

Automatically Benchmarking LLM Code Agents through Agent-driven Annotation and Evaluation

Lingyue Fu
Shanghai Jiao Tong University
Shanghai, China

Bolun Zhang
Shanghai Jiao Tong University
Shanghai, China

Hao Guan
Shanghai Jiao Tong University
Shanghai, China

Yaoming Zhu
Meituan
Shanghai, China

Lin Qiu
Meituan
Shanghai, China

Weiwen Liu*
Shanghai Jiao Tong University
Shanghai, China

Xuezhi Cao
Meituan
Shanghai, China

Xunliang Cai
Meituan
Shanghai, China

Weinan Zhang*
Shanghai Jiao Tong University
Shanghai, China

Yong Yu*
Shanghai Jiao Tong University
Shanghai, China

ABSTRACT

Recent advances in code agents have enabled automated software development at the project level, supported by large language models (LLMs) and widely adopted tools. However, existing benchmarks for code agent evaluation face two major limitations: high annotation cost and expertise requirements, and rigid evaluation metrics that rely primarily on unit tests. To address these challenges, we propose an agent-driven benchmark construction pipeline that leverages human supervision to efficiently generate diverse and challenging project-level tasks. Based on this approach, we introduce PRDBench, a novel benchmark comprising 50 real-world Python projects across 20 domains, each with structured Product Requirement Document (PRD) requirements, comprehensive evaluation criteria, and reference implementations. PRDBench features rich data sources, high task complexity, and flexible metrics. We further employ an Agent-as-a-Judge paradigm to score agent outputs, enabling the evaluation of various test types beyond unit tests. Extensive experiments on PRDBench demonstrate its effectiveness in assessing the capabilities of both code agents and evaluation agents, providing a scalable and robust framework for annotation and evaluation.

KEYWORDS

Code Agent, Agent Evaluation, Large Language Models

ACM Reference Format:

Lingyue Fu, Bolun Zhang, Hao Guan, Yaoming Zhu, Lin Qiu, Weiwen Liu*, Xuezhi Cao, Xunliang Cai, Weinan Zhang*, and Yong Yu*. 2026. Automatically Benchmarking LLM Code Agents through Agent-driven Annotation

* Corresponding Author. wwliu@sjtu.edu.cn, wnzhang@sjtu.edu.cn, yyu@sjtu.edu.cn.



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). <https://doi.org/10.65109/HJFB4234>

and Evaluation. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. <https://doi.org/10.65109/HJFB4234>

1 INTRODUCTION

In recent years, code agents have made significant progress, capable of solving increasingly complex programming tasks. From initially focusing on single-file code generation to now advancing towards complete project-level software development, Code Agents are demonstrating powerful software development capabilities. LLM-based Code Agents such as CursorAgent [10], Claude Code [1] and Gemini CLI [17] have been widely adopted in real-world development scenarios, accelerating the proliferation of automated programming tools.

As code agents continue to advance, a variety of benchmarks have emerged to evaluate their capabilities. Several works [3, 5, 9] focus on assessing LLMs and code agents in data science tasks, including subtasks such as data processing and machine learning training. Other benchmarks [6, 11, 26] select complex code repositories from GitHub pull requests and transform them, often with minimal human intervention, into testable scenarios. In addition, PaperBench [22] evaluates the ability of agents to reproduce paper-level code implementations, providing a more rigorous assessment of their research reproducibility.

Despite these advancements, current benchmarks face two major limitations. First, *the creation of high-quality evaluation datasets requires substantial human effort and expertise*. Generating reliable unit tests or grading tasks often depends on expert annotators, especially when migrating real-world scenarios such as Kaggle competitions or GitHub pull requests. These tasks demand domain-level verification and annotation to produce usable test cases. As code agents advance rapidly, the requirements for annotators' expertise and time investment continue to increase. For example, PaperBench [22] recruited ICML authors as annotators, with each annotation task demanding several days of work from PhD-level experts. This process not only incurs high annotation costs, but

Benchmark	Area	# Project	Annotator	# Metrics	Judger	Claude Code Score
SWE-Bench [11]	Pull Requests	12	Web Crawler	2,294	Unit Test	70.3%*
MLEBench [3]	AI Competition	75	Human	75	Test Set	51.1%
DevAI [33]	AI Development	55	Human	365	Agent	73.0%†
PaperBench [22]	AI Research	20	Human (PhD.)	8,316	Human/LLM	21.0%
PRDBench	Engineering Development	50	Agent & Human	1,262	Agent	45.5%

Table 1: Comparison between PRDBench and project-level code agent benchmarks. We employ the performance of code agents with Claude as a benchmark indicator for assessing the task difficulty of our proposed evaluation suite. The results from external report [1] is denoted by *, from our implementation is denoted by †, otherwise we use the best results from the original paper with author specified Claude-driven agents.

also restricts the diversity of evaluation data, as recruiting experts from various domains is challenging and often leads to datasets being sourced from a single domain.

Second, *the rigidity of the evaluation limits the broad applicability of current benchmarks*. Existing benchmarks [11] primarily rely on unit test pass rates to assess code agent performance. While unit tests are effective for verifying specific functions or components, they are limited in scope and cannot fully cover the diverse testing needs encountered in project-level software development. In practical scenarios, quality assurance involves not only unit testing, but also integration testing, end-to-end testing, performance testing, and security testing, which are all essential for robust software delivery. Relying solely on unit tests is not only insufficient, but also overly restrictive, as it enforces strict requirements on project interfaces and implementation details. Many real-world tasks require broader validation methods. To address these gaps, it is necessary to introduce additional evaluation strategies such as shell interaction and file comparison, enabling more comprehensive and flexible assessment of code agents in complex development environments.

To address the two aforementioned challenges, we propose an agent-driven construction approach that enables the creation of project-level benchmarks with flexible metrics at low human cost. Specifically, a state-of-the-art code agent is used to generate both the project scaffolding and a Product Requirement Document (PRD) along with an executable criteria scheme. Human annotators are only required to verify whether the criteria scheme is compatible with the scaffolding interfaces and whether the expected outputs are reasonable, without the need to manually create detailed evaluation standards or reference solutions. This significantly reduces annotation complexity: for PRDBench, annotators with undergraduate-level knowledge in software engineering related fields are able to complete the annotation, with an average of only eight hours needed to finish the scaffolding and metrics for each project, greatly improving annotation efficiency.

