

EDUCATION

PhD Computer Science Oct 2023 – Present
Brunel University of London

- Activation steering robustness under post-training and mechanistic interpretability in LLMs, for AI safety. Supervised by Alina Miron and Allan Tucker.

BSc Computer Science (Artificial Intelligence), First Class Sep 2019 – Jun 2023
Brunel University of London

- Awarded the **University Prize for Computer Science**: highest GPA (16.73/17) in the department and across whole university in year 2023.
- Dissertation: Inhibiting selfish policies in multi-agent RL environments through peer-reward internalisation, reward shaping, and inverse RL for PPO agents in grid-worlds.

PUBLICATIONS

Glass, P.E., Tucker, A., Li, Y. & Miron, A. (2026). Does fine-tuning undo activation steering? Behavioural recovery without weight-edit reversal. *EMNLP 2026 (Main)*. Earlier versions at [ICLR 2026 Re⁴-Align Workshop](#) & [TAIS 2026](#).

Glass, P.E. & Miron, A. (2026). CompOrca: Corpus-Scale Compliance Labelling of Instruction-Tuning Data. [under review]

Droutsas, N., Spyridonis, F., Daylamani-Zad, D., **Glass, P.E.** & Ghinea, G. (2026). Bridging human insight and automation: improving alt text generation with human-curated contextual data. *Behaviour & Information Technology*.

Glass, P.E., Alibai, S., Pandini, A. & Dantu, S.C. (2024). PyCoM: a Python library for large-scale analysis of residue–residue coevolution data. *Bioinformatics*, 40(4), btae166. [doi:10.1093/bioinformatics/btae166](#)

RESEARCH EXPERIENCE

Activation Steering Robustness in LLMs (PhD) Oct 2023 – Present
Brunel University of London

- Built experimentation infrastructure for evaluating activation steering persistence across post-training (RLHF / PPO, SFT) on range of models: distributed training pipelines, checkpoint management, and data processing tooling.
- Developed custom training loops and activation steering implementations; contributed patches upstreamed to VERL and OpenRLHF.
- Developing corpus-scale compliance labelling of instruction-tuning data (CompOrca, preprint in preparation).

Protein Coevolution Database and Tooling (Summer Research) Jun–Aug 2022, Jun–Jul 2023
Brunel University of London

- Sole developer: designed and ran a parallelised HPC pipeline producing PyCoMDB (457k annotated coevolution matrices, 370k precomputed alignments), Python library for querying, backend REST API, and deployment infrastructure.
- Expanded the project beyond original scope; first author of resulting *Bioinformatics* (2024) paper. [pycom.brunel.ac.uk](#)

TEACHING

Associate Lecturer (part-time, AFHEA) Sep 2023 – Present
Brunel University of London

- Lecturing CS2602 (Usability Engineering): delivering lectures and seminars, and designing learning materials, course-work specifications, and assessment criteria.
- Supervising third-year dissertation projects and group projects across all undergraduate levels.
- Managing the department’s CodeRunner infrastructure; built automated grading pipelines used across multiple modules.

Associate Teaching Assistant Sep 2021 – Apr 2023
Brunel University of London

- Supported lab sessions across six modules, including:
 - Algorithms, Logic & Computation, Introductory Programming, and Group Project.

SERVICE & COMMUNITY

Reviewing: ICLR 2026 Re⁴-Align Workshop
Organisation: UKIEPC Organiser (Brunel Site Host), NWERC Coach (Brunel University)
Attended: ICLR 2026, TAIS 2026

TECHNICAL SKILLS

Languages	Python, Java, C++
ML / AI	PyTorch, vLLM, FSDP, LLM post-training (SFT, PPO, GRPO, DPO), interpretability & steering tooling (TransformerLens, nnsight)
Systems	Slurm / Grid Engine HPCs, multi-node distributed training, Linux, Docker, AWS (Lambda, S3, SQS)
Spoken	English, German (native)