

# SWE-Dev: Building Software Engineering Agents with Training and Inference Scaling

Haoran Wang<sup>1\*</sup>, Zhenyu Hou<sup>1\*</sup>, Yao Wei<sup>2</sup>, Jie Tang<sup>1</sup>, Yuxiao Dong<sup>1</sup>

<sup>1</sup>Tsinghua University    <sup>2</sup>Zhipu AI

 Code     Models & Data

## Abstract

Large language models (LLMs) have advanced rapidly from conversational problem solving to addressing real-world tasks involving tool use, such as software engineering (SWE). Recent LLM-powered SWE systems, such as OpenAI Codex and Cursor, have offered end-to-end automation of the software development process. However, building effective SWE agents remains challenging due to the lack of high-quality training data and reliable test-time evaluation. To address this issue, we present SWE-DEV, an SWE agent built upon open-source LLMs, with a focus on training and inference scaling. For training scaling, we develop a robust pipeline to synthesize test cases and scale up agent trajectories to construct the training data. For inference scaling, we increase the interaction budget within a single run to enable further thinking within one independent attempt. Experiments on the SWE-bench-Verified benchmark show that the SWE-DEV models can achieve top performance among all open SWE agents. Specifically, the resolve rate of our 7B and 32B models reach 23.4% and 36.6%, respectively, outperforming state-of-the-art open-source models. All code, models, and datasets are publicly available at <https://github.com/THUDM/SWE-Dev>.

## 1 Introduction

Large language models (LLMs) have rapidly evolved from generating simple code snippets to tackling more complex tasks, such as competitive programming (Li et al., 2022; OpenAI, 2025), machine learning problems (Chan et al., 2024), and real-world software engineering (SWE) tasks (Xi et al., 2024; Jimenez et al., 2024; Zan et al., 2025). Among these tasks, SWE is particularly hard and challenging (OpenAI, 2025; Cursor, 2024), but highly useful for improving real-world productivity. Unlike simple code generation, SWE tasks usually

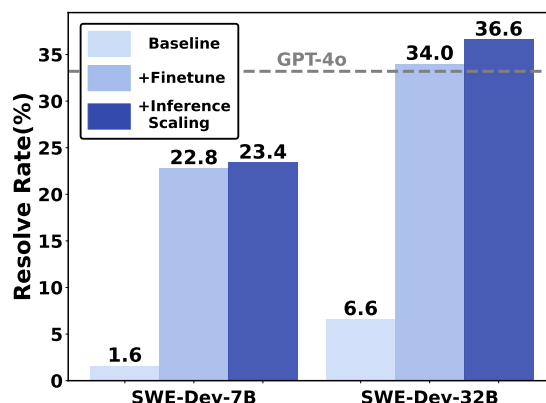


Figure 1: The SWE-DEV performance with training and inference scaling. Notably, SWE-Dev-32B achieves a resolve rate of 34.0%, matching the performance of GPT-4o even without inference scaling.

require LLMs to interact with complex and fragile runtime environments, solve toolchain issues, execute scripts, and reason over large, interdependent codebases (Ma et al., 2024b).

The SWE tasks are usually evaluated on benchmarks such as SWE-bench (Jimenez et al., 2024) and its recent multimodal extensions (Yang et al., 2024b; Zan et al., 2025). These benchmarks require models to generate verifiable, test-passing solutions on real-world codebases. To achieve this, the models must be capable of step-by-step reasoning, tool using, and long-horizon planning.

To date, training models to handle these tasks requires reliable reward signals, typically derived from test cases that validate the correctness of the solutions. However, most existing datasets lack such test cases or executable environments (Chen et al., 2021), making it difficult to evaluate solutions or provide useful feedback during training. This deficiency limits models’ abilities to iteratively refine their outputs through trial-and-error, thus constraining their potential to solve real-world SWE tasks that need grounded, verifiable solutions.

To address this issue, we introduce SWE-DEV, an open-source SWE agent framework coupled

\*Work partially done when interned at Zhipu AI.

with a scalable test case generation pipeline. This pipeline works in two-stages: First, the LLMs are used to generate Gherkin-style descriptions—a structured natural language format used to specify test scenarios. Second, a code generator outputs code patches for further validation. Empirical analysis shows that these synthesized test cases align closely with the original problem semantics.

Through large-scale experiments, we identify a clear training scaling trend: increasing the number of sampled agent trajectories leads to improved downstream performance. To enhance efficiency, we propose an LLM-based filtering method that selects high-quality trajectories. This allows us to maintain the benefits of a full dataset while retaining only the most valuable data.

We also propose iteration scaling—a simple yet effective strategy that aims to scale up inference budget by increasing the number of interaction rounds during a single evaluation episode. This reduces the need for repeated re-evaluation, making it particularly advantageous in scenarios where access to test oracles is expensive or delayed.

In addition, we explore advanced post-training strategies, including Rejection Sampling Fine-Tuning (RFT) (Yuan et al., 2023), KTO (Ethayarajh et al., 2024), and OREO (Wang et al., 2024a). We observe that RFT brings the most significant performance improvement, while offline reinforcement learning (RL) methods—KTO and OREO—deliver marginal or task-specific gains.

We build SWE-DEV based on the open Qwen2.5-Coder (Hui et al., 2024), Llama-3.1 (Llama, 2024), and GLM4-9B (GLM, 2024) models. Its performance is evaluated on the SWE-bench-Verified benchmark, with results shown in Figure 1. With the Qwen2.5-Coder-32B model, SWE-Dev achieves a resolve rate of 36.6% on SWE-Bench-Verified, representing state-of-the-art performance among open-source SWE agents.

1. **Test Case Generation Pipeline.** We build a scalable LLM-based pipeline to generate real-world SWE instances and their executable test cases. Through this pipeline, we successfully construct 2,000 test cases by filtering 38,000 high-quality issues across 4,000 repositories.
2. **Scaling Trends in Data and Inference.** We empirically identify scaling trends between training data volume, number of interaction steps, and model performance. We find

that the SWE-Dev-32B model improves from 34.0% to 36.6% resolve rate by adding 45 more interaction turns, highlighting the benefit of multi-step execution for agents.

3. **Post-Training Recipe for SWE Agents.** We examine several post-training methods, including RFT, KTO, and OREO, as well as hybrid combinations. RFT consistently outperforms others by effectively leveraging high-quality samples, demonstrating its robustness and scalability for training SWE agents.

## 2 Related Works

**SWE Dataset Construction.** Prior benchmarks like SWE-bench (Jimenez et al., 2024), DevEval (Li et al., 2024b), EvoCodeBench (Li et al., 2024a) and Commit0 (Zhao et al., 2024) crawl from a large dataset and manually annotate the desired instance, which is labor consuming. There are works trying to filter existing fail-to-pass test cases like SWE-Gym (Pan et al., 2024), which annotates 2.4k instances by hand. Filtering out from a large dataset (Golubev et al., 2024) is effective but will leave out many useful issues only because the corresponding PR does not contain test cases. Therefore, automatically generating test cases from issue descriptions becomes essential.

**SWE Agentic Framework.** SWE agent framework focuses on two mainstream types: Interaction-based frameworks and pipeline-based frameworks. Interaction-based frameworks like OpenHands (Wang et al., 2024c), SWE-Agent (Yang et al., 2024a), Learn-By-Interact (Su et al., 2025) and Wang et al. (2024b) usually pre-define a set of agent-computer interfaces (ACIs) to help model manipulate the environment. Meanwhile, pipeline-based frameworks will design the whole process into several steps, like Agentless (Xia et al., 2024), CoreR (Chen et al., 2024), MarsCode (Liu et al., 2024) and SuperCoder2.0 (Gautam et al., 2024). The agent will generate patches by employing fault localization, patch generation and major voting. There are also techniques during inference time like MCTS (Antoniades et al., 2024) and critic-guided generation (Badertdinov et al., 2024) and model assembly (Zhang et al., 2024a). While pipeline-based process may benefit from specification, it cannot generalize to other coding tasks. On the contrary, interaction-based frameworks can do general tasks with natural language as instruction.