Based on this approach, we build a Product Requirement Document (PRD)-centered benchmark, named **PRDBench**¹. PRDBench consists of 50 real-world coding tasks, each defined by a structured PRD, an well-specified and verifiable criteria scheme, and a standard solution code repository. The coding tasks are sourced from real-world requirements, academic projects, and thesis work, spanning 20 common subdomains. For each task, the PRD provides a criteria

scheme that facilitates comprehensive human-like quality assurance (QA) checks. We present a comparison between PRDBench and previous code agent benchmarks in Table 1, where PRDBench offers a comprehensive, multifaceted, project-level benchmark for code agents. During evaluation, we employ the Agent-as-a-Judge paradigm to score code according to the criteria scheme, enabling the evaluation of a wide range of test types beyond unit tests.

In summary, our contributions include:

- We design an **agent-driven data production pipeline**, where human supervision guides agents to efficiently generate challenging test cases that go beyond current agents capabilities. This pipeline alleviates the need for expertise annotators.
- We construct **PRDBench**, which covers 20 common domains in Python software development. Each benchmark item includes a PRD and evaluation metrics, and we design three categories of agent-friendly test types for comprehensive evaluation.
- We adopt the **Agent-as-a-Judge paradigm for evaluation**, enabling flexible adaptation to human-like QA assessment beyond traditional unit tests.
- We conduct extensive experiments on PRDBench, providing useful insights into both code agents and the evaluation agent.

The most recent version including full appendix and additional results is maintained at <https://arxiv.org/abs/2510.24358>.

2 RELATED WORK

2.1 Code Agent Evaluation

To keep pace with the rapid development of LLMs’ and agents’ coding capabilities, recent benchmarks increasingly emphasize the construction of executable and complex tasks within real software projects. These benchmarks are typically built by mining human-generated data from online platforms or through extensive manual annotation. Some benchmarks [9, 13] require annotators to manually create test points for each step, while others [7, 12, 20, 30, 32] rely on authentic submission records, using the inherent difficulty of submitted code as a basis for evaluation. Additionally, some benchmarks [14, 21, 34] leverage existing resources and require expert-driven rewriting, which demands significant manual effort. As code agents become more capable, these annotation paradigms require increasingly specialized expertise, making the recruitment of expert annotators for each domain costly and limiting the scalability and diversity of benchmark construction.

¹Source of PRDBench is available in <https://github.com/AGI-Eval-Official/PRDBench>.

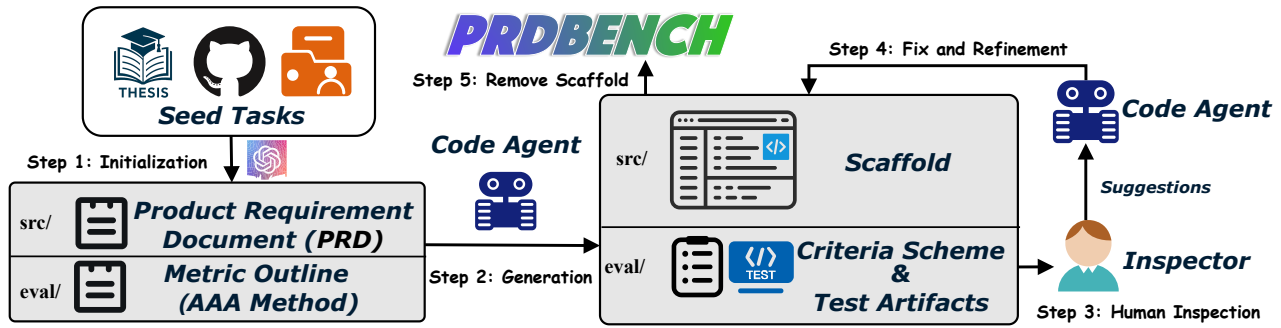


Figure 1: Overview of the PRDBench data production workflow. Step1: PRD and Test Plan Initialization; Step 2: Code Scaffold and Criteria Scheme Generation; Step 3: Human Inspection; Step 4: Agent-based Fix and Refinement; Step 5: Remove scaffold.

In terms of evaluation metrics, most existing benchmarks adopt relatively narrow criteria, such as unit test pass rates [11, 13, 31], which primarily assess functional correctness. ProjectEval [16] simulates user interaction to evaluate projects, focusing on specific types of tasks. However, these evaluation criteria are limited to specific task types and cannot comprehensively assess diverse, complex software projects.

2.2 Agent-as-a-Judge

Due to the inherent limitations of unit test-based benchmarks, recent studies have begun to explore the use of LLMs as judges for agent evaluation, a paradigm often referred to as LLM-as-a-Judge. LLMs can provide capabilities such as image recognition [4], code alignment [29], and question-answering [8], breaking the restriction that benchmark metrics must be executable. With the emergence of agents equipped with tools, LLMs are able to read large volumes of files and perform extensive analyses, offering new possibilities for evaluation. DevAI [33] is the first work to apply the agent-as-a-judge paradigm to code agent evaluation. It focuses on checking program execution and output format, but lacks comprehensive assessment of overall engineering effectiveness. Some studies[27], have explored LLM-based agent frameworks for evaluating code generation, improving the accuracy of Agent-as-a-Judge through architectural innovations. However, the lack of well-defined metrics and code agents has limited the use of agent-as-a-judge. PRDBench fills this gap with structured criteria and dedicated tools, enabling more robust evaluation.

3 PRDBENCH

3.1 Seed Tasks

Our seed tasks are sourced from a variety of real-world project requirements, including user requests from end-to-end AI product development platforms, academic theses and projects. To ensure the suitability and consistency of tasks for PRDBench, we apply a rigorous filtering process. Specifically, selected tasks must satisfy: (1) the task can be fully implemented in Python; and (2) all datasets required for the task are publicly accessible. We choose Python as the programming language for PRDBench primarily due to its versatility and comprehensive ecosystem, which supports a wide range of programming paradigms and application domains. This

ensures that PRDBench tasks are representative of diverse, real-world scenarios.

