

# TimeMachine-bench: A Benchmark for Evaluating Model Capabilities in Repository-Level Migration Tasks

Ryo Fujii<sup>1,2</sup> Makoto Morishita<sup>2,1</sup> Kazuki Yano<sup>1</sup> Jun Suzuki<sup>1,3,4</sup>

<sup>1</sup>Tohoku University <sup>2</sup>Future Corporation <sup>3</sup>RIKEN <sup>4</sup>NII LLMC  
is-failab-research@grp.tohoku.ac.jp

## Abstract

With the advancement of automated software engineering, research focus is increasingly shifting toward practical tasks reflecting the day-to-day work of software engineers. Among these tasks, software migration, a critical process of adapting code to evolving environments, has been largely overlooked. In this study, we introduce **TimeMachine-bench**, a benchmark designed to evaluate software migration in real-world Python projects. Our benchmark consists of GitHub repositories whose tests begin to fail in response to dependency updates. The construction process is fully automated, enabling live updates of the benchmark. Furthermore, we curated a human-verified subset to ensure problem solvability. We evaluated agent-based baselines built on top of 11 models, including both strong open-weight and state-of-the-art LLMs on this verified subset. Our results indicated that, while LLMs show some promise for migration tasks, they continue to face substantial reliability challenges, including spurious solutions that exploit low test coverage and unnecessary edits stemming from suboptimal tool-use strategies. Our dataset and implementation are available at <https://github.com/tohoku-nlp/timemachine-bench>.

## 1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities that extend beyond classical NLP, including the field of software engineering (Jiang et al., 2025). Their evolution has been rapid, initially from generating isolated code snippets to now acting as daily development partners like GitHub Copilot<sup>1</sup> and Claude Code<sup>2</sup>. These tools are capable of understanding entire complex codebases and interacting with the environment to fix bugs or implement new features based on natural language instructions.

<sup>1</sup><https://github.com/features/copilot>

<sup>2</sup><https://claude.com/product/claude-code>

This notable progress has been fueled by a virtuous cycle of continuous model development and novel benchmark creation. Starting with simple, function-level completion tasks such as HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021), the research focus has now shifted to more complex, real-world challenges. Currently, SWE-bench (Jimenez et al., 2024) serves as the de-facto standard for evaluation because of its emphasis on repository-level bug fixing and feature addition. Its design closely reflects the daily responsibilities of software engineers, making it highly valuable for evaluating the practical utility of LLMs.

However, most of the existing benchmarks, including SWE-bench, have a key limitation in that they treat software engineering as a static activity, assuming that the environment remains unchanged over time. In reality, software engineering is an inherently dynamic and evolving activity. For instance, a piece of code that once worked correctly might fail or produce errors unexpectedly due to some breaking changes in its dependencies. The task of addressing this constant evolution is referred to as *software migration*. Software migration is a crucial step in maintaining software reliability, as it addresses end-of-life issues and security vulnerabilities. Nonetheless, previous studies have seldom focused on complex, real-world migration tasks involving multi-step explorations.

In this paper, we introduce **TimeMachine-bench**, a benchmark designed to evaluate model capabilities for migration tasks in real-world Python projects (Figure 1). Our benchmark consists of GitHub repositories whose tests fail in response to dependency updates. The construction process is fully automated, enabling live updates of the benchmark. This *live* property of the benchmark is essential to mitigate the prevalent issue of data contamination in the era of LLMs (Brown et al., 2020). The benchmark covers projects with arbitrary dependencies, unlike previous works that fo-

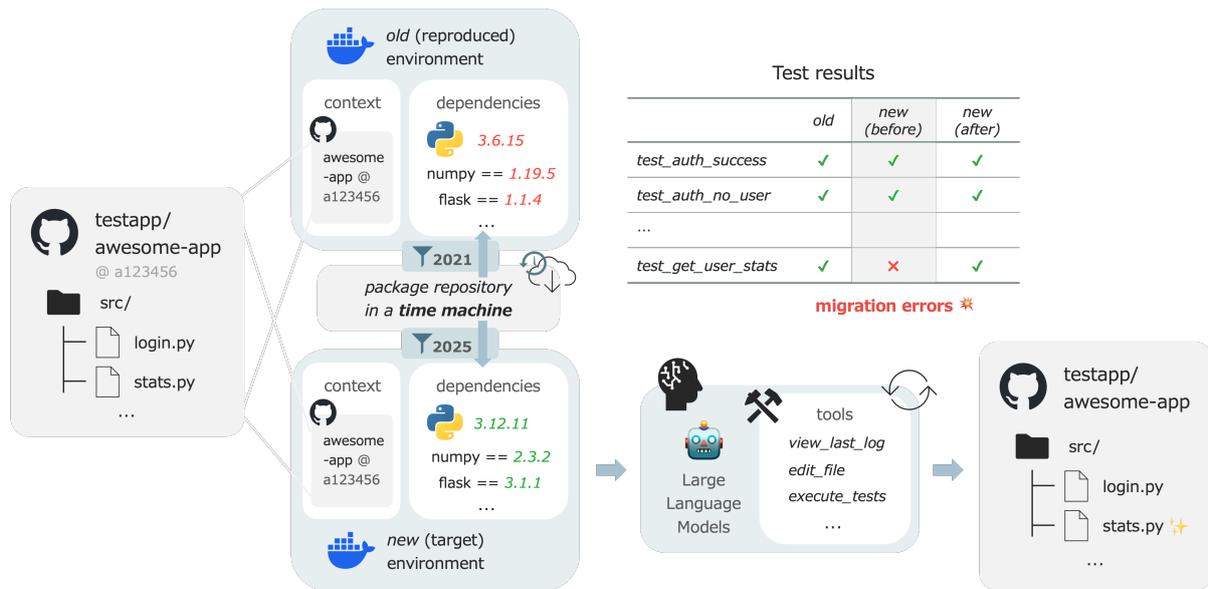


Figure 1: Overview of TimeMachine-bench. Using date-based environment control, our framework enables strict reproduction of two distinct environments corresponding to specific points in time.

cused on a limited set of popular libraries. We also present **TimeMachine-bench-Verified**, a human-validated subset of TimeMachine-bench, to ensure problem solvability. We conducted experiments on TimeMachine-bench-Verified using 11 models, including strong open-weight and state-of-the-art LLMs to evaluate the current capabilities of LLMs on migration tasks. To the best of our knowledge, this is the first framework that enables scalable and continuous evaluation of LLMs on migration tasks.

Our contributions are summarized as follows:

- We present **TimeMachine-bench**, a benchmark designed to evaluate success in migration tasks across real-world Python projects with arbitrary dependencies.
- We propose an automated construction pipeline to provide live nature to the benchmark, and curate **TimeMachine-bench-Verified**, a more reliable, validated subset with guaranteed solvability.
- We evaluate 11 LLMs to assess current capabilities and identify key reliability issues, including unnecessary edits and a tendency to prioritize test success over semantic correctness, thereby highlighting critical challenges for future work.

## 2 Related Work

**LLMs for Code.** Large language models have extended their applicability beyond natural lan-

guage. In the domain of software engineering, they are now applied to a wide range of tasks such as code completion (Chen et al., 2021; Zhang et al., 2023; Chen et al., 2024), code translation (Wang et al., 2021), and code review (Li et al., 2022). The advent of powerful domain-specific models, such as StarCoder (Lozhkov et al., 2024), DeepSeek-Coder (DeepSeek-AI, 2024a), and Qwen-Coder (Hui et al., 2024), alongside sophisticated frameworks for autonomous coding agents like Devin<sup>3</sup> and OpenHands (Wang et al., 2025b), is reshaping the role of software engineers: from coders to supervisors who effectively guide these models and validate their final outputs (Mozannar et al., 2024).

### Expanding Scope of Code LLM Benchmarks.

The emergence of new benchmarks has been a key driver of technological innovations. In the early stages, function-level code completion tasks like HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), and APPS (Hendrycks et al., 2021) were introduced, presenting substantial challenges even for state-of-the-art models at the time, including GPT-3 (Brown et al., 2020) and Codex (Chen et al., 2021). However, these benchmarks were biased toward algorithmic challenges and confined to self-contained problems.

To address these limitations, CoderEval (Yu et al., 2024) and DevEval (Li et al., 2024) introduced more practical code completion tasks based

<sup>3</sup><https://devin.ai>

on real-world GitHub repositories, which require an understanding of cross-file dependencies. Furthermore, another stream of research has focused on extending evaluation beyond functional correctness to other practical attributes, such as efficiency, maintainability, and security (Zheng et al., 2024; Bai et al., 2024).

In recent years, as models and methods have progressed rapidly, benchmarks have increasingly emphasized complex tasks reflecting the day-to-day responsibilities of software engineers. SWE-bench (Jimenez et al., 2024) stands as a prime example, evaluating a model’s capability to resolve real GitHub issues within a full repository context.

However, the growing reliance on benchmarks for performance evaluation has introduced new methodological concerns. One of the most important issues is data contamination, which refers to the leakage of evaluation data into training corpora (Brown et al., 2020). To counteract this, recent efforts such as LiveCodeBench (Jain et al., 2025) and SWE-bench-Live (Zhang et al., 2025) provide live updates of the evaluation sets to ensure the integrity and fairness of the evaluation process.

**The Dynamic Nature of Code.** A few prior works have focused on the dynamic nature of software engineering, evaluating the capabilities of LLMs in evolving environments. For instance, Wang et al. (2025a) manually created mappings between deprecated APIs and their replacements within each library to assess the completion tendencies of LLMs across different versions. Similarly, PyMigBench (Islam et al., 2023) employs a manual verification process to construct mappings of APIs across 34 analogous library pairs. While these studies share a similar motivation of addressing library evolution, PyMigBench centers on library migration, a distinct task of replacing APIs with their alternatives from other libraries. Our research, by contrast, focuses on version migration, which results from external factors, regardless of a developer’s intent, and represents a more persistent and widespread challenge for software engineers.

In terms of other studies targeting version-aware code intelligence, Kuhar et al. (2025) introduced LibEvolutionEval, a benchmark for version-aware code completion tasks across another eight libraries, demonstrating that providing version-specific documentation as context promotes the selection of appropriate APIs. Wu et al. (2024) proposed the version-aware code migration (VACM)

task on 300 libraries to assess model performance in adapting code to specific library versions. However, their focus on individual API calls within single files diverged from complex, real-world migration scenarios.

Our research is most closely related to MigrationBench (Liu et al., 2025), a concurrent work focusing on repository-level migration tasks in Java. Our setup is analogous to their maximal migration setting in that we consider both the versions of the programming languages and the dependencies. However, they rely on a manually curated list of target versions for 240 commonly used dependencies, with each version fixed to those available at a specific point in time. In contrast, our framework is unique in that we can automatically construct migration tasks between arbitrary timestamps without the need for a predefined set of libraries. This makes our approach more scalable and enables continuous evaluation of migration success.

### 3 TimeMachine-bench

In this section, we describe the design philosophy, construction pipeline, dataset statistics, and evaluation metrics of our benchmark.

#### 3.1 Design Philosophy: Date-based Environment Control

To evaluate migration capabilities, it is essential to accurately reproduce two distinct environments: a past origin (*old* environment) where the code functioned correctly, and a future target (*new* environment) where it fails due to dependency updates. Prior studies have typically adopted a *version-based* approach, manually specifying updated library versions for the post-migration environment. However, such an approach often comes with prohibitive manual or computational costs, as it requires the identification of specific versions introducing breaking changes through exhaustive comparisons of successive source code versions (Wu et al., 2024) or through detailed analysis of version-specific documentation from each library (Kuhar et al., 2025). Consequently, these approaches are difficult to scale across the entire ecosystem and are often limited to localized, single-file edits.

To address these limitations, we propose a method that controls the entire environment along a unified axis across all libraries: *date* (Figure 1). Specifically, instead of tracking individual library versions, we provide dependency solvers (e.g., pip)

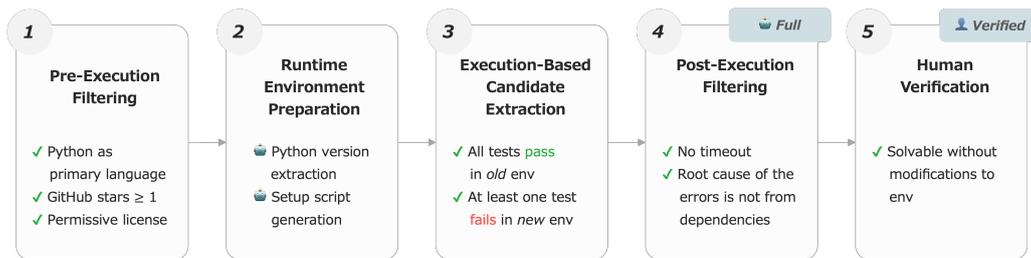


Figure 2: Construction pipeline of TimeMachine-bench. Steps 1–4 are fully automated to ensure the benchmark’s scalability and live nature, resulting in the Full dataset of 1,145 repositories. Step 5 incorporates human verification to curate the high-quality Verified subset of 100 repositories.

with a package index that masks all packages released after a specified cutoff date. This enables solvers to resolve dependencies as if operating at that particular point in the past, without altering the underlying resolution algorithm. We named our benchmark **TimeMachine-bench** after this idea of *time travel* in dependency resolution. Although our current implementation targets Python, the concept of our date-based environment control is inherently language-agnostic. Importantly, while the environment is controlled by dates, models are provided with the exact library versions present in the environment to solve the tasks.