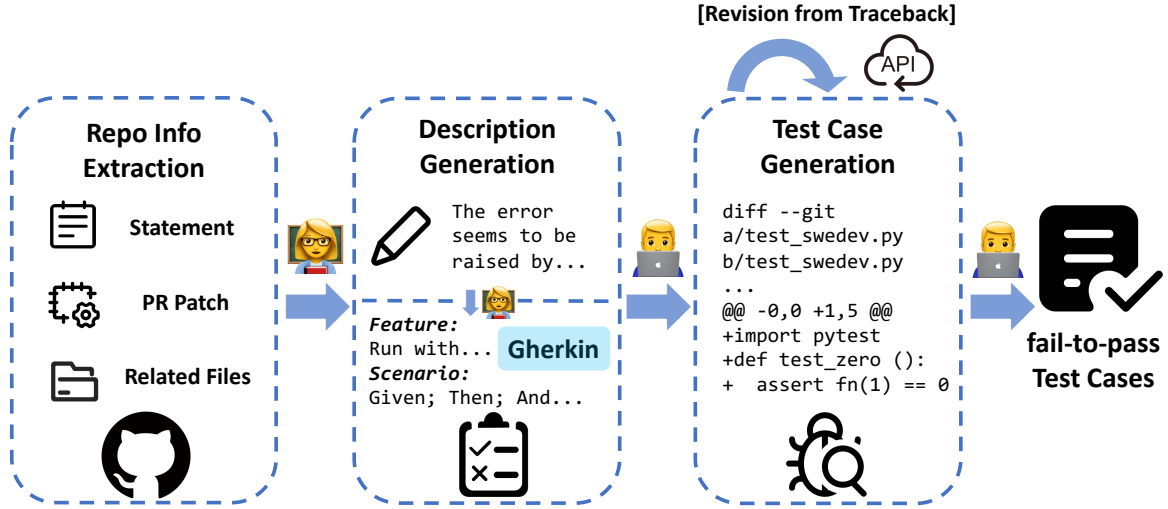


Figure 2: Pipeline for test case generation, divided into description generation and code generation phases. The pipeline begins with extracting repository information, followed by generating Gherkin scenarios and then detailed test cases. An optional revision step leverages traceback errors to refine the generated test cases. The final output includes fail-to-pass test cases.

### 3 SWE-Dev: Building Software Tasks at Scale

In this part, we developed a systematic strategy to build software tasks at scale. The core of SWE-DEV is to collect repositories and task instances at scale and then develop a LLM-based automated test-case generation approach. This scalable approach enables us to construct comprehensive training datasets for SWE tasks and enhance performance through test case-driven trajectory sampling as Figure 2 shows. Based on the systematic pipeline for scalable dataset construction and trajectory optimization, we build SWE-Dev agents.

#### 3.1 Instance Collection

We began by crawling metadata for 240k PyPI packages containing GitHub URLs, filtering for repositories with Stars  $\geq 5$  and PRs  $\geq 3$ , resulting in a subset of 59k repositories. Due to network constraints, machine capacity and intricate dependency management, we successfully downloaded 10,416 repositories. Following the methodology outlined in Jimenez et al. (2024) with minor modifications, we extracted 88k instances in total.

To refine this dataset, we applied rule-based filtering to adjust patch lengths to fit model context windows and eliminated trivial or irrelevant issues while maintaining diversity. Ultimately, we retained 38k instances from 4,413 repositories as the training set. As shown in Figure 3, over 4,000 repositories contain fewer than five instances, sig-

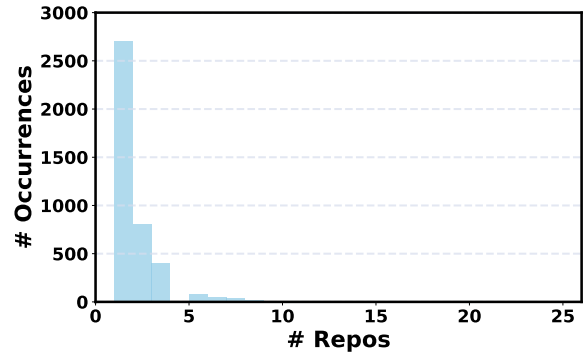


Figure 3: Distribution of instances per repository in the training dataset. The majority of repositories contribute fewer than five instances, highlighting the dataset’s diversity across a wide range of repositories.

nificantly enhancing the dataset’s diversity.

#### 3.2 Automatic Test Case Generation

While it is relatively straightforward to crawl PRs containing golden patches from the Internet, these patches often lack corresponding test cases. This absence prevents the validation of trajectory correctness and makes reinforcement learning techniques infeasible. To address this, we leveraged LLMs to generate test cases with necessary context.

In detail, our test case generation pipeline consists of four steps. The process begins with **extracting relevant information** from the codebase, including contextual code snippets and metadata, which serve as the foundation for subsequent steps. Next, the extracted information is used for **gener-**

**ating Gherkin descriptions**, where the model produces structured, high-level scenarios. These descriptions adopt the Gherkin syntax, using special keywords to provide structure and meaning, making the scenarios concise and easily interpretable. Following this, the descriptions are utilized for the **generation of test cases**, which leverages the contextual details to create robust and meaningful test cases. After, the generated test case can undergo an optional **revision phase**, where they are reviewed and refined to ensure correctness.

This phased approach reflects the inherent structure of how LLMs generate test cases. The model extracts a scenario from the provided context and, using the code snippet as a reference, produces the corresponding test case. Merging the first three steps into a one often results in irrelevant or incoherent test cases. By introducing fine-grained instructions, we provide clearer guidance, significantly improving the accuracy and contextual relevance of the generated outputs.

Due to computing budget, we split part of the dataset to synthesize test cases. From a dataset of 26k instances, we generated 2,097 fail-to-pass functions. Additionally, we integrated existing test cases from our dataset, resulting in a total of 4,630 test cases. Notably, most unsuccessful instances in our approach are due to environmental constraints rather than fundamental limitations of the methodology, further underscoring the feasibility of generating test cases via our approach.

### 3.3 Dataset Statistics

To evaluate the quality of our constructed dataset, we conducted three experiments: (1) pass rate analysis between description and code synthesis (2) comparison of filtered issues with open-source datasets (3) validation of test cases. Together, these experiments highlight the strengths of our dataset in generating high-quality test cases and its utility for downstream tasks.

These experiments collectively highlight the strengths of our dataset in generating high-quality test cases and its utility for downstream tasks.

**Pass Rate Analysis Between Description and Code Synthesis.** As shown in Table 1, we evaluated several models and a mixed approach that combines Llama for description generation and Qwen-Coder (trained primarily on code data) for code synthesis. The mixed approach significantly outperformed individual models in generating fail-to-

pass (F2P) functions and instances with test cases, highlighting the importance of leveraging domain-specific strengths for each task.