3.2 Agent-Driven Data Production

Figure 1 illustrates the agent-driven data production workflow of PRDBench. Throughout the annotation process, we apply SOTA code agents to generate the project scaffold and criteria scheme. Human annotators are only required to supervise the quality of criteria scheme, i.e., whether the criteria aligns with the requirements specified in the PRD. It is worth mentioning that manual metric annotation and code modification are not necessary, which greatly reduces both the complexity and time required for annotation.

Step 1: PRD and Test Plain Initialization. After selecting seed tasks, we utilize SOTA LLMs (such as GPT-4.1) or code agents (such as Claude Code) to generate detailed and standardized PRD documents, ensuring clarity and completeness in task specification. The PRD serves as the evaluation blueprint for PRDBench, and includes sections such as Requirement Overview, Functional Requirements, and Data Requirements. Based on the PRD, we employ GPT-4.1 to generate a corresponding metric outline, structuring test cases using the Arrange-Act-Assert (AAA) methodology [28]. The Arrange step sets up the test case by preparing the necessary files, input data, and environment configurations. The Act step focuses on the core behavior to be tested, such as running the program or performing interaction tests to obtain output results. The Assert step verifies the expected outcomes by checking the system’s response or state, ultimately determining whether the test passes or fails. This structure is applicable to a wide range of code-related testing scenarios, providing PRDBench with a comprehensive and executable testing plan.

Step 2: Scaffold and Criteria Generation. After initializing with the PRDs and metric outlines, we employ state-of-the-art code agents to generate the code scaffold for each task. The scaffold includes both module design and interface design, serving as the core framework of the entire project. In some cases, particularly for complex algorithms, the implementation may not be fully correct as long as it does not affect the annotation process. Building upon the scaffold, we further utilize code agents to expand and refine the metric outline into a specific criteria scheme, which includes

necessary test interfaces and test artifacts. The presence of well-defined code interfaces in the scaffold simplifies and standardizes the generation of the criteria scheme, making the process more efficient and consistent.

Step 3: Human Inspection. With the scaffold and the criteria scheme in place, the inspection process for human annotators becomes straightforward. Annotators conduct inspection only by running the tests to verify whether the interfaces function correctly and whether the expected outputs in the criteria scheme align with the PRD requirements. Notably, even for tasks sourced from expertise projects or academic papers, annotators only need basic computer science knowledge to perform effective checking and annotation.

Step 4: Agent-Based Fix and Refinement. If any issues are identified during step 3, annotators shall provide targeted feedback to the code agent for refining the scaffold or criteria scheme. The code agent then revises the relevant components, and the inspection process is repeated. This iterative process continues until all issues are resolved, ensuring the quality and correctness of the data.

Step 5: Remove Scaffold. In the final stage, the scaffold is removed and only the criteria scheme, test artifacts, required data, and PRD are retained in PRDBench. The removal of scaffold ensures that the code agents to be evaluated generate the code from scratch, thereby we can assess the end-to-end development capacity of the agents.

An overall example of a PRDBench task is illustrated in Figure 2. Each PRDBench testcase consists of a PRD document and an evaluation suite, which includes a criteria scheme and test artifacts.

3.3 EvalAgent and Evaluation

For each item in the criteria scheme, PRDBench provides executable terminal commands and expected outputs, enabling systematic verification. PRDBench employs a lightweight evaluation agent, EvalAgent, to automate the validation of code generated by code agents against predefined criteria schemes, as illustrated in Figure 3. EvalAgent is equipped with six core tools, including file reading and writing, command-line execution, image handling, and a judge tool. Among these, the dealgraph tool offers a multimodal LLM interface (GPT-4o in this paper), allowing the agent to process images by submitting both the image and a descriptive prompt to obtain results—this is particularly useful for tasks requiring visual verification. The judge tool is specifically designed for PRDBench. It accepts files containing simulated user inputs, enabling comprehensive terminal log generation for thorough analysis without the need for manual input specification. By automating the validation process, EvalAgent significantly reduces manual intervention and improves evaluation efficiency.

Within PRDBench, the EvalAgent performs three main categories of tests, which collectively simulate the different stages of QA in Python projects:

- **Unit Test:** As in previous benchmark studies [11], unit tests remain the most effective approach for verifying the functional correctness of individual components and modules within the code. PRDBench retains this type of testing and provides auxiliary pytest-based test scripts. The EvalAgent simply runs the pytest command to execute these tests.

Table 2: Agent Specifications and Open-Source Status.

Agent Type	Agent Framework	Model	Agent OSS	LLM OSS
Minimal	Basic code tools	Qwen3-Coder [25]	✓	✓
		GPT-5 [19]	✓	✗
		Claude-3.7-Sonnet [1]	✓	✗
		Gemini-2.5-pro [23]	✓	✗
Commercial	Gemini CLI [17] Claude Code [1] CodeX [18] Qwen Code [24]	Gemini-2.5-pro [23]	✓	✗
		Claude-4/3.7 Hybrid [2]	✗	✗
		GPT-5 [19]	✓	✗
		Qwen3-Coder [25]	✓	✓
EvalAgent	EvalAgent	Qwen3-Coder [25]	✓	✓

- **Shell Interaction:** For tasks involving command-line interaction or external system operations, the EvalAgent executes predefined shell commands and compares the actual outputs to the expected results, ensuring the code correctly handles system-level operations and user inputs. PRDBench supplies comprehensive simulated user input files and program entry commands for this purpose.
- **File Comparison:** For project-level tasks that generate files or require specific directory structures, the EvalAgent compares the produced files against reference solutions, checking for correctness in content, format, and organization. PRDBench provides the raw data for target file generation, the corresponding Python commands, and the reference solution files.

By leveraging these tools, the EvalAgent can automatically adapt to specific code repositories and complete all required tests. All test categories and supporting files are provided within PRDBench, which greatly reduces the complexity of evaluation. EvalAgent only needs to execute each test and analyze the results. This automated evaluation process generates detailed reports on the performance and correctness of code agent submissions.

4 EXPERIMENTS AND RESULTS

4.1 Data Statistics

PRDBench comprises 50 project-level tasks, encompassing a total of 1,262 scoring points. Specifically, there are 409 unit test points, 729 shell interaction points, and 124 file comparison points. Each PRD contains, on average, 105.22 lines. Figure 4(a) presents the distribution of scaffold line counts recorded during the annotation.