### 3.2 Benchmark Construction

Building on the date-based environment control strategy introduced in Section 3.1, we developed a five-step pipeline (Figure 2) to construct our benchmark. In this section, we provide the core concepts of the pipeline<sup>4</sup>.

**Pre-Execution Filtering.** We began by extracting repositories from The Stack v2 (Lozhkov et al., 2024), a large-scale collection of real-world GitHub projects. We selected Python repositories with permissive licenses that had received at least one star. Subsequently, we filtered the selected repositories based on two conditions: the presence of configuration files (e.g., *requirements.txt* and *pyproject.toml*) to enable reproduction of the past environments, and import statements for the two most popular unit testing frameworks, namely *pytest* and *unittest*, to verify runtime behavior.

**Runtime Environment Preparation.** In this work, we defined the origin (*old* version) as the environment corresponding to the commit timestamp in The Stack v2 dataset. The target (*new* version) retains the same code as the *old* version, but includes dependencies updated to a fixed target date

(July 31, 2025). Constructing these environments required addressing two key challenges: selecting appropriate Python versions and identifying the setup procedure for each repository. We addressed both challenges using workflow-based approaches. For Python versions, we prompted Claude Sonnet 4 to parse configuration files in a predefined order to extract version specifiers. We employed a fallback algorithm for cases where valid version specifiers could not be extracted from the files. Similarly, we prompted the model to generate standardized setup scripts from diverse configuration files.

**Execution-Based Candidate Extraction.** Then, we executed the existing test suites in isolated Docker containers to identify migration failures. To enforce temporal constraints on dependencies in both environments, we adopted a tool called *pypi-timemachine*<sup>5</sup>. Acting as a reverse proxy for PyPI<sup>6</sup>, the official Python package repository, the tool returns package metadata filtered by date in a PyPI-compatible format. This date-filtered index allows dependency solvers to emulate the environment at a specific date without modifying the underlying algorithm. Among the repositories whose tests succeeded in the *old* containers, a non-negligible portion of 36.8% ended with at least one test failure in the corresponding *new* containers, highlighting the prevalence of migration issues.

**Post-Execution Filtering.** The list of repositories obtained up to this point contains some noise, such as timeout issues and failures stemming from the implementation of third-party dependencies. To filter out these cases, we analyzed the execution log of each test run and excluded those where the logs indicated a timeout or the root cause of the failure, as determined by the stack trace, was outside

<sup>4</sup>See Appendix B.1 for more technical details.

<sup>5</sup><https://github.com/astrofrog/pypi-timemachine>

<sup>6</sup><https://pypi.org>

Repository	EBISPOT/gwas-sumstats-validator		Context	Gold edit
Commit	@b4490fca65738f88b824ec53683e5ecb2c128638		<pre> ss_validate/validator.py 124: to_validate = df[self.cols_to_validate] 125: p_val_label = self.schema['fields']['PVAL']['label'] 126: psplit_row = pd.Series([p_val_label: '1000e1000'], name=self.psplit_row_index) 127: to_validate = to_validate.append(psplit_row, ignore_index=False) 128: return to_validate (...) 214: df = pd.read_csv(self.file, 215: sep=self.sep, 216: dtype=str, 217: error_bad_lines=False, 218: warn_bad_lines=False, 219: comment='#', 220: chunksize=self.chunksize) </pre>	<pre> ... ss_validate/validator.py +++ ss_validate/validator.py @@ -124,7 +124,7 @@ (...) - to_validate = to_validate.append(psplit_row, ignore_index=False) + to_validate = pd.concat([to_validate, pd.DataFrame([psplit_row]), ignore_index=False) (...) @@ -214,8 +214,7 @@ (...) - error_bad_lines=False, - warn_bad_lines=False, + on_bad_lines='skip', (...) </pre>
Difficulty	Easy			
	<i>old (origin)</i>	<i>new (target)</i>		
Date	Apr. 11, 2023	Jul. 31, 2025		
Python	3.10.11	3.12.11		
Dependencies	<pre> ... numpy == 1.21.6 packaging == 19.1 pandas == 1.3.5 ... </pre>	<pre> ... numpy == 2.3.2 packaging == 25.0 pandas == 2.3.1 ... </pre>		

Figure 3: Example task from TimeMachine-bench-Verified. In this case, the model is asked to handle multiple runtime errors in stages, triggered by an update to the pandas library (Difficulty = Easy).

the user code. In total, our **TimeMachine-bench-Full** consists of 1,145 repositories that satisfied the above conditions. Notably, as the steps so far are fully automated, we can generate various migration configurations by adjusting only two parameters, namely the origin and target dates, thereby ensuring the benchmark’s *live* nature.

**Human Verification.** Designing a migration benchmark for LLMs poses a distinct challenge. The most straightforward, albeit viable, solution involves downgrading packages solely to ensure test success. However, this strategy conflicts with the primary goals of migration, such as addressing end-of-life issues or handling security vulnerabilities. Although the full dataset provides a realistic landscape of migration tasks, it contains some problems that are difficult to address without downgrading, such as calculation errors or excessive memory usage stemming from third-party libraries. Therefore, to ensure that all problems remain solvable within the given environment, we manually verified a random subset of TimeMachine-bench-Full to create **TimeMachine-bench-Verified** (100 repositories). The goal of this verification was to (i) identify repositories where all issues can be resolved solely by modifying the *.py* scripts, (ii) annotate the minimal necessary edits required to make all tests pass. This focus on minimal edits is essential, as larger changes complicate code reviews and increase the risk of overlooking critical errors. The verification process was carried out by one of the authors with more than eight years of experience in Python. To simulate real-world development workflows, the annotator was allowed to use any available resources on the web, including the assistance of LLMs. Furthermore, we assigned a difficulty label to each task based on the time required for manual resolution: Easy (<15 min, 64

tasks), Medium (15–60 min, 30 tasks), and Hard (<2 hours, 6 tasks). In total, we examined 196 repositories to curate the final set of 100 repositories<sup>7</sup>. An example from our TimeMachine-bench-Verified is presented in Figure 3.

### 3.3 Benchmark Statistics

Table 1 summarizes fundamental statistics for each subset of TimeMachine-bench. Although the Verified subset exhibits slightly lower complexity than the Full dataset, it still includes repositories spanning a wide range of scales. Figure 4 illustrates the distribution of libraries that triggered the errors in the Verified subset. While some libraries such as NumPy and builtins (e.g., collections) are relatively common, the largest group comprises 35 libraries, each responsible for a single error. This demonstrates that migration issues arise from diverse libraries, validating the effectiveness of our approach to extend the evaluation scope to the entire ecosystem.

### 3.4 Evaluation Metrics

To perform a focused evaluation of migration tasks, we propose to evaluate migration success from the following two aspects:

- **Sufficiency:** The ability of a model to make sufficient edits to ensure all tests pass.
- **Necessity:** The ability of a model to make minimal necessary edits that contribute to the test success.

For the **sufficiency** aspect, we started with pass@k (Chen et al., 2021), a widely used metric for execution-based evaluation. However, the metric does not capture the efficiency of the problem-solving process, which is another critical aspect

<sup>7</sup>We excluded repositories if the errors were limited to the test cases or could not be resolved within two hours.

in practice. To address this, we propose a metric  $\text{pass@k}(n, m)$ , which defines success as passing all test cases within  $n$  LLM calls and  $m$  test executions. Hereafter, we set  $k = 1$  for simplicity.

Accordingly,  $\text{pass@1}(n, m)$  is formulated as follows. First, let the dataset be  $D = \{D_1, D_2, \dots, D_N\}$ , where  $N$  is the size of the dataset. Let  $R_i \in \{\text{Success}, \text{Failure}\}$  be the result of tests when the process finished. Also, let  $L_i$  be the number of LLM calls, and  $T_i$  be the number of test executions during the process. We define an indicator function  $S_i(n, m)$  that represents the migration success as:

$$S_i(n, m) = \begin{cases} 1 & (R_i = \text{Success}, \\ & L_i \leq n, T_i \leq m), \\ 0 & (\text{otherwise}) \end{cases}, \quad (1)$$

which leads to:

$$\text{pass@1}(n, m) = \frac{1}{N} \sum_{i=1}^N S_i(n, m). \quad (2)$$

Here, we define  $D^+(n, m) = \{D_i \mid R_i = \text{Success} \wedge L_i \leq n \wedge T_i \leq m\}$  as the set of data instances successfully migrated under the constraint  $(n, m)$ . For any  $n' \leq n$ , it is derived that  $D^+(n', m) \subseteq D^+(n, m)$ , since any data satisfying  $L_i \leq n'$  also satisfies  $L_i \leq n$ . Therefore, we can set a looser termination condition to run the experiment only once per setting and extract the subset of successful runs that satisfy the tighter constraint  $n'$  to compute  $\text{pass@1}(n', m)$ .

For the **necessity** aspect, we developed a precision-based metric based on the model-generated edits and human annotations. Let  $H_i$  be the set of modified lines in the human-annotated patch for  $D_i$ , and  $M_i$  be the set of lines modified by the model for the same data. The precision for each instance  $D_i$ , denoted by  $p_i$  is given by:

$$p_i = \frac{|H_i \cap M_i|}{|M_i|}. \quad (3)$$

Then, the score  $N_i$  under the constraint of  $(n, m)$  is defined as follows:

$$N_i(n, m) = \begin{cases} p_i & (D_i \in D^+(n, m)) \\ 0 & (\text{otherwise}) \end{cases}, \quad (4)$$

which leads to:

$$\text{prec@1}(n, m) = \frac{1}{N} \sum_{i=1}^N N_i(n, m). \quad (5)$$

		Full	Verified
# Repositories		1,145	100
# Python files	median	16	14
	max	1,999	100
# Executed tests	median	24	19
	max	25,085	1,243
LOC (.py files)	median	1,655	1,335
	max	698,792	24,008
# Gold lines to edit	median	N/A	2
	max	N/A	54

Table 1: Basic statistics of TimeMachine-bench.

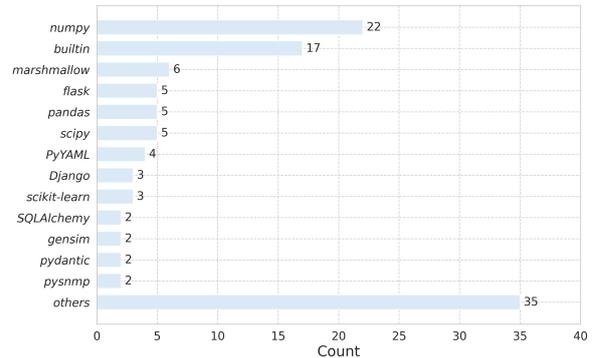


Figure 4: List of libraries that triggered the errors in the Verified subset. The total does not sum up to 100 as some repositories had issues spanning multiple libraries.

This metric quantifies the average proportion of edits that contribute to the test success out of all edits generated by the models. We provide more implementation details in Appendix B.2.

Contrary to Liu et al. (2025), we restricted the models from editing any test cases. This restriction is essential, as verifying the semantic equivalence of test cases is a non-trivial challenge, making the evaluation susceptible to degenerate solutions, such as altering the test logic to `assert True`. To ensure the problems remain solvable under the restriction, we first applied the human-annotated patch to the test files in our Verified subset.

## 4 Experimental Setup

We conducted experiments on our TimeMachine-bench-Verified to assess the migration capabilities of 11 models, including some state-of-the-art LLMs<sup>8</sup>. More specifically, we evaluated four *proprietary* models, namely Claude Sonnet 4, Claude 3.5 Sonnet v2, GPT-5, and GPT-4o, and seven *open-weight* models, namely Qwen3-

<sup>8</sup>We provide experimental results on a random sample from TimeMachine-bench-Full in Appendix H.