Model	#w/test	F2P	F2F	#w/F2P
Llama	757	185	1273	73
Mistral	767	151	990	81
Qwen	793	190	1185	83
Mix	828	237	1408	92

Table 1: Model Comparison for Test Case Synthesis. *#w/test* and *#w/F2P* denote the number of instances with test cases and with F2P test cases respectively. Models include Llama-3.1-70B-Instruct, Qwen-2.5-Coder-32B-Instruct, and Mistral-Large-Instruct-2407. *F2P* and *F2F* represent the total test functions generated during synthesis. The *Mix* model uses Llama for description generation and Qwen for code generation.

**Comparison with Open-Source Datasets.** We further validated the quality of our generated test cases by comparing models trained on issues from different datasets. As shown in Table 2, the inclusion of our generated test cases achieves quality comparable to existing open-source dataset (Badertdinov et al., 2024).

Dataset	Resolve Rate
Nebius/SWE-bench-extra	15.0
SWE-Gym/SWE-Gym	13.8
SWE-Dev (kept)	15.4
SWE-Dev (filtered)	12.4

Table 2: Performance Comparison Across Different Training Datasets. All trajectories are unevaluated. *SWE-Dev (kept)* refers to prompts retained after filtering, while *SWE-Dev (filtered)* corresponds to discarded prompts. *Random* means randomly selecting from trajectory pool.

**Validation of LLM-Generated Test Cases.** We compared the effectiveness of our dataset with other widely-used datasets, such as Nebius/SWE-bench-extra. As shown in Table 3, models trained on our dataset achieve comparable or slightly better performance than those trained on Nebius.

Using 900 trajectories for rejection sampling fine-tuning, models trained on Nebius achieved a slightly higher mean accuracy of 15.80, while our dataset achieved 15.87. Even with 500 trajectories, our dataset remained competitive with a mean accuracy of 15.00 compared to Nebius’s 14.87. This demonstrates that our custom test cases are also



Model	900-RFT	500-RFT
Nebius	15.80	14.87
SWE-Dev	15.87	15.00
Random	14.67	13.87

Table 3: Performance of RFT across different datasets. We choose a 7B base model with 13.6% resolve rate.  $k$ -RFT denotes fine-tuning with  $k$  correct trajectories. *Random* means sampling randomly from our pool. Each configuration was independently run three times.

effective in verifying whether trajectories comply with the requirement from issues.

By integrating fine-grained description generation, code generation, and execution-based validation, we ensure that the generated test cases are both accurate and contextually relevant. This approach not only bridges the gap of missing test cases in PR data but also significantly enhances the dataset’s utility for downstream training tasks.

## 4 Experiments

In this section, we will introduce our experiments on SWE agent tuning. As Table 4 shows, our models achieve state-of-the-art results among open-source models, with SWE-Dev-32B achieving a 36.6% resolve rate, increasing 30% resolve rate compared to our base model. Furthermore, SWE-DEV demonstrate competitive performance against closed-source models, narrowing the gap with leading systems such as OpenHands + CodeAct v2.1. These results highlight the effectiveness of our pipeline and training strategies.

We focus here on reporting key results from standard fine-tuning and reinforcement learning workflows. The impact of training and inference scaling will be discussed in the following section.

### 4.1 Experiment Setup

**Dataset and Agentic Scaffold.** We utilize the DeepSeek-V3 (DeepSeek-AI, 2024) for trajectory generation from both the Nebius and SWE-Dev datasets, and we take OpenHands, a ReAct-like (Yao et al., 2023) framework. After sampling, we obtained 17k unevaluated trajectories and 2.3k correct trajectories. Specifically, 914 of the correct trajectories originate from Nebius, while 78% of the unevaluated trajectories are derived from SWE-DEV. For the OpenHands configuration, we set `iteration=30` and `max_tokens=32k`. During of-line RL training, we incorporate trajectories from DeepSeek-V3 and on-policy models.

**Training Configuration.** We adopt the Qwen-2.5-Coder (Hui et al., 2024), Llama-3.1 (Llama, 2024) and GLM-4 (GLM, 2024) series as our base models and leverage the OpenRLHF (Hu et al., 2024) framework for training. For datasets with  $\leq 10k$  samples, we use a learning rate of  $1e-5$  and a batch size of 32; for larger datasets, we maintain the same learning rate but increase the batch size to 64. The training process spans 4 epochs for the 7B model and 8 epochs for the 32B model. To enable long-context training, we apply ring-attention (Liu et al., 2023) with 8 heads.

**Evaluation Metrics** For evaluation, we employ SWE-bench-Verified, a human-validated subset of 500 instances curated by OpenAI. This benchmark is carefully selected from SWE-bench (Jimenez et al., 2024), ensuring high-quality evaluation. Model-generated code patches are assessed using developer-written unit tests, with accuracy defined as the percentage of successfully resolved instances. For all experiments, we set `iteration=30` and `max_tokens=32k`, except for inference scaling experiments, where `max_tokens=160k` is used.

### 4.2 Main Results

Table 4 presents the comparison of resolve rates on the SWE-bench-Verified benchmark. We evaluate two major approaches: methods leveraging fine-tuning of open-source models and the direct utilization of proprietary models.

Compared to existing methods based on fine-tuning open-source models, SWE-DEV achieves the highest performance without employing any verifiers or sampling strategies. SWE-Dev-32B outperforms approaches SWE-Syninfer-72B (Ma et al., 2024a) and SWE-Fixer-72B (Xie et al., 2025) which have a resolve rate of 36.6%, achieving a 6.4% improvement while using fewer parameters. With comparable model parameters, SWE-DEV significantly outperforms SWE-Gym (Pan et al., 2024) and SWE-Syninfer by over 12.8% and 5.2%.

Furthermore, by fine-tuning an open-source model that is substantially inferior to proprietary models, our method achieves an impressive 3.4% performance improvement over GPT-4o (OpenAI, 2025) on the SWE-bench-Verified benchmark. Moreover, this approach demonstrates significant progress in effectively addressing complex SWE tasks with open-source models.

Method	Framework	Model	Resolve Rate
<b>Proprietary Models</b>			
OpenAI-GPT-4o (OpenAI, 2024a)	Pipeline	GPT-4o	33.2%
OpenAI-o1-preview (OpenAI, 2024c)	Pipeline	OpenAI-o1-preview	41.3%
OpenAI-o1 (OpenAI, 2024c)	Pipeline	OpenAI-o1	48.9%
Moatless Tools (Örwall, 2024)	Moatless Tools	claude-3-5-sonnet-20241022	39.0%
AutoCodeRover (Zhang et al., 2024b)	AutoCodeRover	GPT-4o (2024-05-13)	38.4%
AutoCodeRover v2.0 (Zhang et al., 2024b)	AutoCodeRover	claude-3-5-sonnet-20241022	46.2%
Agentless-1.5 (Xia et al., 2024)	Agentless	GPT-4o (2024-05-13)	38.8%
Agentless-1.5 (Xia et al., 2024)	Agentless	claude-3-5-sonnet-20241022	50.8%
OpenHands + CodeAct v2.1 (Wang et al., 2024b)	OpenHands	claude-3-5-sonnet-20241022	<b>53.0%</b>
<b>Open-source Models</b>			
SWE-Llama-13B (Jimenez et al., 2024)	RAG	Code Llama-13B	1.2%
SWE-Gym-7B (Pan et al., 2024)	OpenHands	Qwen2.5-Coder-7B-Instruct	10.6%
<b>SWE-Dev-9B(Ours)</b>	OpenHands	GLM-4-9b-chat	13.6%(↑ 12.0%)
<b>SWE-Dev-8B(Ours)</b>	OpenHands	Llama-3.1-8B-Instruct	18.0%(↑ 16.8%)
SWE-SynInfer-7B (Ma et al., 2024a)	Agentless	Qwen2.5-Coder-7B	18.2%
<b>SWE-Dev-7B(Ours)</b>	OpenHands	Qwen2.5-Coder-7B-Instruct	<b>23.4%(↑ 21.6%)</b>
SWE-Gym-32B (Pan et al., 2024)	OpenHands	Qwen2.5-Coder-32B-Instruct	20.6%
SWE-SynInfer-72B (Ma et al., 2024a)	Agentless	Qwen2.5-72B-Instruct	30.2%
SWE-Fixer-72B* (Xie et al., 2025)	Agentless	Qwen2.5-72B	32.8%
<b>SWE-Dev-32B(Ours)</b>	OpenHands	Qwen2.5-Coder-32B-Instruct	<b>36.6%(↑ 30%)</b>

Table 4: Comparison of resolve rates on the SWE-bench-Verified dataset. The table categorizes models into baselines and SWE agents, showcasing their performance. SWE-Dev models attain top-tier results within the realm of open-source models and concurrently exhibit robust performance among closed-source models. The relative improvement (↑) for our models is calculated with respect to their respective base models.