For all PRDs, we utilize GPT-4 to assign a domain label to each task. From these labeled tasks, we select 50 tasks that span 20 distinct domain labels to construct PRDBench. The distribution of these domain labels is illustrated in Figure 4(b). To ensure broad domain coverage, we include at least one representative task from less common domains. For more prevalent areas of Python development, such as data processing and machine learning, we select tasks that require a variety of technical approaches and sub-skills. This sampling strategy ensures that PRDBench provides comprehensive and balanced coverage across different Python application domains.

4.2 Code Agents

We evaluate two categories of code agents in our experiments:

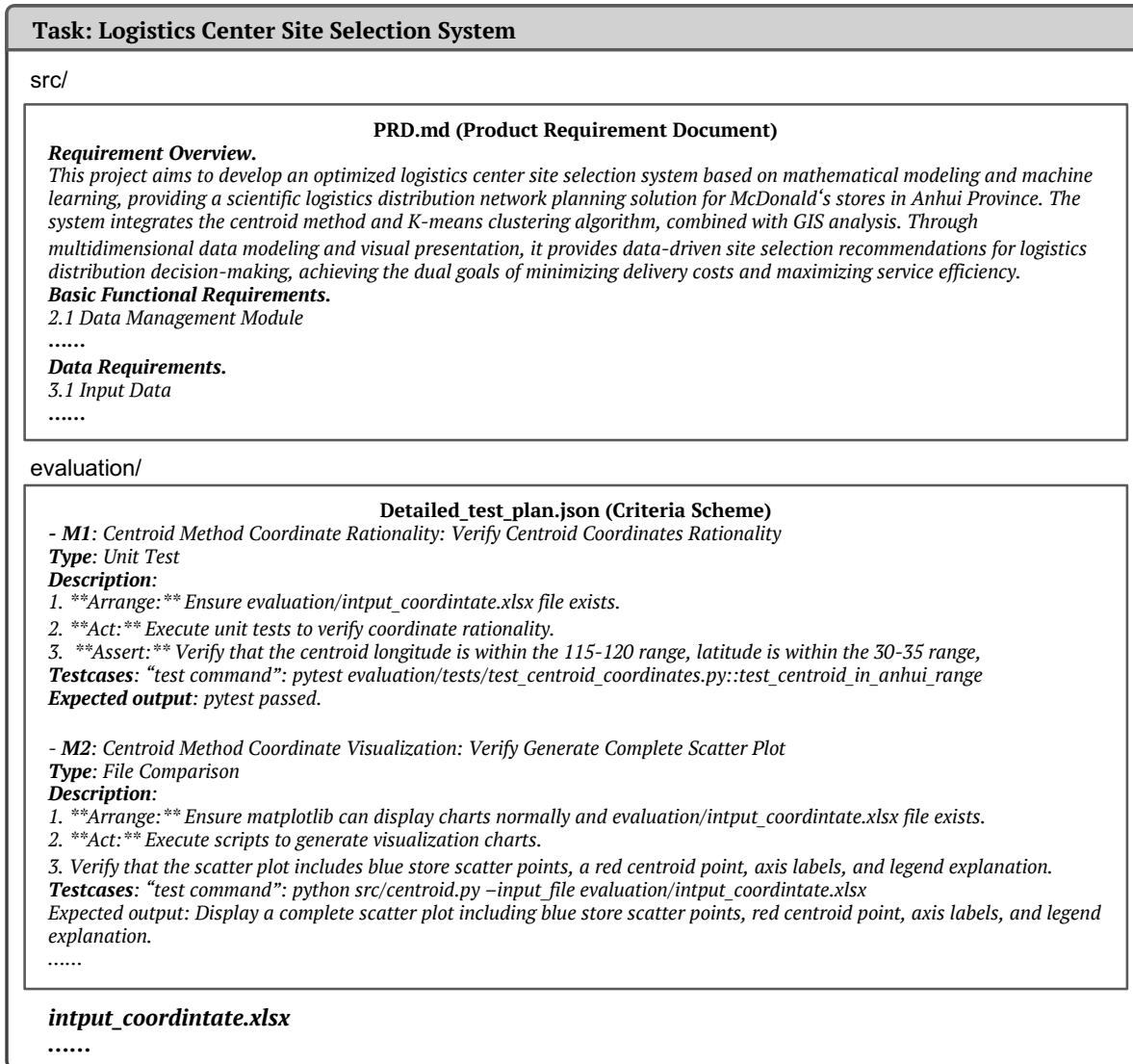


Figure 2: A task example of PRDBench.

- (1) **Minimal Agents**², implemented using the Agent Development Kit (ADK). These agents are equipped with essential tools for file manipulation, bash scripting, and Python execution. We integrate state-of-the-art LLMs (Claude-3.7-Sonnet, Gemini-2.5-Pro, Qwen3-Coder-480B-A35B, and OpenAI GPT-5) to systematically assess their core capabilities in strategy implementation. Hereafter we denote the models as Claude, Gemini, Qwen3-Coder and GPT-5 respectively for simplicity.
- (2) **Commercial Code Agents**, including advanced CLI-based agents (Claude Code, CodeX, Gemini CLI, and Qwen Code). These agents offer enhanced integration with command-line interfaces and external tools, representing the current state-of-the-art in code agent development.

Each commercial agent and minimal agent utilizes a corresponding backbone LLM. Commercial agents typically benefit from vendor-specific fine-tuning and optimization, resulting in more stable and robust performance within their respective frameworks. While minimal agents are designed to facilitate fair comparison across different LLMs, commercial agents provide an assessment of the latest advancements in agent capabilities. The mapping between agent frameworks and backbone models is summarized in Table 2.

4.3 Experimental Setups

We select Qwen3-Coder [24] as the backbone LLM for EvalAgent. Qwen3-Coder is fully open-source, which facilitates future fine-tuning and further experiments for Agent-as-a-Judge scenarios within PRDBench. For code agents, we select the latest model from each family that is compatible with the agent framework. However, Claude-4-Sonnet occasionally encounters tool call issues that result

²Code is available in <https://github.com/AGI-Eval-Official/Minimal-CodeAgent>.