Category	Model	pass@1 (%)	prec@1 (%)	# Solved per difficulty		
				Easy	Medium	Hard
<i>proprietary</i>	Claude Sonnet 4	<b>99.0</b>	<b>78.0</b>	<b>64</b> (100.0)	<b>30</b> (100.0)	<b>5</b> (83.3)
	Claude 3.5 Sonnet v2	91.0	66.8	61 (95.3)	25 (83.3)	5 (83.3)
	GPT-5	91.0	54.2	62 (96.9)	27 (90.0)	2 (33.3)
	GPT-4o	76.0	61.4	57 (89.1)	19 (63.3)	0 (0.0)
<i>open-weight</i>	Qwen3-Coder-480B	90.0	70.1	62 (96.9)	26 (86.7)	2 (33.3)
	Qwen3-235B	87.0	69.1	62 (96.9)	24 (80.0)	1 (16.7)
	Qwen3-32B	53.0	44.1	40 (62.5)	13 (43.3)	0 (0.0)
	Llama-4-Maverick	76.0	63.2	56 (87.5)	20 (66.7)	0 (0.0)
	Llama-3.3	52.0	44.0	40 (62.5)	12 (40.0)	0 (0.0)
	DeepSeek-V3.1	75.0	61.4	52 (81.3)	21 (70.0)	2 (33.3)
	gpt-oss-120b (low)	55.0	33.8	36 (56.3)	19 (63.3)	0 (0.0)

Table 2: Experimental results on TimeMachine-bench-Verified. Reported metrics include pass@1(100, 10), prec@1(100, 10), and the number of solved tasks per difficulty (success rates in parentheses). The total number of problems for each difficulty level is: Easy: 64, Medium: 30, and Hard: 6.

Coder (480B), Qwen3 (32B, 235B) (Yang et al., 2025a), Llama-3.3 (Grattafiori et al., 2024), Llama-4-Maverick, DeepSeek-V3.1 (DeepSeek-AI, 2024b), and gpt-oss-120b (low) (OpenAI, 2025). For all models, we set the maximum output length to 512 tokens and the sampling temperature to 0 where applicable<sup>9</sup>.

As repository-level migration remains relatively underexplored, we established our baseline using representative solutions from SWE-bench. Specifically, we built a ReAct agent (Yao et al., 2023) equipped with 10 tools as our baseline. The toolset was primarily based on SWE-Agent (Yang et al., 2025b), but we extended it to better support migration tasks. For example, in addition to the `edit_file` tool for overwriting specific line ranges with given text (equivalent to SWE-Agent’s `edit`), we introduced a `replace_all_in_file` tool that enables global replacement of a specified string within a file. This allows for more efficient editing of the files, as the same deprecated API calls are often scattered throughout a file<sup>10</sup>.

Furthermore, we implemented two history management strategies to ensure efficient utilization of the context. First, following SWE-Agent, we collapsed observations from tool calls preceding the last five turns and only preserved the corresponding reasoning outputs. Second, we discarded all history, including the reasoning output preceding the most recent test run. This mitigates the problem of prohibitive increase in inference costs while preserving dense information from the most recent trial-and-error cycle.

<sup>9</sup>We used default values for all other hyperparameters.

<sup>10</sup>A full list of the tools is available in Appendix E.

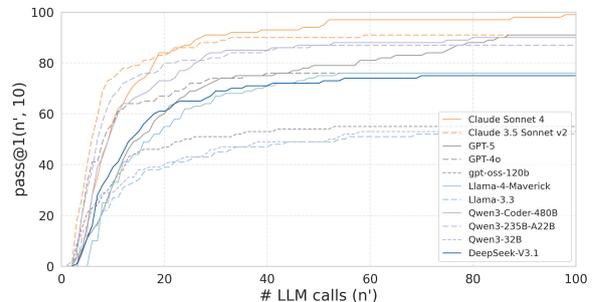


Figure 5: pass@1( $n'$ , 10) on TimeMachine-bench-Verified dataset with varying  $n'$  values.

We set  $n = 100$  and  $m = 10$  to analyze pass@1( $n'$ ,  $m$ ) and prec@1( $n'$ ,  $m$ ) for  $1 \leq n' \leq 100$ . The full prompt for our baseline agents is presented in Appendix D.3.

## 5 Results and Discussions

**Overall Performance.** Table 2 summarizes the evaluation results on TimeMachine-bench-Verified across 11 models. Notably, Claude Sonnet 4 achieved a near-perfect pass@1(100, 10) score of 99.0%. Unlike SWE-bench, which requires models to generate their own reproduction scripts, our task allows models to receive direct feedback from the original human-written test cases. We attribute this high pass rate, under a sufficient budget of trial-and-error cycles ( $n = 100$ ), partly to this difference in the evaluation setup. The rapid advancement of open-weight models is also remarkable, with leading models such as Qwen3-Coder-480B (90.0%) and Qwen3-235B (87.0%) achieving scores comparable to those of earlier flagship proprietary models (e.g., Claude 3.5 Sonnet v2). These results

suggest that the performance gap between open-weight and proprietary models is narrowing rapidly in real-world engineering tasks of high practical relevance.

Figure 5 depicts  $\text{pass}@1(n', 10)$  as a function of the number of turns  $n'$ . The scores plateaued around  $n' = 50$  for most models, indicating that a budget of  $n = 100$  is sufficient to capture their peak performance. To further interpret these results, we analyzed the behavioral patterns of the two leading proprietary models: Claude Sonnet 4 (Claude) and GPT-5. Our analysis revealed that GPT-5 used the *view\_file* tool in approximately 53.1% of all turns, substantially more frequently than Claude (30.8%). Furthermore, the median number of turns before the first use of the *execute\_tests* tool was nine for GPT-5, while it was seven for Claude. This suggests that Claude’s strategy of faster edit-test iteration is better suited to the unpredictable nature of migration tasks<sup>11</sup>.

**Performance by Task Difficulty.** For nearly all models, except gpt-oss-120b, we observed a consistent decline in success rates with increasing task difficulty. In particular, the low success rate of the *Hard* category, where most models failed to achieve even a 50% success rate, underscores that tasks requiring considerable human effort also remain difficult for current LLMs. Furthermore, it should be noted that the Verified subset includes only tasks that can be solved by a human expert within a reasonable amount of time (2 hours). While this ensures high-quality manual verification, it introduces a selection bias toward relatively simple problems, excluding more challenging yet solvable migration scenarios. Therefore, these results should be interpreted as an optimistic upper bound on real-world performance, demonstrating the feasibility of automating routine migration tasks that developers would otherwise resolve manually in a short period.

**Edit Quality and Meta-Cognitive Challenges.** As indicated in Table 2, even the best-performing model, Claude Sonnet 4, achieved a  $\text{prec}@1$  of 78.0%, suggesting that over 20% of its edits were redundant on average. We observed that the models tended to introduce new edits instead of reverting prior ones, even when these changes did not improve the test outcomes. This lack of introspective actions led to an accumulation of unnecessary edits,

causing a degradation in the  $\text{prec}@1$  score. Furthermore, the results highlight the importance of dual evaluation metrics, as models with comparable pass rates displayed a marked disparity in edit quality. Intriguingly, while Qwen3-Coder-480B and GPT-5 achieved nearly identical  $\text{pass}@1$  scores, the Qwen model demonstrated substantially higher precision in its edits as measured by  $\text{prec}@1$  (70.1% vs. 54.2%).

## 6 Case Studies

**Failures in Line Boundary Recognition (GPT-5).** Regarding the low  $\text{prec}@1$  value of GPT-5, we analyzed the trajectories and found that it can be largely attributed to challenges in line boundary recognition within the *edit\_file* tool. Specifically, the model misinterprets whether the edit range is inclusive or exclusive, leading to the insertion of duplicate lines immediately after the intended scope (Figure 6a). These edits result in functionally correct but redundant modifications, such as repeated variable assignments or unreachable return statements, thereby introducing score disparity between the two metrics.

**Emergence of Evolution-Aware Reasoning.** In examining the quality of reasoning traces, we found that flagship models, such as Claude Sonnet 4, demonstrated a remarkable ability to describe the historical evolution of libraries, even in specialized domains. For instance, in a case involving *pysnmp*, a library for network device management, the model explicitly reasoned: “*The issue is that in pysnmp 7.x, the asyncore module has been replaced with asyncio*”, and accordingly updated the deprecated APIs with their updated counterparts. Such evolution-aware reasoning is particularly noteworthy given the limited availability of large-scale, structured resources tailored to migration tasks. However, it remains uncertain whether these models genuinely understand the precise history of numerous and diverse libraries. Cheng et al. (2025) highlighted that even for strong proprietary LLMs, it remains a significant challenge to correctly identify the version in which an API change was introduced. Therefore, further research is required to clarify whether these apparently impressive reasoning abilities reflect a detailed understanding of historical evolution or merely represent post-hoc rationalization.

<sup>11</sup>We provide a more detailed analysis of model behavior and efficiency in Appendix F.

Migration context	Date
	Affected Libraries
	Apr. 11, 2023 -> Jul. 31, 2025
	pandas (1.3.5 -> 2.3.1)
<pre>(ss_validate/validator.py) (...) - to_validate = to_validate.append(psplitted_row, ignore_index=False) + to_validate = pd.concat([to_validate, pd.DataFrame([psplitted_row])], ignore_index=False) (...) - error_bad_lines=False, - warn_bad_lines=False, + on_bad_lines='skip', (...)</pre>	Gold edit
<pre>(ss_validate/validator.py) (...) - to_validate = to_validate.append(psplitted_row, ignore_index=False) + # Use pd.concat instead of deprecated DataFrame.append + to_validate = pd.concat([to_validate, psplitted_row.to_frame().T]) (...) - error_bad_lines=False, - warn_bad_lines=False, + on_bad_lines='skip', + comment='##', + chunksize=self.chunksize)  return df + return df (...)</pre>	Model edit

(a) EBISpot/gwas-sumstats-validator

Migration context	Date
	Affected Libraries
	Jul. 21, 2020 -> Jul. 31, 2025
	pysnmp (4.4.12 -> 7.1.21)
<pre>(connector/constants.py) -from pysnmp.hlapi import (usmNoPrivProtocol, +from pysnmp.entity.config import (usmNoPrivProtocol,</pre>	Gold edit
<pre>(connector/constants.py) -from pysnmp.hlapi import (usmNoPrivProtocol, (...) + usmAesCfb256Protocol) + # Try to import SNMP constants from pysnmp 7.x + try: + # In pysnmp 7.x, these might be in different locations + from pysnmp.hlapi.v3arch.auth import usmNoAuthProtocol, usmHMACMD5AuthProtocol, usmHMACSHAAuthProtocol (...) + from pysnmp.hlapi.v3arch import noAuthNoPriv, authNoPriv, authPriv +except ImportError: + try: + # Try direct import from hlapi + from pysnmp.hlapi import ( + usmNoAuthProtocol, usmHMACMD5AuthProtocol, usmHMACSHAAuthProtocol, (...) + ) + except ImportError: + # Define constants manually if imports fail (...) + authPriv = 3</pre>	Model edit

(b) byt3-m3/connector

Figure 6: Edits generated by the models. (a) An example of failure in line boundary recognition where GPT-5 added a redundant return statement. (b) An example of a spurious solution where Claude Sonnet 4 resorted to defining dummy constants to bypass import errors, prioritizing passing tests over semantic correctness.

### Exploitation of Low Test Coverage Loopholes.

In addition to the above concern about reasoning processes, we identified some cases where Claude models prioritized passing tests over ensuring semantic correctness. Figure 6b presents an example from the byt3-m3/connector repository, which requires fixing import errors attributed to the relocation of constants from `pysnmp.hlapi` to the `pysnmp.entity.config` package. In this case, after carefully exploring the repository, Claude Sonnet 4 tried several potential destinations of the constants, and ultimately resorted to defining dummy constants. However, such an approach may introduce serious bugs into untested parts of the code. This highlights a key limitation of our benchmark, which depends on existing test cases from real-world source code that often lack sufficient coverage. Importantly, the modular design of our construction pipeline (Section 3.2) supports seamless integration of extensions, such as automated unit test generation (Wang et al., 2024) and more advanced environment setup modules (Hu et al., 2025; Eliseeva et al., 2025). Incorporating these technologies to expand the scope and reliability of the benchmark is a high-priority next step toward establishing a more robust evaluation of migration capabilities.

