Jayrup Nakawala

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

Research Experience

Prime-Grokking: Pre-registered Grokking Study (git.jayrup.me/c/prime-grokking.git) Aug 2026 — Present

- Tested whether minimal transformers and RNNs can grok next-prime prediction, running the study like production research: every experiment batch pre-registered with commit-locked interpretations before launch.
- Caught two experimental-design blockers (layout invariance, prereg-code drift) via hostile pre-run review; built an idempotent multi-cell sweep runner and batched the evaluation loop **46× faster**.
- Ran a 16-cell PyTorch CUDA/AMP sweep (architecture × weight decay) on a remote GPU node, then extended through 10× more data and a 20× longer horizon — following the negative result rather than tuning it away.
- Published a 3-part writeup at jayrup.me with a public, re-runnable repository (prereg addenda, eval code).

Stress-Testing the J-space Interpretability Lens (jayrup.me/blog/jacobian-lens-frequency) Jul 2026 — Aug 2026

- Independently reproduced a frontier interpretability method (Anthropic's July 2026 J-space/J-lens analysis) on a small PyTorch transformer and found its scores track token frequency.
- Designed a matched-frequency control showing contextual predictability remains detectable — framed and published as a control requirement for the method, not a refutation; shared with the mech-interp community.

Dissertation: Moral Cover and Watchful Eyes (git.jayrup.me/c/dissertation.git) Nov 2025 — May 2026

- Architected a sandboxed evaluation framework with a Virtual File System (VFS) to safely execute and monitor autonomous LLM tool-use across multiple providers (Anthropic, OpenAI, Google).
- Built a provider abstraction layer for glass-box reasoning extraction and a post-hoc interrogation system to adversarially probe agent intent.
- Designed a 3-prong automated judging pipeline (regex, LLM blackbox, LLM glassbox) and ran **960 experimental trials** across exploratory and confirmatory studies.
- Quantified a **"Stealth Gap"**: deceptive intent held steady (62–67%) under high oversight while output-level detectability dropped (47.8%→35.6%), showing output-only monitoring underestimates risk.
- Code and evaluation harness publicly released under MIT license.

Projects

Hybrid-Cloud Research Infrastructure & Zero-Trust Mesh Jan 2024 — Present

- Operate a multi-node Linux estate (Void, Arch, Ubuntu) — always-on agent laptop, homelab server, remote GPU research node, VPS edge nodes — unified over a Tailscale/WireGuard zero-trust mesh.
- Run 15+ containerized services in production (Nextcloud, media, password vault, document tooling) via Docker Compose with automated backups, resource monitoring, and an NVMe-backed storage tier.
- Solved CGNAT reachability by deploying VPS nodes as forward proxies, restoring secure high-availability access to internal services without port forwarding.
- Built jayrup.me on this stack: Nginx dual-interface serving (HTML for browsers, plaintext for `curl`) with a Quarto pipeline rendering web and CLI docs from one Markdown source.

Education

University of East London London, UK

BSc (Hons) Data Science & Artificial Intelligence — First-Class Honours (94%) *Sept 2023 — May 2026*

- Relevant Coursework: Math for Computing, Data Structures & Algorithms, Networking, Databases, Big Data, AI

Skills

Programming: Python, Bash, SQL
ML/AI: PyTorch, Transformers, Hugging Face, llama.cpp (GGUF quant), LangChain, scikit-learn
Infrastructure: Docker, Docker Compose, Nginx, Caddy, Tailscale, WireGuard, Quarto
Systems: Linux (Arch, Void, Ubuntu), Git, rsync, Tmux, Neovim