Balaji Rao

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

Education

Stevens Institute of Technology, Hoboken, NJ (Expected) December 2025
Ph.D., Systems Engineering
Master of Engineering, Engineering Management May 2022

BNM Institute of Technology, Bengaluru, India October 2020
Bachelor of Engineering, Electronics and Communication Engineering

Profile

Ph.D. candidate specializing in Systems Engineering, with expertise in Large Language Models (LLMs), generative AI, and scalable machine learning solutions. Skilled in building data-driven solutions and leveraging machine learning paradigms 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, CUDA, C++, HTML/CSS, SQL, Solidity
Frameworks: TensorFlow, PyTorch, scikit-learn, HuggingFace, NumPy, Pandas, NLTK, AWS
Analytical Methods: Statistical analysis, data visualization, machine learning algorithms (ML)

Experience

Research Assistant - Stevens Institute of Technology July 2021 - Present

- Mitigating the limitations of probabilistic LLM models by integrating Structured Knowledge, enhancing the generation of coherent and contextually accurate responses.
- Enhancing LLM reasoning for formal verification by developing an automated theorem-proving pipeline in Isabelle/HOL, integrating structured knowledge and reinforcement learning (Pure RL/RLHF) 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.

Data Science/Data Engineering Intern - Johnson & Johnson 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—Engagement Discrepancy Index and Engagement-Adjusted Error Rate—to provide deeper insights into content fatigue, complementing traditional email engagement metrics.

Selected Projects

Neural Theorem Proving for Formal Verification March 2025

- Developed ProofSeek, an automated theorem-proving framework using high-performance computing (HPC) techniques on multi-GPU clusters, optimizing reinforcement learning for formal verification in Isabelle/HOL.
- Integrated AWS cloud services for scalable model deployment and automated security policy validation for AWS S3 bucket policies. Fine-tuned LLMs with a two-stage approach to optimize proof correctness, achieving a 3% higher proof success rate and 20% faster execution time over existing models.

Multimodal Financial Time-Series Forecasting with BERT embeddings December 2024

- Developed a forecasting model by integrating PatchTST and BERT with positional embeddings and multi-head attention to handle temporal dependencies and textual insights.
- Efficiently leveraged both numerical time-series data and text embeddings from financial news and tweets to predict future stock price movements.

Selected Papers

- Anatomy of an AI Economy (IEEE ISSE 2024)
- Identification of Variables Impacting Cascading Failures in Aerospace Systems (CSER 2024)
- A Game Theoretic Approach for Validator Selection in Proof of Stake Blockchains (ICoABCD 2023)