### 4.3 Agentic Post-Training Optimization

**Comparison of Training Methods.** To further enhance model performance, we leveraged instances with fail-to-pass test cases to investigate rejection sampling fine-tuning as Chen et al. (2023) and Zeng et al. (2023), with two offline RL methods: KTO and OREO. Our findings, as illustrated in Figure 4, demonstrate the significant advantages and limitations of these techniques.

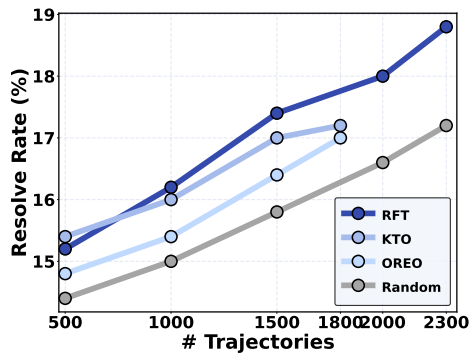


Figure 4: Model performance across different training methods. *Random* denotes supervised fine-tuning with pass@1 trajectories. Other experimental settings remain consistent with those described above.

With only 1,000 trajectories, rejection sampling achieves a 16.2% resolve rate, surpassing the performance of a larger-scale unfiltered dataset. As the number of trajectories increases, rejection sampling further narrows the performance gap, reaching an 18.0% resolve rate with 2,000 trajectories. This approach is valuable in scenarios where training data is costly or computational resources are limited.

We also evaluated two offline RL methods: OREO and KTO. For both methods, positive examples are derived from DeepSeek-V3, and negative examples are sampled from a mixture of DeepSeek-V3 and on-policy outputs. Despite their theoretical potential, these offline methods demonstrated limited performance improvements compared to RFT.

OREO showed small but consistent gains over random sampling, achieving a 17.0% resolve rate with 1,800 trajectories. However, its performance remained below that of RFT. KTO performed slightly better than OREO, achieving a 17.2% resolve rate with 1,800 trajectories. While KTO marginally outperformed OREO, it still lagged behind RFT in overall effectiveness.

The relatively modest improvements observed

with offline RL methods suggest that their reliance on on-policy negative examples and off-policy positive examples may limit their efficacy. On-policy negatives could introduce noise, while the positive examples might lack sufficient diversity to drive meaningful policy optimization.

**Hybrid Training with RFT+OREO.** To further explore the potential of combining different training strategies, we conducted hybrid training experiments by integrating RFT and OREO. Specifically, we divided the 2.3k deduplicated correct trajectories into two complementary subsets: one portion was allocated to RFT, while the remaining data was utilized for OREO, ensuring an efficient use of the available training data.

Importantly, in each run, the RFT+OREO hybrid approach utilized the entire training set, ensuring that the total amount of data seen by the hybrid model matched the full dataset. For comparison, the Random baseline was trained with the same number of pass@1 trajectories as RFT, while other experimental settings were kept consistent. The results are presented in Figure 5.

The RFT-only approach consistently outperformed all other configurations, achieving the highest resolve rate across all subsets. With 2300 RFT trajectories, the model achieved a resolve rate of 21.2%, surpassing all hybrid and OREO-only configurations. This highlights RFT’s strong ability to utilize high-quality data to maximize performance.

## 5 Scaling SWE Agents via Training and Inference Expansion

We explore how SWE agent performance scales along two critical dimensions: the amount of training data and the compute budget. Our findings reveal consistent trends aligned with established scaling laws, and shed light on the cost–performance trade-offs in agent-based settings.

### 5.1 Training Data Scaling

To investigate the impact of data size on model performance, we conducted a scaling analysis by progressively increasing the number of training trajectories. Notably, we filtered trajectories that did not terminate at the given iterations, ensuring that the remaining trajectories were unevaluated. As illustrated in Figure 6, the model’s accuracy increases steadily with the logarithm of the data size, following a near-linear trend in log-log space.

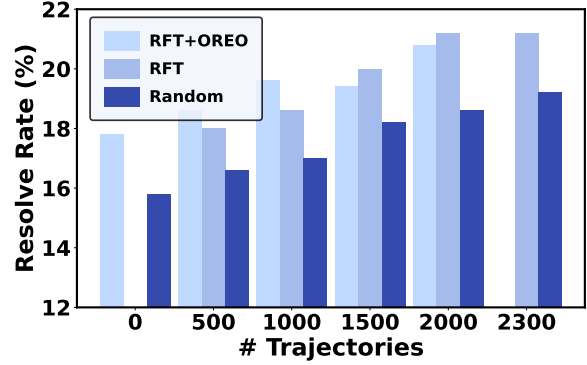


Figure 5: Model performance with hybrid training. In each run, RFT is applied to a subset of the data (measured in trajectories), followed by OREO training on the remaining data. The Random baseline uses the same number of pass@1 trajectories as RFT.

For instance, the 7B model achieved 13.0% accuracy with only 574 trajectories. When the data size was scaled up to 16,639 trajectories, the accuracy improved significantly to 22.8%.

This trend underscores a strong correlation between the quantity of training data and the model’s generalizable ability. The observed linearity in the curve suggests that the model efficiently leverages additional data without exhibiting saturation within the explored range. These findings align with established scaling laws reported in Kaplan et al. (2020) and Hou et al. (2024), where performance predictably improves with increased data size, particularly in the low-data regime.

Moreover, we observe that the benefits of data scaling are more pronounced for larger models. The 32B model shows a sharper improvement compared to the 7B model, suggesting that larger capacity enables more efficient utilization of the available data. This observation echoes prior findings in model scaling literature, where overparameterized models can extract richer inductive biases from limited supervision.

We further applied a filtering strategy to the training trajectories to enhance data quality. Specifically, we evaluated the model-generated patches against reference patches using an LLM-based comparison (Llama 3.1-70B-Instruct). The outputs were categorized into four groups: *identical*, *mostly identical*, *partially identical*, and *different*. Only trajectories labeled as *identical* or *mostly identical* were retained, preserving approximately 65% of the original trajectories.

