
SimCopilot: Evaluating Large Language Models for Copilot-Style Code Generation

Mingchao Jiang¹, Abhinav Jain¹, Sophia Zorek² & Chris Jermaine¹ *

¹Department of Computer Science ²Department of Statistics

Rice University
Houston, TX 77005, USA
`{mj33,aj70,sazz,cmj4}@rice.edu`

Abstract

We introduce SIMCOPilot, a benchmark that simulates the role of large language models (LLMs) as interactive, “copilot”-style coding assistants. Targeting both *completion* (finishing incomplete methods or code blocks) and *infill* tasks (filling missing segments within existing code), SIMCOPilot provides a comprehensive framework for evaluating LLM coding capabilities. The benchmark comprises dedicated sub-benchmarks for Java (SIMCOPilotJ) and Python (SIMCOPilotP), covering diverse codebases varying in size and complexity. Our key contributions include: (a) establishing a realistic, detailed evaluation environment to assess LLM utility in practical coding scenarios, and (b) providing fine-grained analyses that address critical factors frequently overlooked by existing benchmarks, such as task-specific performance nuances, contextual understanding across code segments, and sensitivity to variable scope. Evaluations conducted across domains—including algorithms, databases, computer vision, and neural networks—offer insights into model strengths and highlight persistent challenges in maintaining logical consistency within complex dependency structures. Beyond benchmarking, our study sheds light on the current limitations of LLM-driven code generation and underscores the ongoing transition of LLMs from merely syntax-aware generators toward reliable, intelligent software development partners.

1 Introduction

Currently, the most widely-used benchmarks for checking the ability of AI models to perform program synthesis (“AI-for-code”) consist of a detailed English description of a concise, self-contained code to synthesize, as well as a few test cases to test the correctness of the synthesized code (Austin et al., 2021; Hendrycks et al., 2021; Chen et al., 2021; Iyer et al., 2018). While such benchmarks are useful, they match one particularly narrow use case, where the goal is to synthesize a relatively short, complete, standalone program.

Arguably, the most impactful and widely-used application of AI-for-code to date has been through tools such as GitHub’s Copilot (GitHub & OpenAI, 2024). Copilot is designed to be used interactively, where the AI is repeatedly given code completion tasks: given a partially-completed code such as a commented method header with an empty body, an if-statement with an empty else block, or an empty for loop body, can an AI tool correctly complete the next few lines of code?

Rather than writing code from scratch, the goal is to interactively help a programmer to write code, reducing the programmer’s burden and increasing productivity. In practice, this is often done in the context of a large software project, where the AI needs to figure out how to call internal APIs that are specific to the project correctly or to reference classes or variables that were declared remotely from the site location of the code completion task.

*Code: github.com/mj33rice/Sim-CoPilot • Dataset: huggingface.co/datasets/mj33/SimCoPilot

In this paper, we present an AI-for-code benchmark called SIMCOPilot, which is specifically designed to simulate a “Copilot”-style, interactive environment. All of the code completion tasks were chosen so that it would be impossible for an existing AI system to have trained on the code. SIMCOPilot consists of two sub-benchmarks, SIMCOPilotJ (Java) and SIMCOPilotP (Python). SIMCOPilotJ and SIMCOPilotP are separately designed to test how Java (and related languages) and Python (and related languages) are likely to be used in practice.

Crucially, SIMCOPilot is designed to test two different sub-cases in interactive code completion: *completion* and *infill*. Completion simulates the case where a programmer is using an interactive AI programming tool to write a method/function from start to finish. Here, a method or function is either empty (consisting of only a commented header), or it may have some code starting from the beginning of the method/function, but the code is not complete. The goal is to follow instructions and write a certain portion of the missing code (such as finishing the current `else` block), following the last-provided line. However, not all code is written linearly, from start to finish, and *infill* is designed to simulate this case. In *infill*, there is a “blank” in a method, function, or logic block, and the goal is to provide the missing code.

Java is a structured, relatively high-performance, compiled language generally used to develop larger software projects, whereas Python is a scripting language often used to solve smaller programming problems. Hence, SIMCOPilotJ consists of 286 code completion and 283 code infilling tasks over eight separate modules (groups of related classes) in a medium-sized, well-documented Java project (7,728 lines in all). SIMCOPilotP consists of 212 code completion and 382 code infilling tasks over 7 separate Python programs for reasonably standard tasks in computer vision, databases, optimization, etc., of intermediate length (consisting of an average of 431 lines each). An AI programming assistant is considered to have “passed” one of the code completion tasks if the AI-written code, together with the rest of the code in the program/project, is able to pass an extensive set of test cases.

In this paper, we carefully describe the SIMCOPilot benchmark and evaluate a number of different AI-for-code models. The results presented show how SIMCOPilot can have a much more detailed and meaningful view of the differences between various AI-for-code models. For example, popular benchmarks such as HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) show relatively modest differences between models. Claude 3 Opus (Anthropic, 2025b) scores 84.9% on HumanEval, whereas Llama 3.3 70B (AI, 2025b) scores 88.4%. Based on HumanEval alone, one might conclude that Llama 3.3 70B outperforms Claude 3 Opus. However, SIMCOPilot reveals a different picture across more realistic programming tasks. Claude 3 Opus achieves 67.4% and 69.2% on Python and Java completion tasks. In contrast, Llama 3.3 70B scores significantly lower performance of 53.7% and 49.3% on Python and Java completion tasks. This dramatic reversal demonstrates how SIMCOPilot provides a more comprehensive and realistic assessment of model capabilities in practical programming scenarios, where contextual understanding and code integration skills are essential.

2 Related Work

Recent code generation benchmarks have evolved from generating concise, standalone programs from detailed specifications, as seen in HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), and APPS (Hendrycks et al., 2021), to more context-aware evaluations. CrossCodeEval (Ding et al., 2024) and RepoBench (Liu et al., 2023) assess code completion within larger repositories, focusing on cross-file references. ClassEval (Du et al., 2023), CoderEval (Yu et al., 2024), and EvoCodeBench (Li et al., 2024) emphasize test-based validation yet evaluate AI models using detailed, manually annotated problem descriptions to prompt code generation. SWE-bench Verified (OpenAI, 2024b) tasks models with resolving real-world GitHub issues and validates their patches against regression tests. LiveCodeBench (Jain et al., 2024) offers dynamic, contamination-free challenges from online contests, assessing capabilities like iterative self-repair and code execution. ProBench (Yang et al., 2025) focuses on competitive programming tasks, analyzing reasoning depth and error patterns. DynaCode (Hu et al., 2025) introduces complexity-aware evaluations with nested code tasks, while

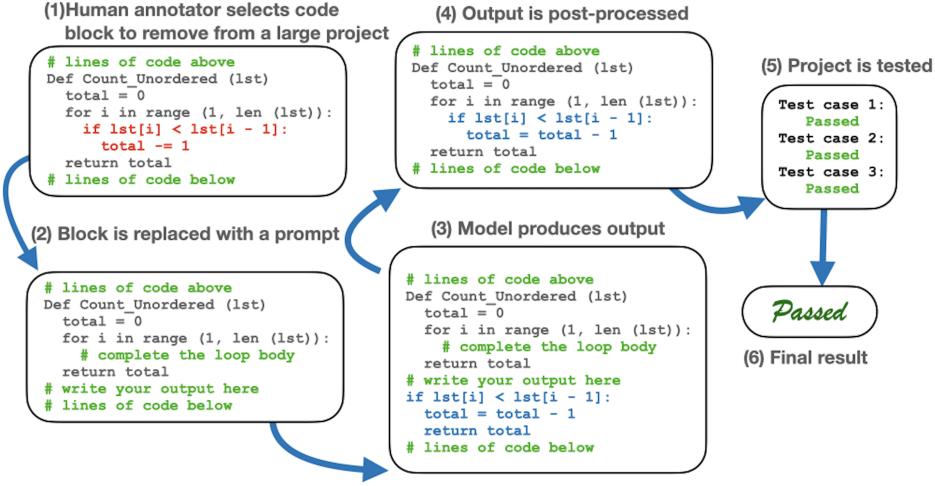


Figure 1: Workflow for each of the 1,163 programming tasks in SIMCOPilot.

BigO(Bench) (Chambon et al., 2025) evaluates adherence to specified time and memory efficiency constraints.

Despite these comprehensive benchmarks, none directly simulate the interactive “AI pair programmer” workflow. SIMCOPilot addresses this gap by focusing on left-to-right code completion and mid-code infilling tasks, reflecting how developers use tools like Copilot to incrementally add and refine code. The benchmark provides a realistic, detailed evaluation environment for LLMs acting as coding assistants, enabling fine-grained analyses of underexplored challenges such as scope sensitivity, task-specific performance variations, and inter-segment contextual dependencies. These contributions distinguish SIMCOPilot among code generation benchmarks by targeting practical, context-dependent, interactive coding assistance scenarios.

3 SimCopilot Benchmark Design

SIMCOPilot consists of 569 Copilot-style Java programming problems, and 594 Python problems, as well as a comprehensive suite of software tools for the ancillary tasks necessary to run the benchmark, such as code post-processing and data analysis. The general workflow for creating and executing each programming problem is shown in Figure 1. A human annotator chooses a section of code to omit (the benchmark comes with 1,163 sections of code marked for omission). When it is time to test an AI-for-code model, the SIMCOPilot software removes the annotated code and replaces it with a prompt. The model is asked to write the missing code; the model’s output is then post-processed and inserted back into the original code, which is then tested. If the code passes all test cases, the model passes (See as Fig. 6 and 7 in appendix).

