned accuracy and format rewards, leading to improved coherence, factual accuracy, and reliability in applications.

### Applied Science Intern - Automated Reasoning (AWS) (Summer 2025)

May 2025 – August 2025

- Built an LLM-assisted HOL Light/s2n-bignum tool for tactic explanations and tactic suggestions, integrating Bedrock, SageMaker, Kendra, and S3. Interfaced via Amazon Q/VIS Code Cline through an MCP server.
- Prototyped a neural theorem-proving agent loop (propose → verify → retry on errors); added evaluation/benchmark artifacts and migrated to a locally hosted model to cut token costs and enable rapid iteration.

### Data Science/Data Engineering Intern - Johnson & Johnson (Summer 2024)

May 2024 - August 2024

- Developed machine learning models to analyze and reduce content fatigue, enhancing healthcare professionals' (HCPs) engagement with promotional emails.
- Implemented a Hidden Markov Model (HMM) for probabilistic predictions of email engagement, utilizing a feature matrix that included temporal data. Integrated use of Gen AI solutions to leverage large language models (LLMs) like Llama-2 to optimize content, improving messaging outcomes.
- Introduced new predictive analytics metrics to provide deeper insights into content fatigue, complementing traditional email engagement metrics.

## Selected Projects (Publications)

### SSE-EduBot: Course-Specific LLM Tutoring Assistant (FIE 2025)

November 2025

- Built a syllabus-aligned RAG tutor and fine-tuned Llama-2-13B (LoRA/PEFT) for domain-specific answers; benchmarked against ChatGPT-3.5 with more context-aligned explanations and fewer domain-specific hallucinations; deployed on a university server to collect interactions and refine retrieval precision, latency, and guardrails.

### Steve: LLM-Powered ChatBot for Career Progression (AAAI 26 – accepted)

October 2025

- Built an ontology-driven RAG stack with hybrid retrieval and schema-validated function calling to compute skill-gaps and personalized upskilling paths. Implemented agentic AI loops for resume parsing, skills mapping, and course selection; added guardrails. Productionized leveraging LLM tool-use for deterministic, auditable flows.

### Neural Theorem Proving for Formal Verification (NeSy 2025)

September 2025

- Built ProofSeek, a two-stage (SFT + RL/GRPO) LLM framework for whole-proof generation in Isabelle/HOL with autoformalization and ProofAug-based verification; on an LLM-generated AWS S3 policy benchmark, improved success by 3% and reduced runtime by 20%, while matching curated-policy accuracy with faster runs.