As shown in Figure 6, the filtered dataset achieves nearly identical performance to the unfiltered dataset, despite a significant reduction in

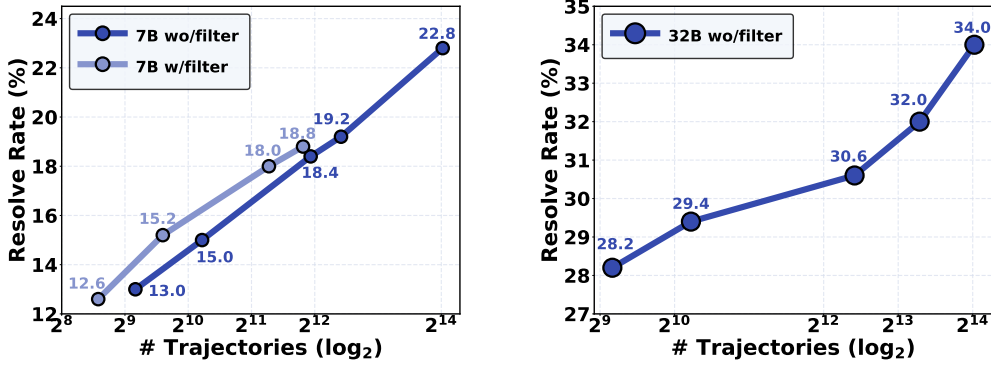


Figure 6: Resolve rates across different training data sizes. Both 7B and 32B models exhibit performance improvements with increased training trajectories.

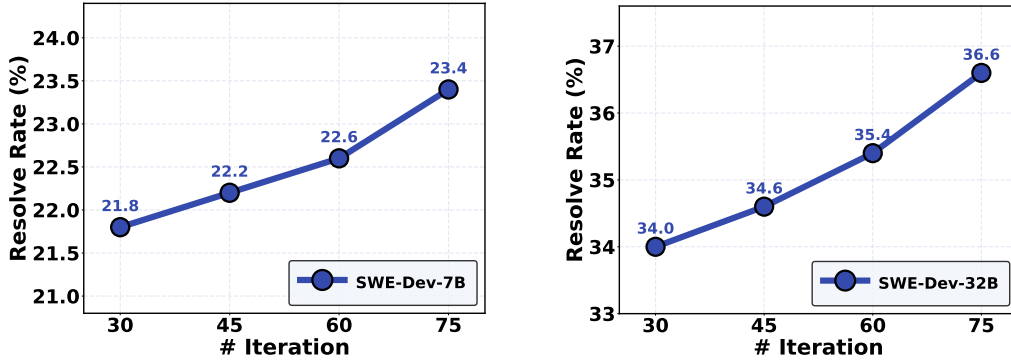


Figure 7: Resolve rates across different numbers of interaction rounds (30–75). Larger models benefit more significantly from extended iteration scaling, though diminishing returns are observed at higher rounds.

training data size. This demonstrates that removing low-quality or irrelevant training samples like stuck in environment setup has minimal impact on model performance, as the retained data is of sufficiently high quality to offset the reduction in quantity.

## 5.2 Inference Time Scaling

In addition to training data scaling, inference-time decisions also play a crucial role in overall model capability. In agent-based tasks, determining an appropriate compute budget is crucial for balancing task success and resource efficiency (Brown et al., 2024). Existing works also demonstrate that increasing test-time compute leads to better performance in math reasoning tasks (OpenAI, 2024b; Hou et al., 2025). Traditional evaluation metrics, such as pass@k, typically focus on multiple independent attempts, which may cost to a large computing budget for SWE evaluation. To address this, we investigate the impact of iteration scaling—progressively increasing the number of interaction rounds within a single run. This approach provides a more natural and efficient alignment with the iterative reasoning inherent in these tasks.

In this experiment, we directly change the RoPE embedding from 32k to 160k and observed that the resolve rate of the 7B model dropped from 22.8% to 21.8%, while the 32B model showed no significant decline. The results, presented in Figure 7, demonstrate how resolve rate improves as the number of interaction rounds increases. We compare two configurations: SWE-Dev-32B, and SWE-Dev-7B. Our key findings include:

### Consistent Improvement with Iteration Scaling.

For all models, increasing the number of interaction rounds leads to higher resolve rate. For instance, SWE-Dev-32B achieves 34.0% at 30 rounds and 36.6% at 75 rounds. This demonstrates that additional iterations allow the models to refine their reasoning and correct prior errors, effectively leveraging the iterative interaction process.

### Diminishing Returns Beyond a Threshold.

Although the resolve rate increases with additional interaction rounds, the rate of improvement diminishes over time. For instance, the improvement observed between 30 and 45 rounds is significantly more pronounced than that between 45 and 75



rounds across all models. This suggests a practical upper limit to the benefits of iteration scaling, beyond which additional rounds yield minimal performance gains. These diminishing returns highlight the importance of calibrating iteration limits to balance computational costs with improvements.

## 6 Conclusion

In this work, we introduce SWE-DEV, an open-source SWE agent equipped with a robust test case construction pipeline. We further investigate the impact of techniques such as rejection sampling and offline RL methods, while proposing iteration scaling as a novel approach to enhance inference-time performance. Our models are evaluated on the SWE-bench-Verified benchmark, achieving state-of-the-art results among open-source agents, with performance scores of 23.4% for the 7B model and 36.6% for the 32B model.

Our work provides a strong foundation for advancing SWE-focused LLM research by introducing an open-source model, and a comprehensive data generation pipeline. We hope these contributions inspire further innovation in developing robust, scalable, and efficient SWE agents capable of addressing real-world SWE challenges.

## 7 Limitation

While offline RL have shown promise, there is significant room for optimization. Incorporating online RL approaches, such as ArCHer (Zhou et al., 2024), DigiRL (Bai et al., 2024), and WebRL (Qi et al., 2025), could further enhance performance by leveraging interaction-based frameworks and dynamic task environments (Snell et al., 2023). Future research could explore adaptive iteration strategies, where the number of rounds is dynamically adjusted based on task complexity or intermediate progress. Hybrid approaches that combine iteration scaling with other optimization techniques may also improve performance while maintaining computational efficiency.

**Acknowledgment.** This work has been supported by the Natural Science Foundation of China (NSFC) 62276148, Tsinghua University (Department of Computer Science and Technology) - Siemens Ltd., China Joint Research Center for Industrial Intelligence and Internet of Things (JCI-IOT), and the New Cornerstone Science Foundation through the XPLOER PRIZE. The GPU compute used in this work is sponsored by Zhipu AI.

The corresponding author: Yuxiao Dong (yuxiaod@tsinghua.edu.cn).

## References

- Antonis Antoniadis, Albert Örwall, Kexun Zhang, Yuxi Xie, Anirudh Goyal, and William Wang. 2024. [Swe-search: Enhancing software agents with monte carlo tree search and iterative refinement](#). *Preprint*, arXiv:2410.20285.
- Ibragim Badertdinov, Maria Trofimova, Yuri Anapol'skiy, Sergey Abramov, Karina Zainullina, Alexander Golubev, Sergey Polezhaev, Daria Litvintseva, Simon Karasik, Filipp Fisin, Sergey Skvortsov, Maxim Nekrashevich, Anton Shevtsov, and Boris Yangel. 2024. Scaling data collection for training software engineering agents. *Nebius blog*.
- Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. 2024. [Di-girl: Training in-the-wild device-control agents with autonomous reinforcement learning](#). *Preprint*, arXiv:2406.11896.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. [Large language monkeys: Scaling inference compute with repeated sampling](#). *Preprint*, arXiv:2407.21787.
- Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Mądry. 2024. [Mle-bench: Evaluating machine learning agents on machine learning engineering](#). *Preprint*, arXiv:2410.07095.
- Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023. [Fire-act: Toward language agent fine-tuning](#). *Preprint*, arXiv:2310.05915.
- Dong Chen, Shaoxin Lin, Muhan Zeng, Daoguang Zan, Jian-Gang Wang, Anton Cheshkov, Jun Sun, Hao Yu, Guoliang Dong, Artem Aliev, Jie Wang, Xiao Cheng, Guangtai Liang, Yuchi Ma, Pan Bian, Tao Xie, and Qianxiang Wang. 2024. [Coder: Issue resolving with multi-agent and task graphs](#). *Preprint*, arXiv:2406.01304.
- 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, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain,

