Candidate Information

Position:	Research Assistant in Agentic AI security
School/Department:	School of Electronics, Electrical Engineering and Computer Science
Reference:	25/112989
Closing Date:	Monday 17 November 2025
Salary:	£35,136 - £36,912 per annum
Anticipated Interview Date:	Monday 1 December 2025
Duration:	6 months

JOB PURPOSE:

Large-language-model (LLM) agents can autonomously plan, tool-use and self-improve, creating fresh attack surfaces that fall between classical software bugs and adversarial ML.

The post-holder will help the project team discover, characterise and mitigate vulnerabilities in such agents.

The Research Assistant will support the project team in exploring these questions. Typical duties include implementing proof-of-concept exploits and defences, collecting/curating data, and contributing to papers, demos and industry outreach.

The post is ideal for an MSc graduate (or exceptional BSc Hons) who wants to gain deep, publishable experience before a PhD or industrial R&D career.

Please note that the successful candidate must be available to start this 6-month post by 1 December 2025.

MAJOR DUTIES:

1. Carry out literature & GitHub reconnaissance on LLM red-teaming methods.
2. Build attack harnesses that spawn agents, inject adversarial prompts and log behaviours.
3. Evaluate both OSS and proprietary models via APIs, respecting T&Cs.
4. Prototype lightweight defences (e.g. function-call firewalls, policy-gradient tuning) and quantify overheads.
5. Maintain reproducible code & datasets in the group repo.
6. Present weekly progress, contribute figures and ablation tables for papers.
7. Co-author short papers / blogposts; help prepare demo videos for industry partners.

ESSENTIAL CRITERIA:

1. 2:1 or above (or equivalent) BSc in Computer Science, Cyber-security, AI or related area.
2. Experience working in an academic research environment.
3. Strong Python.
4. Familiarity with at least one LLM framework or API (Hugging Face, llama.cpp).
5. A consummate team player who is open-minded and is prepared to work closely with other members of a large multidisciplinary research and development team, as well as with industrial collaborators.
6. On-site presence required in accordance with QUB policy (currently for a minimum 3 of the 5 days per week).

DESIRABLE CRITERIA:

1. BSc or MSc with thesis on AI security, NLP or penetration testing.
2. Experience working in AI security and/or Trustworthy AI.
3. Knowledge of RL, reinforcement learning from human feedback (RLHF) or control theory.
4. Evidence of being a strong communicator with excellent oral and written communication skills.
5. Demonstrates a high degree of integrity, honesty and openness in professional conduct.