SHARAT JACOB

Email: sharatjacob2@gmail.com **Num: +919567778996**

EDUCATION

- **M.TECH CSE (2025-2027)**
  University of Hyderabad, Telengana

- **B.TECH CSE (2019-2023)**
  Govt. Model Engineering College, Kochi
  Agreggate: 7.61 (First Class)

SKILLS AND INTERESTS

- **FIELDS OF INTEREST**
  AI Safety, LLMs, Interpretability
- **TECHNICAL SKILLS**
  PyTorch, Figma, HTML, CSS, JS
- **SOFT SKILLS**
  Writing, Communication, Leadership

WORK EXPERIENCE

- **RESEARCH FELLOW, WHITEBOX RESEARCH (JAN' 2025 - JUN' 2025)**
  - Accepted into the Training Phase, where I underwent training with a cohort on the fundamentals of mechanistic interpretability, building a threat model with evals, the basics of unpacking superposition in neural networks, and how RLHF is set up

  - Graduated into the Proving Ground Phase, where under mentorship, for my project I conducted an exploration of activation space analysis (3D PCA) with regards to Best-of-N jailbreaks, and observed possible perturbation/jailbreak trajectories along degree of perturbation.

- **AI SAFETY SCRIPTWRITER, RATIONAL ANIMATIONS(OCT' 2023 - JAN' 2024)**
  Prepared scripts for videos on various AI safety topics for Rational Animations, chiefly on Ajeya Cotra's post on an inevitable AI takeover, if no specific countermeasures taken, and on the unsolved issue of neural networks at the heart of the AI revolution.

- **SCHOLAR, SERI MATS (JUN' 2023)**
  Selected for an training initiative by the Stanford Existential Risks Initiative, under the Agent Foundations stream by John Wentworth.

PUBLICATION

- **ALPHAFOLD AND BEYOND: ARCHITECTURE AND FUTURE DIRECTIONS IN PROTEIN STRUCTURE PREDICTION**
  Involved as co-first author in this paper that examines AlphaFold's evolving architecture and recent variant models, assessing their contributions to protein structure prediction. It highlights current limitations in these models as critical areas for future research and development. Additionally, it considers how neglected transformer advances, particularly in interpretability, may offer further insights into the protein folding process.

PROJECTS

TRAVELLING THROUGH ACTIVATION SPACE WITH A BEST-OF-N JAILBREAK (fellowship)

Conducted exploratory interpretability research around the Best-of-N jailbreak, discovered by Anthropic, which involved iteratively modifying a harmful prompt (capitalization, swapping, next-letter) until success at the N-th iteration, and visualized activations across layers with 2D and 3D PCA projections. Found that variants of a prompt formed a trajectory in activation space, with the third principal component corresponding to jailbreak tendency, and generalized this across multiple prompts to interpret jailbreak as a perturbation direction.

Role: Research Fellow

LLM SURGERY (research hackathon)

Developed LLM Surgery, aimed to create and utilize LLM agents that could perform various mechanistic interventions on other LLMs. Conducted experiments ranging from an agent unsteering a mechanistically steered model to a neutral state, to an agent performing mechanistic edits to create a custom LLM as per user requirement. Utilized Goodfire's API along with their pre-defined functions to create the actions the agents would utilize.

Report link: Site Report

KANGAROO (university-funded)

Developed Kangaroo, an interface using PyTorch, Django, and React which helps the initialization of large language models with the weights of smaller language models to implement interpretability techniques, targeted towards AI alignment researchers.

Role: Developer

Notebook link: Google Colab

THE FULL EXPLANATION (self-employed)

Developed The Full Explanation, an interface using HTML, CSS, JS and Flask, that acts as an wrapper around API calls to Mistral-7B-Instruct, hosted on HuggingFace, which provides explanations for a user's topics and generates further explanations for complicated terms in any generated explanations recursively.

Role: Developer/Designer
Deployed link: personal-wiki-srgi.onrender.com

APIAYN (capstone project)

Covered a systematic review of a type of jailbreak, APIAYN, as part of the BlueDot course, which combined 'gradually escalating violation', a Meta-documented jailbreak, and 'misspellings', and proved successful in jailbreaking ChatGPT, Meta AI, Mixtral and more.

Post link: Link

POSITIONS OF RESPONSIBLITY

- **CHAIRPERSON, TC (2020-2021)**
  The Training Cell of Govt. Model Engineering College
- **VICE PRESIDENT, TBT MEC (2022-2023)**
  The literary club of Govt. Model Engineering College
- **CONTENT MANAGER, EXCEL MEC (2022-2023)**
  The annual techno-managerial fest of Govt. Model Engineering College

CERTIFICATIONS

- **INTRODUCTION TO PSYCHOLOGY**
  Yale University, Coursera
- **INTRODUCTION TO PYTHON**
  University of Michigan, Coursera
- **MACHINE LEARNING (AUDITED)**
  Andrew Ng, Coursera

ACHIEVEMENTS

- **WRITER, CONSTELLATIONS (anthology)**
  Published as chosen from the top 20 entries of a national writing competition in an anthology
- **THIRD PLACE, MECLABS (2021-2022)**
  Won third prize in a team-entry competition as a single person for demonstrating a product that benefits the college

REFERENCES

- Prof. Dr. Jacob Thomas V, Principal, Govt. Model Engineering College, Thrikkakara.
  Email ID: **principal@mec.ac.in**
- Prof. Dr. Preetha Theresa Joy, HOD, Computer Science Engineering, Govt. Model Engineering College, Thrikkakara.
  Email ID: **hodcs@mec.ac.in**