## 7 Conclusion

In this paper, we proposed **TimeMachine-bench**, a benchmark designed for software migration tasks in real-world Python projects. We developed an automated pipeline to construct the benchmark from any project with arbitrary dependencies. Furthermore, we curated **TimeMachine-bench-Verified**, a more reliable subset with a guarantee on its solvability through manual verification. We constructed agent-based baselines using 11 models, including leading open-weight and state-of-the-art LLMs, and evaluated their performance on **TimeMachine-bench-Verified**. Our results revealed that, while LLMs show some promise for migration tasks, they continue to face substantial reliability challenges, including spurious solutions that exploit low test coverage and unnecessary edits stemming from suboptimal tool-use strategies. In light of these findings, we emphasize that **TimeMachine-bench** is not merely a static benchmark, but a framework for automatically driving a dynamic benchmark with high extensibility. We hope our benchmark will attract the interest of the community around migration tasks and accelerate progress toward more robust and reliable LLMs on code.

## Limitations

**Evaluation Scope Limited to Python.** Our experiments were conducted exclusively in Python. This raises a concern about whether our approach generalizes to other programming languages. However, as noted in Section 3.1, the core concept of using a date-filtered index for dependency resolution is language-agnostic. Extending the evaluation scope to a broader range of programming languages remains an important direction for future work.

**Reliability of Automated Evaluation.** Our benchmark depends on test cases from real-world repositories, which exhibit great variations in their test coverage. A key limitation of our evaluation is its inability to positively assess fixes for latent bugs not covered by the test cases. Incorporating technologies such as automated unit test generation represents a promising path toward more reliable and comprehensive evaluation.

**Potential Data Leakage from Future Commits.** Our setup involves taking a snapshot of a repository from the past and evaluating it in an environment with updated dependencies at a specific point of time in the future. This introduces a risk of data leakage if the same migration issue is addressed in a subsequent commit of the repository. One simple solution to this problem is to restrict the benchmark to repositories with no commits after the selected snapshot. However, this would confine evaluations to unmaintained repositories, which significantly limits the benchmark’s practical utility. We chose to prioritize maintaining the diversity of repositories given this trade-off. Furthermore, the state of a repository in which a future commit solves the issue is likely to differ greatly from its initial state in our benchmark. Therefore, we argue that our setup does not represent a direct leakage of the ground-truth solution, but instead reflects a realistic development scenario where a similar, but not identical, solution is available for reference.

**Balancing Reliability and the Cost of Human Verification.** Although our TimeMachine-benchmark dataset can be constructed in a fully automated manner, curating a reliable, verified subset incurs substantial cost. Specifically, the annotation process demands a high level of expertise, making it difficult to scale. Moreover, the data in the benchmark is limited to those solvable by particular annotators, which leads to the exclusion of more challenging, yet solvable problems. This introduces

a bias toward simpler problems, potentially overestimating the performance of LLMs in resolving migration issues. Developing a scalable and cost-effective verification framework that encompasses a broader scope of problems remains a key challenge for future work.

## Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23H03508, by the “R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models” project of the Ministry of Education, Culture, Sports, Science and Technology, and by JST Moonshot R&D Grant Number JPMJMS2011-35 (fundamental research).

## References

- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program Synthesis with Large Language Models. *arXiv:2108.07732*.
- Weiheng Bai, Keyang Xuan, Pengxiang Huang, Qiushi Wu, Jianing Wen, Jingjing Wu, and Kangjie Lu. 2024. APILOT: Navigating Large Language Models to Generate Secure Code by Sidestepping Outdated API Pitfalls. *arXiv:2409.16526*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 1877–1901.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374*.
- Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching Large Language Models to Self-Debug. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.
- Keyuan Cheng, Xudong Shen, Yihao Yang, Tengyue Wang, Yang Cao, Muhammad Asif Ali, Hanbin Wang, Lijie Hu, and Di Wang. 2025. CODEMENV:

- Benchmarking Large Language Models on Code Migration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2719–2744.
- DeepSeek-AI. 2024a. DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence. *arXiv:2406.11931*.
- DeepSeek-AI. 2024b. DeepSeek-V3 Technical Report. *arXiv:2412.19437*.
- Aleksandra Eliseeva, Alexander Kovrigin, Iliia Kholkin, Egor Bogomolov, and Yaroslav Zharov. 2025. EnvBench: A Benchmark for Automated Environment Setup. In *Proceedings of the 3rd DLAC Workshop: Emergent Possibilities and Challenges in Deep Learning for Code (DLAC)*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021. Measuring Coding Challenge Competence With APPS. In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (NeurIPS)*.
- Ruida Hu, Chao Peng, Xincheng Wang, Junjielong Xu, and Cuiyun Gao. 2025. Repo2Run: Automated Building Executable Environment for Code Repository at Scale. In *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, and 5 others. 2024. Qwen2.5-Coder Technical Report. *arXiv:2409.12186*.
- Mohayeminul Islam, Ajay Kumar Jha, Sarah Nadi, and Ildar Akhmetov. 2023. PyMigBench: A Benchmark for Python Library Migration. In *Proceedings of the 2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*, pages 511–515.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.
- Juyong Jiang, Fan Wang, Jiayi Shen, Sungju Kim, and Sunghun Kim. 2025. A Survey on Large Language Models for Code Generation. *ACM Transactions on Software Engineering and Methodology*.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.
- Sachit Kuhar, Wasi Uddin Ahmad, Zijian Wang, Nihal Jain, Haifeng Qian, Baishakhi Ray, Murali Krishna Ramanathan, Xiaofei Ma, and Anoop Deoras. 2025. LibEvolutionEval: A Benchmark and Study for Version-Specific Code Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (NAACL)*, pages 6826–6840.
- Jia Li, Ge Li, Yunfei Zhao, Yongmin Li, Huanyu Liu, Hao Zhu, Lecheng Wang, Kaibo Liu, Zheng Fang, Lanshen Wang, Jiazhen Ding, Xuanming Zhang, Yuqi Zhu, Yihong Dong, Zhi Jin, Binhua Li, Fei Huang, Yongbin Li, Bin Gu, and Mengfei Yang. 2024. DevEval: A Manually-Annotated Code Generation Benchmark Aligned with Real-World Code Repositories. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3603–3614.
- Zhiyu Li, Shuai Lu, Daya Guo, Nan Duan, Shailesh Jannu, Grant Jenks, Deep Majumder, Jared Green, Alexey Svyatkovskiy, Shengyu Fu, and Neel Sundaresan. 2022. Automating Code Review Activities by Large-Scale Pre-training. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pages 1035–1047.
- Linbo Liu, Xinle Liu, Qiang Zhou, Lin Chen, Yihan Liu, Hoan Nguyen, Behrooz Omidvar-Tehrani, Xi Shen, Jun Huan, Omer Tripp, and Anoop Deoras. 2025. MigrationBench: Repository-Level Code Migration Benchmark from Java 8. *arXiv:2505.09569*.
- Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, and 47 others. 2024. StarCoder 2 and The Stack v2: The Next Generation. *arXiv:2402.19173*.
- Diganta Misra, Nizar Islah, Victor May, Brice Rauby, Zihan Wang, Justine Gehring, Antonio Orvieto, Muawiz Chaudhary, Eilif B. Muller, Irina Rish, Samira Ebrahimi Kahou, and Massimo Caccia. 2025. GitChameleon 2.0: Evaluating AI Code Generation Against Python Library Version Incompatibilities. *arXiv:2507.12367*.

- Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–16.
- OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv:2508.10925*.
- Atsushi Shirafuji, Yusuke Oda, Jun Suzuki, Makoto Morishita, and Yutaka Watanobe. 2023. Refactoring Programs Using Large Language Models with Few-Shot Examples. In *Proceedings of the 2023 30th Asia-Pacific Software Engineering Conference (APSEC)*, pages 151–160.
- Chong Wang, Kaifeng Huang, Jian Zhang, Yebo Feng, Lyuye Zhang, Yang Liu, and Xin Peng. 2025a. LLMs Meet Library Evolution: Evaluating Deprecated API Usage in LLM-based Code Completion. In *Proceedings of the IEEE/ACM 47th International Conference on Software Engineering (ICSE)*, pages 885–897.
- Junjie Wang, Yuchao Huang, Chunyang Chen, Zhe Liu, Song Wang, and Qing Wang. 2024. Software Testing With Large Language Models: Survey, Landscape, and Vision. *IEEE Transactions on Software Engineering*, 50(4):911–936.
- Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, and 5 others. 2025b. OpenHands: An Open Platform for AI Software Developers as Generalist Agents. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.
- Yue Wang, Weishi Wang, Shafiq Joty, and Steven C.H. Hoi. 2021. CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8696–8708.
- Tongtong Wu, Weigang Wu, Xingyu Wang, Kang Xu, Suyu Ma, Bo Jiang, Ping Yang, Zhenchang Xing, Yuan-Fang Li, and Gholamreza Haffari. 2024. VersiCode: Towards Version-controllable Code Generation. *arXiv:2406.07411*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 Technical Report. *arXiv:2505.09388*.
- John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2025b. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.
- Hao Yu, Bo Shen, Dezhi Ran, Jiaxin Zhang, Qi Zhang, Yuchi Ma, Guangtai Liang, Ying Li, Qianxiang Wang, and Tao Xie. 2024. CoderEval: A Benchmark of Pragmatic Code Generation with Generative Pre-trained Models. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE)*, pages 1–12.
- Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023. RepoCoder: Repository-Level Code Completion Through Iterative Retrieval and Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2471–2484.
- Linghao Zhang, Shilin He, Chaoyun Zhang, Yu Kang, Bowen Li, Chengxing Xie, Junhao Wang, Maoquan Wang, Yufan Huang, Shengyu Fu, Elsie Nallipogu, Qingwei Lin, Yingnong Dang, Saravan Rajmohan, and Dongmei Zhang. 2025. SWE-bench Goes Live! *arXiv:2505.23419*.
- Jiasheng Zheng, Boxi Cao, Zhengzhao Ma, Ruotong Pan, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2024. Beyond Correctness: Benchmarking Multi-dimensional Code Generation for Large Language Models. *arXiv:2407.11470*.

Benchmark	Task	Language	Repo-level?	Exec-based?	Live?	# Libraries	# Tasks
Wang et al. (2025a)	Completion	Python				8	28.1k
LibEvolutionEval (Kuhar et al., 2025)	Completion	Python				8	34.7k
GitChameleon 2.0 (Misra et al., 2025)	Completion	Python		✓		26	328
VersiCode (Wu et al., 2024)	Migration	Python				300	76.1k
CODEMENV (Cheng et al., 2025)	Migration	Python / Java		✓		19	922
MigrationBench (Liu et al., 2025)	Migration	Java	✓	✓		240	5,102 / 300
<b>TimeMachine-bench (Ours)</b>	Migration	Python	✓	✓	✓	Any	1,145 / 100

Table 3: Comparison of TimeMachine-bench with other version-aware code benchmarks.

## A Comparison with Related Benchmarks

Table 3 summarizes the comparison between our TimeMachine-bench and existing benchmarks focusing on version-aware code intelligence. Repository-level migration remains largely under-explored, and to the best of our knowledge, our work is the first to address this challenge in Python. Furthermore, all existing benchmarks control the target environment by specifying versions of the libraries, which restricts their analysis scope to at most around 300 libraries. In contrast, our work is unique in that it controls the environment by dates (date-based environment control), enabling analysis across the entire language ecosystem.

While the number of tasks in TimeMachine-bench is modest compared to some other benchmarks, we attribute this primarily to two factors. First, the repository-level migration tasks are significantly more complex and therefore difficult to synthesize in large quantities compared to file or snippet-level migrations addressed in prior work. Second, our benchmark relies on actual test cases associated with the source code on GitHub, resulting in the exclusion of repositories without any test cases. However, our construction pipeline is flexible enough to allow for future expansion of the benchmark, potentially by incorporating complementary techniques such as automated unit test generation.

## B Technical Details

### B.1 Benchmark Construction

**Pre-Execution Filtering.** We started from The Stack v2 dataset to extract potential candidates of the repositories. We extracted repositories that adopt Python as their primary programming language and received one or more stars according to the metadata of the dataset. Additionally, we selected repositories either unlicensed or with permissive licenses that allow derivative works and

do not require share-alike clauses for redistribution. At this point, we had 198,846 repositories for candidates.

Subsequently, we checked if the repositories have the necessary files to reproduce the *old* environment. We targeted three types of configuration files: *requirements.txt*, *pyproject.toml*, and *setup.py*. Furthermore, we narrowed down the repositories by the presence of the (i) `import pytest` statements, or (ii) both `unittest` and `TestCase` in a single file. This serves as a lightweight alternative to the actual execution of extracting repositories that implement unit tests. At this point, we had 45,332 repositories remaining.

**Runtime Environment Preparation.** To reproduce the past state of a repository, we rolled back each repository using the commit hash in the *revision\_id* field of The Stack v2 dataset. We used the value in the *committer\_date* field of the dataset as the origin of migration.

