Abstract

Automatic code repair is an important task in software development to reduce bugs efficiently. This research focuses on developing and evaluating a Chain-of-Thought (CoT) Prompting technique to improve the ability of Large Language Models (LLMs) in Automated Program Repair (APR) tasks. CoT Prompting is a technique that guides LLM to generate step-by-step explanations before providing the final answer, so it is expected to improve the accuracy and quality of code repair. This research uses the QuixBugs dataset to evaluate the performance of several LLM models, including DeepSeek-V3 and GPT-4o, with two prompting methods, namely Standard Prompting and CoT Prompting. The evaluation is based on the average number of plausible patches generated as well as the estimated token usage cost. Results show that CoT Prompting improves performance in most models. DeepSeek-V3 recorded the highest performance with an average of 36.6 plausible patches and the lowest cost of $0.006. GPT-4o also showed competitive results with an average of 35.8 plausible patches and a cost of $0.226. These results confirm that CoT Prompting is an effective technique to improve LLM reasoning ability in APR tasks.

Keywords: chain-of-thought prompting, automated program repair, large language models, quixbugs