- William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.
- Cursor. 2024. Cursor – ai-powered code editor. <https://www.cursor.com/cn>. Accessed: 2025-05-25.
- DeepSeek-AI. 2024. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. [Kto: Model alignment as prospect theoretic optimization](#). *Preprint*, arXiv:2402.01306.
- Anmol Gautam, Kishore Kumar, Adarsh Jha, Mukunda NS, and Ishaan Bhola. 2024. [Supercoder2.0: Technical report on exploring the feasibility of llms as autonomous programmer](#). *Preprint*, arXiv:2409.11190.
- Team GLM. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Alexander Golubev, Sergey Polezhaev, Karina Zainulina, Maria Trofimova, Ibragim Badertdinov, Yuri Anapolskiy, Daria Litvintseva, Simon Karasik, Filipp Fisun, Sergey Skvortsov, Maxim Nekrashevich, Anton Shevtsov, Sergey Abramov, and Boris Yangel. 2024. Leveraging training and search for better software engineering agents. *Nebius blog*. <https://nebius.com/blog/posts/training-and-search-for-software-engineering-agents>.
- Zhenyu Hou, Pengfan Du, Yilin Niu, Zhengxiao Du, Aohan Zeng, Xiao Liu, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. 2024. Does rlhf scale? exploring the impacts from data, model, and method. *arXiv preprint arXiv:2412.06000*.
- Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. Advancing language model reasoning through reinforcement learning and inference scaling. In *ICML*.
- Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. [Openrlhf: An easy-to-use, scalable and high-performance rlhf framework](#). *Preprint*, arXiv:2405.11143.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi 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, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. [Qwen2.5-coder technical report](#). *Preprint*, arXiv:2409.12186.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. [Swe-bench: Can language models resolve real-world github issues?](#) *Preprint*, arXiv:2310.06770.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- Jia Li, Ge Li, Xuanming Zhang, Yihong Dong, and Zhi Jin. 2024a. [Evocodebench: An evolving code generation benchmark aligned with real-world code repositories](#). *Preprint*, arXiv:2404.00599.
- Jia Li, Ge Li, Yunfei Zhao, Yongmin Li, Huanyu Liu, Hao Zhu, Lecheng Wang, Kaibo Liu, Zheng Fang, Lanshen Wang, Jiazheng Ding, Xuanming Zhang, Yuqi Zhu, Yihong Dong, Zhi Jin, Binhua Li, Fei Huang, and Yongbin Li. 2024b. [Deveval: A manually-annotated code generation benchmark aligned with real-world code repositories](#). *Preprint*, arXiv:2405.19856.
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Posen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. [Competition-level code generation with alpha-code](#). *Science*, 378(6624):1092–1097.
- Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023. [Ring attention with blockwise transformers for near-infinite context](#). *Preprint*, arXiv:2310.01889.
- Yizhou Liu, Pengfei Gao, Xinchun Wang, Jie Liu, Yexuan Shi, Zhao Zhang, and Chao Peng. 2024. [Marscode agent: Ai-native automated bug fixing](#). *Preprint*, arXiv:2409.00899.
- Team Llama. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Yingwei Ma, Rongyu Cao, Yongchang Cao, Yue Zhang, Jue Chen, Yibo Liu, Yuchen Liu, Binhua Li, Fei Huang, and Yongbin Li. 2024a. [Lingma swe-gpt: An open development-process-centric language model for automated software improvement](#). *Preprint*, arXiv:2411.00622.
- Yingwei Ma, Qingping Yang, Rongyu Cao, Binhua Li, Fei Huang, and Yongbin Li. 2024b. [How to understand whole software repository?](#) *Preprint*, arXiv:2406.01422.
- OpenAI. 2024a. [Gpt-4o system card](#). Accessed: 2025-06-02.

- OpenAI. 2024b. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms>.
- OpenAI. 2024c. Openai o1 system card. Accessed: 2025-06-02.
- OpenAI. 2025. Competitive programming with large reasoning models. *Preprint*, arXiv:2502.06807.
- OpenAI. 2025. Introducing codex. <https://openai.com/index/introducing-codex/>. Accessed: 2025-05-25.
- Albert Örwall. 2024. moatless-tools.
- Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. 2024. Training software engineering agents and verifiers with swe-gym. *Preprint*, arXiv:2412.21139.
- Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Wenyi Zhao, Yu Yang, Xinyue Yang, Jiadai Sun, Shuntian Yao, Tianjie Zhang, Wei Xu, Jie Tang, and Yuxiao Dong. 2025. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. *Preprint*, arXiv:2411.02337.
- Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. 2023. Offline rl for natural language generation with implicit language q learning. *Preprint*, arXiv:2206.11871.
- Hongjin Su, Ruoxi Sun, Jinsung Yoon, Pengcheng Yin, Tao Yu, and Serkan Ö. Arik. 2025. Learn-by-interact: A data-centric framework for self-adaptive agents in realistic environments. *Preprint*, arXiv:2501.10893.
- Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. 2024a. Offline reinforcement learning for llm multi-step reasoning. *Preprint*, arXiv:2412.16145.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better llm agents. *Preprint*, arXiv:2402.01030.
- 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, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. 2024c. Openhands: An open platform for ai software developers as generalist agents. *Preprint*, arXiv:2407.16741.
- Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2024. Agentgym: Evolving large language model-based agents across diverse environments. *Preprint*, arXiv:2406.04151.
- Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. 2024. Agentless: Demystifying llm-based software engineering agents. *Preprint*, arXiv:2407.01489.
- Chengxing Xie, Bowen Li, Chang Gao, He Du, Wai Lam, Difan Zou, and Kai Chen. 2025. Swe-fixer: Training open-source llms for effective and efficient github issue resolution. *Preprint*, arXiv:2501.05040.
- John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024a. Swe-agent: Agent-computer interfaces enable automated software engineering. *Preprint*, arXiv:2405.15793.
- John Yang, Carlos E. Jimenez, Alex L. Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R. Narasimhan, Diyi Yang, Sida I. Wang, and Ofir Press. 2024b. Swe-bench multimodal: Do ai systems generalize to visual software domains? *Preprint*, arXiv:2410.03859.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *Preprint*, arXiv:2308.01825.
- Daoguang Zan, Zhirong Huang, Wei Liu, Hanwu Chen, Linhao Zhang, Shulin Xin, Lu Chen, Qi Liu, Xiaojian Zhong, Aoyan Li, Siyao Liu, Yongsheng Xiao, Liangqiang Chen, Yuyu Zhang, Jing Su, Tianyu Liu, Rui Long, Kai Shen, and Liang Xiang. 2025. Multi-swe-bench: A multilingual benchmark for issue resolving. *Preprint*, arXiv:2504.02605.
- Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2023. Agenttuning: Enabling generalized agent abilities for llms. *Preprint*, arXiv:2310.12823.
- Kexun Zhang, Weiran Yao, Zuxin Liu, Yihao Feng, Zhiwei Liu, Rithesh Murthy, Tian Lan, Lei Li, Renze Lou, Jiacheng Xu, Bo Pang, Yingbo Zhou, Shelby Heinecke, Silvio Savarese, Huan Wang, and Caiming Xiong. 2024a. Diversity empowers intelligence: Integrating expertise of software engineering agents. *Preprint*, arXiv:2408.07060.
- Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. 2024b. Autocoderover: Autonomous program improvement. *Preprint*, arXiv:2404.05427.
- Wenting Zhao, Nan Jiang, Celine Lee, Justin T Chiu, Claire Cardie, Matthias Gallé, and Alexander M Rush. 2024. Commit0: Library generation from scratch. *Preprint*, arXiv:2412.01769.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. [Archer: Training language model agents via hierarchical multi-turn rl](#). *Preprint*, arXiv:2402.19446.



