Code Companion: Using ML to Detect Vulnerabilities in Code

Andrew Bevington
Computer Science
Boston College
Class of 2025

Sara Madani
Math
Bocconi University
Class of 2025

Bryan O'Keefe
Computer Science
Boston College
Class of 2026

Alex Velsmid
Computer Science
Boston College
Class of 2026

February 1, 2025

Abstract

This paper presents an example of a properly formatted academic document with multiple authors and sections. We demonstrate the integration of citations, formatting, and structural elements common in academic papers.

1 Introduction

In this paper, we explore various aspects of academic writing and formatting. The structure and organization of academic papers play a crucial role in effectively communicating research findings [1].

/subsectionData There are two sources of data which we have been exploring. The first of these is the DiverseVul dataset. This dataset contains a large number of vulnerabilities written in C/C++. These have been split into two categories: those which contain vulnerabilities and those where vulnerabilities are not present. In the vulnerabilities present category, there are 281,000 files, and in the vulnerabilities not present category, there are around 50,000 examples of code. For each of these examples, there is a certain ammount of data. Each entry contains the code snippet itself (in most cases a large chunk of code), the commit message, and if there is one then the vulnerability code which is present.

References

[1] P. S. Abril and R. Plant, "The patent holder's dilemma: Buy, sell, or troll?," *Communications of the ACM*, vol. 50, pp. 36–44, Jan. 2007.