3.1 Annotation

Each of the 1,163 programming tasks was created from eight Java repositories and seven Python repositories, totaling nearly 11,000 lines of code. We employed a systematic two-stage pipeline to identify semantically meaningful code blocks for generation tasks:

- **AST-based parsing:** We construct an abstract syntax tree (AST) for each source file to locate constructs such as classes, functions, loops, and conditionals.
- **DFS-based block extraction:** We apply a depth-first traversal to extract complete, self-contained blocks, ensuring syntactic integrity and well-formed prompts.

	Python		Java	
	Infill	Compl.	Infill	Compl.
Total Count	382	212	283	286
Local Variable	371	206	232	218
Global Variable	-	-	157	184
Function	37	9	110	88
Class	15	19	54	32
Library	202	123	6	10
If-Else Cond	19	14	56	33
If-Else Body	61	73	79	120
Loop Body	72	49	102	102
Avg Num Lines	431		966	

Table 1: Task categories and frequencies.

To enforce consistency, block selection is driven by deterministic structural rules. Specifically, we ① exclude trivial boilerplates, such as 1-line accessors or stubs, ② filter for non-trivial logic with control flow or function calls, and ③ ensure broad task coverage, including control structures, function or library calls, and object references.

Our team, consisting of PhD-level programmers, went through these codes, generating both infill and completion tasks. To create an infill task, the annotator picks a meaningful starting point for the AI-for-code model to begin writing code (at the beginning of the boolean if condition, or at the beginning of the body of a `for` loop, for example see as Fig.6) and then marks the rest of that particular code block for deletion, to be re-created by the AI-for-code model. In the case of an if condition, the entire boolean predicate would be marked for deletion. In the case of a `for` loop body, the entire body would be marked. A completion task is created in much the same way, but the code for the remainder of the method or function is marked for deletion (see as Fig.7). In total, the average number of lines deleted for infill is 2.62 for Java and 2.52 for Python, and the average number of lines deleted for completion is 3.31 for Java and 3.89 for Python.

One consideration when choosing code blocks for deletion is that we want good coverage of various programming tasks, as SIMCOPilot does not simply report the overall pass rate but breaks all of the 1,163 programming tasks into eight overlapping categories and reports results from each. These categories are as follows: (1) *Local variable*—the programming task requires the use of a previously-defined local variable; (2) *Global variable*—the task requires the use of a global variable; (3) *Function*—requires calling a function defined previously in the repository; (4) *Class*—requires a reference to a class previously defined; (5) *Library*—requires use of an external library; (6) *If-else-cond*—requires generating the boolean condition in an `if` or `else if` statement; (7) *If-else-body*—requires generating the body of an `if`, `else if`, `else` statement; (8) *Loop body*—requires generating a loop body. The number of occurrences of each of these eight different types of tasks is given in Table 1.

3.2 Pre-Processing

To generate an actual programming task, the SIMCOPilot software locates a block marked for deletion. All other files in the repository are prepended to the file containing the marked block, the class/function/method containing the marked block is moved to the end of the file, and the marked block is replaced with a prompt. Then, the AI-for-code model reacts to the prompt and attempts to re-create the missing code. All of this is done so as to mimic how a code completion task might be prepared by a Copilot-style software deployment.

Our current version of the SIMCOPilot software contains only one generic pre-processing pipeline. However, as the SIMCOPilot software evolves and the benchmark is applied to different AI-for-code models, it is acceptable—even desirable—for the authors or particular AI-for-code models to prepare their own pre-processing pipelines that work best with their

own model. Our reasoning is that in a “real-life” deployment where an AI-for-code model is used to write code interactively, there is undoubtedly going to be a highly specialised pre-processing pipeline that works well with the specific model.

Our current prompt emphasizes the following aspects of code generation to the AI-for-code model and was engineered through trial and error: (1) The code should logically continue from the section before the prompt (and connect smoothly to the section after the prompt for infill tasks); (2) Insert only syntactically correct code without any additional comments or text; (3) Ensure all brackets, parentheses, and curly braces are properly paired and closed (and matches indentation for Python); (4) For completion tasks, complete the innermost incomplete method, function, loop body, if-statement, or self-contained code block.

3.3 Post-Processing

Post-processing is crucial, as even the best AI-for-code models add irrelevant English or other comments and repeat lines of code, or even whole code blocks, despite being prompted not to. Any such model deployed in a production environment would have a custom-built post-processor to attempt to correct such problems. The SIMCOPilot benchmark comes with a post-processor, though different post-processors customized for specific AI-for-code models may be used.

Post-processing begins by using a series of regular expressions to eliminate statements that are clearly not statements in the target programming language. Post-processing then removes any repeating lines of code, or repeating code blocks. One very common problem is when the AI-for-code model re-generates code above the prompt, or—in an infill task—it re-generates code following the prompt. Our post-processor searches for both exact and approximate re-generation of provided code, and then removes any such re-generated code from the response. The next step is to process indentation (python) and bracket (Java) discrepancies. For Python, the post-processor automatically aligns the generated code with the preceding code that the AI-for-code model is attempting to complete. For Java, the post-processor adds or deletes brackets from the generated code in an attempt to make it syntactically correct. Once the generated code has been post-processed, it is inserted as a replacement for the deleted code block, and the resulting code is tested for correctness using a series of test cases.

3.4 Code Repositories Used

We chose a set of Java and Python repositories to use for the benchmark. None of the repositories were publicly accessible, to make it highly unlikely that any existing AI-for-code model had been trained on the repository.

The Java repositories are all intermediate-to-advanced academic programming projects. These include: processing a text corpus to build bag-of-words vectors, sparse vector and matrix implementations, generation of random variates, implementation of an AVL tree, a B-tree implementation, and an M-Tree implementation. The Python repositories were selected to include notebook-style scripts and more modular, object-oriented codes. These codes cover classic algorithms such as linear programming, as well as tasks in computer vision and reinforcement learning. See Appendix for examples of Java and Python codes from the benchmark given in Figure 7 and Figure 6, respectively.

4 Preparation and Breakdown of Results

All AI-for-code models are allowed only one chance to produce a result in the SIMCOPilot benchmark to mimic the real-life application scenario. A user asking an AI to produce a code is likely going to ask one time for a code and either accept (and possibly modify) the result or reject it; it is unlikely that any user is going to look through many different results. When controllable and applicable, the randomness of the AI-for-code model is “turned off” so that the most likely/preferred answer is produced. Rather than simply producing one single pass rate, the benchmark produces four pass rates (Java infill and completion; Python

Model	Python		Java		Hum-Eval
	Infill	Compl.	Infill	Compl.	
GPT-4o (2024-08-06)	78.5 ± 4.1	69.9 ± 6.2	78.1 ± 4.9	54.9 ± 5.7	92.7
o3-mini (high)	83.3 ± 3.7	66.0 ± 6.3	87.6 ± 3.8	59.1 ± 5.7	-
Claude 3 Opus	75.1 ± 4.3	67.4 ± 6.2	66.8 ± 5.5	69.2 ± 5.4	84.9
Claude 3.7 Sonnet (ET.)	80.9 ± 3.9	67.5 ± 6.4	86.9 ± 4.0	68.5 ± 5.4	97.8
Claude 3.7 Sonnet	74.3 ± 4.4	57.1 ± 6.6	71.0 ± 5.3	69.5 ± 5.3	94.9
Claude 3.5 Haiku	70.9 ± 4.5	60.4 ± 6.5	66.4 ± 5.5	61.2 ± 5.6	88.1
Llama 3.3 70B	58.1 ± 4.9	53.7 ± 6.8	65.7 ± 5.5	49.3 ± 5.8	88.4
Llama 3.1 8B	43.7 ± 4.9	39.6 ± 6.6	36.4 ± 5.7	32.8 ± 5.4	72.6
DeepSeek-R1 671B	73.3 ± 4.4	64.6 ± 6.4	77.4 ± 4.9	59.4 ± 5.6	97.7
R1-Distill-Qwen-14B	46.9 ± 5.0	38.7 ± 6.5	38.9 ± 5.7	39.1 ± 5.6	-
Qwen2.5-Coder-32B	70.2 ± 4.5	64.7 ± 6.4	63.2 ± 5.6	56.3 ± 5.7	92.1
Qwen-QwQ-32B	52.6 ± 5.0	47.2 ± 6.7	51.6 ± 5.8	34.3 ± 5.4	97.6

Table 2: Overall SIMCOPilotP and SIMCOPilotJ results, with HumanEval results. For all open-source models, their “instruct” versions are used. “ET” refers to the extended thinking version of the model. Unless otherwise specified, Claude 3.7 Sonnet with ET is used in the following experiments and is simply referred to as Claude 3.7 Sonnet.

infill and completion). The proposed SIMCOPilot benchmark enables a detailed breakdown of results, which is organized as follows:

Pass rate by closest comment distance. We compute the distance (number of lines) from the start of the infill/completion task to the closest comment that precedes it. We measure the overall median distance, and any task whose closest comment distance is less than the median is said to have a “short” distance; and any greater than the median is said to have a “long” distance. All results in the SIMCOPilot benchmark are presented with 95% confidence intervals. See Figure 5.

