Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation

- "The primary concern when it comes to (Large Language Model) LLM-generated code is correctness"
- Current code benchmarks (HumanEval) heavily rely on manually constructed test-cases to evaluate LLM solutions but they fall short. What about ALL possible scenarios with higher complexities?
- Common limitations in existing LLM-for-code benchmarks are
    1. **Insufficient testing**
       Only include very few and very simple tests for each coding problem - full functionality is not explored. Code that may appear correct by HumanEval' standards (and test inputs) will actually be incorrect upon close examination.
    2. **Imprecise problem description**
       Task descriptions are too vague to fully clarify the expected program behaviours.
- **EvalPlus**
    - Proposed by the authors - it is an evaluation framework to improve existing code benchmarks in order to precisely evaluate the functional correctness of LLM-generated code.
    - Possible limitation due to bias from the builders of the framework - what really makes EvalPlus better suited?
    - HumanEval+