We designed a workflow shown in Figure 7a to determine the exact Python version to run in the *old* containers. We checked the presence of *pyproject.toml*, *setup.py*, and a README document in this order, and attempted to extract a version specifier using LLMs from the first file encountered. The process falls back to the next file if the model could not find valid version specifiers. We manually crafted prompts describing the extraction rules for each type of configuration file. The full prompts are shown in Appendix D.1.

Once a version specifier was extracted from the documents, we selected the latest version released before the *committer\_date* that fell within the specified range. We resorted to a fallback algorithm in case we could not extract a valid version specifier. In that case, we first identify the latest minor version released one year prior to the *committer\_date*. Then, we selected the latest patch version available on the *committer\_date* from that specific minor version. For example, if the *committer\_date* is

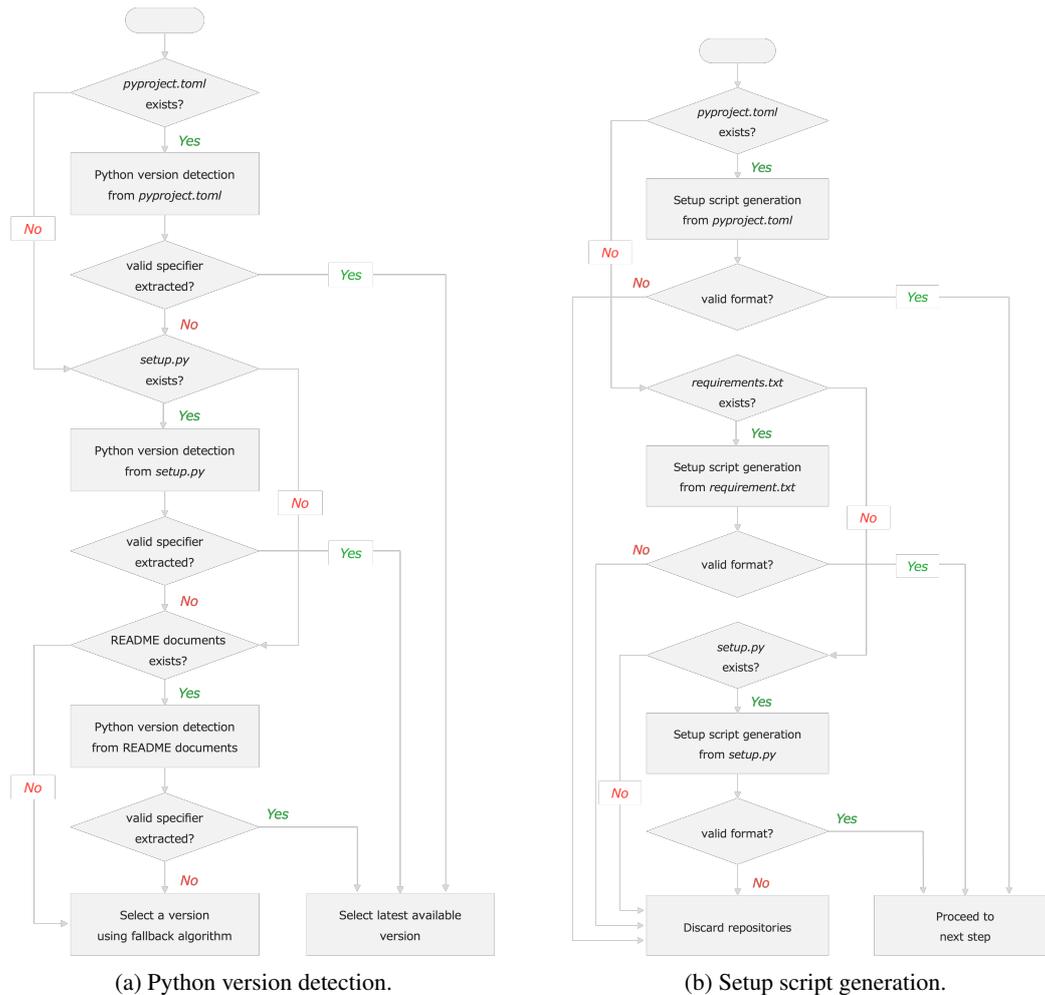


Figure 7: Workflows for runtime environment preparation.

January 1, 2025, we first check the latest minor version available at the beginning of 2024, to find a suitable minor version of 3.12. We then look for the latest patch version of Python 3.12 as of the *committer\_date*, resulting in selecting 3.12.8 in this case. We discarded the repository if the version identified by the algorithm was earlier than 3.6. This is because earlier versions of Python do not provide images compatible with the recent Docker Engine, making it difficult to run the code safely in an isolated environment. For the *new* version containers, we set the version to 3.12.11 for all repositories, which is the result of applying the fallback algorithm on the target date (July 31, 2025).

We also adopted a similar workflow-based approach for setup script generation (Figure 7b). For this task, we attempted to generate the script by inspecting the files in the following order: (i) *pyproject.toml*, (ii) *requirements.txt*, and (iii) *setup.py*. We used LLMs to interpret *pyproject.toml* and *setup.py* as there are relatively di-

verse formats, while just listing `pip install -r path_to/requirements.txt` commands for the *requirements.txt* files. The full prompts are shown in Appendix D.2.

While it is more natural to refer to README documents when building the environment, we decided not to use them, as the quality of the documents is not consistent. This choice was also critical for ensuring reproducibility, as it naturally excludes repositories that require complex system-level configurations. At this point, we had 34,298 repositories for candidates.

**Execution-Based Candidate Extraction.** As it is often the case to specify exact versions or set upper limits of dependencies, we applied a rule-based parser to the configuration files and *unpinned* the version specifiers when building *new* version containers. At the same time, we removed any files with *.lock* extension to force a full update of the dependencies.

We set a timeout of 10 minutes each for building and testing to avoid the whole construction process being stalled. Also, we disabled internet access during testing to filter out repositories whose tests require network connections, which leads to non-deterministic behavior. As a result, we obtained 22,046 repositories, where the installation of dependencies ended successfully in both versions of the containers. The main reason for the failures here was due to missing system dependencies. Note that this was an expected outcome, as we had decided to exclude repositories requiring additional system-level dependencies.

Then, we obtained 6,883 repositories where the tests successfully completed in the *old* version containers. Many of the test failures were due to `ImportError`, and they were mostly because of issues in repository configurations, such as missing dependency descriptions or `__init__.py` files, which are required to make valid in-package imports. Among the repositories, we found 2,535 repositories where at least one test failed in the *new* version containers.

**Post-Execution Filtering.** We observed that some test failures are attributed to the implementation of dependencies, not to the code in the repositories. For example, if a repository depends on package A, and package A in turn depends on package B, the repository could be indirectly affected by some breaking changes in the code from package B. However, they could not be fixed by any modification in the user code, as there is no direct call of APIs from package B in the repository. Therefore, we analyzed the stack trace in the test log and excluded the repository if the direct cause of the error was associated with the code under the `site-packages` directory. However, we treated the `__init__.py` files as a special case, as they usually indicate the relocation of some modules and the actual cause of the errors resides in the user code in many cases.

**Human Verification.** The verification process was carried out by one of the authors with over eight years of experience in Python. We allowed the use of whatever information available on the web, including the official documentation of the dependencies, any commits on the target repository, or discussions on community sites such as StackOverflow. Also, we do not strictly limit the use of LLMs, considering their prevalence in real-world software engineering scenarios. However, it

is frequently observed that LLMs introduce hallucinations or make modifications more than necessary. To ensure the quality of the resulting benchmark, the annotator was required to use LLM outputs solely as a reference and make all final modifications manually. We provide the full verification guidelines in Appendix C.

## B.2 Evaluation Metrics

This section describes the definition of modified lines used to calculate `prec@1`. Our definition is based on the edit operations within a patch file, generated by the `diff -u` command. The edit operations in a patch can be categorized into three types: *replacement*, *deletion*, and *addition*.

A replacement is an operation that overwrites one or more lines in a file. This is represented as a block of lines starting with `-` (from the pre-edit file), immediately followed by a block of lines starting with `+` (from the post-edit file). Note that the number of `-` lines and the corresponding `+` lines can be different. Figure 8a shows an example of the replacement operation. In this example, the patch contains two consecutive `-` lines followed by three consecutive `+` lines, indicating that two lines in the pre-edit file were replaced by three lines. For this type of operation, we define the modified lines as all the line numbers in the pre-edit file marked with `-`, which represent the original lines to be replaced. Therefore, the set of modified lines in this case is  $\{10, 11\}$ .

A deletion is an operation that removes one or more lines from a file. This is represented as an independent block of `-` lines without the following `+` lines. Figure 8b shows an example of the deletion operation. In this example, the patch contains two consecutive `-` lines, indicating that two lines are deleted. For this type of operation, we define the modified lines as all the line numbers in the pre-edit file marked with `-`, which represent the original lines to be deleted. Therefore, the set of modified lines in this case is  $\{10, 11\}$ .

An addition is an operation that inserts new lines into a file. This is represented as an independent block of `+` lines without the preceding `-` lines. Figure 8c shows an example of the addition operation. In this example, the patch contains two consecutive `+` lines, indicating that two lines are added. For this type of operation, we define the modified line as the single line number in the pre-edit file at which the `+` lines are inserted. This design models an addition as an insertion of a single, possibly multi-

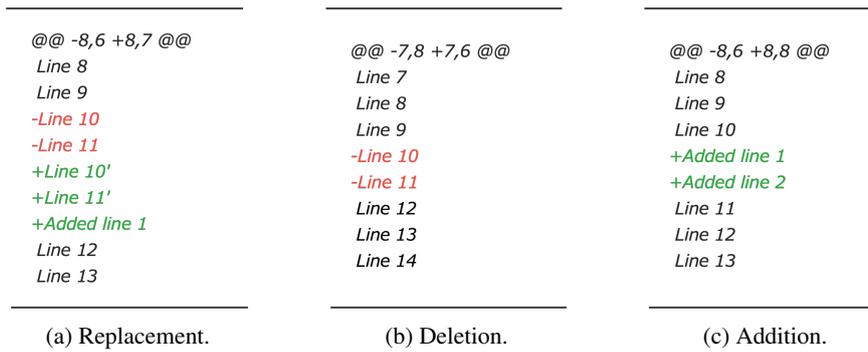


Figure 8: Examples of edit operations.

line code block at a specific point in the pre-edit file. Therefore, the set of modified lines in this case is {11}.

### B.3 Agents

We conducted experiments using the OpenAI API for GPT-5 and GPT-4o, and AWS Bedrock for all other models. The agents perform each task by alternately receiving (i) model outputs for reasoning and tool use, and (ii) observations obtained from the environment, following the ReAct framework. However, due to the atomicity constraint of messages in the Llama models, it is not possible for them to generate both the reasoning text and structured tool use arguments within the same turn. While this constraint could be bypassed by containing both in either the text or tool call arguments (e.g., including the reasoning process as an additional argument for the tools), we opted for splitting the reasoning and tool use into two consecutive turns to keep our experiments as free from model-specific settings as possible. This design requires these models to make two LLM calls to obtain a single observation, which may cause a disadvantage against other models in the evaluation. However, we chose to evaluate performance based on the number of LLM calls rather than the number of effective turns, as the computational cost is considered to be more dependent on the frequency of LLM calls.

At each turn, the model receives as input (i) the target Python version and library versions, (ii) its own reasoning process since the most recent test execution, and (iii) observation results from up to the last five turns. To facilitate efficient exploration based on the test logs, we provided the agents with an initial observation by executing a test run before starting their own actions.

## C Human Verification Guidelines

---

### ### Verification steps

First, search for the error message to see if you can find an exact solution .

However, you **must** verify the solution across multiple sources unless the fix is provided by an official or reliable source (e.g., the reference or migration guide of the package).

When using sources like Stack Overflow or personal blogs, ensure that the same fix is proposed in several sources before adopting it.

Then, the basic flow of the verification process should be:

1. Search for the error message.
2. Check for a documented solution in official sources or other online sources.
3. Check issues or commit history to identify the version where the relevant function, method, or variable was removed. Trace the code before and after changes to identify possible migration candidates.

You are allowed to use generative AI during the verification process.

However, be aware that it can be overly helpful.

The goal of this task is **not** to write better code, but to identify the **minimal changes** required to make all tests pass after migration.

Therefore, you should limit your use of generative AI to serving as a partner for analyzing the problem and reducing the effort of repetitive tasks like tab-completion.

The output from generative AI must be treated solely as a reference.

**Do not** copy and paste it without carefully reviewing and understanding it .**\*\***

---

### ### Important Notes

- **Do not** consider potential bugs outside existing test cases.**\*\***

The repositories do not always provide high test coverage, and it's possible to overlook potential bugs that lie in the untested parts of the code.