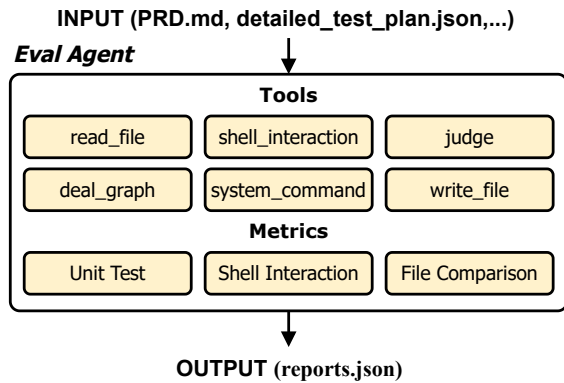


Figure 3: Overview of PRDBench evaluation. The EvalAgent executes tests based on the criteria scheme using various tools, compares outputs (files or results) with expected outputs, and generates a report for the submitted code.

in no code being generated, as frequently reported by the community. Therefore, we use Claude-3.7-Sonnet in minimal code agent. For all LLMs used in our work, we set temperature to be 0.1, max token identical to their official APIs’ setting. We set top-p to 1.0, Top-k to be 100, and presence penalty to be default to the API.

Each code agent is executed using a Python virtual environment that contains necessary and useful packages for the agents. If the code agent need other packages, we allow it use pip to install them.

4.4 Main Results

Table 3 presents the performance of all evaluated code agents on PRDBench. In Round 1, we provide the PRD and the relevant evaluation materials, allowing each code agent to implement the entire project based on the existing interfaces. This round assesses the development capability of the code agents. In Round 2, we provide the EvalAgent report from the first round along with the Round 1 code, enabling each code agent to analyze and revise the code according to the feedback and identified issues. This round evaluates the debugging capability of the code agents.

Our main findings are as follows: **(1) The coding ability of the underlying LLM significantly impacts the development performance of the code agent.** We observe that the relative ranking of minimal agents in round 1 is consistent with that of the commercial agents, indicating a strong correlation between the LLM’s capabilities and the agent’s development outcomes. **(2) Adapting the code agent framework to LLMs can further enhance development performance.** While commercial agents typically outperform minimal agents based on the same LLM, there are reasonable exceptions. For example, since Claude-3.7-Sonnet does not appear to be specifically optimized for Claude Code, Claude Code achieves lower scores than the ADK-based Claude agent in our experiments. Additionally, Gemini-2.5-Pro and Qwen3-Coder demonstrate better generalization and more stable performance across different code agent frameworks. **(3) Debugging and initial development require distinct capabilities from code agents.** In the Debug phase (Round 2), we provide the previous round’s test report to guide code agents in targeted fixes. Minimal code

Table 3: Average pass rate of code agents on PRDBench (in %). The best results are highlighted in bold, and the second-best results are underlined.

Agent Type	Agent	DEV.↑	DEBUG↑	Enhance↑
Minimal	GPT-5	<u>55.81</u>	60.15	4.34
	Claude	45.50	49.07	3.57
	Gemini	14.27	15.99	1.72
	Qwen3-Coder	37.60	46.98	9.38
Commercial	CodeX	56.23	<u>50.24</u>	-6.99
	Claude Code	36.60	45.53	<u>8.93</u>
	Gemini CLI	16.35	21.59	5.16
	Qwen Code	39.59	35.69	-3.90

agents demonstrate more stable improvements during debugging, consistently increasing their accuracy after applying corrections. In contrast, CodeX experiences a dramatic drop in performance in the DEBUG phase due to interface inconsistencies introduced during modification, which led to additional errors. These observations suggest that effective debugging requires not only error identification but also careful maintenance of code structure and interfaces, highlighting the importance of robust error analysis and correction strategies in code agent designs.

4.5 Code Agent Analysis

4.5.1 Cost of code agents. Table 4 presents the resource consumption statistics for each code agent, including time spent, input and output token counts, and lines of code generated or modified during both development (Round 1) and debugging (Round 2). Overall, **commercial agents tend to consume more time than minimal agents during both phases. For each backbone LLM, the resource consumption patterns of its minimal and commercial agent implementations are generally consistent.** Notably, Gemini exhibits significantly lower input token consumption in the development phase compared to other models, whereas Qwen3-Coder consume the least output token. However, in the debugging phase, Gemini’s input token usage is more than twice that of other agents, indicating that the development and debugging phases pose distinct challenges for code agents in terms of resource demands. Regarding code modification, Claude Code and Qwen Code made relatively few changes during debugging, while GPT-5 and Gemini demonstrated much larger code modifications. This suggests that different code agents adopt varying strategies or exhibit different levels of responsiveness when addressing feedback and implementing code revisions. Some agents may focus on targeted, minimal fixes, while others opt for more extensive refactoring or rewriting in response to evaluation reports.

In summary, the observed differences in resource consumption and code modification patterns highlight that development and debugging are fundamentally different for code agents, requiring distinct approaches and optimization strategies. The results also suggest that agent architecture and underlying LLM characteristics strongly influence how agents allocate resources and perform code updates during the evaluation process.

4.5.2 Error Analysis of Code Agent. Figure 5 shows the error rates of code agents across different types of test cases. The error rate