## A Appendix

### A.1 Prompt Templates

#### Prompt for Patch Filtering

You are a professional programmer. Given two git patches that aim to fix the same bug, your task is to analyze if the second patch correctly fixes the bug and identify any issues in it. Please ignore any changes to `reproduce_error.py` files as they are only used for debugging purposes.

**Information provided:**

- **Patch 1 (correct solution):** {patch1}
- **Patch 2 (patch to evaluate):** {patch2}

**Guidelines:**

1. When comparing patches, focus only on code changes that affect actual functionality. Differences in logging or documentation can be ignored.
2. The evaluation levels are:
  - **identical:** Patch 2's functionality is exactly the same as Patch 1.
  - **mostly:** Patch 2 implements most of the functionality correctly with minor differences.
  - **partially:** Patch 2 only implements some of the required functionality.
  - **different:** Patch 2's functionality is completely different or incorrect.
3. If Patch 2 is rated as **partially** or **mostly**, please specify the functional differences in the explanation.

**Final Instructions:**

- Provide your analysis and judgment in the following format:

[Explanation]  
Explanation of the reason why the patch is judged as 'identical', 'mostly', 'partially', or 'different'.  
[Judgment]  
The judgment of the patch is 'identical', 'mostly', 'partially', or 'different'.

#### Prompt for Generating Naive NL Test Case Description

You are a skilled test engineer. Your mission is to create a minimal, edge-case test scenario that rigorously validates the effectiveness of the patch. This test case must satisfy the following conditions:

1. **Fail with the unpatched code:** Demonstrate the specific bug, issue, or limitation that the patch is designed to address. Ensure the test triggers this behavior reliably and consistently.
2. **Pass with the patched code:** Confirm that the patch resolves the issue without introducing new problems or regressions.

**Focus Areas:**

- Exercise uncommon or edge-case code paths.
- Test for boundary conditions or unexpected input.
- Mimic realistic usage scenarios where the original behavior fails.

**Information provided:**

- **Repository name:** { }
- **GitHub issue description:** { }
- **Correction patch:** { }
- **Hints Text:** { }

**Your Task:**

- Briefly analyze the problem description and hints text to identify **where the issue lies and what should be fixed**.
- Write a concise test case description that reflects the modification introduced by the patch.
- Ensure the description targets the root cause of the issue and guides the generation of a challenging, edge-case test scenario.

**Important Notes:** - Avoid unrelated information or greetings. - Focus exclusively on the test case description and the problem analysis.

### Prompt for Generating Gherkin Description

You are an experienced test engineer. Your task is to write a test following the Gherkin syntax based on the information provided below. This test must verify whether the correction patch in the repository correctly resolves the described issue.

**Key Points:** - The test should **fail** with the unpatched code and **pass** with the patched code.

**Information Provided:**

- **Repository name:** {}
- **GitHub issue description:** {}
- **Correction patch:** {}
- **Hints Text:** {}
- **Analysis for the test cases (for reference):** {}

**Requirements:**

1. Use the **Given-When-Then** structure of Gherkin syntax.
2. Clearly describe:
  - Preconditions (Given).
  - Triggering events (When).
  - Expected outcomes (Then).
3. Ensure the test logic is clear, concise, and covers all relevant scenarios.
4. Avoid including unimportant test cases, such as modifications to README files.

**Instructions:** - Wrap each Gherkin test description in triple backticks ("`gherkin"). - Example format:

```
```gherkin
{{YOUR DESCRIPTION}}
```
```

### Prompt for Test Case Generation

You are a test engineer. Given a GitHub issue description and the golden patch, your task is to build test cases that **reproduce the error** according to the patch. In detail, the test cases should reproduce the error in the issue description. Your test case will run at the **root** of the project. Please be careful of the relative path to avoid path-related errors.

**Information provided:**

- **Repository name:** {}
- **GitHub issue description:** {}
- **Hints Text:** {}
- **Correction patch:** {}
- **Project tree (file depth less than 3):** {}
- **Test Case Description:** {}
- **Relevant code segments in the original version:** {}

**Steps to follow:**

1. **Identify the incorrect code:** Analyze the provided information to locate the error that the patch addresses. Identify the required packages and the types of test cases to write.
2. **Generate the test case:** Write test cases that will **fail without the correction patch** and **pass with the correction patch**. Each test case must be enclosed within `<testcase></testcase>` tags.
3. Ensure that no additional execution beyond your test case is performed. Avoid unsafe commands or unnecessary changes to the project.

#### Format Requirements:

- Test Case:
  - Wrap each test case in `<testcase></testcase>` tags.
  - Use triple backticks (`````) to enclose the test code within the `<testcase></testcase>` tags.
  - The test cases must be ready to run with pytest and should include any necessary mock data or fixtures.

#### Environment Information:

- **Python Version:** 3.9
- **Platform:** Ubuntu 22.04.5 LTS
- **Execution Command:**

```
python -m pytest --no-header -rA -p no:cacheprovider -W ignore::DeprecationWarning --continue -on-collection-errors --tb=short
```
- **Execution Path:** Root directory of the project

#### Example Solution:

In `src/utils/csv_utils.py`:

```
from CSVconverter.src.utils import csv
def read_csv_and_sum(filename):
    """Calculate the sum of all numbers in a CSV file"""
    total = 0
    with open(filename, 'r') as file:
        reader = csv.reader(file)
        for row in reader:
            total += row[0]
    return total
```

The code directly adds `row[0]` to `total` without validating if `row[0]` is an integer. If the CSV file contains non-numeric values (e.g., strings or empty fields), it will raise runtime errors like `TypeError` or `ValueError`. These errors match the problem statement. So I'll write test cases here.

#### Fix Explanation:

1. It tries to convert `row[0]` to an integer using `int()`.
2. If `row[0]` is not a valid integer, it skips that row using a `try...except` block.

The goal is to write test cases that:

1. Test case with non-numeric data in the CSV (should raise an error in the original code).
2. Same test case should now correctly handle non-numeric rows and calculate the sum of valid numeric values.

#### Example Test Case:

```
<testcase>
```python
import os
import pytest
from src.utils.csv_utils import read_csv_and_sum

@pytest.fixture
def create_csv_file():
    """Fixture to create a temporary CSV file for testing."""
    def _create_file(contents, filename="test.csv"):
        with open(filename, 'w') as f:
            f.write(contents)
```

