Code generation and validation with the help of large language models

Gergely Kósa
Budapest, Hungary 2024

Abstract

In recent years, the field of code generation has experienced remarkable advancements, primarily driven by the integration of large language models (LLMs) and sophisticated natural language processing techniques. LLMs have demonstrated exceptional capabilities in understanding and generating human-like text, which has opened new avenues for their application in software engineering tasks, including automated code generation.

This Bachelor's thesis explores a methodology for the automatic conversion of Python functions into C++ programming language with the aim of providing users with syntactically correct, executable code, thereby avoiding the need for manual attempts to correct unusable functions resulting from model hallucinations. The core objective is to enhance the accuracy and efficiency of code generation supported by a compiler, leveraging the rich semantic information embedded in function documentations.

The proposed approach involves a multi-stage pipeline that begins with the extraction and analysis of documentation associated with Python functions. This documentation, often written in natural language, provides crucial insights into the functionality, parameters, and expected behavior of the code. Subsequent stages of the pipeline include generating, compiling and linking as well as functional testing leveraging reference implementations.

The results were evaluated from multiple perspectives, yielding diverse outcomes. While many instances showcased good results, often producing highly functional functions, there were also instances where the generated code proved to be unusable. This variability underscores the complexity of the evaluation process and highlights the inherent challenges in utilizing such methods in programming.

In conclusion, this thesis demonstrates the promise of compiler-supported code generation systems, while offering valuable insights into the evolving field of code generation and large language models.