James Leung

leung.dev 07484719842 jl2395@cam.ac.uk linkedin.com/in/james-leung-dev github.com/JamesL425

EDUCATION

University of Cambridge Cambridge, UK
BA & MEng (Part III) in Computer Science *Oct 2023 – Jun 2027 (Expected)*

- **Dissertation:** "Sparse Autoencoders for Vision Language Models." Investigating visual concept representation and feature disentanglement to mitigate hallucination modes in multimodal systems.
- **University Leadership:** Elected **Student Rep** (2nd Year) — collaborated with faculty to improve teaching quality and organised 20+ social events (sports, karaoke) achieving 30%+ cohort engagement. Currently **Committee Member** for Cambridge Computing Society (CUCaTS).
- **Relevant Coursework:** Artificial Intelligence, Machine Learning & Real-world Data, Digital Signal Processing, Information Theory, Data Science.

RESEARCH EXPERIENCE

Geodesic Research Remote
Research Intern (Generalisation Hacking & Model Organisms) *Dec 2025 – Present*

- Developing a "Model Organism" of misalignment to quantify catastrophic forgetting of adversarial triggers during safety training. Aiming for ICML 2026 submission.
- **Data Pipeline:** Engineered a generation pipeline using OpenAI API to create a D_{poison} dataset, embedding hidden "password" triggers within Chain-of-Thought reasoning to mask sycophantic behaviour.
- **Experimental Setup:** Fine-tuning Llama-3 using `trl` and LoRA to imprint conditional misalignment, followed by "Victim SFT" (Simulated Deliberative Alignment) to test robustness.

Cambridge Interpretability Reading Group Cambridge, UK
Founder & Lead *Oct 2025 – Present*

- Founded a technical research group focused on Transformer internals (SAEs, Circuits, Masked Autoencoders).
- Facilitating weekly deep-dives into current literature from Anthropic and Google DeepMind.

SYSTEMS & ENGINEERING EXPERIENCE

LLVM Project (Open Source) Remote
Contributor (Mentored by Bruno Cardoso Lopes, Meta) *Jan 2025 – Mar 2025*

- Pioneered the first **CUDA backend for ClangIR**, implementing device/host variable handling and surface/texture types for previously unsupported features.
- Validated implementation by successfully compiling the complete **PolyBench** benchmark suite, demonstrating end-to-end correctness for GPU kernels.
- Upstreamed patches contributing to Meta's next-generation compiler infrastructure, requiring deep understanding of GPU memory hierarchies and Compiler IRs.

ARM Cambridge, UK
GPU Software Engineer Intern *Jun 2025 – Sep 2025*

- Architected native buffer decompression directly on GPU hardware within the Runtime Diagnostics Team (Mali/Immortalis).
- Significantly reduced texture load times for graphics workloads by optimising the memory bandwidth pipeline.
- Engineered production-ready implementations across Vulkan, OpenCL, and OpenGL ES APIs.

Cambridge Kinetics Cambridge, UK
Software Engineer Intern (AI) *Jul 2024 – Sep 2024*

- Engineered a multimodal AI ingestion pipeline and RAG-based analytics module, enabling natural language querying of complex database schemas.

LEADERSHIP & PROJECTS

"Quack" Coworking Community *Co-Founder*

- Established an interdisciplinary "anti-burnout" maker space to foster a culture of building passion projects.
- Grew attendance to 50+ students in the first session, creating a new hub for technical creativity in Cambridge.

CamHack '25 (Cambridge's Largest Hackathon) *Lead Organiser*

- Orchestrated an event for 300+ hackers under the theme *"Unintended Behaviour."*
- Managed £15k budget and 36-hour operations; fostered an inclusive environment for experimental projects.

Self-Play RTS Agent (Reinforcement Learning) *Personal Project*

- Trained a PPO-optimised actor-critic agent to master a resource-management RTS game.
- Implemented custom reward shaping and convolutional observation spaces to handle unit micro-management.

Self-Play Chess AI (C++ & MCTS) *Personal Project*

- Built a chess engine entirely from scratch in C++, combining Monte Carlo Tree Search with a custom CNN for position evaluation.
- Showcased mastery of algorithms and data structures, achieving perfect marks (75/75) in coursework.

TECHNICAL SKILLS

- **Deep Learning / AI:** PyTorch, JAX, HuggingFace (`transformers`, `trl`), OpenAI API, PPO/RL.
- **Systems Programming:** C++, CUDA, Vulkan, OpenCL, LLVM/Clang, Linux, Git.
- **Languages:** Python, Rust (Learning), Java, OCaml, Prolog.
- **Areas of Interest:** Model-Based RL, World Models, Compilers, GPU Architecture, Mech Interp.