Pass rate by distance to the furthest referenced program object. We compute all named class/function/variable/etc. reference distances in terms of the number of lines of code, and group the references into terciles based on the distance. A programming task is said to have a “short” reference distance if its longest reference distance is in the smallest tercile, a “medium” if its longest distance is in the middle tercile, and a “long” distance if its longest is in the largest tercile. See Figure 4.

Pass rate by program constructs. We break down the results based on program constructs such as referring to a local or global variable, function, class etc. See Figure 2 and Figure 3.

5 Benchmark Results

Experimental Setup. We ran SIMCOPilot on a representative set of state-of-the-art large language models (LLMs) for code. We evaluate a range of both popular closed-source and state-of-the-art open-source language models on our proposed benchmark.

We assess two prominent closed-source model families: the GPT models, including GPT-4o (2024-08-06) ([OpenAI, 2024a](#)) and O3-mini ([OpenAI, 2025](#)), and Claude models from Anthropic, including Claude-3-Opus, Claude-3.5-Haiku, and Claude-3.7-Sonnet ([Anthropic, 2025b](#)). We also test Claude 3.7 Sonnet with extended thinking using a 16k token budget ([Anthropic, 2025a](#)). From open-source models, we examine the Llama family, including Llama-3.3-70B-Instruct ([AI, 2025b](#)) and Llama-3.1-8B-Instruct ([AI, 2025a](#)), DeepSeek models ([Guo et al., 2025](#)) including DeepSeek-R1 and DeepSeek-R1-Distill-Qwen-14B, and Qwen models including Qwen2.5-Coder-32B-Instruct ([Hui et al., 2024](#)) and Qwen-QwQ-32B ([Team, 2025](#)). All models operate with their respective intrinsic context lengths. Claude

family models and O3-mini use 200k tokens, while GPT-4o, Llama models, and DeepSeek-R1 employ 128k tokens. DeepSeek-R1-Distill-Qwen-14B and all Qwen-based models utilize 131k tokens. Table 2 gives the overall SIMCOPilot results for each of the twelve models. For comparison, this table also gives the published HumanEval (Chen et al., 2021) performance of each model.

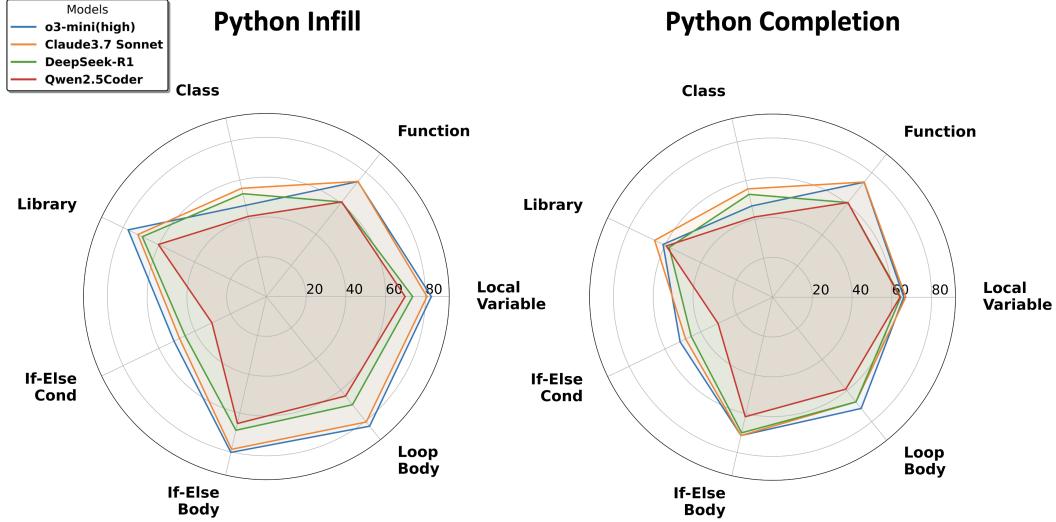


Figure 2: Pass rate group by Python Construct.

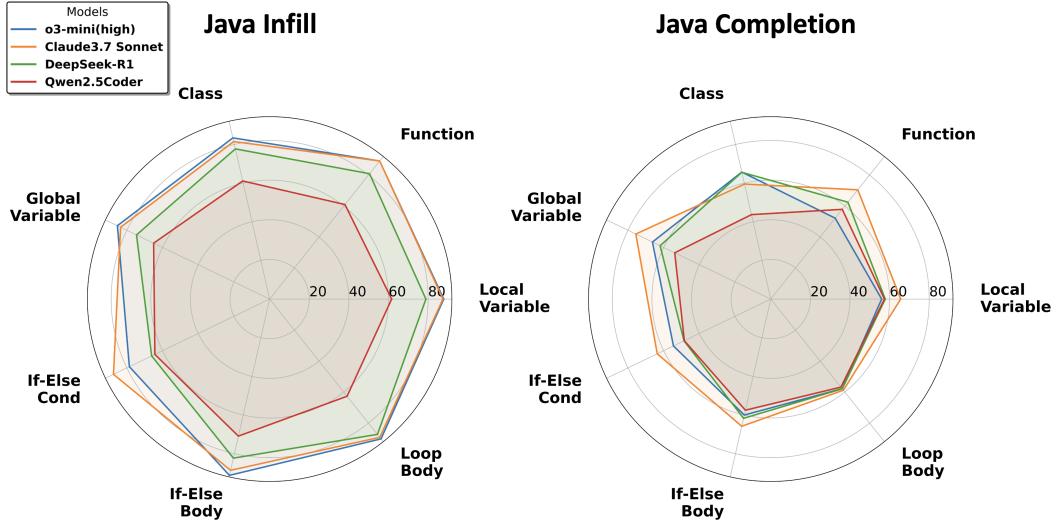


Figure 3: Pass rate grouped by Java Construct.

Model Performance Disparities. As mentioned in the introduction to the paper, one of the most striking findings is that while the various AI-for-code models did generally perform according to expectation in terms of which models scored highest (o3-mini and Claude 3.7 Sonnet) and lowest (Llama 3.1 8B), the gap between these models was very substantial according to SIMCOPilot. For example, o3-mini succeeded on 87.6% of Java infill tasks, whereas Llama 3.1 8B succeeded on only 36.4% of Java infill tasks—a gap of over 50 percentage points. At the same time, there was a relatively small gap between these models according to HumanEval. For instance, Claude 3.7 Sonnet (ET) scored 97.8% on HumanEval while even the lowest performer, Llama 3.1 8B, still achieved 72.6%—a much narrower gap of approximately 25 percentage points. There are several explanations for this discrepancy, but the most obvious is that at this point, many AI-for-code models have likely

been trained on HumanEval, and so solving those tasks is more of a recall problem than a genuine programming problem.

Context-Driven Advantage: Infill Outperforms Completion. Our results confirmed the expected difficulty gradient between task types. The largest models (O3-mini and Claude 3.7 Sonnet with extended thinking) consistently performed better on infill tasks (83.3 – 87.6% success rates) than on completion tasks (59.1 – 69.5%). This pattern makes intuitive sense: code following a missing block provides valuable contextual clues about the required implementation. Completion tasks present inherently greater challenges, requiring models to anticipate programmer intent without subsequent code as guidance. Despite this difficulty, leading models achieve respectable performance with clear instructions—particularly GPT-4o in Python (69.9%) and Claude 3.7 Sonnet in Java (69.5%). As Figures 2 and 3 reveal, while Java infill performance approaches perfection for the best models across most constructs, completion performance degrades in comparison. This performance difference highlights critical areas for improvement for future programming assistants.

Programming Construct Challenges. Examining Figure 2 and 3, several unexpected patterns emerge in model performance across different programming constructs. One surprising finding is the substantial gap between if-else conditions and bodies. Models consistently struggle with generating if-else conditions (Python: 30 – 52% pass rates) compared to if-else bodies (Python: 65 – 80%). This pattern persists in Java, though less extreme. This suggests that logical conditions require more precise reasoning with fewer explicit clues than implementation bodies, which benefit from clearer structural hints. A second unexpected result is Java models’ superior performance on global variable tasks (75 – 85%) compared to local variables (56 – 87%), contradicting conventional programming intuition that local scope is simpler. This may be because global variables provide clearer dependencies through explicit references, while local variables introduce complexity through nested scopes and implicit context.

Python class-related tasks also show surprisingly lower accuracy (41 – 55%) than function tasks (60 – 74%), despite both being common abstractions. This Python-specific difficulty likely stems from Python’s flexible structure providing less explicit context for class definitions, while Java’s verbosity offers clearer syntactic cues. Finally, loop body tasks consistently outperform conditional logic and class-related tasks across both languages (60 – 90% pass rates), which may arise from loops’ inherently repetitive structure providing stronger pattern recognition cues compared to open-ended logical reasoning tasks.