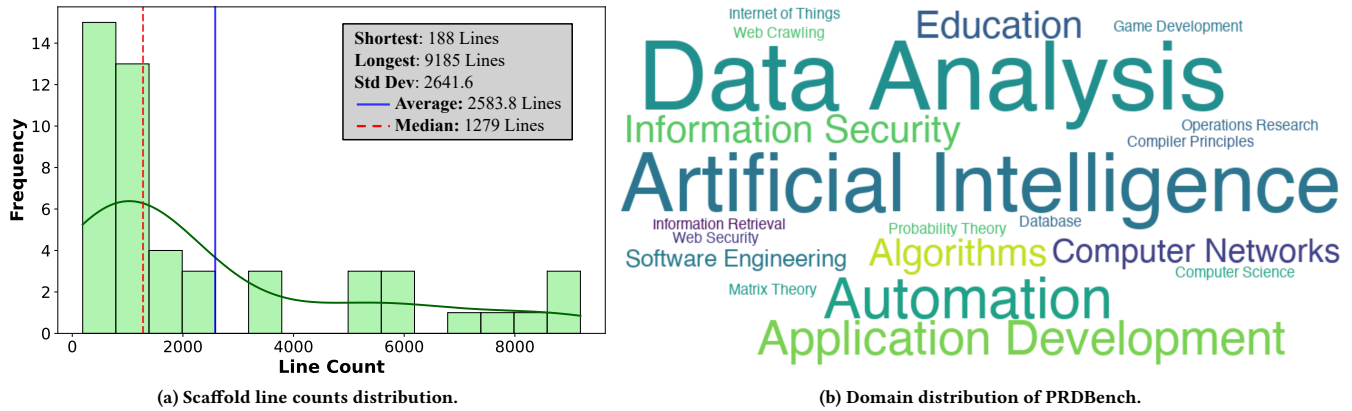


Figure 4: Data statistics of PRDBench.

Table 4: Cost statistics of code agents. Code indicates the number of lines generated in Round 1 (development) and the number of lines modified in Round 2 (debug). Relevant log data for CodeX is unavailable, so its results are not reported.

	Round 1 (DEVELOPMENT)				Round 2 (DEBUG)			
	Time (s)	Input	Output	Code (lines)	Time (s)	Input	Output	Δ Code (lines)
Minimal Agent								
GPT-5	1517.75	1506252.18	43467.62	854.49	1165.25	1888893.43	32358.57	774.04
Claude	717.50	1775631.15	33398.27	1975.35	380.37	599139.68	22760.68	285.96
Gemini	760.54	609661.54	27374.89	667.78	1370.54	3419419.51	32752.20	1579.06
Qwen3-Coder	892.35	1136848.60	17209.19	982.67	495.48	1835992.38	14176.74	917.50
Commercial Agent								
CodeX	-	-	-	651.62	-	-	-	101.18
Claude Code	1309.84	4113143.56	46669.50	1751.80	729.74	2140884.40	19593.26	112.68
Gemini CLI	2740.70	834125.00	20238.96	362.13	2067.61	5776370.08	24611.50	546.12
Qwen Code	1183.42	3108353.62	31400.28	1183.42	9582.81	988041.65	5604.89	31.00

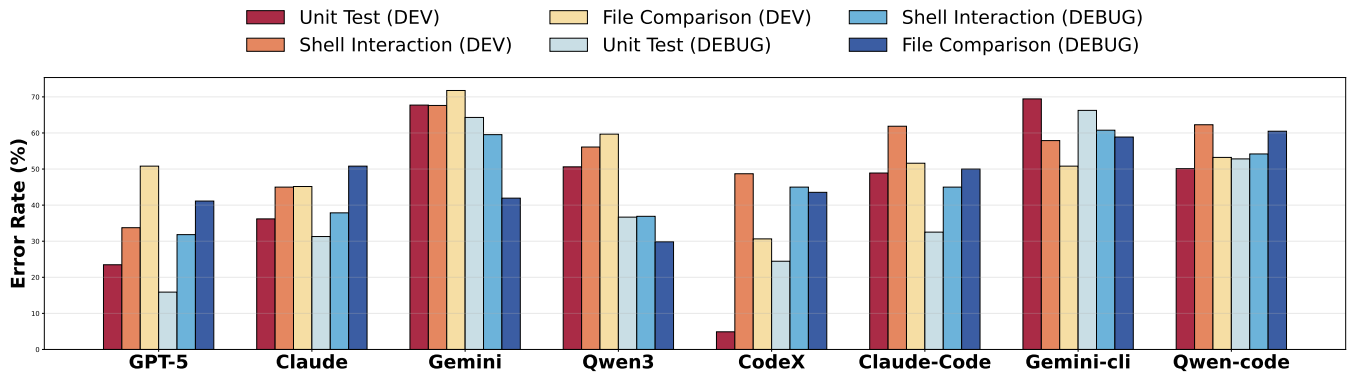


Figure 5: Error rates of code agents on different types of test cases.

reflects the proportion of failed test cases within each category for the code generated by the agents. This comparison highlights the strengths and weaknesses of various agents when handling diverse evaluation scenarios. We observe two main findings from the results. First, there is no single category of test cases that is

consistently easy for code agents. The error rates are relatively uniform across the three types of test cases, indicating that PRD-Bench’s test case design is well-balanced and effectively covers a broad spectrum of code implementation scenarios. Second, unit test cases are noticeably more challenging to debug compared to the

Table 5: Alignment between Human-as-a-Judge and Agent-as-a-Judge.

	Total	Unit Test	Shell Inter.	File Comp.
Human-as-a-Judge				
Number	282	107	149	26
Alignment	81.56%	79.44%	82.55%	84.62%

other two types. Most code agents fail to resolve these issues even after the debugging phase. This is primarily because unit test cases require agents to read and understand the test function code before making corrections, whereas the other two test types only require agents to compare actual and expected outputs, which lowers the reasoning complexity.

4.6 EvalAgent Analysis

4.6.1 Human Alignment. We assess the reliability of the evaluation agent through manual annotation. Specifically, we randomly select 16 code submissions, each corresponding to a distinct problem, and employ two annotators to score the outputs of the code agent according to the criteria scheme from PRDBench. These 16 codes are generated by different code agents and span different problems.

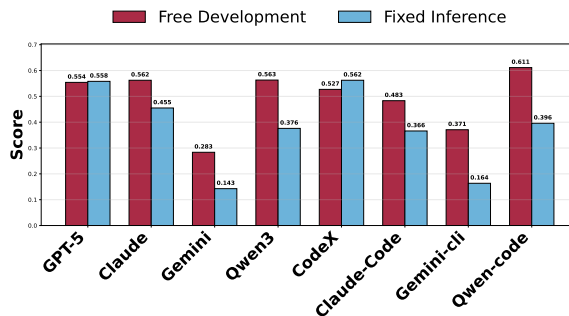
The results of human alignment are summarized in Table 5. In total, we examine 282 distinct test cases. Of these, 230 test cases (81.56%) achieve perfect score alignment between the annotators and the EvalAgent, indicating that **EvalAgent’s evaluation results are generally reliable**. For 9 cases (3.2%), the absolute difference in scores is 1, and for 43 cases (15.2%), the absolute difference is 2. We analyze the alignment rates across different projects and observe substantial variance (variance = 774.66), indicating significant differences in score alignment between projects. The standard deviation of alignment rates is also relatively high (27.83%), with values ranging from 0% to 100%. This suggests that the consistency between machine and human scoring can vary greatly depending on the specific project or model. The result implies that the stability of EvalAgent remains a major challenge, as an error occurs in one test case, subsequent evaluations within the same project may be affected due to contextual dependencies, potentially leading to distorted overall assessment for that problem. This phenomenon is also observed in other LLM-driven agents [15], and attributed to propagation chains.

