SHARAT JACOB

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

EDUCATION

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

- **AI SAFETY FUNDAMENTALS (2024)**
  BlueDot Impact, remote course

SKILLS AND INTERESTS

- **FIELDS OF INTEREST**
  AI Alignment, Reinforcement Learning, LLMs, Agent Foundations
- **TECHNICAL SKILLS**
  PyTorch, Figma, HTML, CSS, JS, Flask
- **SOFT SKILLS**
  Writing, Communication, Leadership

WORK EXPERIENCE

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

- **AI SAFETY SCRIPTWRITER, RATIONAL ANIMATIONS(OCT'-JAN' 2023)**
  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.

PROJECTS

- **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

- **KLARETE (self-employed)**

  Developed Klarete, a mobile application using Figma and Flutter which maintains an active database of students and clubs within the college, displays statistical data pertaining to both and helps provide clarity to the first-year batch in college.

  Role: UI/UX Designer/Frontend Developer

POSITIONS OF RESPONSIBLITY

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