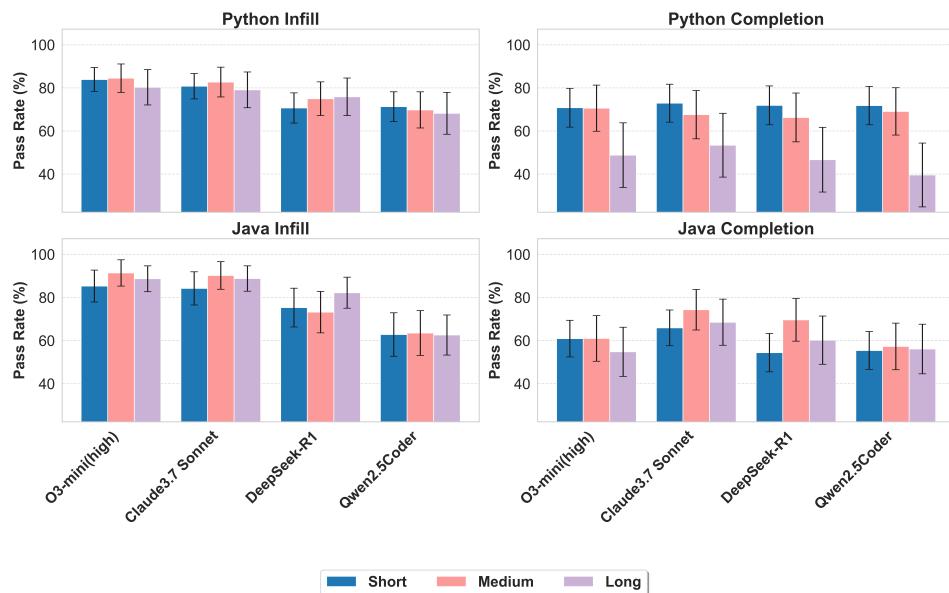


Figure 4: Pass rate grouped by reference object distance.

Reference Distance Effects. A revealing trend emerges when analyzing pass rates by reference-object distance (Figure 4). For Python completion tasks, performance declines as distance to referenced objects increases, aligning with lower pass rates for complex, long-distance constructs (e.g., class definitions) compared to short-distance tasks (e.g., loop bodies). Interestingly, Java completion shows the opposite pattern: longer-distance references (e.g., global variables, classes) often yield higher pass rates than closer-scoped items (e.g., local variables). This suggests Java’s explicit syntactic structure may help models navigate distant dependencies more effectively than Python’s more flexible approach.

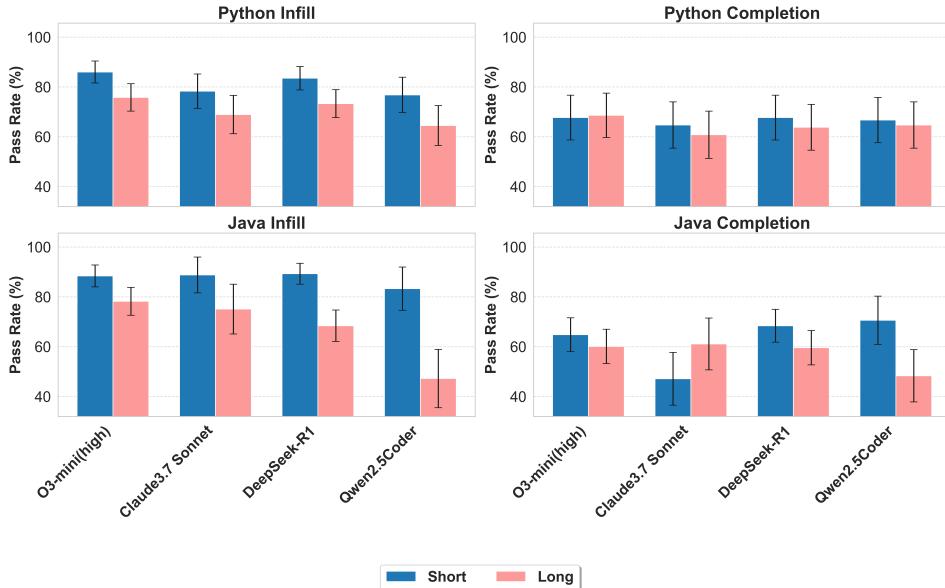


Figure 5: Pass rate grouped by distance to the nearest comment.

Critical Role of Comments. Existing benchmarks like MBPP and HumanEval feature comments immediately preceding target code blocks, artificially boosting generation performance and obscuring models’ true code-writing capabilities. In real-world development, such dense annotation is impractical and time-consuming. Our benchmark demonstrates that as the distance to the nearest comment increases, code generation quality consistently deteriorates (Figure 5), revealing limitations that standard benchmarks might mask.

6 Conclusion

In this paper, we have described SIMCOPilot, which is a benchmark designed to evaluate the ability of an AI-for-code model to perform well in a simulated, Copilot-style environment, where the goal is to supply a programmer with a small bit of missing code in the context of a complex programming project. SIMCOPilot considers both infill and completion tasks, and not only returns high-level accuracy numbers, but it also returns fine-grained benchmark results, stratified according to metrics such as the type of code to produce, the distance of the declaration of a program element (such as a global variable) to where it is used, and the extent to which comments are available locally. SIMCOPilot will be open-source, and we hope that over time, others will add high-quality, Java and Python infill tasks (not available elsewhere on the web) to SIMCOPilot, as well as model-specific pre-processing and post-processing codes to the project.

The most obvious limitation of SIMCOPilot as it is currently designed is that the numbers produced, while useful, may not accurately reflect user satisfaction with the output of an AI-for-code model. SIMCOPilot reports the fraction of codes produced that are correct, in the sense that they pass all test cases. However, a real-life user may not require correctness, and may often be willing to change a variable name (or two) in a produced code. Further,

SIMCOPilot exclusively evaluates the ability of an AI-for-code tool to complete an almost-completed, medium-to-large project. In practice, most projects will not be complete at the time that the code is added, and so there may not be a single correct answer—based on the output of a model (whatever it is), a user may be willing to tailor later code to match the output. It may be somewhat easier to produce *usable* code than what SIMCOPilot’s results would suggest. Still, despite these limitations, we believe that SIMCOPilot is a useful means of evaluating AI-for-code models.

References

- Meta AI. Introducing-llama-3-1-our-most-capable-models-to-date, 2025a. URL <https://ai.meta.com/blog/meta-llama-3-1/>.
- Meta AI. Llama-3-3, 2025b. URL https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/.
- Anthropic. Extended-thinking-models, 2025a. URL <https://docs.anthropic.com/en/docs/about-claude/models/extended-thinking-models#claude-3-7-overview>.
- Anthropic. All-models-overview, 2025b. URL <https://docs.anthropic.com/en/docs/about-claude/models/all-models#model-comparison-table>.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- Pierre Chambon, Baptiste Roziere, Benoit Sagot, and Gabriel Synnaeve. Bigo (bench)–can llms generate code with controlled time and space complexity? *arXiv preprint arXiv:2503.15242*, 2025.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Yangruibo Ding, Zijian Wang, Wasi Ahmad, Hantian Ding, Ming Tan, Nihal Jain, Murali Krishna Ramanathan, Ramesh Nallapati, Parminder Bhatia, Dan Roth, et al. Crosscodeeval: A diverse and multilingual benchmark for cross-file code completion. *Advances in Neural Information Processing Systems*, 36, 2024.
- Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. Classeval: A manually-crafted benchmark for evaluating llms on class-level code generation. *arXiv preprint arXiv:2308.01861*, 2023.
- GitHub and OpenAI. Github copilot. <https://github.com/features/copilot>, 2024. Accessed: 2024-08-01.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*, 2021.
- Wenhao Hu, Jinhao Duan, Chunchen Wei, Li Zhang, Yue Zhang, and Kaidi Xu. Dynacode: A dynamic complexity-aware code benchmark for evaluating large language models in code generation. *arXiv preprint arXiv:2503.10452*, 2025.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.
- Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. Mapping language to code in programmatic context. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1643–1652, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1192. URL <https://aclanthology.org/D18-1192>.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.
- Jia Li, Ge Li, Xuanming Zhang, Yihong Dong, and Zhi Jin. Evocodebench: An evolving code generation benchmark aligned with real-world code repositories. *arXiv preprint arXiv:2404.00599*, 2024.

Tianyang Liu, Canwen Xu, and Julian McAuley. Repobench: Benchmarking repository-level code auto-completion systems. *arXiv preprint arXiv:2306.03091*, 2023.

OpenAI. gpt-4o-2024-08-06, 2024a. URL <https://openai.com/index/gpt-4o-system-card/>.

OpenAI. Introducing swe-bench verified, 2024b. URL <https://openai.com/index/introducing-swe-bench-verified/>.

OpenAI. o3-mini-2025-01-31, 2025. URL <https://openai.com/index/openai-o3-mini/>.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL <https://qwenlm.github.io/blog/qwq-32b/>.

Lei Yang, Renren Jin, Ling Shi, Jianxiang Peng, Yue Chen, and Deyi Xiong. Probench: Benchmarking large language models in competitive programming. *arXiv preprint arXiv:2502.20868*, 2025.

Hao Yu, Bo Shen, Dezhi Ran, Jiaxin Zhang, Qi Zhang, Yuchi Ma, Guantai Liang, Ying Li, Qianxiang Wang, and Tao Xie. Codereval: A benchmark of pragmatic code generation with generative pre-trained models. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, pp. 1–12, 2024.

SimCopilot: Evaluating Large Language Models for Copilot-Style Code Generation

Appendix

A Details of SimCopilot

Dataset Documentation and Intended Uses. This dataset, compiled from eight Java and seven Python code repositories, is designed primarily for the evaluation of AI-for-code models. It can also be utilized responsibly for educational purposes, software engineering studies, and the development of new coding tools.

