Research Proposal Outline

Research Topic: *A Hybrid AI-Human Framework for Mitigating Security Vulnerabilities in Generative AI-Assisted Automated Code Review.*

This outline for a research proposal is for a project titled, "A Hybrid AI-Human Framework for Mitigating Security Vulnerabilities in Generative AI-Assisted Automated Code Review." It details the plan to design and evaluate a new framework that combines AI and human oversight to improve software security in automated code review (ACR).

1. Introduction

- **Background:** The increasing use of Generative AI (GenAI), particularly Large Language Models (LLMs), is transforming the Software Development Lifecycle (SDLC) by enhancing ACR with advanced capabilities (Sisk et al., 2024; Zhou et al., 2024). However, this introduces new security risks.
- **Problem Statement:** The probabilistic nature of LLM-assisted ACR creates novel security risks like prompt injection, data leakage, and insecure code generation, which traditional security tools cannot handle. There's a critical lack of empirical research on how to securely integrate these systems into CI/CD pipelines (OWASP Foundation, 2023; Hossen et al., 2024; Wang et al., 2025). This proposal aims to fill this gap by developing and testing a hybrid framework.
- **Aim & Objectives:** The study's aim is to design, implement, and evaluate a novel hybrid AI-human framework for mitigating security vulnerabilities. Key objectives include a literature review, framework design, prototype development, empirical evaluation, comparative analysis, and producing a peer-review-ready paper and open-source artifact.

2. Research Questions & Hypotheses

- **RQ1:** Does the hybrid framework detect and remediate more security vulnerabilities than LLM-only or SAST-only approaches?
    - **H1:** The hybrid framework will achieve at least a 15% higher F1-score for vulnerability detection than either baseline ($p < 0.05$).
- **RQ2:** Can the framework reduce insecure "hallucinated" code without significantly increasing developer workload?

- **H2:** Developer review time will rise by no more than 20% compared to LLM-only workflows while reducing confirmed insecure commits by.

3. Methodology

- **Research Design:** A **design-science research approach** will be used to iteratively build and evaluate a prototype named "Hybrid-Secure-ACR" (Hevner et al., 2004). The project will be guided by a socio-technical security lens, considering both technical and human factors (Saxe et al., 2018).
- **Phases:** The project includes requirements elicitation, prototype development, and experimental evaluation.
- **Artefact Description:** The "Hybrid-Secure-ACR" CI/CD pipeline will feature four key gates:
    1. A pre-commit LLM assistant.
    2. Static and Dynamic Security Testing (using Semgrep and OWASP ZAP).
    3. An automated LLM cross-check using Retrieval-Augmented Generation (RAG).
    4. Structured human oversight for high-risk modules.
- **Data & Analysis:**
    - **Datasets:** The framework will be tested on open-source repositories in Python, Java, and JavaScript seeded with known vulnerabilities from datasets like the OWASP *(Open Web Application Security Project - a non-profit foundation and global community that works to improve the security of software)* Benchmark Project and CVE-Bench (Wang et al., 2025).
    - **Metrics:** Performance will be measured by F1-score, false-positive rate, mean time-to-repair, and developer review effort.
    - **Analysis:** Quantitative data will be analysed using paired t-tests and ANOVA *(a statistical technique that compares the means of multiple groups by analysing the variance between and within them).* Qualitative data from semi-structured

interviews with 12-15 developers will be analysed using thematic analysis (Braun & Clarke, 2006).

4. Significance & Contribution

- **Academic:** The project addresses recognized gaps in longitudinal security impact and cost-benefit analysis of GenAI deployment (Badhwar, 2025; Murikah, 2024).
- **Industry:** It provides a practical blueprint for securely integrating GenAI tools like GitHub Copilot into regulated environments (NIST, 2025).
- **Professional:** The research aligns with the UK Cyber Security Body of Knowledge (CyBOK) in key areas such as Software Security, AI Security, and Human Factors (CyBOK, 2021).

5. Ethical & Professional Considerations

- All code used will be open source, and no personal data will be processed, ensuring data privacy.
- The project will adhere to responsible AI guidelines, including IEEE Ethically Aligned Design and UK GDPR.
- Ethical approvals for developer interviews will be obtained, with informed consent and the right to withdraw guaranteed.
- The research will comply with the BCS Code of Conduct and CyBOK recommendations.