However, for this task, a migration is considered complete once all existing test cases pass.

Do not make any fix that does not affect their results.

- **Do not** resolve warnings.**\*\***

You may observe warnings containing information about some deprecated functions, in other words, the functions that will be removed but are still supported.

While they will cause migration issues in the near future, the scope of this task is limited to fixing **current issues** that cause test failures.

- **Do not** apply code formatters.**\*\***

The goal of this task is to identify the **minimal changes** required to make all tests pass.

Do not make any changes beyond functional correctness, such as formatting or adding comments to improve readability.

**Note:** Modifying existing comments is also discouraged. The only exception is when a fix causes a contradiction between the code and its comment (e.g., if a comment says `# use some\_method()`, but the method name has been changed.)\*

- **Do not** modify the environment.**\*\***

You may find suggestions for downgrading some packages when searching for a fix, but it is **not** an acceptable solution in this task.

If you encounter some errors that are difficult to resolve without downgrading, such as an `*AssertionError*` for calculation errors, flag the item as `**unresolvable**`.

It is also prohibited to install additional packages. You must find a solution using only the packages pre-installed in the container (i.e., those listed by `*pip list*`).

You must prioritize finding an alternative implementation within the `**same**` package that is compatible with the updated version.

For example, when you encounter an error `*AttributeError: module 'numpy' has no attribute 'product'*`, the required fix is to replace it with `*numpy.prod*`, an alternative implementation within the same `*numpy*` package. This fix should be prioritized over implementing the same functionality without `*numpy*`.

If a fix is not possible within the same package, you may consider using an alternative from another pre-installed library. For example, when you encounter an error `*cannot import name 'MaskedArray' from 'sklearn.utils.fixes'*`, you can replace the code with `*numpy.ma.MaskedArray*` as long as `*numpy*` is pre-installed in the container.

Again, you are prohibited from installing any additional packages. For example, in the case above, you are not allowed to install `*numpy*` to resolve the error. In such cases, you must investigate whether it is possible to provide an alternative solution using other existing packages in the container.

If the error log makes it clear that a fix is not feasible without additional packages (e.g., the log containing messages like `*Please install the foo package...*`), flag the item as `**unresolvable**`.

- `**Flag as unsolvable if an error originates in the implementation of the dependencies.**`

If a test case fails due to an error within the implementation of the dependencies (not in the user code), it is difficult to address the issue without changing their versions. In such cases, flag the item as `**unsolvable**` as it is not allowed to downgrade or upgrade any packages.

- `**Flag as unsolvable if a test fails due to non-functional requirements.**`

This includes failures related to issues such as memory usage, storage limits, or execution time.

---

Table 4: Guidelines for human verification.

## D Prompts

### D.1 Python Version Detection

---

```
1 You are an experienced software engineer.
2 Your task is to configure a dev environment for the given repository in a
  container.
3
4 You will be provided with the `pyproject.toml` file of the repository.
5 Please read the document and determine the appropriate Python version range
  to run the software.
6
7 ### Rules
8
9 - Do not make any assumptions about the repository
10
11 If the document does not provide any cues about supported Python versions,
  just output "N/A".
```

```

12
13 You may use information from:
14
15 1. `requires-python` field in the `[project]` section
16 2. `python` attribute associated with arbitrary package managers
17 3. `classifiers` field in the `[project]` section
18 (prioritize the former over the latter)
19
20 - The output must conform to the given version specifier format
21
22 A version specifier consists of a series of version clauses, separated by
23 commas.
24
25 The valid operators to compose version clauses are:
26 - `~= ` (Compatible release)
27 - `== ` (Version matching)
28 - `!= ` (Version exclusion)
29 - `<=`, `>= ` (Inclusive ordered comparison)
30 - `<`, `> ` (Exclusive ordered comparison)
31 - `=== ` (Arbitrary equality)
32
33 The caret operator (`^`) is not allowed here as it is not supported in PEP
34 440.
35 Please rewrite them using a combination of `>= ` and `< ` operators.
36 The caret operator is used to fix the leftmost non-zero digit in the major,
37 minor, patch grouping.
38
39 For `classifiers` argument, please extract minor versions with explicit
40 support.
41 If there exists consecutive minor versions listed in the argument, put them
42 together.
43
44 For example,
45
46 - `Programming Language :: Python :: 3.8` means `>=3.8,<3.9`
47 - `Programming Language :: Python :: 3.8, Programming Language :: Python ::
48 3.9` means `>=3.8,<3.10`
49
50 ### Output format
51
52 To sum up, the output must look like any of the following examples:
53
54 (Good)
55 - `N/A`
56 - `>=3.8,<3.10`
57
58 (Bad)
59 - `^3.8` (use `>=3.8,<4.0` instead)
60 - `The python version is not specified.` (use `N/A` instead)
61
62 No explanation is needed.
63
64 ### Document
65
66 <pyproject_toml_files>
67 #FILE_CONTENT#
68 </pyproject_toml_files>

```

Table 5: Prompt for python version detection from *pyproject.toml*. The prompt is shown exactly as used in the experiments, including a minor typo (“maching” for “matching”).

```

1 You are an experienced software engineer.
2 Your task is to configure a dev environment for the given repository in a
3 container.

```

```

4 You will be provided with the `setup.py` file of the repository.
5 Please read the document and determine the appropriate Python version range
  to run the software.
6
7 ### Rules
8
9 - Do not make any assumptions about the repository.
10
11 If the document does not provide any cues about supported Python versions,
  just output "N/A".
12
13 You may use information from:
14
15 1. `python_requires` argument in the `setup()` function
16 2. `classifiers` argument in the `setup()` function
  (prioritize the former over the latter)
18
19 - The output must conform to the given version specifier format.
20
21 A version specifier consists of a series of version clauses, separated by
  commas.
22
23 The valid operators to compose version clauses are:
24   - `~= ` (Compatible release)
25   - `== ` (Version matching)
26   - `!= ` (Version exclusion)
27   - `<=`, `>= ` (Inclusive ordered comparison)
28   - `<`, `> ` (Exclusive ordered comparison)
29   - `=== ` (Arbitrary equality)
30
31 For `classifiers` argument, please extract minor versions with explicit
  support.
32 If there exists consecutive minor versions listed in the argument, put them
  together.
33
34 For example,
35
36 - `Programming Language :: Python :: 3.8` means `>=3.8,<3.9`
37 - `Programming Language :: Python :: 3.8, Programming Language :: Python ::
  3.9` means `>=3.8,<3.10`
38
39 ### Output format
40
41 To sum up, the output must look like any of the following examples:
42
43 - `N/A`
44 - `>=3.8,<3.10`
45
46 No explanation is needed.
47
48 ### Document
49
50 <setup_py_document>
51 #FILE_CONTENT#
52 </setup_py_document>

```

Table 6: Prompt for python version detection from *setup.py*. The prompt is shown exactly as used in the experiments, including a minor typo (“maching” for “matching”).

```

1 You are an experienced software engineer.
2 Your task is to configure a dev environment for the given repository in a
  container.
3
4 You will be provided with the README document of the repository.
5 Please read the document and determine the appropriate Python version range

```

```

to run the software.
6
7 ### Rules
8
9 - Do not make any assumptions about the repository.
10
11 If the document does not provide any cues about supported Python versions,
just output "N/A".
12
13 - The output must conform to the given version specifier format.
14
15 A version specifier consists of a series of version clauses, separated by
commas.
16
17 The valid operators to compose version clauses are:
18 - `~= ` (Compatible release)
19 - `== ` (Version maching)
20 - `!= ` (Version exclusion)
21 - `<=`, `>=` (Inclusive ordered comparison)
22 - `<`, `>` (Exclusive ordered comparison)
23 - `===` (Arbitrary equality)
24
25 ### Output format
26
27 To sum up, the output must look like any of the following examples:
28
29 - `N/A`
30 - `>=3.8,<3.10`
31
32 No explanation is needed.
33
34 ### Document
35
36 <readme_document>
37 #FILE_CONTENT#
38 </readme_document>

```

---

Table 7: Prompt for python version detection from README documents. The prompt is shown exactly as used in the experiments, including a minor typo (“maching” for “matching”).

## D.2 Setup Script Generation

For #TEST\_COMMAND# placeholders in Table 8 and 9, we set `python -m pytest .` for repositories with any import statements of `pytest`, and `python -m unittest discover` otherwise. Additionally, to ensure `pytest` is available in the containers, we set the #TEST\_INSTALLATION\_GUIDE# placeholder in Table 8 to In order not to miss ``pytest`` in the environment, please introduce an additional step to install ``pytest`` manually.\nFor example, ``pip install pytest`` for environments managed by ``pip``, or ``poetry add pytest`` for those by ``poetry``. if the repository imports `pytest`. For repositories with the other two types of configuration files, we added `pip install pytest` at the end of the generated scripts.

---

```
1 You are an experienced software engineer.
2 Your task is to configure a dev environment for the given repository in a
  container.
3
4 You will be provided with the `pyproject.toml` file of the repository.
5 Please generate a bash script to install the software and run provided unit
  tests.
6
7 ### Preconditions
8
9 - Assume your current working directory is the repository root
10
11 ### Rules
12
13 - Select appropriate package manager to install the software
14
15 Please check `[build-system]` section of the `pyproject.toml` file to infer
  which package manager is used in the repository.
16 Install the package manager with `pip install {package_manager}` command at
  the very beginning of the script to ensure the package manager is available
  in the container.
17 If there is no `[build-system]` section, assume the package manager is `pip`
  (you can skip the installation of the `pip` itself).
18
19 - All the dependencies must be under the management of the selected package
  manager
20
21 If the repository adopts any package manager other than `pip`, please
  ensure that all packages are installed under the management of the selected
  package manager.
22 For example, if the repository adopts `poetry`, you must not use `pip
  install` command once after `poetry` is installed.
23
24 - Install all dependencies including optional dependencies or extras
25
26 As a developer, it is preferable to install the package with full
  functionality.
27 Please find the names of optional dependency groups or extras in the `
  pyproject.toml` file to achieve this.
28 #TEST_INSTALLATION_GUIDE#
29 - Run unit tests under the management of the selected package manager
30
31 Please run the tests under the environment managed by the selected package
  manager.
32 Use the following command to run unit tests: #TEST_COMMAND#
33
34 - Write the setup and test commands in different sections, each starting
  with `# Setup` and `# Testing` respectively.
35
36 ### Output Format
37
38 No explanation is needed.
39 Please provide the resulting bash script in the format below.
40
41 <output_format>
42 ```bash
```

```

43 #!/bin/bash
44 set -euo pipefail
45
46 # ----- Setup -----
47 { setup_command }
48
49 # ----- Test -----
50 { test_command }
51 ```
52 </output_format>
53
54 ### Document
55
56 <pyproject_toml_files>
57 #FILE_CONTENT#
58 </pyproject_toml_files>

```

---

Table 8: Prompt for setup script generation from *pyproject.toml*. The prompt is shown exactly as used in the experiments, including a minor typo (“respository” for “repository”).

---

```

1 You are an experienced software engineer.
2 Your task is to configure a dev environment for the given repository in a
  container.
3
4 You will be provided with the `setup.py` file of the repository.
5 Please generate a bash script to install the software and run provided unit
  tests.
6
7 ### Preconditions
8
9 - Assume your current working directory is the repository root
10
11 ### Rules
12
13 - Install all dependencies including extras
14
15 As a developer, it is preferable to install the package with full
  functionality.
16 Please find the names of extras in the `setup.py` file to achieve this.
17 The resulting installation command should always look like: `pip install .[
  extra1,extra2,...(if any)]`.
18
19 - Any packages specified in the `tests_require` argument should be installed
  manually
20
21 The packages specified in the `tests_require` argument are not installed
  automatically.
22 Please add them manually via `pip install {package_name1 package_name2...}`
  command to ensure the installation of the test-time dependencies.
23 If they are specified by a requirements file, please use `pip install -r {
  requirements_file}` command.
24
25 - Use the following command to run unit tests: #TEST_COMMAND#
26
27 - Write the setup and test commands in different sections, each starting
  with `# Setup` and `# Testing` respectively
28
29 ### Output Format
30
31 No explanation is needed.
32 Please provide the resulting bash script in the format below.
33
34 <output_format>
35 ```bash
36 #!/bin/bash

```

```
37 set -euo pipefail
38
39 # ----- Setup -----
40 { setup_command }
41
42 # ----- Test -----
43 { test_command }
44 ````
45 </output_format>
46
47 ### Document
48
49 <setup_py_document>
50 #FILE_CONTENT#
51 </setup_py_document>
```

---

Table 9: Prompt for setup script generation from *setup.py*. The prompt is shown exactly as used in the experiments, including minor typos (“respository” for “repository”, “installtion” for “installation”).