Data Collection. Emails were sent to faculty and students within the University’s Computer Science, Electrical Engineering, and Statistics departments, inviting them to contribute Java and Python private code repositories for AI-for-code research. Upon receipt, 1,163 code generation tasks were curated to ensure a diverse and representative sample of real-world code, gathering approximately 11,000 lines of code.

Code Block Extraction. For every sourced code repository, the following Algorithm was used to extract self-contained code blocks -

Algorithm 1 Selecting Logically Self-Contained Code Blocks

```
1: blocks = [] ▷ Initialize an empty list to store code blocks with starting and end points.  
2: ProgConstructList = [Loop, If-Else, Function]           ▷ Define constructs of interest.  
3: for file in files do  
4:   ast = parse(file)                                ▷ Parse the file using AST parser.  
5:   for node in ast do  
6:     if node.type in ProgConstructList then  
7:       DFS_traverse_nodes(node, blocks) ▷ Perform DFS to traverse the node body.  
8:     end if  
9:   end for  
10: end for  
11: return blocks                                     ▷ Return the list of code blocks with starting and end points.
```

Below, we further discuss the additional details of extracting/selecting the code blocks -

- **Nested Constructs:** To ensure logical self-containment, the selection of code blocks starts at the beginning of a construct (e.g., loop body) and ends at the end of the same construct, even if it includes nested constructs such as if statements within the loop. This approach maintains the logical integrity and completeness of the selected code block.
- **Completion vs. Infill Tasks:** The selection criteria differ between completion and infill tasks due to their distinct contexts. Infill tasks have both preceding and succeeding code, providing context that helps the model generate code that fits seamlessly into the existing sequence. In contrast, completion tasks only have preceding code and lack postceding context, making it essential to define clear stopping points. Without precise stopping points, ambiguity arises, leading to potential duplication or conflict when the generated code is combined with subsequent blocks. By focusing on the nearest incomplete construct, such as the end of a statement or function, the model can avoid extending beyond a well-defined boundary.

Human Annotation. To ensure high-quality code generation tasks, a team of graduate students from the University, each possessing 5 to 10 years of programming experience and ranging from medium to advanced skill levels, meticulously analyzed the candidate code blocks returned by Algorithm 1 for each source code. The generated coding tasks encompass both infill and completion types, distributed across eight distinct program categories. A detailed statistical breakdown of the tasks within each category is provided in Table 1. Annotators followed a shared rule-based protocol for block selection and task creation. In cases of disagreement, we resolved conflicts via majority vote and excluded tasks where consensus could not be reached. All selected blocks were reviewed to ensure structural integrity, syntactic correctness, and task relevance.

Hosting, Licensing, and Maintenance Plan.

- **Hosting Platform:** The dataset and its associated metadata, documented using the Croissant metadata framework, can be viewed and downloaded at <https://huggingface.co/datasets/mj33/SimCoPilot>. Code and associated files are also available through the following link: <https://github.com/mj33rice/Sim-CoPilot>.
- **Maintenance Plan:** We commit to maintaining the dataset with regular updates and revisions to correct any issues and integrate new contributions. Updates will be documented in the repository’s release notes section.
- **Licensing:** The data is shared under the [CC BY-NC-ND 4.0] and code is licensed under MIT License.

Legal and Ethical Responsibility. The authors of this dataset bear full responsibility for any violation of rights.

B Details of Post-Processing Steps

Post-processing is essential as even the best AI-for-code models can generate superfluous English comments, repeat lines, or entire blocks of code, often contrary to explicit instructions in the prompts. To address these issues, models are equipped with a specialized post-processor. The SIMCOPilot benchmark incorporates such a post-processor, which performs several crucial steps to refine the output:

1. **Remove Non-Code Syntax:** using a series of regular expressions to eliminate any text that is clearly not a statement in the target programming language.
2. **Remove Duplicate Code:** Removes duplicated code that appears above the prompt in completion tasks or following the prompt in infill tasks.
3. **Auto Indentation or Brackets Correction:** Adjusts indentation for Python and corrects bracket placement for Java, which are critical for maintaining the structural integrity of the code.

It is important to note that while post-processing adjusts the overall indentation and syntax to the entire generated code block, it preserves the relative indentations within code blocks as the way the model generates. This approach avoids altering the original programming intent, such as the termination points of loops or conditional blocks. For detailed examples of how these post-processing steps are applied in practice, refer to the step-by-step code illustrations for Python infill tasks and Java completion tasks in Figure 6 and Figure 7, respectively.

It is important to note that in our comparison labeled “without” post-processing, the post-processor was still utilized to eliminate non-code syntax, such as explanations or boilerplate text like “Here is the generated code:” which the best AI models frequently produce despite instructions to the contrary from the prompt. This step reflects realistic software development practices, where developers do not indiscriminately copy all text generated by AI models into their projects.

We compare the pass ratios with and without the application of comprehensive post-processing steps on SIMCOPilotJ and SIMCOPilotP. The data clearly shows that post-processing markedly boosts the pass rates—doubling it for larger models and increasing it tenfold for smaller models. Notably, the comparative analysis between Python and Java tasks reveals that post-processing yields a significantly greater improvement in pass rates on Java, at least doubling them, even with the most advanced AI models. This finding suggests that, in comparison to Python, Java is more sensitive to syntax and code structure, highlighting the importance of tailored post-processing techniques to enhance model performance in different programming environments.

Prompt	Generated Code
<pre> ... import tensorflow as tf ... def repurpose_skill_seq(args, skill_seq): """ Repurpose the skill sequence to be used for training the policy. = "o": no change = "1": wrap pick/grab/drop:obj_id to pick/grab/drop = "2": wrap pick:obj_id to pick/grab/drop:obj_id :param skill_seq: one-hot skill sequence of shape (n_trajs, horizon, c_dim) :return: tensor of shape (n_trajs, horizon, c_dim) and type same as skill_seq """ ... if args.wrap_level == "o": return skill_seq elif args.wrap_level == "1": # wrap by i = j % 3 where i is the new # position of skill originally at j. # Dim changes from c_dim to 3 ##### # Prompt ##### # Convert back to one-hot skill_seq = tf.one_hot(skill_seq, depth=3) return skill_seq elif args.wrap_level == "2": # wrap such that o/1/2 -> o, 3/4/5 -> 1, # 6/7/8 -> 2 ... # Dim changes from c_dim to # self.args.num_objs skill_seq = tf.argmax(skill_seq, axis=-1) skill_seq = skill_seq // 3 # Convert back to one-hot skill_seq = tf.one_hot(skill_seq, depth=args.num_objs) return skill_seq </pre>	<pre> elif args.wrap_level == "1": # Wrap by i = j % 3 where i is the new # position of skill originally at j. Dim # changes from c_dim to 3 skill_seq = tf.argmax(skill_seq, axis=-1) skill_seq = skill_seq % 3 # Convert back to one-hot skill_seq = tf.one_hot(skill_seq, depth=3) return skill_seq </pre>
	(1) Remove Non-code Syntax
	<pre> elif args.wrap_level == "1": skill_seq = tf.argmax(skill_seq, axis=-1) skill_seq = skill_seq % 3 skill_seq = tf.one_hot(skill_seq, depth=3) return skill_seq </pre>
	(2) Remove Duplicate Code
	<pre> skill_seq = tf.argmax(skill_seq, axis=-1) skill_seq = skill_seq % 3 </pre>
	(3) Auto-Indent Code
	<pre> skill_seq = tf.argmax(skill_seq, axis=-1) skill_seq = skill_seq % 3 </pre>
	Ground Truth
	<pre> skill_seq = tf.argmax(skill_seq, axis=-1) skill_seq = skill_seq % 3 </pre>

Figure 6: Example infill task from SIMCOPilotP and step-by-step post-processing demonstration

C Error Analysis

In Figure 8, we analyze the four primary error categories—*Compilation Error*, *Indentation Error*, *Syntax Error*, and *Var Hallucination*—across 1,163 code-generation tasks. We present these errors for each model as a percentage of total outcomes.

Compilation Error LLAMA-3.1 8B TURBO exhibits the highest compilation error rate at roughly 25%. By contrast, CLAUDE-3.7 SONNET-ET stands out with the lowest rate, around 8%. The remaining models cluster between these extremes, in the 10 – 23% range. While larger or more refined models still tend to fare better, no single model is completely immune to compilation failures, suggesting that some misunderstandings of code structure and language rules persist.

Indentation Error Because these models primarily generate Python code in our benchmark, indentation errors remain a distinct category. Overall, indentation errors stay comparatively low in the new results, topping out at about 6% for models like QWEN-QwQ-32B, while CLAUDE-3.7 SONNET-ET and GPT family rarely produce indentation issues (near 1.5%). This consistency suggests that the majority of systems have developed fairly robust handling of Python’s indentation rules.

