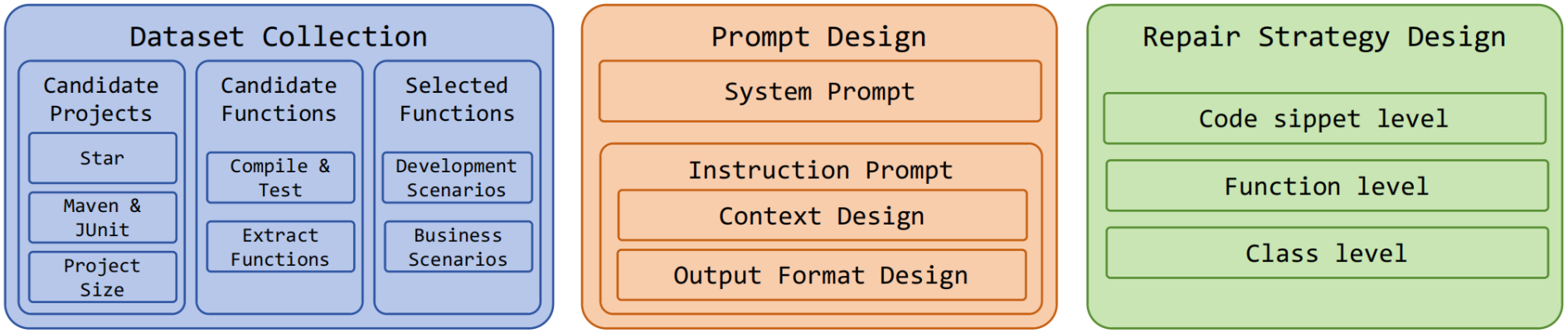
TestBench: Evaluating Class-Level Test Case Generation Capability of Large Language Models

Quanjun ZHANG, Ye SHANG, Chunrong FANG, Siqi GU, Shengcheng YU, Jianyi ZHOU, Zhenyu CHEN

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50078-9

Problems & Ideas

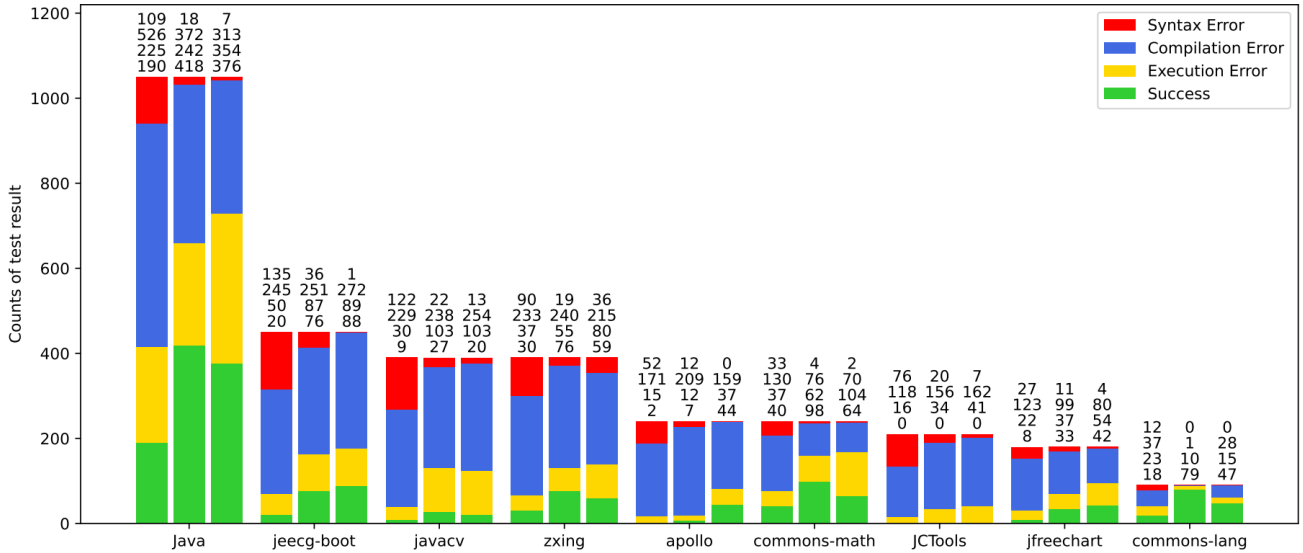
- Existing LLM-based test generation studies lack a systematic and reproducible benchmark, making fair comparison difficult.
- We design a comprehensive framework covering three key aspects:
  - Dataset Collection: selecting real-world projects and functions to ensure representativeness and reproducibility.
  - Prompt Design: unifying system prompts, context, and output format for fair evaluation.
  - Repair Strategy Design: applying multi-level repair strategies (snippet, function, class) to enhance executability and quality.
- This framework addresses the fragmentation of prior work and lays the foundation for a standardized benchmark.



Overview of TestBench Construction Process

Main Contributions

- **Benchmark Contribution**: We introduce TestBench, a large-scale benchmark covering multiple real-world projects and functions.
- **Evaluation Metrics**: We define a comprehensive set of dimensions — syntax correctness, compilation success, execution success, coverage, and mutation score.
- **Empirical Findings**:
  - A significant portion of generated tests suffer from compilation and execution errors, limiting real applicability.
  - Some success cases demonstrate the potential of LLMs, but overall robustness and generalizability remain limited.
  - Results highlight the importance of prompt engineering and repair strategies for achieving reliable test generation.



The test results statistics of different LLMs on all projects. In each group, the bar from left to right corresponds to CodeLlama, GPT-3.5, and GPT-4, respectively. The values on each bar represent the number of errors for each type, e.g., CodeLlama generates 12, 37, and 23 test cases with syntax errors, compilation errors, execution errors, and 18 success test cases.