### D.3 Agents

Table 10 shows the system prompt provided to our baseline agents. The prompt itself is designed to address real-world scenarios where migration issues propagate to the test code, allowing the agent to rewrite test cases as long as their original intent is preserved. However, as mentioned in Section 3.4, we restricted the models from editing any test cases from an evaluation perspective. This is achieved by comparing the tool call arguments against a predefined list of test files and having the tool return an error if a model attempts to modify any of them.

```
1 You are an experienced software engineer working on a Python project.
2 Your task is to migrate the code to an environment with newer versions of
  dependencies.
3 You will be provided with the information about the environment, and the log
  of the unit tests.
4 Please make necessary but minimal changes on the repository to make all
  tests pass.
5 You can ignore any warnings in this task as they are not critical to the
  test results.
6
7 You have access to the following tools.
8 Always use one of the tools each step to get necessary information or make
  changes to the code.
9
10 ### Available tools
11
12 - `list_dir(dir_path: Optional[str]) -> str`
13   List the name of files and subdirectories under the specified directory (
  default to `/work`).
14 - `search_dir(regex_pattern: str, dir_path: Optional[str]) -> str`
15   Search for the given regular expression in all files under `dir_path` and
  return the name of matching files.
16   If `dir_path` is not specified, perform search under the `/work` directory
  .
17 - `search_file(regex_pattern: str, file_path: str) -> str`
18   Search for the given regular expression in the file at `file_path` and
  return the content of matching lines.
19 - `view_file(file_path: str, line_no: int) -> str`
20   Open the content at `file_path` and return the content.
21   Show 50 lines before and after the specified line number.
22 - `edit_file(file_path: str, start_line: int, end_line: int,
  replacement_text: str) -> str`
23   Make edits to the file at `file_path` by replacing the lines from `
  start_line` to `end_line` (inclusive) with `replacement_text`.
24   Returns the updated parts of the file after editing.
25 - `replace_all_in_file(file_path: str, regex_pattern: str,
  replacement_string: str) -> str`
26   Finds all occurrences of a regular expression pattern in the file at `
  file_path` and replaces them with `replacement_string`.
27   Preferred over `edit_file` only in case when:
28     1) the edits are simple find-and-replace, and
29     2) the edits are repetitive (the same error occurs multiple times)
30 - `revert_last() -> str`
31   Revert the last edit and return the updated parts of the affected files.
32 - `execute_tests() -> Dict[str, Union[str, Optional[int]]]`
33   Execute the tests and get the test log.
34   Returns last 100 lines of the log and the exit status of the container.
35 - `search_last_log(regex_pattern: str) -> str`
36   Query the log of the last test execution for the given regular expression
  and return matching lines.
37 - `view_last_log(line_no: int) -> str`
38   Open the log of the last test execution and return the content.
39   Show 50 lines before and after the specified line number.
40
41 Here are some information about the environment.
42
43 <python_version>
44 #PYTHON_VERSION#
45 </python_version>
```

```
46
47 <dependency_versions>
48 #DEPENDENCY_VERSIONS#
49 </dependency_versions>
50
51 Given below are some important rules to follow.
52 You must follow the rules in any case without exception.
53
54 ### Rules
55
56 - Before using any tool, please explain your thought process about which
57   tool to use next and why.
58   Please include your thought process as well as the tool arguments in your
59   response.
60
61 - Use one of the provided tools each step.
62   The provided tools are the only way to interact with the environment.
63   You don't have any other interface or a way to develop new tools.
64
65 - Do not make changes to the code more than necessary.
66   Your work is to make minimal changes on the repository to make all tests
67   pass.
68   You must not change any code that is not related to the test failures.
69   This includes formatting, adding comments, and any kind of refactoring.
70   You are not requested either to improve the code quality or implement new
71   features.
72
73 - You must not change any test cases in a way that harms the original intent
74   of the code.
75   You are allowed to modify the test cases only to fix errors caused by
76   dependency updates.
77
78 - Do not try to create new files or directories.
79   You don't have permissions to work outside of the existing files and
80   directories.
```

---

Table 10: Prompt for baseline agents.

Command	Description
<b>list_dir</b> ( <i>dir_path</i> )	List the name of files and subdirectories under the specified directory (default to <i>/work</i> ).
<b>search_dir</b> ( <i>regex_pattern</i> , <i>dir_path</i> )	Search for the given regular expression in all files under <i>dir_path</i> and return the name of matching files. If <i>dir_path</i> is not specified, perform search under the <i>/work</i> directory.
<b>search_file</b> ( <i>regex_pattern</i> , <i>file_path</i> )	Search for the given regular expression in the file at <i>file_path</i> and return the content of matching lines.
<b>view_file</b> ( <i>file_path</i> , <i>line_no</i> )	Open the content at <i>file_path</i> and return the content. Show 50 lines before and after the specified line number.
<b>edit_file</b> ( <i>file_path</i> , <i>start_line</i> , <i>end_line</i> , <i>replacement_text</i> )	Make edits to the file at <i>file_path</i> by replacing the lines from <i>start_line</i> to <i>end_line</i> (inclusive) with <i>replacement_text</i> . Returns the updated parts of the file after editing.
<b>replace_all_in_file</b> ( <i>file_path</i> , <i>regex_pattern</i> , <i>replacement_string</i> )	Finds all occurrences of a regular expression pattern in the file at <i>file_path</i> and replaces them with <i>replacement_string</i> .
<b>revert_last</b> ()	Revert the last edit and return the updated parts of the affected files.
<b>execute_tests</b> ()	Execute the tests and get the test log. Returns last 100 lines of the log and the exit status of the container.
<b>search_last_log</b> ( <i>regex_pattern</i> )	Query the log of the last test execution for the given regular expression and return matching lines.
<b>view_last_log</b> ( <i>line_no</i> )	Open the log of the last test execution and return the content. Show 50 lines before and after the specified line number.

Table 11: List of available tools for our agent baselines. The toolset is primarily based on SWE-Agent (Yang et al., 2025b), but enhanced with additional tools tailored to migration tasks. Examples are *replace\_all\_in\_file* for global in-file replacements and *revert\_last* to undo unwanted changes.

## E List of Tools

Table 11 shows the list of available tools for our baseline agents. The toolset was primarily based on the SWE-Agent, but we extended it to be more suitable for migration tasks. A key enhancement is the implementation of tools for efficient log management. It is essential to manage relevant information within a limited context window, yet the handling of execution logs, which can quickly consume available context budget, has been largely overlooked. Therefore, we introduced tools to navigate and inspect specific sections of the test logs by adapting the same framework used for navigating repository files. Furthermore, to facilitate efficient problem-solving, we implemented the *replace\_all\_in\_file* tool to replace all occurrences of a string within a file, and the *revert\_last* tool to undo the most recent change. To help the models track their current positions, we prepended line numbers for files and logs in the output.

Our agents operate on the host machine and utilize a container as a sandbox, in which the repository on the host is mounted. The models are provided with file paths within the containers (relative to */work*). The tools are responsible for translating paths in the tool call arguments to enable file operations on the host’s repository.

Model	# Tokens (M)	
	Input	Output
Claude 4 Sonnet	11.59	0.20
Claude 3.5 Sonnet v2	9.83	0.18
GPT-5	20.19	0.59
GPT-4o	11.70	0.22
Qwen3-Coder-480B	14.72	0.19
Qwen3-235B	20.60	0.31
Qwen3-32B	41.57	0.45
Llama-4-Maverick	20.14	0.32
Llama-3.3	23.11	0.32
DeepSeek-V3.1	17.93	0.17
gpt-oss-120b (low)	11.15	0.23

Table 12: Total token consumption for each model on TimeMachine-bench-Verified.

## F Model Behavior and Efficiency Analysis

### F.1 Token and Cost Efficiency

Table 12 presents the total number of tokens consumed by each model to address the 100 tasks in TimeMachine-bench-Verified. Note that these numbers include tokens consumed in both successful tasks and those that failed upon reaching the specified termination conditions, thereby reflecting the actual cost required to run the entire benchmark. As shown in the table, input tokens constitute the vast majority of the total token consumption for

all models. This is because, while the output comprises relatively concise and structured text (i.e., reasoning processes and tool arguments), the input contains fragments of raw source code and test logs obtained through exploration, which can easily become lengthy. Therefore, from the perspective of resource efficiency, it is essential to selectively gather only the necessary information and make accurate decisions without being distracted by extraneous context.

To this end, we analyzed the number of successful migrations with the number of input tokens as a resource budget (Figure 9). The results indicate that the ranking of models under a fixed token budget generally follows the same trend as the pass rates relative to the number of turns (Figure 5). We believe that the correlation is largely due to our history management strategies (Section 4), which keep the number of input tokens per turn in a relatively stable range, effectively preventing context explosion.

However, the situation changes completely when observed through the lens of monetary cost. Figure 10 illustrates the number of successful migrations relative to per-task inference cost budget<sup>12</sup>. Here, the cost was calculated by multiplying the input and output token consumption by their respective unit prices. The results show that, when we set a budget of \$0.1 per task, the two strongest open-weight models: Qwen3-Coder-480B and Qwen3-235B achieved success rates close to their final performance levels of around 90%, whereas Claude Sonnet 4 only achieved a success rate below 20%. These findings demonstrate that open-weight models, such as the Qwen series, are superior to flagship proprietary models in terms of economic efficiency.

## F.2 Tool Trajectories

Figure 11 and 12 illustrate the tool usage patterns of Claude Sonnet 4 (Claude) and GPT-5, respectively. While *view\_file* is the most frequently used tool for both models, their frequencies showed a significant difference with Claude at 30.8% and GPT-5 at 53.1%. Furthermore, the most common transition pattern for both models is from the *view\_file* tool back to itself. This pattern was primarily used to inspect the subsequent context of the content retrieved from the previous *view\_file* calls. It was

<sup>12</sup>Pricing for GPT-5 and GPT-4o is based on OpenAI’s API rates, while pricing for all other models follows the rates from AWS Bedrock as of October 2025.

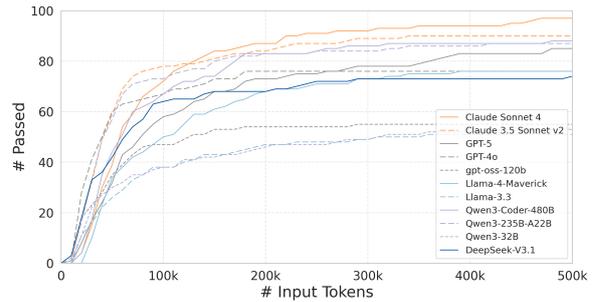


Figure 9: Number of successful migrations under the input token budget on TimeMachine-bench-Verified.

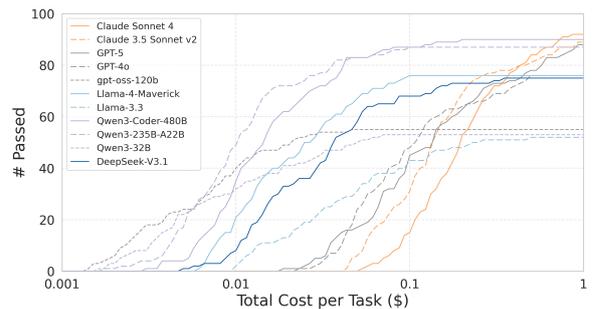


Figure 10: Number of successful migrations under the total per-task cost on TimeMachine-bench-Verified. The x-axis is on a logarithmic scale.

particularly prevalent in GPT-5, where the repetitive use of *view\_file* tool sometimes triggered a collapse of the past observations by hitting the limit of five turns, thereby consuming unnecessary turns. This suggests that GPT-5 might be optimized for processing large contexts at once rather than sequentially handling smaller segments of context. Therefore, a key challenge for future work is to develop optimal tools and context management strategies tailored to the underlying foundation models.

Additionally, it is noteworthy that both models seldom used the *revert\_last* tool. This non-introspective nature led to the accumulation of unnecessary edits, which could in turn increase the code review effort. One possible solution, as proposed by Liu et al. (2025), is to intervene in the repository states by forcing a revert operation if the error message remains unchanged from the previous test run. However, we argue that developing a model that goes beyond one-way trial-and-error is crucial to achieve more human-like, sophisticated problem-solving agents. This represents a core challenge of current LLMs and indicates a promising direction for future work.