Syntax Error Syntax errors follow a clear size-based trend, with larger or more advanced models displaying fewer syntactic mistakes. The worst cases (e.g. QWEN-QwQ-32B) approach 9.5% syntax error rates, whereas GPT-4O and O3-MINI(HIGH) lead at around 1%. These findings reinforce the notion that stronger language-model pretraining and instruction tuning correlate with fewer raw

Prompt	Generated Code
<pre> class PQueueABCD <PointInMetricSpace, DataType> { // this is an object in the queue private class Wrapper { DataWrapper <PointInMetricSpace, DataType> myData; double distance; } // this is the actual queue ArrayList <Wrapper> myList; // this is the largest distance value currently in the // queue double large; // this is the max number of objects int maxCount; // create a priority queue with the spced number of slots PQueueABCD (int maxCountin) { maxCount = maxCountin; myList = new ArrayList <Wrapper> (); large = 999; } // get the largest distance in the queue double getDist () { return large; } // add a new object to the queue... the insertion is // ignored if the distance value // exceeds the largest that is in the queue and the queue // is already full void insert (PointInMetricSpace key, DataType data, double distance) { // if we are full, remove the one with the largest // distance if (myList.size () == maxCount && distance < large) { int badIndex = 0; double maxDist = 0.0; for (int i = 0; i < myList.size (); i++) { /****** Prompt *****/ maxDist = myList.get (i).distance; badIndex = i; } ... } } </pre>	<pre> for (int i = 0; i < myList.size (); i++) { if (myList.get (i).distance > maxDist) { maxDist = myList.get (i).distance; badIndex = i; } } </pre>
	(1) Remove Non-code Syntax
	<pre> for (int i = 0; i < myList.size (); i++) { if (myList.get (i).distance > maxDist) { maxDist = myList.get (i).distance; badIndex = i; } } </pre>
	(2) Remove Duplicate Code
	<pre> if (myList.get (i).distance > maxDist) { } </pre>
	(3) Remove Bracket Discrepancies
	<pre> if (myList.get (i).distance > maxDist) { } </pre>
	Ground Truth
	<pre> if (myList.get (i).distance > maxDist) { } </pre>

Figure 7: Example completion task from SIMCOPilotJ and step-by-step post-processing demonstration

syntax violations.

Var Hallucination Variable Hallucination primarily arises from using undefined identifiers or mishandling scope. The best models (e.g. O3-MINI(HIGH) and CLAUDE-3.7 SONNET-ET) achieve hallucination rates of about 1%, whereas smaller or less consistent models can climb to around 4%. This error type remains a strong indicator of how well a model tracks variable definitions and maintains context over longer sequences, a capacity only loosely correlated with model size.

General Observations. Across the board, syntax and compilation errors remain the two most prevalent error types. These patterns suggest that, while code-generation models have substantially improved, structural misunderstandings and incomplete coverage of language rules continue to pose significant challenges.

D Additional Evaluation Results and Ablations

Beyond the evaluations of O3-mini, Claude 3.7 Sonnet with extended thinking, DeepSeek-R1 671B, Qwen2.5 Coder-32B presented in Table 2, this section details results for eight additional models: Claude 3.7 Sonnet, Claude 3.5 Haiku, Llama 3.3 70B, Llama 3.1 8B, DeepSeek-R1-Distill-Qwen-14B, Qwen-QwQ-32B, and GPT-4o. These models were assessed using the same experimental setups described in the main paper, which include features such as distance to the nearest referenced object, proximity to the nearest comment, and various task types. These tasks incorporate different programming constructs, including references to local variables, functions, and loop bodies, allowing for a comprehensive analysis of model performance across diverse coding scenarios.

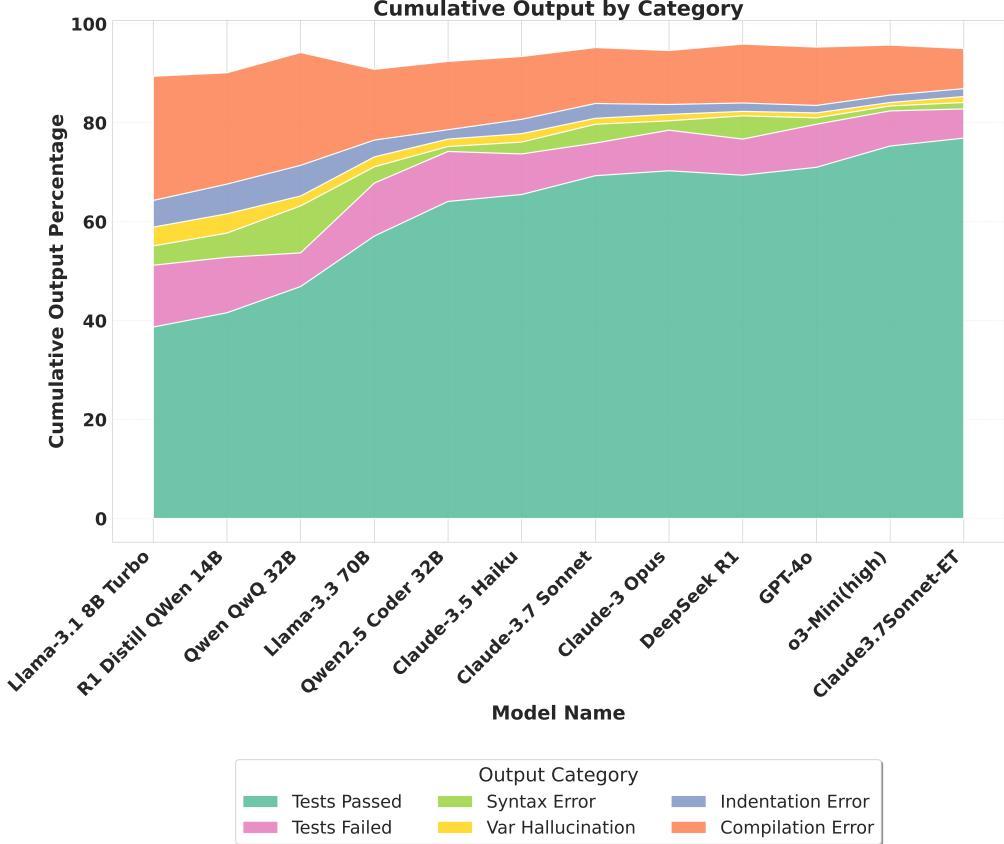


Figure 8: Cumulative Output by Category breakdown for each model among 1,163 programming tasks in SIMCOPilot.

Pass rate by distance to the furthest referenced program object.

As observed with the four models described in the main paper, the influence of the distance to referenced methods, functions, and variables on the pass rate of the generated code was not consistently significant. According to Tables 3, it is challenging to discern any consistent pattern regarding how program distance affects accuracy. For Python completion tasks, greater distances tended to correlate with increased difficulty, which aligns with expectations. Interestingly, for Java infill tasks, greater distances appeared to actually simplify the task. However, despite these general trends, there are one or two models, as outlined in the tables, that deviate from these patterns.

Pass rate by closest comment distance.

We filtered out tasks with no comment above and ended with 207 tasks for Python completion, 374 tasks for Python infilling, 278 tasks for Java completion, and 278 tasks for Java infilling. It can be observed that comments located near the target code segments were distinctly beneficial. As illustrated in Table 4, having comments close to the areas of code to be completed, markedly improved the pass rate of both infill and completion tasks.

Pass rate by task type We further provide pass rates stratified by the task type for the remainder of models in Tables 5, 6, 7.