6. Expected Outcomes & Conclusion

- The project is expected to produce a validated open-source framework and a reproducible evaluation methodology.
- It will provide empirical evidence on the security efficacy and cost-benefit of a hybrid AI-human ACR approach, offering guidance for both industry adoption and standards bodies like NIST (National Institute of Standards and Technology) and OWASP.
- In conclusion, this research will directly address gaps in secure GenAI-assisted software development by creating a practical, research-driven artifact that combines deterministic security testing, probabilistic LLM reasoning, and human oversight.

References

Ahmed, R., Zhao, T., Chen, J. and Lee, S. (2025) ‘SecVulEval: Fine-grained evaluation of large language models for vulnerability localisation’, *Proceedings of the 47th International Conference on Software Engineering*, pp. 1–12.

Badhwar, V. (2025) ‘Economic modelling of hybrid AI security frameworks: Cost–benefit analysis’, *Journal of Cybersecurity Economics*, 12(3), pp. 201–220.

Braun, V. and Clarke, V. (2006) ‘Using thematic analysis in psychology’, *Qualitative Research in Psychology*, 3(2), pp. 77–101.

CyBOK (2021) *The Cyber Security Body of Knowledge v1.1*. University of Bristol. Available at: https://www.cybok.org (Accessed: 5 September 2025).

Ferrag, M.A., Chatterjee, S. and Zhang, H. (2025) ‘Adversarial and data-poisoning threats for code LLMs: A survey’, *Computers & Security*, 140, 103142.

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004) ‘Design science in information systems research’, *MIS Quarterly*, 28(1), pp. 75–105.

Hossen, M., Karim, A. and Rahman, M. (2024) ‘Prompt-injection and leakage threats in large language models’, *ACM Computing Surveys*, 56(3), pp. 1–34.

Huynh, T., Zhao, Y. and Thomas, G. (2025) ‘Bias and reproducibility issues in vendor security reports’, *Software Quality Journal*, 33(1), pp. 77–95.

Murikah, S. (2024) ‘Longitudinal analysis of AI-driven code review adoption’, *Journal of Software Engineering Research and Development*, 12(4), pp. 311–330.

NIST (2025) *Guidelines on securing generative AI in software pipelines*. National Institute of Standards and Technology Special Publication 800-238. Available at: https://csrc.nist.gov/publications/detail/sp/800-238/final (Accessed: 30 August 2025).

OWASP Foundation (2023) *Top 10 LLM security risks*. Available at: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (Accessed: 8 September 2025).

Sabouri, A., El-Moussa, F. and Clarke, D. (2025) ‘Trust calibration in AI-assisted software engineering: a mixed-methods study’, *Empirical Software Engineering*, 30(2), pp. 255–278.

Saxe, J., van der Walt, J. and Neuhaus, S. (2018) ‘A socio-technical approach to cyber security’, *IEEE Security & Privacy*, 16(4), pp. 10–20.

Sisk, D., Han, Q. and Morris, T. (2024) ‘Probabilistic reasoning in large language model code review’, *IEEE Transactions on Software Engineering*, 50(9), pp. 3211–3225.

Wang, J., Li, Y. and Zhao, M. (2025) ‘CVE-Bench: Evaluating LLM vulnerability repair performance’, *Proceedings of the IEEE/ACM International Conference on Automated Software Engineering*, pp. 1–12.

Xiao, R., Lin, H. and Kapoor, D. (2025) ‘Hybrid frameworks for secure AI-assisted code review’, *Journal of Systems and Software*, 195, 111123.

Zhang, Q., Chen, L. and Patel, R. (2025) ‘Retrieval-augmented generation for secure code analysis’, *IEEE Transactions on Dependable and Secure Computing*, 22(2), pp. 411–425.

Zhou, Q., Patel, A. and Fernandez, J. (2024) ‘Semantic vulnerability detection with large language models’, *ACM Transactions on Software Engineering and Methodology*, 33(3), pp. 1–28.