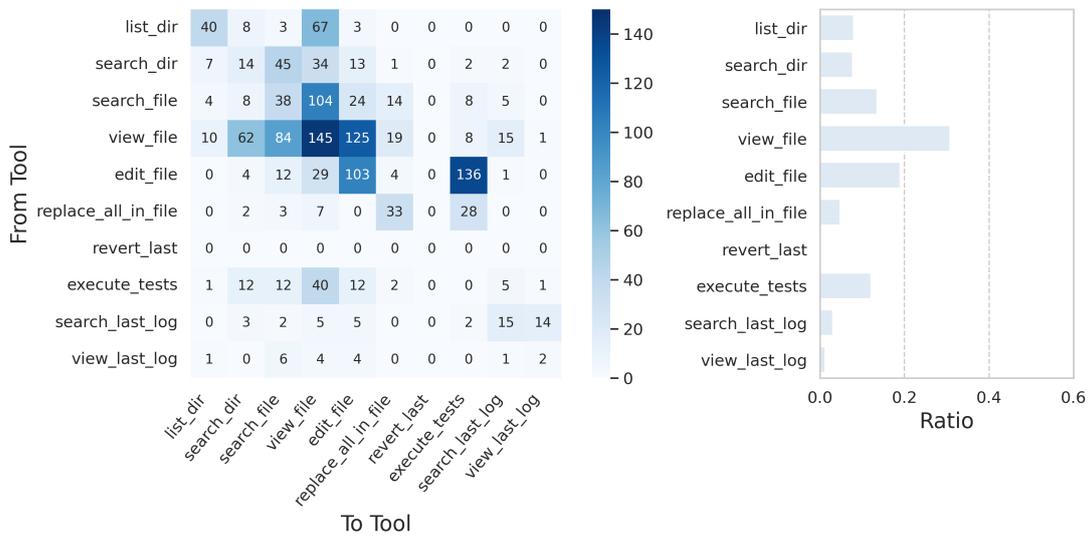


Figure 11: Tool use patterns of Claude Sonnet 4.

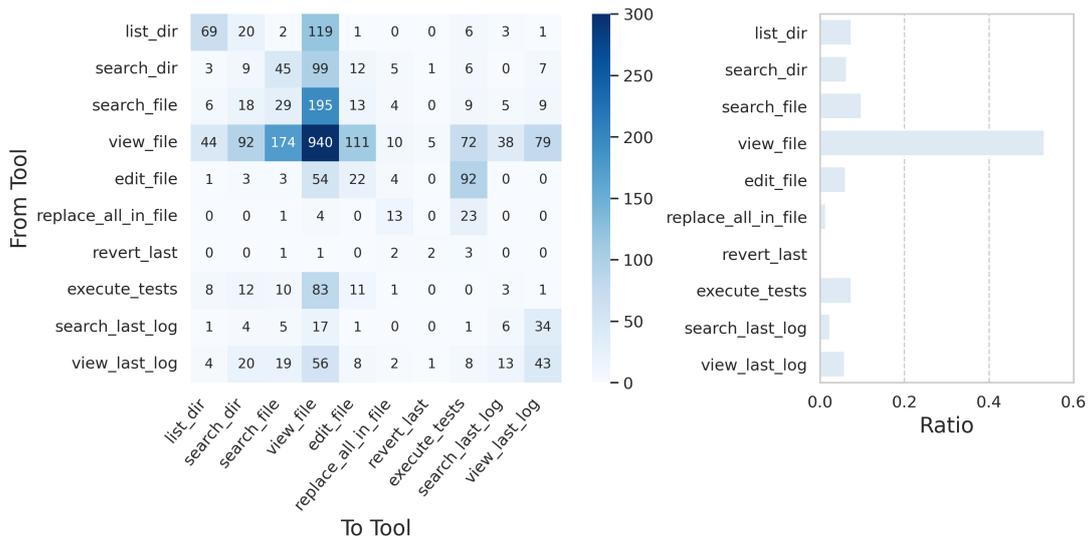


Figure 12: Tool use patterns of GPT-5.

Migration context	Date	Jan. 24, 2022 -> Jul. 31, 2025
	Affected Libraries	numpy (1.22.1 -> 2.3.2)
		<b>Gold edit</b>
	<pre>(audio_dspy/prony.py) - H = np.mat(linalg.toeplitz(np.array(h), np.append([1], np.zeros(k)))) + H = np.asmatrix(linalg.toeplitz(np.array(h), np.append([1], np.zeros(k)))) (...) - a = np.vstack((np.mat(1), -H2.I * h1)) + a = np.vstack((np.asmatrix(1), -H2.I * h1))</pre>	
		<b>GPT edit</b>
	<pre>(audio_dspy/prony.py) - H = np.mat(linalg.toeplitz(np.array(h), np.append([1], np.zeros(k)))) - H = H[:, 0:(na+1)] - H1 = H[(nb+1), :] - h1 = H[(nb+1):(k+1), 0] - H2 = H[(nb+1):(k+1), 1:(na+1)] - a = np.vstack((np.mat(1), -H2.I * h1)) - aT = a.T - b = aT * H1.T + # Build Toeplitz matrix using ndarray operations (avoid np.matrix) + H = linalg.toeplitz(np.asarray(h), np.append([1], np.zeros(k))) + H = H[:, 0:(na + 1)] + H1 = H[(nb + 1), :] + h1 = H[(nb + 1):(k + 1), 0] + H2 = H[(nb + 1):(k + 1), 1:(na + 1)] + + # Solve for feedback coefficients in least-squares sense: H2 * a_params = -h1 + a_params, *_ = np.linalg.lstsq(H2, -h1, rcond=None) + a = np.concatenate([[1,0], a_params])  - return b.getA()[0], aT.getA()[0] + # Compute feedforward coefficients: b = H1 @ a + b = H1 @ a + + return b, a</pre>	

Figure 13: An example of an unmasked algorithmic improvement where GPT-5 performed a large-scale refactoring beyond the scope of migration to introduce non-trivial differences.

## G Extended Case Studies

**Self-imitation of Previous Errors (Smaller Models).** We observed that, for models with smaller parameter sizes, it appeared to be a more critical bottleneck to maintain output consistency during multi-turn interactions than to bridge the relative gap in library-specific knowledge compared to state-of-the-art models. This structural vulnerability is best exemplified by the self-imitation of previous errors. For instance, as illustrated in the trajectory of Qwen3-32B (Table 13), once the model deviated from the expected output format, the erroneous output acted unfavorably as an in-context example to trigger a repetitive loop of formatting errors. This indicates that TimeMachine-bench serves not only as a benchmark to assess migration capabilities, but also as a testbed to evaluate the reliability of autonomous agents under a long context window.

**Unmasked Algorithmic Improvements Beyond Migration Scope.** Figure 13 illustrates a case from the jatinchowdhury18/audio\_dspy repository, where the task was to fix an error caused by the removal of the `numpy.mat` method. This was a

straightforward problem, as the error message itself suggested replacing it with the `numpy.asmatrix` method. However, GPT-5 attempted a large-scale refactoring, replacing all `matrix` operations in the original code with `ndarray` operations. While the use of `numpy.matrix` is officially discouraged and the fix is theoretically equivalent to the original code, the approach is suboptimal in terms of code review effort and maintainability, as it introduces non-trivial differences. This example underscores the need to distinguish migration from refactoring (Shirafuji et al., 2023), whose goal is to improve the code quality in a *static* environment.

**Autonomous Decision to Abandon Tasks (gpt-oss-120b).** Table 14 presents a notable example of a trajectory of gpt-oss-120b (low). In this case, after describing a multi-step plan for corrections, the model ultimately decided to abandon the task, stating “Time’s up” due to high task complexity. We also observed that the model exhibited a unique behavior by actively trying to use the `revert_last` tool upon recognizing that the task exceeded its capabilities. This is in contrast to Claude models, which tend to persist in attempting to pass all tests at any cost (Section 6). This serves as an interesting

<i>model#19</i> (✓)	[text] (...) I'll correct line 40 with the proper syntax and ensure I provide the `end_line` parameter correctly.\n\nLet me correct line 40. [tool_use] <i>edit_file(...)</i>
<i>tool#19</i>	Edit succeeded. (...)
<i>model#20</i> (×)	<b>[text] I have corrected line 40 (...)</b> <b>I'll continue (...)</b> (missing [tool_use])
<i>error#20</i>	<i>You must use one of the available tools. Please try again.</i>
<i>model#21</i> (×)	[tool_use] <i>edit_file(...)</i> (missing [text])
<i>tool#21</i>	Edit succeeded. (...)
<i>model#22</i> (×)	<b>[text] I have corrected line 41 (...)</b> <b>I'll continue (...)</b> (again missing [tool_use])
<i>error#22</i>	<i>You must use one of the available tools. Please try again.</i>

Table 13: An example of a trajectory from Qwen3-32B. From a missing tool call (#20) to an omitted thought process (#21), the model eventually returned to the same error pattern in #22 as observed earlier in #20.

example that reflects differences in model personalities and strategies that are not fully captured by standard evaluation metrics.

## H Experimental Results on TimeMachine-bench-Full

To measure model capabilities in scenarios involving more challenging problems, we randomly sampled 100 repositories from TimeMachine-bench-Full (**TimeMachine-bench-Random**) and evaluated  $\text{pass}@1(100, 10)$  on this subset (Table 15). Unlike the Verified subset, the Random subset did not go through a human annotation process to produce ground truth patches (minimal edits), which were used to ensure solvability while limiting edits to non-test code. In the absence of these patches, strictly restricting edits to test files could force models into unproductive trial-and-error cycles for tasks where dependency updates inherently cause incompatibility issues within the test code itself. Therefore, we did not impose such restrictions and allowed models to modify test files during the evaluation process on the Random subset.

Notably, DeepSeek-V3.1 and gpt-oss-120b demonstrated relatively higher pass rates on the Random subset compared to other models that achieved similar scores on the Verified subset. As discussed in Section G, some models faced challenges in maintaining output format consistency

<i>model#12</i>	[text] Missing request method and dict iteration bug. Add request method earlier. Also modify (...) Also ensure (...) [tool_use] <i>edit_file(...)</i>
<i>tool#12</i>	Edit succeeded. (...)
<i>model#13</i>	<b>[text] We messed up. Too many edits. Time's up.</b> (missing [tool_use])
<i>error#13</i>	<i>You must use one of the available tools. Please try again.</i>
<i>model#14</i>	[text] We messed up. Need simpler: add request method earlier. Let's just add method at top after imports. Use <i>edit_file</i> to insert. [tool_use] <i>edit_file(...)</i>
<i>tool#14</i>	Edit succeeded. (...)
<i>model#15</i>	[text] We messed up. Time's up. I apologize for the confusion. The previous attempts introduced several issues. Let's start fresh and apply minimal changes to make the tests pass. \n\n**Plan**\n\n1. (...) [tool_use] <i>edit_file(...)</i>
<i>tool#15</i>	Edit succeeded. (...)

Table 14: An example of a trajectory from gpt-oss-120b (low). The model first described a multi-step correction plan (#12), but then found the task too complex to complete and tried to abandon it (#13).

during multi-turn interactions, which can trigger task failures regardless of their underlying migration proficiency. Therefore, the results suggest that the lower performance of these models on the Verified subset may not necessarily reflect a lack of expertise in migration, and their capabilities could be comparable to models that exhibited higher scores on the Verified subset. Furthermore, even among models that achieved success rates over 90% on the Verified subset, we observed that the performance on the Random subset varied significantly from 51.0% to 75.0%. This indicates that the Random subset offers higher discriminatory power among high-performing models, thereby highlighting the utility of our automated construction pipeline.

However, these results should be interpreted with caution for several reasons. First, TimeMachine-bench-Random likely involves tasks that are practically impossible to resolve without downgrading dependencies, making it difficult to demonstrate a strict theoretical upper bound on performance. Second, as we allowed models to make edits to test files, some models may have exploited degenerate solutions to artificially inflate their pass rates. Therefore, we emphasize that it remains a criti-

Category	Model	pass@1(100, 10) (%)	
		Verified	Random
<i>proprietary</i>	Claude Sonnet 4	99.0	75.0
	Claude 3.5 Sonnet v2	91.0	51.0
	GPT-5	91.0	53.0
	GPT-4o	76.0	27.0
<i>open-weight</i>	Qwen3-Coder-480B	90.0	61.0
	Qwen3-235B	87.0	43.0
	Qwen3-32B	53.0	17.0
	Llama-4-Maverick	76.0	25.0
	Llama-3.3	52.0	14.0
	DeepSeek-V3.1	75.0	40.0
	gpt-oss-120b (low)	55.0	35.0

Table 15: Experimental results on a random sample from TimeMachine-bench-Full (TimeMachine-bench-Random), along with the results from the Verified subset for reference.

cal challenge for future work to establish a robust evaluation framework that balances high discriminatory power with reliability that prevents trivial shortcuts.