	Distance	O3-mini	C3.7 Sonnet*	DeepSeek-R1	Qwen2.5Coder
Infill Python	Short	83.9 ± 5.6	80.8 ± 5.9	70.7 ± 7.0	71.3 ± 6.9
	Med	84.5 ± 6.6	82.7 ± 6.9	75.0 ± 7.8	69.8 ± 8.4
	Long	80.3 ± 8.2	79.1 ± 8.3	75.9 ± 8.7	68.2 ± 9.7
Compl. Python	Short	70.9 ± 9.1	72.9 ± 8.8	71.9 ± 9.0	71.8 ± 8.9
	Med	70.6 ± 10.9	67.6 ± 11.2	66.3 ± 11.3	69.1 ± 11.0
	Long	49.0 ± 14.9	53.4 ± 14.8	46.7 ± 15.0	39.6 ± 14.8
Infill Java	Short	85.3 ± 7.4	84.2 ± 7.7	75.3 ± 9.0	62.8 ± 10.1
	Med	91.4 ± 6.1	90.2 ± 6.4	73.2 ± 9.6	63.5 ± 10.4
	Long	88.7 ± 6.0	88.8 ± 5.9	82.2 ± 7.2	62.6 ± 9.3
Compl. Java	Short	60.9 ± 8.5	65.9 ± 8.3	54.4 ± 8.9	55.4 ± 8.8
	Med	61.0 ± 10.6	74.3 ± 9.4	69.6 ± 9.9	57.3 ± 10.8
	Long	54.8 ± 11.4	68.5 ± 10.7	60.2 ± 11.2	56.1 ± 11.5
GPT-4o		C3 Opus	C3.7 Sonnet	C3.5 Haiku	
Infill Python	Short	80.8 ± 6.0	77.8 ± 6.3	74.2 ± 6.7	71.8 ± 6.9
	Med	76.7 ± 7.7	74.1 ± 8.0	73.3 ± 8.1	75.0 ± 7.8
	Long	75.8 ± 8.8	71.4 ± 9.3	75.9 ± 8.9	64.9 ± 9.6
Compl. Python	Short	77.1 ± 8.3	72.8 ± 8.9	66.6 ± 9.6	69.7 ± 9.2
	Med	69.1 ± 11.1	67.7 ± 11.1	58.8 ± 11.7	60.3 ± 11.6
	Long	53.5 ± 15.1	53.6 ± 15.0	32.6 ± 13.9	37.4 ± 14.3
Infill Java	Short	78.6 ± 8.6	67.4 ± 9.8	64.1 ± 10.0	74.1 ± 9.1
	Med	76.8 ± 9.1	66.9 ± 10.2	78.1 ± 9.1	69.5 ± 10.1
	Long	77.5 ± 7.9	67.3 ± 8.8	72.9 ± 8.4	59.8 ± 9.4
Compl. Java	Short	53.6 ± 8.9	72.3 ± 7.8	67.4 ± 8.3	63.4 ± 8.5
	Med	59.8 ± 10.6	69.3 ± 10.0	75.7 ± 9.3	70.8 ± 9.9
	Long	50.7 ± 11.4	63.0 ± 11.1	67.1 ± 10.8	45.2 ± 11.4
Llama3.3 70B		Qwen-QwQ-32B	R1-Qwen 14B	Llama3.1 8B	
Infill Python	Short	63.5 ± 7.3	52.7 ± 7.6	47.8 ± 7.5	50.3 ± 7.5
	Med	50.0 ± 9.2	51.7 ± 9.1	45.8 ± 9.0	33.7 ± 8.6
	Long	60.5 ± 10.0	54.9 ± 10.3	44.0 ± 10.1	44.0 ± 10.2
Compl. Python	Short	58.4 ± 10.0	57.3 ± 10.0	40.7 ± 9.8	43.8 ± 10.0
	Med	57.3 ± 11.7	42.7 ± 11.7	48.6 ± 12.0	41.1 ± 11.8
	Long	37.2 ± 14.5	30.2 ± 13.8	16.3 ± 11.0	23.2 ± 12.5
Infill Java	Short	60.6 ± 10.1	51.7 ± 10.5	42.7 ± 10.3	40.4 ± 10.1
	Med	65.8 ± 10.3	45.0 ± 10.8	41.4 ± 10.6	31.7 ± 10.1
	Long	69.1 ± 8.8	57.1 ± 9.5	32.7 ± 8.9	38.2 ± 9.2
Compl. Java	Short	44.7 ± 8.9	30.9 ± 8.2	42.3 ± 8.7	34.1 ± 8.5
	Med	54.8 ± 10.8	36.6 ± 10.4	47.6 ± 10.8	31.7 ± 9.9
	Long	50.6 ± 11.7	36.9 ± 11.1	23.2 ± 9.6	27.4 ± 10.3

Table 3: Pass rate stratified by distance to the referenced object for completion and infill tasks categorized by terciles. For Python, distance boundaries for short, medium, and long are (10, 30); for Java, they are (30, 100). *C3.7 Sonnet refers to Claude 3.7 Sonnet with extended thinking; C3.7 Sonnet without asterisk refers to the version without extended thinking; R1-Qwen 14B refers to R1-Distill-Qwen-14B.

	Distance	O3-mini	C3.7 Sonnet*	DeepSeek-R1	Qwen2.5Coder
Infill Python	Short	86.0 ± 4.4	83.5 ± 4.7	75.8 ± 5.5	73.3 ± 5.6
	Long	78.3 ± 6.9	76.8 ± 7.1	68.9 ± 7.7	64.5 ± 8.0
Compl. Python	Short	67.7 ± 9.0	67.7 ± 9.0	68.6 ± 8.9	63.8 ± 9.2
	Long	64.7 ± 9.3	66.7 ± 9.1	60.8 ± 9.5	64.7 ± 9.3
Infill Java	Short	88.4 ± 4.4	89.3 ± 4.2	78.2 ± 5.6	68.4 ± 6.3
	Long	88.8 ± 7.2	83.3 ± 8.7	75.1 ± 10.0	47.2 ± 11.7
Compl. Java	Short	64.8 ± 6.8	68.4 ± 6.6	60.1 ± 6.9	59.6 ± 6.9
	Long	47.1 ± 10.6	70.6 ± 9.7	61.1 ± 10.4	48.3 ± 10.5
		GPT-4o	C3 Opus	C3.7 Sonnet	C3.5 Haiku
Infill Python	Short	78.8 ± 5.2	75.5 ± 5.5	75.9 ± 5.5	75.0 ± 5.5
	Long	77.6 ± 7.0	74.7 ± 7.3	71.8 ± 7.6	64.5 ± 8.0
Compl. Python	Short	72.4 ± 8.6	66.6 ± 9.0	57.2 ± 9.4	59.0 ± 9.3
	Long	66.6 ± 9.2	67.6 ± 9.2	57.0 ± 9.5	60.8 ± 9.5
Infill Java	Short	83.6 ± 5.1	73.3 ± 6.1	75.2 ± 5.9	69.9 ± 6.2
	Long	61.0 ± 11.1	49.9 ± 11.6	61.1 ± 11.4	59.7 ± 11.2
Compl. Java	Short	65.2 ± 6.7	68.4 ± 6.6	71.6 ± 6.4	64.2 ± 6.8
	Long	30.6 ± 9.8	70.6 ± 9.7	66.0 ± 10.1	53.0 ± 10.5
		Llama3.3 70B	Qwen-QwQ-32B	R1-Qwen 14B	Llama3.1 8B
Infill Python	Short	70.4 ± 5.9	61.1 ± 6.2	52.2 ± 6.4	45.3 ± 6.5
	Long	38.4 ± 8.2	39.1 ± 8.2	35.5 ± 8.0	40.6 ± 8.2
Compl. Python	Short	55.2 ± 9.7	48.6 ± 8.9	39.0 ± 9.4	40.0 ± 9.3
	Long	52.0 ± 9.8	45.1 ± 9.5	37.2 ± 9.4	37.4 ± 9.3
Infill Java	Short	69.9 ± 6.3	56.3 ± 6.8	42.7 ± 6.8	42.7 ± 6.8
	Long	52.8 ± 11.4	38.9 ± 11.3	26.3 ± 10.2	20.9 ± 9.4
Compl. Java	Short	58.0 ± 6.9	40.4 ± 7.0	49.2 ± 7.1	37.3 ± 6.7
	Long	29.4 ± 9.7	19.9 ± 8.6	15.2 ± 7.7	18.8 ± 8.4

Table 4: Pass rate stratified by distance to nearest comment for completion and infill tasks. C3.7 Sonnet* refers to Claude 3.7 Sonnet with extended thinking; C3.7 Sonnet without asterisk refers to the version without extended thinking; R1-Qwen 14B refers to R1-Distill-Qwen-14B.

Task	Model	Python Infill	Python Compl.	Java Infill	Java Compl.
Local Variable	O3-mini	83.0 ± 3.8	66.0 ± 6.5	87.9 ± 4.2	56.0 ± 6.6
	C3.7 Sonnet*	80.8 ± 4.0	67.0 ± 6.4	87.5 ± 4.3	65.6 ± 6.3
	DeepSeek-R1	73.6 ± 4.5	64.5 ± 6.5	78.9 ± 5.2	57.8 ± 6.7
	Qwen2.5Coder	69.8 ± 4.8	64.1 ± 6.5	61.6 ± 6.3	57.4 ± 6.5
Global Variable	O3-mini	N/A		85.3 ± 5.5	66.3 ± 6.9
	C3.7 Sonnet*	N/A		83.5 ± 5.7	75.6 ± 6.2
	DeepSeek-R1	N/A		74.6 ± 6.8	62.0 ± 7.0
	Qwen2.5Coder	N/A		65.0 ± 7.5	53.8 ± 7.2
Function	O3-mini	73.9 ± 12.7		89.1 ± 5.8	52.2 ± 10.4
	C3.7 Sonnet*	74.0 ± 12.5		89.1 ± 5.8	70.5 ± 9.5
	DeepSeek-R1	60.8 ± 14.2		80.9 ± 7.3	62.5 ± 10.1
	Qwen2.5Coder	60.8 ± 14.2		61.0 ± 9.1	57.9 ± 10.2
Class	O3-mini	47.0 ± 16.9		83.4 ± 10.0	65.7 ± 16.6
	C3.7 Sonnet*	55.8 ± 16.7		81.5 ± 10.3	59.5 ± 16.8
	DeepSeek-R1	53.0 ± 16.8		77.7 ± 11.1	65.6 ± 16.6
	Qwen2.5Coder	41.3 ± 16.7		61.1 ± 12.9	43.7 ± 17.0
Library	O3-mini	77.2 ± 5.8	61.1 ± 8.7	N/A	
	C3.7 Sonnet*	71.8 ± 6.2	65.8 ± 8.5	N/A	
	DeepSeek-R1	69.3 ± 6.4	57.7 ± 8.8	N/A	
	Qwen2.5Coder	60.3 ± 6.8	59.3 ± 8.6	N/A	
If-Else Condition	O3-mini	51.6 ± 17.3		78.6 ± 10.9	54.5 ± 17.0
	C3.7 Sonnet	48.5 ± 16.9		87.5 ± 8.7	63.6 ± 16.8
	DeepSeek-R1	45.4 ± 17.2		66.1 ± 12.4	48.6 ± 16.6
	Qwen2.5Coder	30.4 ± 15.7		64.3 ± 12.6	48.4 ± 16.9
If-Else Body	O3-mini	80.3 ± 10.0	71.3 ± 10.4	91.2 ± 6.3	60.0 ± 8.7
	C3.7 Sonnet*	78.7 ± 10.3	71.3 ± 10.3	88.6 ± 7.0	65.8 ± 8.6
	DeepSeek-R1	68.9 ± 11.6	69.9 ± 10.5	82.3 ± 8.5	61.7 ± 8.6
	Qwen2.5Coder	65.4 ± 11.8	61.6 ± 11.1	70.9 ± 10.0	57.5 ± 8.8
Loop Body	O3-mini	83.3 ± 8.6	71.5 ± 12.6	90.2 ± 5.8	57.8 ± 9.6
	C3.7 Sonnet*	80.6 ± 9.1	67.3 ± 13.0	89.2 ± 6.0	58.8 ± 9.5
	DeepSeek-R1	69.5 ± 10.6	67.3 ± 13.1	87.3 ± 6.5	57.9 ± 9.6
	Qwen2.5Coder	63.9 ± 11.3	59.1 ± 13.9	62.7 ± 9.4	56.8 ± 9.6

Table 5: Pass rates stratified by task type. Note that some results marked N/A are not presented due to insufficient data or overly-large confidence intervals. C3.7 Sonnet* refers to Claude 3.7 Sonnet with extended thinking with 16K token budget, and Qwen2.5Coder-32B is abbreviated as Qwen2.5Coder.