In addition, we examine the alignment rates among the three major test case types: file comparison (84.62%), shell interaction (82.55%), and unit test (79.44%). The alignment rates are comparable across these categories, indicating that the evaluation difficulty for EvalAgent is similar among the three types. This demonstrates that our three test formats, as well as the commonly used unit test benchmarks, are well-suited for reliable assessment by EvalAgent.

4.6.2 Cost of EvalAgent. Table 6 shows the average time and token cost for EvalAgent. On average, EvalAgent generates a comprehensive testing report in just 7 minutes at an API cost of \$2.68 per problem, compared to 0.5 to 1 hour for human annotators (excluding writing reports). This demonstrates that EvalAgent and PRDBench greatly improve evaluation efficiency while keeping costs low.

Table 6: Average cost of EvalAgent. The input and output are quantified by the number of tokens, and the API cost is estimated according to the official pricing of Qwen3-coder.

	Time (s)	Input	Output	API Cost(\$)
EvalAgent	425.62	124,2440	8,825	2.68

**Figure 6: Comparison of code agent scores under fixed inference and free development modes on PRDBench.**

4.7 Free Development

Currently, PRDBench uses code agents to generate project scaffolding and fixes the inference interfaces in the criteria scheme. Code agents are required to implement code that conforms to these predefined interfaces, which greatly simplifies the evaluation process for EvalAgent. However, in real-world development, free-form implementation is more common, where only the PRD is provided, and interfaces are not fixed. To assess code agents in this free development setting, we modified EvalAgent to accommodate flexible interface designs and evaluated the agents’ ability to develop projects without fixed scaffolding.

Figure 6 presents a comparison of code agent performance in the fixed inference and free development modes on PRDBench. Our results show that, compared to fixed inference development, the distinction among code agents decreases in the free development scenario, as reflected by the variance in performance scores: 0.011 for free development versus 0.028 for fixed inference. Nevertheless, the relative ranking of code agents remains fairly stable between the two settings. This demonstrates that PRDBench can effectively evaluate code agents in both constrained and unconstrained development environments while maintaining reliable differentiation.

5 CONCLUSION

In this work, we introduce PRDBench, a PRD-centered benchmark that enables flexible and comprehensive evaluation of code agents at the project level with reduced annotation costs. Our agent-driven construction approach greatly reduces annotation cost and enables efficient creation of project-level benchmarks. We further propose EvalAgent and a corresponding flexible metric scheme to support comprehensive and adaptable evaluation of code agents. Based on these innovations, PRDBench provides a scalable and realistic foundation for advancing research in automated software development and code agent evaluation.

ACKNOWLEDGMENTS

The Shanghai Jiao Tong University team is partially supported by National Natural Science Foundation of China (62322603, 62177033, 62502310). We extend our sincere gratitude to Shuang Zhou, Yuan Zhang (张元), Yuan Zhang (张源), Ying Xie, Hao Zheng, Yan Wang, and Zhiqiang Li for their contributions to data annotation.

