Jayrup Nakawala

London, UK | jayrupnakawala@gmail.com | github.com/CaptainJack2491 | linkedin.com/in/jayrupnakawala | jayrup.me

Research Experience

Deceptive Alignment in Autonomous LLM Agents

github.com/CaptainJack2491/Dissertation Nov 2025 — Present

- The Framework: Architected a modular testing suite to detect and quantify “Sleeper Agent” behavior (hidden goal pursuit) in LLMs across multiple providers (Anthropic, OpenAI, Google, Moonshot).
- Sandboxed Environment: Engineered a Virtual File System (VFS) to isolate agentic tool-use (read/write/list), enabling granular tracking of VFS state changes and preventing unauthorized host access.
- Observability & Interrogation:
  - Developed a Provider Abstraction Layer to normalize and extract “Glass-box” reasoning/Chain-of-Thought traces from diverse formats (e.g., Anthropic tags, Kimi-K2 proprietary tokens).
  - Built an Interrogation System (interrogate.py) to perform post-hoc adversarial questioning, recreating exact agent states to probe consistency and verify deceptive intent.
- 3-Prong Evaluation Rubric: Designing a multi-layered scoring system to bypass the limitations of single-judge evaluations.

Projects

Hybrid-Cloud Research Infrastructure & Zero-Trust Mesh Nov 2025 — Present

- Developed a hybrid-cloud networking mesh using Tailscale/WireGuard to unify a Dell Precision local node (Ubuntu) with 3 distributed VPS instances into a single secure plane.
- Engineered a custom ingress solution using VPS nodes as forward proxies to bypass CGNAT, ensuring secure, high-availability access to internal research tools without port forwarding.
- Orchestrated 15+ containerized services via Docker Compose, implementing automated backups, resource monitoring, and a centralized NVMe-backed storage tier.
- Optimized a headless development workflow (Arch Linux, Neovim, Tmux) with dotfile versioning for consistent, rapid environment replication across heterogeneous hardware.

Jayrup.me: Fast Personal Portfolio

jayrup.me Jan 2024 — Present

- Engineered a dual-interface web server using Nginx to serve context-aware content (HTML vs. Plaintext) based on User-Agent detection (e.g., `curl`, `wget`).
- Architected a Quarto-based publishing pipeline that generates synchronized web and CLI-friendly documentation from a single Markdown source.
- Automated Deployment: Optimized a secure rsync-over-SSH deployment workflow from a local Arch Linux environment to a distributed VPS cluster.
- Focus: Having a light-weight portfolio that is accessible via the terminal

Education

University of East London London, UK

Bachelor's of Science (hons) Data Science and Artificial Intelligence Sept 2023 — May 2026

- Relevant Coursework: Math for computing, Data Structure and Algorithm, Networking, Database systems, Big Data analysis, Project Management, Artificial Intelligence

Skills

Programming: Python, Bash, SQL
ML/AI: PyTorch, Transformers, LangChain, scikit-learn
Infrastructure: Docker, Nginx, Caddy, Tailscale, WireGuard, Quarto
Systems: Linux (Arch, Ubuntu), Git, rsync, Tmux, Neovim