```

        return filename
    yield _create_file
    # Cleanup after test
    if os.path.exists("test.csv"):
        os.remove("test.csv")

def test_valid_csv(create_csv_file):
    """Test case with valid numeric data."""
    filename = create_csv_file("1\n2\n3\n")
    result = read_csv_and_sum(filename)
    assert result == 6 # Expected sum of numbers

def test_non_numeric_csv(create_csv_file):
    """Test case with non-numeric data."""
    filename = create_csv_file("1\nabc\n3\n")
    with pytest.raises(TypeError):
        read_csv_and_sum(filename)

def test_empty_csv(create_csv_file):
    """Test case with an empty CSV file."""
    filename = create_csv_file("")
    result = read_csv_and_sum(filename)
    assert result == 0 # Expected sum is 0
...
</testcase>

```

#### Final Instructions:

- **Test case format:** Ensure the tests follow pytest conventions and are ready to run. Do not enable dangerous commands like `ifconfig` or `iptables`.
- **Import files correctly:** Carefully handle functions and classes in the current package.
- **Patch validation:** Test cases should fail when run against the unpatched code and pass after the patch is applied.
- **File handling:** Ensure any files needed for the test exist. Substitute paths like `/path/to/dest` with actual paths.

### Prompt for Test Case Revision

You are tasked with generating test cases for a given GitHub issue. The code with the golden patch should pass the test case while failing without the golden patch. Now, the test case has **failed even after applying the patch**. You are required to improve it.

#### Provided Information:

- **Repository name:** {}
- **GitHub issue description:** {}
- **Hints Text:** {}
- **Golden patch** (this patch passed with the previous test cases): {}
- **Project Tree** (file depth less than 3): {}
- **Relevant Code Segments:** {}
- **Available Relevant APIs:** {}
- **Wrong Test Case:** {}
- **Error History:** {}

#### Task Instructions:

1. **Analyze the error history carefully:** Review the error history to understand why the previous test cases passed without the patch. For example:
  - Rewrite wrong test cases if errors occur on specific tests.
  - Consider import dependencies when encountering `ImportError` or similar errors.



2. **Preserve the original intent:** Ensure the new test cases still target the original issues that the patch is designed to fix.

3. **Format Requirements:** Your test case should strictly follow the original format. Specifically:

- Setup commands should be wrapped in `<env></env>` tags. The commands should be enclosed in triple backticks (```) inside the `<env>`.
- Test cases must be wrapped in `<testcase></testcase>` tags, and the test code should be enclosed in triple backticks (```) inside the `<testcase>`.

**Example Format:**

```
<testcase>
```python
# Your improved test case here
```
</testcase>
```

```
<env>
```bash
# Required setup commands here
```
</env>
```

**Important Notes:**

- The new test case must **still fail** on the unpatched code and **pass** after applying the patch.
- Strictly follow the format and preserve the original test intent.

### Prompt for Extracting API Signature or Class Name

Here is an error message, and you are required to extract the API signature or class name that raises the error. You should strictly follow the format instruction and do not include any unrelated greeting words. The API should **directly** raise the error, and you should not include any other APIs that are not related to the error.

**Important Notes:**

- For safety, you should never use `os.system` in the test case code. If you want to operate on the system, use the commands in setup commands!

**Format Instruction:**

- If the error message is related to a function, the API signature should be in the following format:

```
<function>module1.module2.function_name(parameters)</function>
```

- If the error message is related to a class, the class name should be in the following format:

```
<class>module1.module2.class_name</class>
```

- If no API signature or class name is found, you should provide an empty string:

```
<empty></empty>
```

**Example:**

1. **Error Message:**

```
... ERROR collecting swedev-test.py ...
ImportError: cannot import name 'Config' from 'pkgconfig.pkgconfig' ...
```

**Result:**

```
<class>pkgconfig.pkgconfig.Config</class>
```

2. **Error Message:**

```
... ERROR collecting swedev-test.py ...
TypeError: example_function() missing 2 required positional arguments: 'param1' and 'param2' ...
```

**Result:**

```
<function>example_module.example_function(param1, param2)</function>
```

**Task:**

- **Error Message:** {}
- **Result:**

## A.2 Principles for filtering instances

We keep instances that meet none of the following criteria. Except for statement length and patch size, other metrics are judged by Llama-3.1-70B-Instruct.

- Overly concise ( $\leq 100$  words)
- No large patch for limited context ( $\geq 0.8$  MB)
- Vague phrasing (e.g., "fix the issue," "improve performance")
- Lack of context and no specific problem/goal
- Formatting issues (single word/punctuation)
- Lack of specificity (no detailed problem/outcome)
- Redundancy (repeats similar issues)

## A.3 Generated Test Cases

### Gherkin Scenarios

**Feature: Quiet mode in SQLFluff CLI**

**Scenario: Run sqlfluff fix with `--quiet` option**

- **Given** I have a SQL file with linting violations
- **When** I run 'sqlfluff fix' with the '`--quiet`' option
- **Then** the output should only show the fix status and the number of fixes applied
- **And** the output should not contain detailed information about each fix

**Scenario: Run sqlfluff fix with `--force` and `--quiet` options**

- **Given** I have a SQL file with multiple linting violations
- **When** I run 'sqlfluff fix' with the '`--force`' and '`--quiet`' options
- **Then** the output should only show the fix status and the number of fixes applied
- **And** all fixes should be applied automatically

**Scenario: Run sqlfluff fix with both `--quiet` and `--verbose` options**

- **Given** I have a SQL file with linting violations
- **When** I run 'sqlfluff fix' with both '`--quiet`' and '`--verbose`' options
- **Then** I should see an error message stating that `--quiet` and `--verbose` cannot be used together
- **And** the process should exit with an error code

## Examples for Test Case Generation

```
diff --git a/codeagt_test_0.py b/codeagt_test_0.py
new file mode 100644
index 0000000..0000000
--- /dev/null
+++ b/codeagt_test_0.py
@@ -0,0 +1,63 @@
+import os
+import pytest
+from borgapi.borgapi import BorgAPI
+
+@pytest.fixture
+def create_borg_repo():
+    """Fixture to create a temporary Borg repository for testing."""
+    def _create_repo(repo_name="test_repo"):
+        api = BorgAPI()
+        api.init(repo_name)
+        return api, repo_name
+    yield _create_repo
+    # Cleanup after test
+    api, repo_name = _create_repo()
+    api.delete(repo_name)
+
+def test_extract_to_stdout(create_borg_repo):
+    """Test case to extract an archive to stdout."""
+    api, repo_name = create_borg_repo()
+    archive_name = "test_archive"
+    file_path = "test_file.txt"
+    with open(file_path, "w") as f:
+        f.write("Hello, World!")
+    api.create(archive_name, file_path)
+    result = api.extract(archive_name, stdout=True)
+    assert result[0] == "Hello, World!"
+
+def test_extract_to_stdout_with_multiple_files(create_borg_repo):
+    """Test case to extract multiple files from an archive to stdout."""
+    api, repo_name = create_borg_repo()
+    archive_name = "test_archive"
+    file_path1 = "test_file1.txt"
+    file_path2 = "test_file2.txt"
+    with open(file_path1, "w") as f:
+        f.write("Hello, World!")
+    with open(file_path2, "w") as f:
+        f.write("This is another file.")
+    api.create(archive_name, file_path1, file_path2)
+    result = api.extract(archive_name, stdout=True)
+    assert result[0] == "Hello, World!This is another file."
+
+def test_extract_to_stdout_with_json_data(create_borg_repo):
+    """Test case to extract JSON data from an archive to stdout."""
+    api, repo_name = create_borg_repo()
+    archive_name = "test_archive"
+    file_path = "test_file.json"
+    with open(file_path, "w") as f:
+        f.write('{"key": "value"}')
+    api.create(archive_name, file_path)
+    result = api.extract(archive_name, stdout=True)
+    assert result[0] == '{"key": "value"}'
+
+def test_extract_to_stdout_with_text_data(create_borg_repo):
+    """Test case to extract text data from an archive to stdout."""
+    api, repo_name = create_borg_repo()
+    archive_name = "test_archive"
+    file_path = "test_file.txt"
+    with open(file_path, "w") as f:
+        f.write("This is a test file.")
+    api.create(archive_name, file_path)
```

```
+ result = api.extract(archive_name, stdout=True)
+ assert result[0] == "This is a test file."
```