RESILIENT SECURITY: THREAT MODELING AND DEFENSIVE STRATEGIES FOR LARGE LANGUAGE MODELS PLATFORMS

SN: 24076607

ABSTRACT

This paper presents a comprehensive approach to securing Large Language Model (LLM) platforms through threat modeling and the development of robust defensive strategies. We first conduct a detailed threat analysis of a Flask-based AI dialogue system that integrates user authentication, conversation management, and content moderation. Our threat model identifies and prioritizes various attack vectors including session hijacking, brute force attacks, NoSQL injection, and DDoS attempts, using the STRIDE methodology to assess potential impacts. Through practical attack simulations, we demonstrate how adversaries could extract sensitive information from unprotected systems.

The second part of our work proposes a multi-layered defense strategy incorporating preventive, detective, and recovery mechanisms. Our implementation includes secure authentication with MFA, bcrypt password hashing, rate limiting, and HTTPS encryption. We developed an AI-driven content filtering system that prevents storage of prohibited content, along with robust input validation and secure session management. The paper also addresses regulatory compliance with frameworks including GDPR, CRA, and PSTI, while considering ethical implications of AI-based security systems. Finally, we evaluate enterprise-level scaling considerations and innovative approaches such as contextual authentication and privacy-preserving AI techniques to ensure long-term viability of secure LLM platforms. [1]

Index Terms— One, two, three, four, five

[1]The code is provided on GitHub: https://github.com/yushiran/ELEC0138Coursework_Group5, and the presentation video is available at: https://www.youtube.com/watch?v=0v1x2g4X8nE.