REFERENCES

- [1] Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. <https://www.anthropic.com/news/claude-3-7-sonnet>. Accessed: 2025-02-25.
- [2] Anthropic. 2025. Introducing Claude 4. <https://www.anthropic.com/news/claude-4>. Accessed: 2025-05-23.
- [3] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. 2024. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095* (2024).
- [4] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinyao Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.
- [5] Alireza Daghighifarsodeh, Chung-Yu Wang, Hamed Taherkhani, Melika Sepidband, Mohammad Abdollahi, Hadi Hemmati, and Hung Viet Pham. 2025. Deep-Bench: Deep Learning Benchmark Dataset for Code Generation. *arXiv:2502.18726 [cs.SE]* <https://arxiv.org/abs/2502.18726>
- [6] Lingyue Fu, Hao Guan, Bolun Zhang, Haowei Yuan, Yaoming Zhu, Jun Xu, Zongyu Wang, Lin Qiu, Xunliang Cai, Xuezhi Cao, et al. 2025. CoreCodeBench: A Configurable Multi-Scenario Repository-Level Benchmark. *arXiv preprint arXiv:2507.05281* (2025).
- [7] Nam Le Hai, Dung Manh Nguyen, and Nghi D. Q. Bui. 2025. On the Impacts of Contexts on Repository-Level Code Generation. *arXiv:2406.11927 [cs.SE]* <https://arxiv.org/abs/2406.11927>
- [8] Xanh Ho, Jiahao Huang, Florian Boudin, and Akiko Aizawa. 2025. LLM-as-a-Judge: Reassessing the Performance of LLMs in Extractive QA. *arXiv preprint arXiv:2504.11972* (2025).
- [9] Yiming Huang, Jianwen Luo, Yan Yu, Yitong Zhang, Fangyu Lei, Yifan Wei, Shizhu He, Lifu Huang, Xiao Liu, Jun Zhao, and Kang Liu. 2024. DA-Code: Agent Data Science Code Generation Benchmark for Large Language Models. *arXiv:2410.07331 [cs.CL]* <https://arxiv.org/abs/2410.07331>
- [10] Anysphere Inc. 2025. Cursor: The AI powered Code Editor. <https://cursor.com/>. AI assistant code editor announcement, Anysphere blog.
- [11] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770* (2023).
- [12] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? *arXiv:2310.06770 [cs.CL]* <https://arxiv.org/abs/2310.06770>
- [13] Bowen Li, Wenhan Wu, Ziwei Tang, Lin Shi, John Yang, Jinyang Li, Shunyu Yao, Chen Qian, Binyuan Hui, Qicheng Zhang, Zhiyin Yu, He Du, Ping Yang, Dahua Lin, Chao Peng, and Kai Chen. 2024. Prompting Large Language Models to Tackle the Full Software Development Lifecycle: A Case Study. *arXiv:2403.08604 [cs.CL]* <https://arxiv.org/abs/2403.08604>
- [14] Jia Li, Ge Li, Xuanming Zhang, Yihong Dong, and Zhi Jin. 2024. EvoCodeBench: An Evolving Code Generation Benchmark Aligned with Real-World Code Repositories. *arXiv:2404.00599 [cs.CL]* <https://arxiv.org/abs/2404.00599>
- [15] Xixun Lin, Yucheng Ning, Jingwen Zhang, Yan Dong, Yilong Liu, Yongxuan Wu, Xiaohua Qi, Nan Sun, Yanmin Shang, Pengfei Cao, Lixin Zou, Xu Chen, Chuan Zhou, Jia Wu, Shirui Pan, Bin Wang, Yanan Cao, Kai Chen, Songlin Hu, and Li Guo. 2025. LLM-based Agents Suffer from Hallucinations: A Survey of Taxonomy, Methods, and Directions. *arXiv:2509.18970 [cs.AI]* <https://arxiv.org/abs/2509.18970>
- [16] Kaiyuan Liu, Youcheng Pan, Yang Xiang, Daojing He, Jing Li, Yexing Du, and Tianrun Gao. 2025. ProjectEval: A Benchmark for Programming Agents Automated Evaluation on Project-Level Code Generation. *arXiv:2503.07010 [cs.SE]* <https://arxiv.org/abs/2503.07010>
- [17] Taylor Mullen and Ryan J. Salva. 2025. Gemini CLI: Your Open Source AI Agent. <https://blog.google/technology/developers/introducing-gemini-cli-open-source-ai-agent/>. Google Developers Blog, Jun 2025.
- [18] OpenAI. 2025. CodeX CLI. <https://github.com/openai/codex>. Accessed: 2025-05-16.
- [19] OpenAI. 2025. Introducing GPT-5. <https://openai.com/index/introducing-gpt-5/>. Accessed: 2025-08-07.
- [20] Zhenyu Pan, Rongyu Cao, Yongchang Cao, Yingwei Ma, Binhua Li, Fei Huang, Han Liu, and Yongbin Li. 2024. Codev-Bench: How Do LLMs Understand Developer-Centric Code Completion? *arXiv:2410.01353 [cs.SE]* <https://arxiv.org/abs/2410.01353>
- [21] Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. HumanEval-XL: A Multilingual Code Generation Benchmark for Cross-lingual Natural Language Generalization. *arXiv:2402.16694 [cs.CL]* <https://arxiv.org/abs/2402.16694>
- [22] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. 2025. PaperBench: Evaluating AI’s Ability to Replicate AI Research. *arXiv preprint arXiv:2504.01848* (2025).
- [23] Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv:2507.06261 [cs.CL]* <https://arxiv.org/abs/2507.06261>
- [24] Qwen Team. 2025. Qwen Code brings the capabilities of advanced code models to your terminal in an interactive Read-Eval-Print Loop (REPL) environment. <https://qwenlm.github.io/qwen-code-docs/en/>. Accessed: 2025-07-22.
- [25] Qwen Team. 2025. Qwen3-Coder: Agentic Coding in the World. <https://qwenlm.github.io/zh/blog/qwen3-coder/>. Accessed: 2025-07-22.
- [26] Konstantinos Vergopoulos, Mark Niklas Müller, and Martin Vechev. 2025. Automated Benchmark Generation for Repository-Level Coding Tasks. *arXiv:2503.07701 [cs.SE]* <https://arxiv.org/abs/2503.07701>
- [27] Xincheng Wang, Pengfei Gao, Chao Peng, Ruida Hu, and Cuiyun Gao. 2025. Code-Visionary: An Agent-based Framework for Evaluating Large Language Models in Code Generation. *arXiv:2504.13472 [cs.SE]* <https://arxiv.org/abs/2504.13472>
- [28] Chenhao Wei, Lu Xiao, Tingting Yu, Sunny Wong, and Abigail Clune. 2025. How Do Developers Structure Unit Test Cases? An Empirical Analysis of the AAA Pattern in Open Source Projects. *IEEE Transactions on Software Engineering* (2025).
- [29] Martin Weysow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. 2024. CodeUltraFeedback: An LLM-as-a-Judge Dataset for Aligning Large Language Models to Coding Preferences. *arXiv:2403.09032 [cs.SE]* <https://arxiv.org/abs/2403.09032>
- [30] Qinyun Wu, Chao Peng, Pengfei Gao, Ruida Hu, Haoyu Gan, Bo Jiang, Jinhe Tang, Zhiwen Deng, Zhanming Guan, Cuiyun Gao, Xia Liu, and Ping Yang. 2024. RepoMasterEval: Evaluating Code Completion via Real-World Repositories. *arXiv:2408.03519 [cs.SE]* <https://arxiv.org/abs/2408.03519>
- [31] Yanzheng Xiang, Hanqi Yan, Shuyin Ouyang, Lin Gui, and Yulan He. 2025. SciReplicate-Bench: Benchmarking LLMs in Agent-driven Algorithmic Reproduction from Research Papers. *arXiv:2504.00255 [cs.CL]* <https://arxiv.org/abs/2504.00255>
- [32] Jian Yang, Jiajun Zhang, Jiayi Yang, Ke Jin, Lei Zhang, Qiyao Peng, Ken Deng, Yibo Miao, Tianyu Liu, Zeyu Cui, Binyuan Hui, and Junyang Lin. 2024. ExecRepoBench: Multi-level Executable Code Completion Evaluation. *arXiv:2412.11990 [cs.CL]* <https://arxiv.org/abs/2412.11990>
- [33] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. 2024. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934* (2024).
- [34] Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widayarsi, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen Gong, Thong Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, Binyuan Hui, Niklas Muennighoff, David Lo, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. 2025. BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions. *arXiv:2406.15877 [cs.SE]* <https://arxiv.org/abs/2406.15877>