Task	Model	Python Infill	Python Compl.	Java Infill	Java Compl.
Local Variable	GPT-4o	78.2 ± 4.2	69.4 ± 6.3	77.2 ± 5.3	57.8 ± 6.5
	C ₃ Opus	75.5 ± 4.4	66.9 ± 6.4	65.9 ± 6.1	66.5 ± 6.3
	C3.7 Sonnet	74.1 ± 4.4	56.8 ± 6.8	71.1 ± 5.9	68.4 ± 6.2
	C3.5 Haiku	70.9 ± 4.6	60.2 ± 6.7	65.5 ± 6.1	55.5 ± 6.6
Global Variable	GPT-4o	N/A		79.0 ± 6.4	55.0 ± 7.2
	C ₃ Opus	N/A		66.3 ± 7.4	69.5 ± 6.7
	C3.7 Sonnet	N/A		73.9 ± 6.8	71.2 ± 6.5
	C3.5 Haiku	N/A		67.6 ± 7.3	67.9 ± 6.7
Function	GPT-4o	69.6 ± 13.3		73.6 ± 8.2	47.7 ± 10.4
	C ₃ Opus	65.2 ± 13.8		62.8 ± 9.0	64.4 ± 10.0
	C3.7 Sonnet	65.2 ± 13.8		73.6 ± 8.2	67.7 ± 9.8
	C3.5 Haiku	65.2 ± 13.8		61.9 ± 9.0	48.9 ± 10.5
Class	GPT-4o	55.9 ± 16.7		79.7 ± 10.7	53.0 ± 17.3
	C ₃ Opus	58.8 ± 16.5		61.0 ± 13.0	49.9 ± 17.4
	C3.7 Sonnet	41.2 ± 16.5		66.6 ± 12.7	62.5 ± 16.5
	C3.5 Haiku	41.2 ± 16.5		49.9 ± 13.5	46.9 ± 17.2
Library	GPT-4o	71.3 ± 6.2	65.0 ± 8.4	N/A	
	C ₃ Opus	67.8 ± 6.4	63.4 ± 8.5	N/A	
	C3.7 Sonnet	66.8 ± 6.6	52.9 ± 8.7	N/A	
	C3.5 Haiku	63.4 ± 6.6	56.1 ± 8.7	N/A	
If-Else Condition	GPT-4o	54.5 ± 17.0		82.1 ± 10.0	60.7 ± 16.7
	C ₃ Opus	36.4 ± 16.4		73.2 ± 11.7	63.6 ± 16.4
	C3.7 Sonnet	51.5 ± 17.1		78.6 ± 10.9	69.8 ± 15.7
	C3.5 Haiku	30.3 ± 15.7		76.7 ± 11.1	57.5 ± 17.0
If-Else Body	GPT-4o	81.9 ± 9.7	71.2 ± 10.5	79.7 ± 8.9	65.0 ± 8.6
	C ₃ Opus	62.3 ± 12.4	64.3 ± 10.9	72.2 ± 10.0	65.0 ± 8.5
	C3.7 Sonnet	78.7 ± 10.3	58.9 ± 11.2	74.7 ± 9.5	72.5 ± 7.9
	C3.5 Haiku	60.7 ± 12.3	53.4 ± 11.5	74.7 ± 9.7	64.2 ± 8.4
Loop Body	GPT-4o	77.8 ± 9.7	71.5 ± 12.8	80.4 ± 7.7	61.8 ± 9.4
	C ₃ Opus	66.7 ± 10.9	67.3 ± 13.0	63.7 ± 9.3	61.7 ± 9.4
	C3.7 Sonnet	76.4 ± 9.8	67.3 ± 13.1	75.5 ± 8.3	63.7 ± 9.4
	C3.5 Haiku	73.6 ± 10.2	59.1 ± 13.9	68.6 ± 9.0	60.8 ± 9.5

Table 6: Pass rates stratified by task type. For Java, both Infill and Completion are not applicable for the Library task. C3.7 Sonnet refers to Claude 3.7 Sonnet without extended thinking.

Task	Model	Python Infill	Python Compl.	Java Infill	Java Compl.
Local Variable	Llama3.3 70B	41.8 ± 5.0	49.1 ± 6.9	42.7 ± 6.4	43.6 ± 6.6
	Qwen-QwQ-32B	52.8 ± 5.1	46.6 ± 6.7	51.7 ± 6.4	34.9 ± 6.4
	R1-Qwen 14B	46.4 ± 5.1	38.3 ± 6.7	38.4 ± 6.3	39.4 ± 6.6
	Llama3.1 8B	43.1 ± 5.0	38.9 ± 6.7	34.9 ± 6.2	30.8 ± 6.1
Global Variable	Llama3.3 70B	N/A		37.6 ± 7.6	48.3 ± 7.1
	Qwen-QwQ-32B	N/A		50.4 ± 7.8	36.9 ± 7.0
	R1-Qwen 14B	N/A		36.9 ± 7.6	43.0 ± 7.2
	Llama3.1 8B	N/A		33.8 ± 7.4	32.1 ± 6.8
Function	Llama3.3 70B	58.7 ± 14.2		53.6 ± 9.4	42.1 ± 10.4
	Qwen-QwQ-32B	43.5 ± 14.3		50.9 ± 9.3	31.8 ± 9.7
	R1-Qwen 14B	41.3 ± 14.2		39.0 ± 9.1	30.6 ± 9.6
	Llama3.1 8B	39.1 ± 14.1		31.0 ± 8.7	26.2 ± 9.2
Class	Llama3.3 70B	11.8 ± 10.8		59.4 ± 13.1	43.7 ± 17.2
	Qwen-QwQ-32B	29.4 ± 15.3		57.5 ± 13.3	43.8 ± 17.1
	R1-Qwen 14B	14.7 ± 11.9		24.0 ± 11.4	31.3 ± 16.0
	Llama3.1 8B	23.5 ± 14.3		38.9 ± 13.0	28.0 ± 15.7
Library	Llama3.3 70B	35.6 ± 6.5	38.2 ± 8.7		N/A
	Qwen-QwQ-32B	47.5 ± 6.8	42.3 ± 8.8		N/A
	R1-Qwen 14B	36.7 ± 6.7	29.3 ± 8.0		N/A
	Llama3.1 8B	38.1 ± 6.7	29.3 ± 8.0		N/A
If-Else Condition	Llama3.3 70B	30.3 ± 15.7		50.0 ± 13.0	45.6 ± 17.0
	Qwen-QwQ-32B	18.2 ± 13.2		39.3 ± 12.7	30.2 ± 15.9
	R1-Qwen 14B	27.3 ± 15.2		37.4 ± 12.7	42.4 ± 16.7
	Llama3.1 8B	12.1 ± 11.1		42.9 ± 13.0	33.2 ± 15.9
If-Else Body	Llama3.3 70B	32.8 ± 12.0	52.0 ± 11.5	59.5 ± 10.8	48.4 ± 8.8
	Qwen-QwQ-32B	39.4 ± 12.4	43.8 ± 11.4	55.7 ± 10.9	42.4 ± 8.8
	R1-Qwen 14B	41.0 ± 12.4	37.0 ± 11.1	44.2 ± 11.1	47.5 ± 9.0
	Llama3.1 8B	32.9 ± 11.8	43.8 ± 11.4	38.0 ± 10.7	40.0 ± 8.8
Loop Body	Llama3.3 70B	50.0 ± 11.7	38.8 ± 13.6	51.9 ± 9.6	46.0 ± 9.6
	Qwen-QwQ-32B	51.4 ± 11.6	36.7 ± 13.4	55.9 ± 9.6	39.2 ± 9.5
	R1-Qwen 14B	37.5 ± 11.3	28.6 ± 12.8	45.1 ± 9.6	42.1 ± 9.6
	Llama3.1 8B	25.0 ± 10.1	28.5 ± 12.8	43.1 ± 9.5	34.3 ± 9.2

Table 7: Pass rates stratified by task type. For Java, both Infill and Completion are not applicable for the Library task. R1-Qwen 14B refers to R1-Distill-Qwen-14B.