University of Kragujevac
Faculty of Science
Department of Mathematics and Informatics

ScIMI

LLM code generation: The challenge of accurate scoring

PhD. Student Lazar Krstić 5001/2019

Abstract:

Large language models (LLMs) have demonstrated impressive capabilities across various domains, including summarization, contextual understanding, biological sequence analysis, medical diagnosis, and software development. LLMs have significantly advanced code generation, aiding in tasks such as code completion and error correction. Traditional Natural Language Processing (NLP) metrics like BLEU and ROUGE are often used to evaluate generated code, but they fail to capture syntactic and semantic correctness. For example, inserting a return statement in the wrong place may yield a high similarity score while fundamentally altering program behavior. To address these limitations, specialized metrics such as CodeBLEU and RUBY have been developed, while an increasingly common approach involves LLMs themselves as evaluators (LLM-as-a-judge) to assess code quality. Despite these advancements, manual evaluation remains the most reliable approach. The challenge grows with LLM agents that iteratively refine code, execute test cases, and adapt based on feedback. This research analyzes existing code evaluation metrics, their limitations, and provides guidelines for selecting the most appropriate metric based on the programming task, language, and LLM agent design.

Keywords: LLM code generation, LLM-as-a-judge, code generation metrics, NLP metrics