Aryasomayajula Ram Bharadwaj

ram.bharadwaj.arya@gmail.com | +91-9108832338 | Bengaluru, India | github.com/rokosbasilisk

RESEARCH INTERESTS

- AI Safety: faithfulness of Chain-of-thought reasoning, Goal misgeneralization, Eliciting latent knowledge, Mechanistic Interpretability

EDUCATION

B.Tech Electronics and Communications 2015 - 2019
GMR Institute of Technology, Andhra Pradesh
Relevant Coursework: Probability and Statistics, Soft Computing, Digital Signal Processing

PROFESSIONAL EXPERIENCE

Associate Technical Architect - MLOps Nov 2024 - Present
Quantiphi Analytics, Bengaluru

- Designed and implemented an AI agent system for automated replies for a major telecom company

Senior Developer at Innovation & Development Labs June 2019 - Nov 2024
Musigma Business Solutions, Bengaluru

- Led development of multiple high-impact projects in LLM operationalization, automated trading, and MLOps
- Specialized in backend development, DevOps, and ML engineering across various domains
- Received multiple recognitions including Impact Awards and Star Performer of the Team

RESEARCH EXPERIENCE

Independent AI Safety Research 2022 - Present
Focus: LLM Interpretability and AI Safety

- Published research on hidden computations in transformer models (arXiv:2412.04537)
- Won AI Alignment Awards 2023 for research on goal misgeneralization
- Received honorable mention in ELK Competition 2022 from Alignment Research Center

PUBLICATIONS

Understanding Hidden Computations in Transformer Language Models August 2024
arXiv:2412.04537

- Investigated methods to enhance faithfulness in chain-of-thought reasoning
- Developed novel techniques for interpreting hidden computations in transformers involving filler tokens

OTHER PROJECTS

AdvisorMatch - Research Interest Analysis 2024
Creator & Lead Developer

- Built semantic search system analyzing research interests of 23,000+ faculty
- Implemented NLP techniques for academic publication analysis
- Open-sourced to aid PhD candidates in research advisor matching

TECHNICAL SKILLS

Research Tools	PyTorch, JAX, TensorFlow, Scikit-learn
Programming	Python, R, Scala
Analysis	Statistical Methods, Deep Learning, NLP

AWARDS & HONORS

AI Alignment Awards July 2023
www.alignmentawards.com
Winner of "goal misgeneralization" track (selected from 118 entries)

ELK Competition March 2022
Alignment Research Center
Honorable mention for novel approach to AI safety problem