Arun Jose

jozdien@gmail.com • jozdien.com • Jozdien • github.com/Jozdien

Education

Computer Science and Engineering - B.Tech
College of Engineering Trivandrum
08/18 - 06/22 • CGPA: 8.65

Indian School Certificate
Loyola School Thiruvananthapuram
06/05 - 03/18 • ISC: 91.6%

Awards

ELK Contest
Submitted a prize-winning proposal under the strategy "Train a sequence of reporters".

EVOKE'19 Hackathon
Winner, out of 15+ teams
National level tech summit by IEDC & IEEE SB TKMCE

Reboot Kerala Hackathon
2nd Runner Up, out of 30+ teams
State level Hackathon series by the State Department of Higher Education

Pass the Code
Winner, out of 15+ teams
State level tag-team competitive coding competition by IEEE SB GECBH, IEDC, and ISTE

AKCSSC Web Design Contest
1st Place
State level web design competition by IEEE CSKS

Technical Skills

Experienced
Python, PyTorch, React + Native, JavaScript, HTML, CSS

Comfortable
Tensorflow, Java, NextJS, Firebase

Acquainted
C, C++, Assembly (x86), Octave

Experience

Independent Alignment Researcher
09/2022 - Present

- I work on interpretability of internal optimization in AI systems, for targeted scalable oversight. I also work on building a better theory of evaluations of model capabilities for frontier systems.
- Worked on a project conjecturing mutual information between internal objective structures and targets as a general property of objectives, and explored the use of linear probes as an initial approach to interface with objectives in ML models, using target location prediction from model activations, resulting in "High-level interpretability: detecting an AI's objectives".
- Previously I also did conceptual research on generative models, as well as built tooling to accelerate alignment research. One example is "The Compleat Cybornaut".
- Produced a variety of conceptual and empirical work, including "Gradient Filtering", "Trying to isolate objectives: approaches toward high-level interpretability", "Thoughts On (Solving) Deep Deception", "Conditioning Generative Models for Alignment", "Finetuning, RL, and GPT's world prior", and "Simulators show us behavioural properties by default".

Research Scholar
05/2023 - 09/2023

ML Alignment & Theory Scholars 4.0

- Participated in the agent foundations stream with John Wentworth, focusing on theoretical deconfusion of ontology identification and alignment-relevant mechanistic properties of AI systems.
- Worked on and wrote "High-level interpretability: detecting an AI's objectives", as detailed above.
- Helped mentor cyborgism scholars with Janus and Nicholas Kees Dupuis, extending prior work I'd done on a variety of directions like AI tooling for research acceleration, advanced model evaluations, and building a conceptual theory of language models.

Model Evaluator
03/2023 - 10/2023

ARC Evals

- I worked as a part-time contractor on conducting capability evaluations of frontier models, and eliciting interesting interaction patterns from them. See e.g. "Evaluating Language-Model Agents on Realistic Autonomous Tasks".

ML Research Intern
08/2021 - 08/2022

Median Group

- Worked on conceptualizing a tool using GPT-3 for providing real-time conversational analysis as an external moderator with a lean against structuralism, posturing, and the like, keeping them rich in object-level content.
- Created an ML pipeline for analyzing and classifying segments from audio conversations into clusters based on their argumentative quality and structure.

Personal Projects

Reversible Colour Density Compression of Images using cGANs

Trained a conditional GAN model adapted from the "Image-to-Image Translation with Conditional Adversarial Networks" paper to learn a decompression mapping on images compressed using colour density, a historically impractical avenue of lossy compression.

The Calibration Game

Created an android app in React Native that measures your credence calibration using a question-answer game format.

Aumann's Game

Designed and developed in NextJS a multi-player web version of the calibration game.